北京郵電大學
BEIJING UNIVERSITY OF POSTS AND TELECOMMUNICATIONS

Computer Organization and Architecture

Chapter 6

External Memory

School of Computer Science（National Pilot Software Engineering School）
AO XIONG (熊翱)
xiongao@bupt.edu.cn

Preface

We have learned:

- Basic Concepts and Computer Evolution 基本概念和计算机发展历史
- Performance Issues 性能问题
- Top level view of computer function and interconnection 计算机功能和互联结构顶层视图
- Cache Memory cache存储器
- Internal Memory 内部存储器
  - Semiconductor main memory 半导体内存
  - Error correction 纠错
  - DDR DRAM 高级DRAM组织
  - Flash 闪存

Preface

- Internal memory is volatile 内部存储器易失
- Data needs to be saved permanently 数据需要永久保存
- External Memory 外部存储器
- Many types of components undertake such work，such as punch tape 有很多类型的外存，比如打孔纸带
- Disk is the most commonly used external storage 磁盘最常用
- interconnected to system bus through I/O , is called external memory 通过I/O和系统总线互联，称为外部存储器

Preface

We will focus the following contents today:

- External Memory 外部存储器
  - What is the principle of disk operation? 磁盘的工作原理是什么?
  - How to improve the reliability of disk? 如何提高磁盘的可靠性?
  - What external storage is there besides disk? 除了磁盘之外，还有哪些外部存储器?

Outline

- Magnetic Disk 磁盘
- RAID RAID技术
- Solid State Drives 固态硬盘
- Optical Memory 光盘
- Magnetic Tape 磁带

Magnetic disk 磁盘

- Magnetic Read and Write Mechanisms 磁盘读写机制
- Data Organization and Formatting 数据的组织和格式
- Physical Characteristics 物理特性
- Disk Performance Parameters 磁盘性能参数

Composition of magnetic disks 磁盘的构成

- A disk is a nonmagnetic circular platter coated with magnetizable material 磁盘是在圆形非磁性盘片上涂了一层磁性材料
- Data are recorded to or read from the disk using the read/write head 数据通过读写头来写入或者从磁盘中读取
- Hard disks can have one platter, or more 硬盘可能有多个盘片
- Direct access storage 直接访问存储方式

Access method 存储器存取方式

有哪几种存取方式？

Access method 存储器存取方式

- Sequential 顺序存取
- Direct 直接存取
- Random 随机存取
- Associative 关联存取

Access method 存储器存取方式

- **顺序存取**：数据以线性的方式存放在存储介质中。读取数据必须按照顺序的方式，从当前的位置按照顺序移动到数据所在的位置进行读取。典型的顺序存取是磁带。
- **直接存取**：数据按块存储在介质中，并且每个块都有一个唯一的地址。存取时，先按照这个唯一地址到达所在的块，然后在块中，顺序搜索到数据。典型的直接存取是硬盘。
- **随机存取**：每个存取单元都有一个唯一的地址，通过寻址机制可以直接找到这个位置，不依赖于之前的存取操作所在的位置。典型的随机存取是内存。
- **关联存取**：关联存取是通过对字中的部分内容进行比较，如果匹配就进行存取操作。关联存取是对字的内容进行比较，不是地址寻址。典型的关联存取是cache。

Read and write mechanism 读写机制

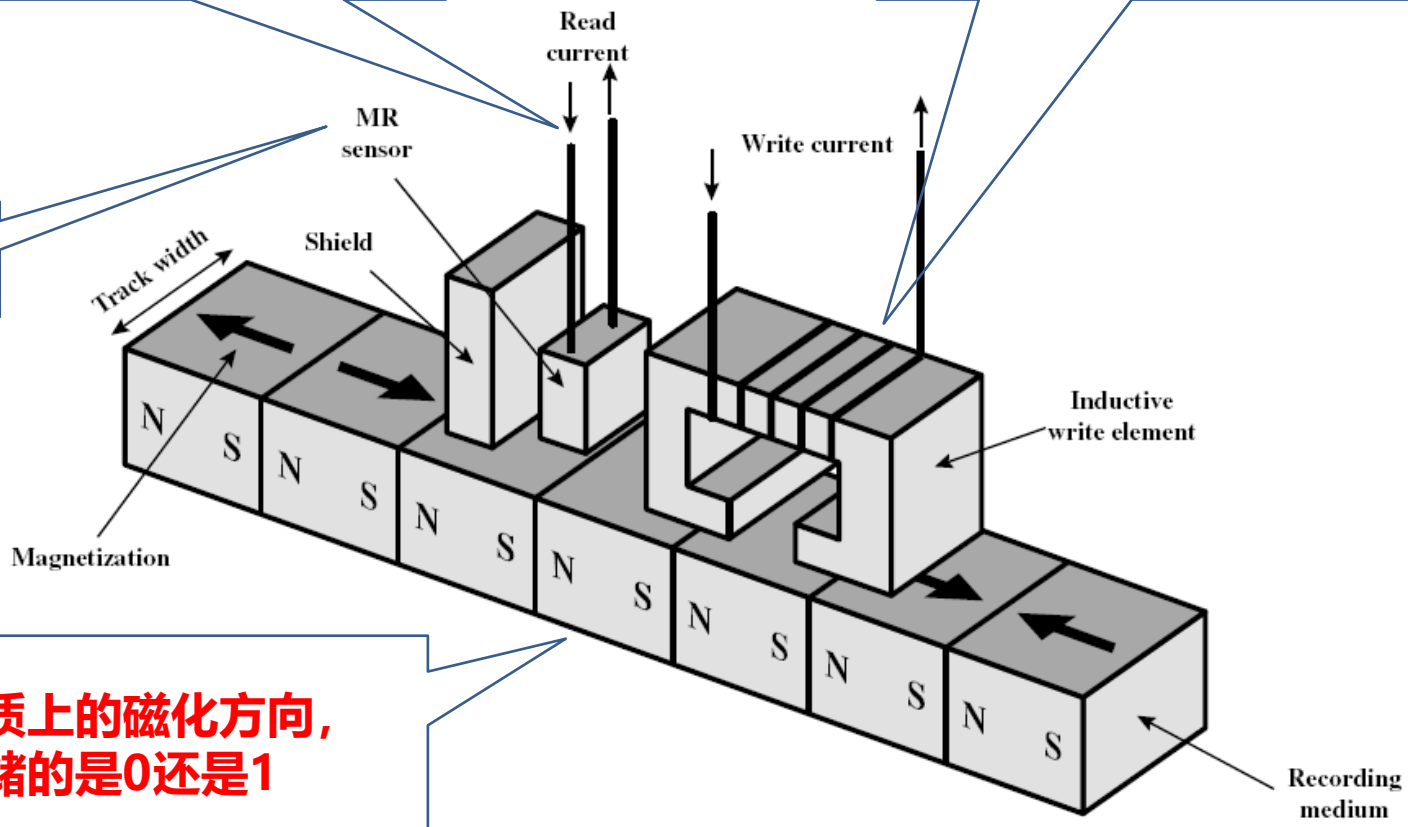
- Recording & retrieval via conductive coil called a head 读写通过一个导电线圈，这个导电线圈称为磁头
  - May be single read/write head or separate ones 可能是读写共用一个磁头，也可以分别有读和写的磁头
  - During read/write, head is stationary, platter rotates 读写的时候，磁头不动，盘片旋转
- Write 写
  - Pulses sent to head 脉冲发送给磁头
  - Current through coil produces magnetic field 通过线圈电流产生磁场
  - Magnetic pattern recorded on surface below 磁性模式记录在磁盘表面

Read and write mechanism 读写机制

读的时候，读磁头读取磁性介质上的磁化方向，经过转换，生成0或者1

写的时候，正或负的电流通过写磁头，产生不同的磁性，对盘上的磁介质进行不同方向的磁化，从而记录0或者1

MR：磁阻感应器



磁性介质上的磁化方向，表示存储的是0还是1

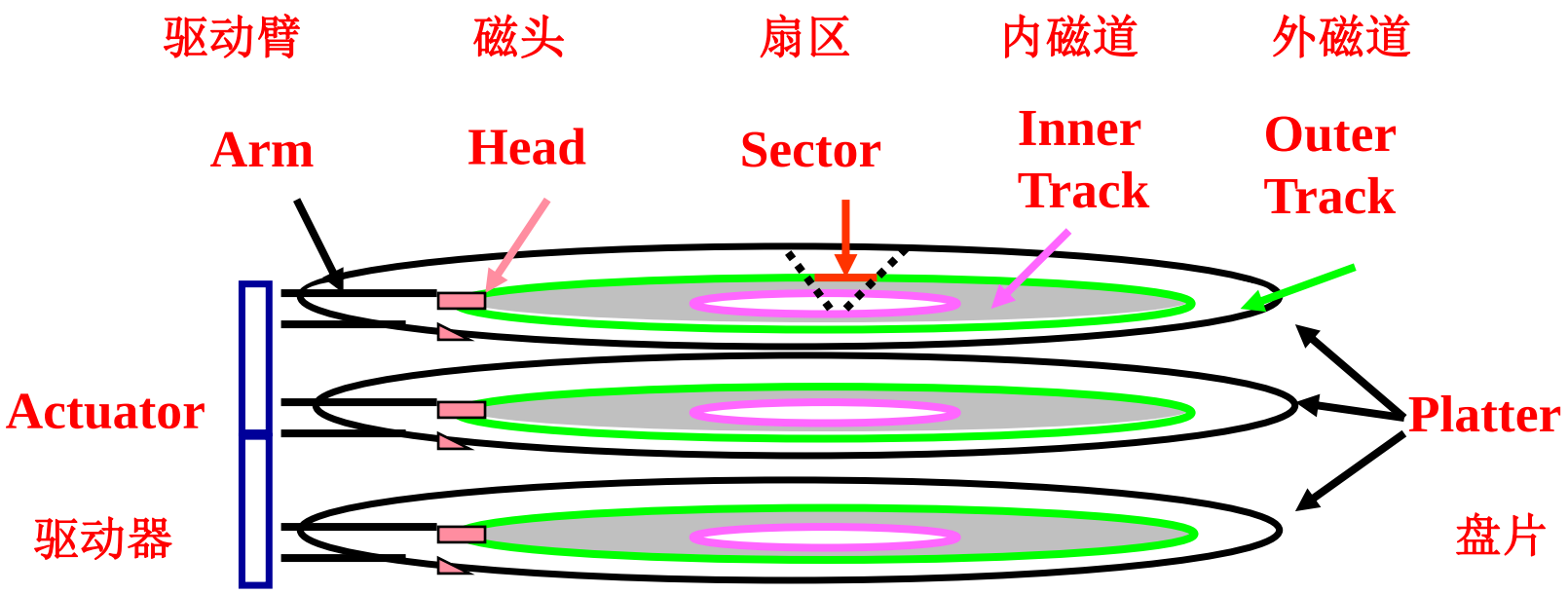
Read and Write Mechanisms 读写机制

- Read (traditional) 传统读机制
  - 磁盘上的磁性物质被磁化成两个方向，表示0和1
  - 磁头上有线圈，读的时候磁头不动，磁盘通过马达在旋转
  - 磁性物质在转动的时候，会在磁头的线圈上产生电流
  - 不同的磁化方向产生的电流方向不一样
  - 通过分析电流方向，就可以得到磁化方向，从而确定存储的是0还是1
  - Magnetic field moving relative to coil produces current 相对于线圈运动的磁场产生电流
  - Coil is the same for read and write 读写磁头相同
  - Slow reading and writing speed 读写速度慢

Read and Write Mechanisms 读写机制

- Read (contemporary) 当前读机制
  - Separate read head, close to write head 独立的读磁头，和写磁头相邻
  - MR（magneto resistive）： it senses the magnetization direction of upper disc 磁阻传感器：感应盘片上位的磁化方向
  - magnetization direction changes the resistance of the magneto resistive sensor 磁化方向改变磁阻传感器的电阻的大小
  - magnetization direction can be obtained by measuring the resistance value to determine the stored information 测量电阻值就可以得到磁化方向，确定存储的信息
  - Set shield to prevent interference 设置屏蔽器防止干扰
  - High frequency operation，Higher storage density and speed 高频操作，更高的存储密度，更快的速度

Terminology 术语



A hard disk may have several platters, with information recorded magnetically on both surfaces 硬盘可能会有多个盘片，信息以磁性记录在盘片的双面

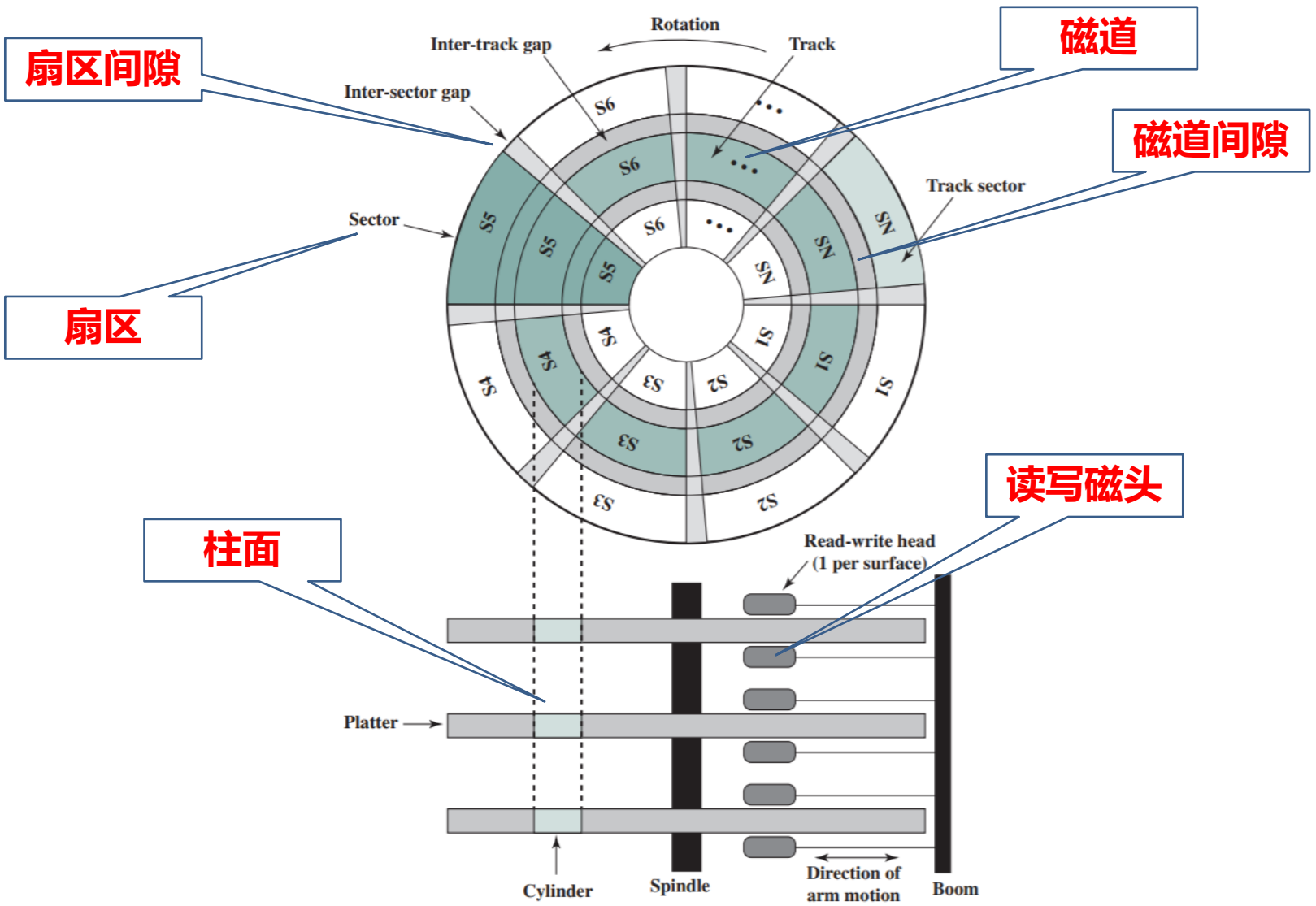
Data Organization and Formatting 1 数据组织和格式1

- Disk is divided into several concentric rings 磁盘划分为若干个同心圆
  - Called track 称为磁道
  - Data is stored in the track 数据存在磁道上
- tracks 磁道
  - Gaps between tracks to prevent interference between different tracks 磁道间有间隙，防止磁道之间的干扰
  - Reduce gap to increase capacity 减小间隙以提高容量
  - Disc rotates at constant angular velocity 盘片以恒定的角速度旋转
  - Same number of bits per track (variable packing density) 每个磁道的位数相同（可变存储密度）

Data Organization and Formatting 数据组织和格式2

- Tracks divided into sectors 磁道划分为扇区
  - The smallest unit of data storage 数据存取的最小单位
  - Each track generally contains several hundred sectors 每个磁道包含几百个扇区
  - Current sector size is 512 bytes 一个扇区一般为512字节
  - Data is written to or read from the disk in sectors 数据以扇区为单位写入磁盘
- Gaps are also left between sectors to avoid interference between sectors 扇区之间有空隙，防止干扰

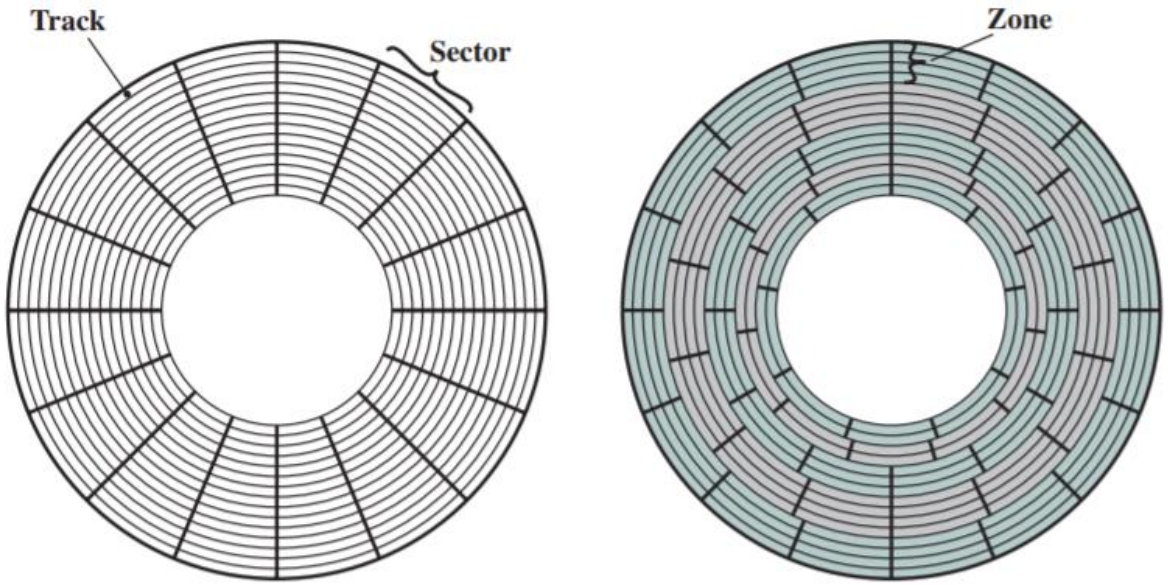
Disk Data Layout 磁盘数据分布



Disk velocity 磁盘转速

- Rotate disk at constant angular velocity (CAV) 以恒定角速度CAV旋转的盘
- Same number of sectors in different tracks 不同的磁道中的扇区数相同
- Different space between bits in different tracks 不同的磁道位之间的空隙不同
- Storage density of the inner ring determines the capacity of disk 内圈的存储密度决定了磁盘的容量
- Waste of storage space on external tracks 靠外磁道的存储空间浪费
- Can use zones to increase capacity 通过使用分区来提高容量

Disk layout methods diagram 磁盘分布方法示意图



(a) Constant angular velocity

(b) Multiple zone recording

- 磁盘以恒定角速度旋转
- 每个磁道存储的数据量相同
- 外圈的存储区域浪费

- 磁盘分为若干个区域，每个区域包含多个磁道
- 区域内所有磁道的扇区数相同
- 存储容量提高，但读取的电路复杂，不同区域需要不同的读写速度

Winchester disk 温盘1

- First disk storage system，305 RAMAC by IBM in 1956, 5MB 1956年IBM发明第一款磁盘存储系统305RAMAC，5MB容量
- Hard disk 3340 Invented by IBM in 1973， 1973年IBM发明3340硬盘
- Has several coaxial metal discs coated with magnetic material 若干个涂有磁性材料的同轴金属盘
- Disk, magnetic head and drive mechanism are sealed in a box 磁盘、磁头和驱动装置都封装在一个盒子里
- The disk has two 30M storage units, and the caliber and charge of "Winchester rifle" are also two 30, so it is named 磁盘有两个30MB的存储单元，“温彻斯特步枪”的口径和装药也是两个30，因此得名

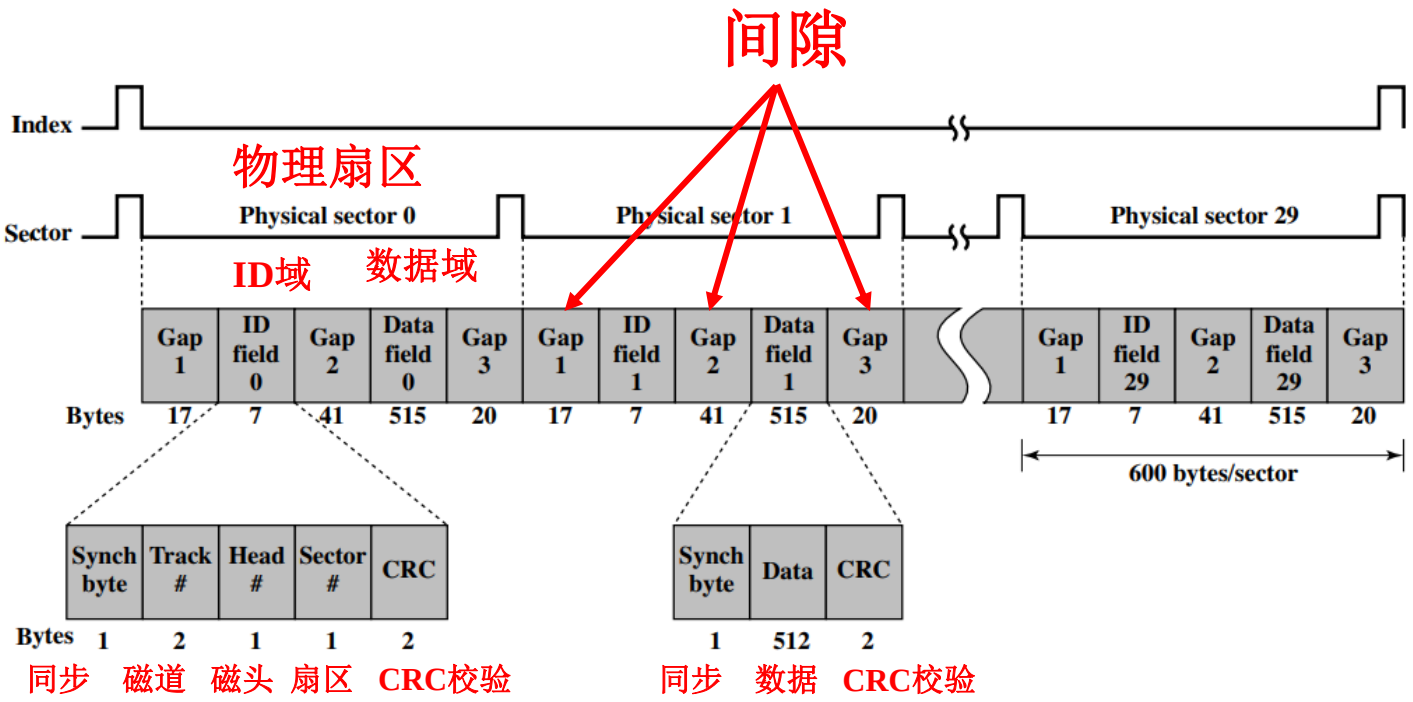
Winchester disk 温盘2

- Magnetic head is not in contact with the disk to improve the durability of the disk 磁头不与磁盘接触，以提高磁盘的耐久性
- In 1980， Seagate manufactured the first Winchester hard disk on a personal computer 1980年，希捷公司制造出了个人电脑上的第一块温彻斯特硬盘
- Invention of high-sensitivity magnetic head makes it possible for high-density storage 高灵敏度的磁头为高密度的存储提供了可能
- Almost all mechanical hard disks are based on Winchester technology 几乎所有的机械硬盘都基于温盘技术
- Although the capacity of hard disk has increased many times, the principle is the same as before 尽管现在硬盘的容量增加了很多倍，但是原理和之前的一样

Hard disk – Winchester 温氏硬盘

- Developed by IBM IBM发明
- Sealed unit 封装单元
- One or more platters (disks) 一个或多个盘片
- Heads fly on boundary layer of air as disk spins 盘片旋转时，磁头在盘片上方贴近飞行
- Very small head to disk gap 盘片和磁头间隙很小
- Getting more robust 高可靠性

Winchester disk format Seagate 希捷硬盘格式



- 磁道划分为30个扇区，每个扇区总共600个字节
- 每个扇区包含ID域和数据域，扇区前后、ID域和数据域之间均有间隙，间隙总共78个字节
- ID域7个字节，分别是1个字节同步，2个字节磁道号、1个字节磁头号，1个字节扇区号。2个字节校验码
- 数据域515个字节，包括1个同步标识字节，512个数据字节，2个校验码字节

Characteristics of disk 磁盘物理特性

- Fixed (rare) or movable head 固定或可移动磁头
- Removable or fixed 盘片可更换还是固定的
- Single or double (usually) sided 单面还是双面
- Single or multiple platter 单片还是多盘
- Head mechanism 磁头机制
  - Contact (Floppy) 接触式，软盘
  - Fixed gap 固定间隙
  - Flying (Winchester) 飞行模式，温盘

Floppy disk 软盘

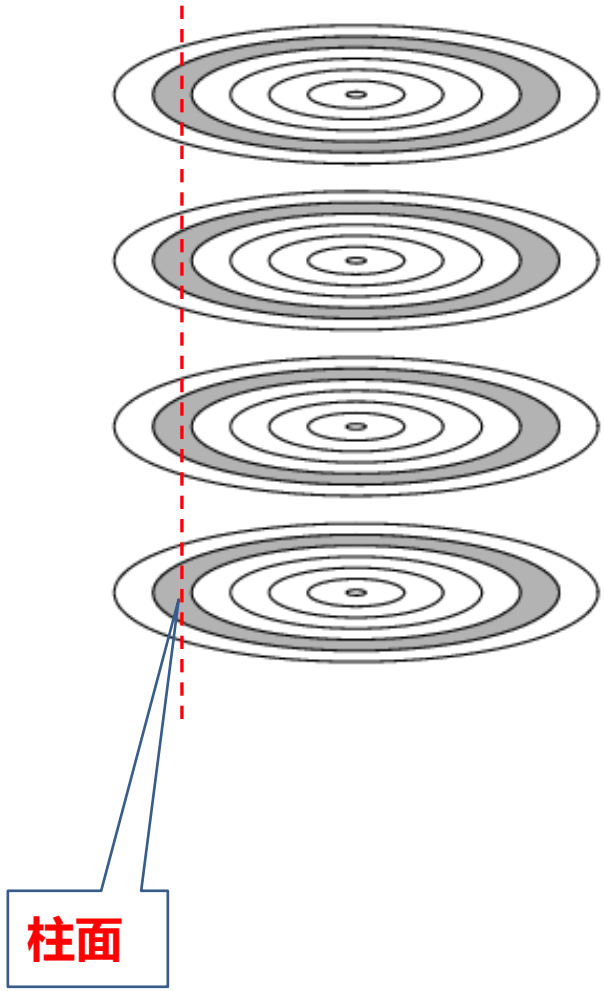
- 8″，5.25″，3.5″
- Small capacity 容量小
  - 5″，600kB or 1.2MB
  - 3.5″，1.44MB（2.88M never popular）
- Slow 慢
- Universal 普及
- Cheap 便宜
- Obsolete 淘汰

Multiple platter 多盘片

- Install multiple discs in vertical direction 垂直方向上安装多个盘片
- One head per side 每个面一个磁头
- Heads are joined and aligned 磁头连接并且对齐
- Aligned tracks on each platter form cylinders 每个盘面上相对应的位置的一组磁道设置为一个柱面
- Data is striped by cylinder 数据按照柱面划分条带
  - Reduces head movement 减少了磁头的移动
  - Increases speed (transfer rate) 提高了传输速度

Cylinders 柱面

- Platter has two heads 每个盘片有2个磁头
- Multiple tracks at the same position is called cylinder 同一位置的磁道称为柱面
- Data is stored according to cylinder rather than sequentially on a certain disk 数据按照柱面存储，而不是在磁盘上依次存储
- Improve the reading and writing speed by reducing the movement of the magnetic head 通过减少磁头的移动来提高读写速度



Typical hard disk drive parameters 典型的硬盘参数

Characteristics	Seagate Barracuda ES.2	Seagate Barracuda 7200.10	Seagate Barracuda 7200.9	Seagate	Hitachi Micro-drive
Application	High-capacity server	High-performance desktop	Entry-level desktop	Laptop	Handheld devices
Capacity	1 TB	750 GB	160 GB	120 GB	8 GB
Minimum track-to-track seek time	0.8 ms	0.3 ms	1.0 ms	—	1.0 ms
Average seek time	8.5 ms	3.6 ms	9.5 ms	12.5 ms	12 ms
Spindle speed	7200 rpm	7200 rpm	7200	5400 rpm	3600 rpm
Average rotational delay	4.16 ms	4.16 ms	4.17 ms	5.6 ms	8.33 ms
Maximum transfer rate	3 GB/s	300 MB/s	300 MB/s	150 MB/s	10 MB/s
Bytes per sector	512	512	512	512	512
Tracks per cylinder (number of platter surfaces)	8	8	2	8	2

应用场景

容量

最小寻道时间

平均寻道时间

转速

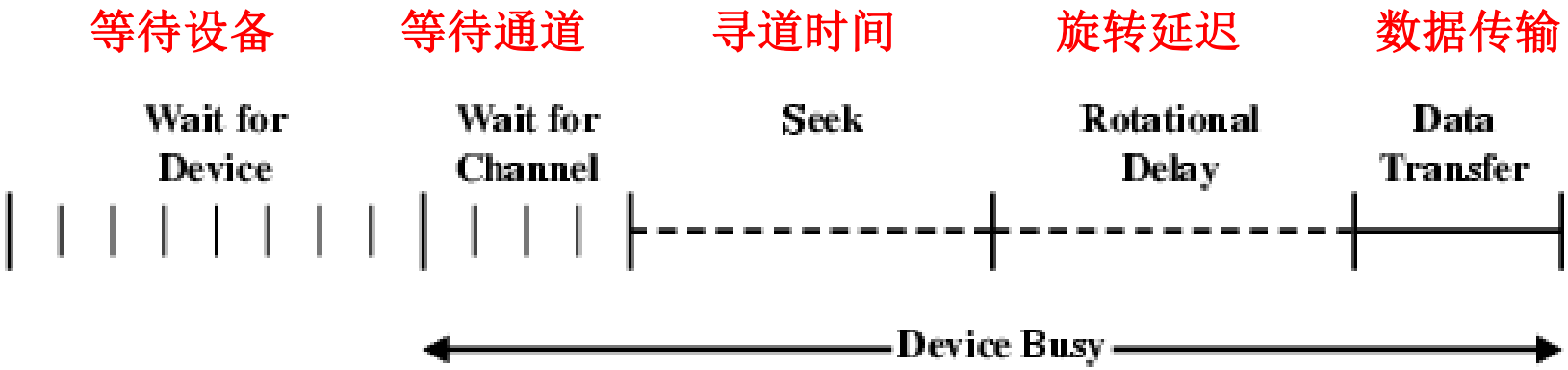
平均旋转延迟

最大传输速率

扇区字节数

柱面磁道数

Timing of disk I/O transfer 磁盘IO传输的时序



- 读写之前，首先要等待设备和通道空闲
- 寻道时间：磁头从当前位置移动到数据所在的磁道的时间
- 旋转延迟：数据所在的扇区旋转到磁头可以读写的位置
- 数据传输：数据实际的传送阶段
- 存取时间：寻道时间+旋转延迟

Speed 传输速度

- Seek time 寻道时间
  - Moving head to correct track 磁头移动到正确的磁道
- Rotational latency 旋转延时
  - Waiting for data to rotate under head 数据旋转到磁头下面
- Access time = Seek + Latency 存取时间=寻道时间+旋转延迟
- Transfer time 传送时间
  - Time for data transfer after track positioning 磁道定位后进行数据实际传送的时间

Transfer time 传送时间

- Transfer time - T 传送时间T
  - T = b/rN
  - b = number of bytes to be transferred 需要传送的字节数
  - N = number of bytes on a track 一个磁道上的字节总数
  - r = rotation speed, in revolutions per second 旋转速度
- The total average access time (including transfer time) 总的平均访问时间，包括传输时间
  - Ta = Ts + 1/2r + b/rN 总的时间=寻道时间+平均旋转延时+传送时间
  - Ts = seek time 寻道时间
  - 1/2r is the average Rotational latency 平均旋转延时

Example

- Suppose we have a such disk
  - rotate speed：15000 rpm 旋转速度为15000转每分钟
  - 500 sectors per track 每个磁道有500个扇区
  - 512 byte per sector 每个扇区512个字节
  - Average seek time ：4ms 平均寻道时间是4ms
- If the file size is 1.28M, how long does it take in total？如果文件大小是1.25M，总共需要多长时间？
  - Files are stored in sector and track order 文件按照扇区和磁道顺序存储
  - File contents are completely stored randomly 文件内容完全随机存储

问题：完全随机存储会带来什么问题?

Example

- Ta = Ts + 1/2r + b/rN 总耗时=寻道时间+平均旋转延时+传送时间
- Ts=4ms
- r=15000rpm, then one revolution need 1/15000 minutes =4ms

  旋转一圈，需要4ms
- Time consumption of a sector 一个扇区的耗时

  Ta = Ts + 1/2r + b/rN

  =4ms+2ms+4ms*1/500=6.008ms
- 1.25M/512B=2500 sectors
- Total time=6.008ms*2500=15.02s **Right ?????**

Example

- If each sector is independent 各个扇区都是独立的
- In fact, files are generally stored in order. Therefore, it will save a lot of time 实际上，文件一般是按顺序存储。所以，中间会省去很多时间
- A file consisting 2500 sectors – 6 tracks occupied 文件包含2500个扇区，占5个磁道
- These five tracks are generally continuous 这5个磁道一般都是连续的
- After the first seek, there is no need to seek again 一次寻道之后，后面就不需要再次寻道时间了
- Similarly, the sectors of the file in the track are also continuous 同理，文件在磁道中的扇区也是连续的
- After one rotation delay, no rotation delay is required later 一次旋转延迟后，后面也不需要旋转延迟

Example

- Average seek time of 4 ms 平均寻道时间4ms
- Each track needs 4ms 每个道旋转一圈需要4ms
- The total time for the first track transfer: 第一个道的总传输时间
  - 4ms + 2ms + 4ms = 10 ms 需要10ms
- Subsequent tracks do not need seek time, but only need an average rotation delay to locate the head to the first data sector
  后续的磁道不需要寻道时间，只需要一个平均旋转延迟，定位磁头到第一个数据扇区
  - 2ms + 4ms=6ms
- Total time 总的时间
  - = 10 ms + 4 x (2+4) ms = 34 ms 首个磁道时间+4个其他磁道时间=34ms

15.02s

Conclusion

- If the data is stored on the disk completely randomly 如果数据在磁盘上完全随机
  - The read of each sector requires seek time + rotation delay + transmission 每个扇区的读取都需要寻道时间+旋转延迟+传输时间
  - Most of the time is spent in seek track and sector, which is very time-consuming 大部分时间都耗在寻道+寻扇区，非常耗时
- Therefore, data is generally stored in adjacent tracks and sectors in sequence 数据一般都顺序存放在相邻的磁道和扇区
  - Only one seek time and several sector seeking times are required 只需要一次寻道时间和若干次寻扇区的时间
  - How data is organized is important 数据的组织方式很重要！！

Outline

- Magnetic Disk 磁盘
- RAID RAID技术
- Solid State Drives 固态硬盘
- Optical Memory 光盘
- Magnetic Tape 磁带

RAID 冗余磁盘阵列

- Redundant Array of Independent Disks 冗余独立磁盘阵列
- Redundant Array of Inexpensive Disks 冗余便宜磁盘阵列
- First proposed by the University of California, Berkeley 加州大学伯克利分校提出
- Store data in multiple disks to improve I/O performance and increasing the disk ‘s capacity 数据存储在多个磁盘上，提高I/O性能，同时提高磁盘容量
- Parity information was used to prevent equipment failures 使用校验信息防止设备错误

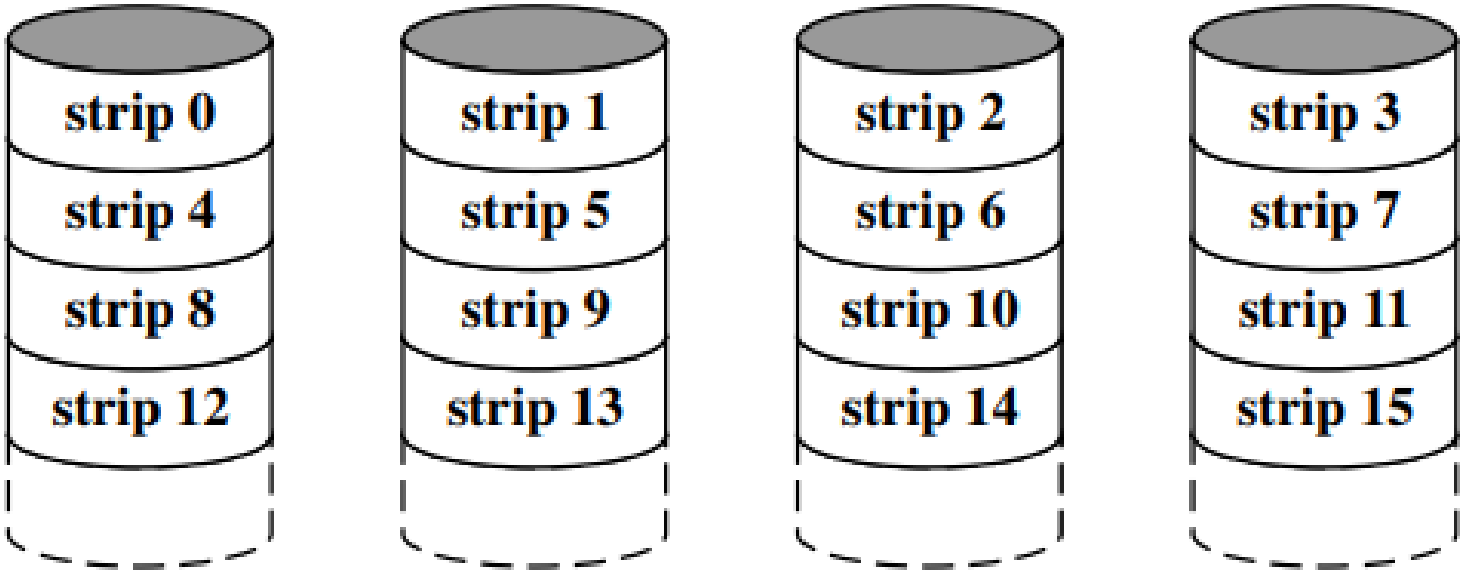
Characteristics 特征

- 7 levels in common use 常用的有7个级别
- Not a hierarchy，but 7 schemes 不是层级结构，7种方案
- common characteristics: 一般特征
  - Set of physical disks viewed as single logical drive by OS 一组物理磁盘，由操作系统驱动形成单一的逻辑硬盘
  - Data distributed across physical drives 数据在不同的物理盘上分布
  - Use redundant capacity to store parity information 使用冗余容量来保存校验信息

RAID0

- Can not be considered as a level in RAID 不能算是RAID的层级
- No redundancy，expanded storage space没有冗余，扩充存储空间
- Disks are striped 磁盘以条带的形式划分
- Data is stored in the adjacent physical disks in strip order 数据按条带顺序保存在相邻的物理盘
- Increase speed 能够提升访问速度
  - Multiple data requests probably not on the same disk 多个数据请求可能不在同一个磁盘
  - Disks seek in parallel 磁盘寻道并行化
  - A set of data is likely to be striped across multiple disks 一组数据的条带跨多个磁盘

Strip of RAID0 RAID0的条带分布

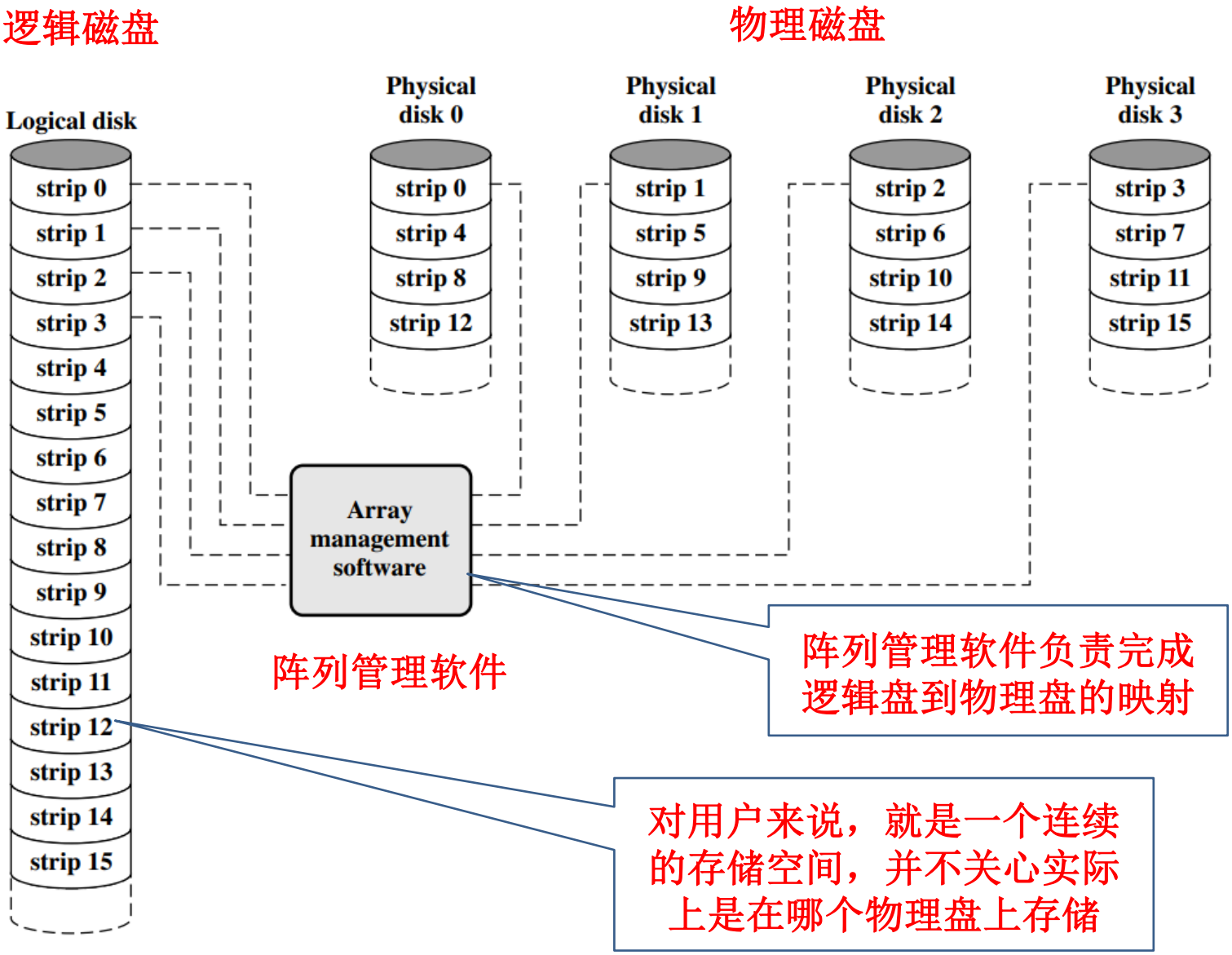


(a) RAID 0 (Nonredundant)

没有冗余

- 条带可以是物理的块、扇区或其他单位
- 条带没有冗余，不能保证数据的可靠性
- 扩充了存储空间
- 可以在一定程度上提高I/O性能

Data mapping for RAID0 RAID0的数据映射



Advantages and disadvantages of RAID0

- Advantages 优点
  - I/O performance is greatly improved by spreading the I/O load across many channels and drives 通过将I/O负载分散到多个通道和驱动器上，I/O性能大大提高
  - Very simple design and easy to implement 设计非常简单，易于实现
- Disadvantages 缺点
  - NOT fault-tolerant: the failure of just one drive will result in all data in an array being lost 没有容错机制，一个驱动器发生故障将导致阵列中的所有数据丢失

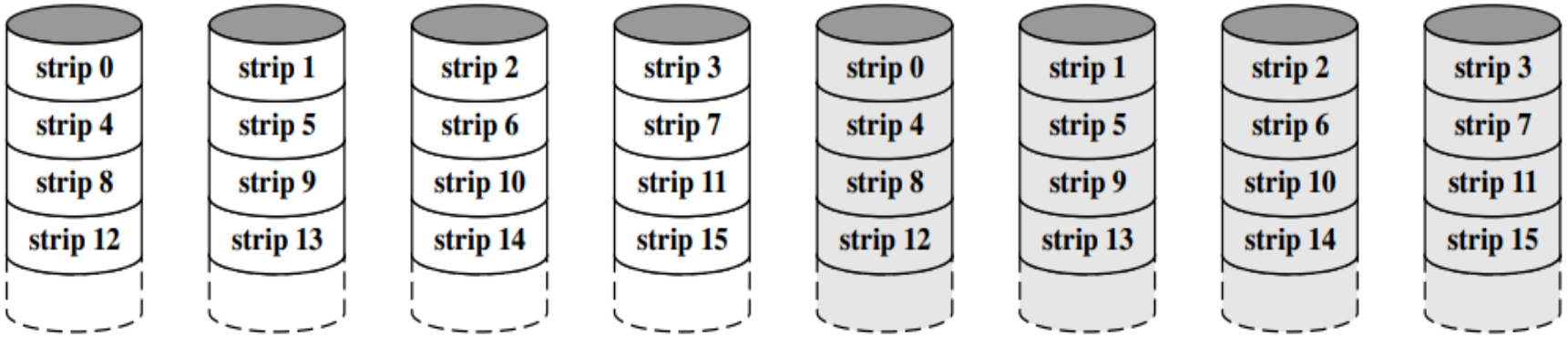
Applications of RAID0 RAID0的应用

- Not every application scenario requires high data reliability 并不是每个应用场景都要求数据的可靠性很高
- In some scenarios，low cost is more important than reliability 低成本比可靠性更重要
- Supercomputers in which 超级计算机
  - Performance and capacity are primary concerns 更关心性能和吞吐量
- Video production and editing 比如视频生产和编辑
  - A few data errors do not affect the overall effect 少量的数据错误并不影响整体效果

RAID1

- Mirrored Disks 镜像磁盘
- Data is striped across disks 数据以条带的形式跨磁盘保存
- 2 copies of each stripe on separate disks 每个条带都有2份，在独立的磁盘上
- Read from either 从任意一个拷贝中读
- Write to both 两个拷贝都要写
- Recovery is simple 恢复简单
  - Swap faulty disk & re-mirror 换掉故障盘，重新镜像
  - No down time 不需要宕机
- High reliability，expensive 高可靠性，价格昂贵

Strip of RAID1 条带分布



(b) RAID 1 (Mirrored)

- 全镜像模式，5~8号磁盘是1~4号的镜像
- 数据以条带的形式存储，每个条带保存在2个独立的磁盘中
- 整体可靠性高
- 读只需要从其中一个镜像读取即可
- 写的时候，需要同时写入2个磁盘

Advantages and disadvantages of RAID1 优缺点

- Advantages 优点
  - 100% redundancy of data: no rebuild is necessary in case of a disk failure, just a copy to the replacement disk 100%的数据冗余，一块磁盘故障后，不需要重构数据，只需要拷贝数据到替换的磁盘
  - Fault recovery is simple. No down time 故障恢复简单，没有宕机时间
  - Simplest implementation 最容易实现
  - May improve read performance 可能提高读的性能
- Disadvantages 缺点
  - Expensive! 贵

Applications of RAID1 应用场景

- Applications： Any applications that request a high reliability

  对于可靠性要求高的应用

  – Accounting 计费

  – Payroll 工资表

  – Financial 金融

  ……

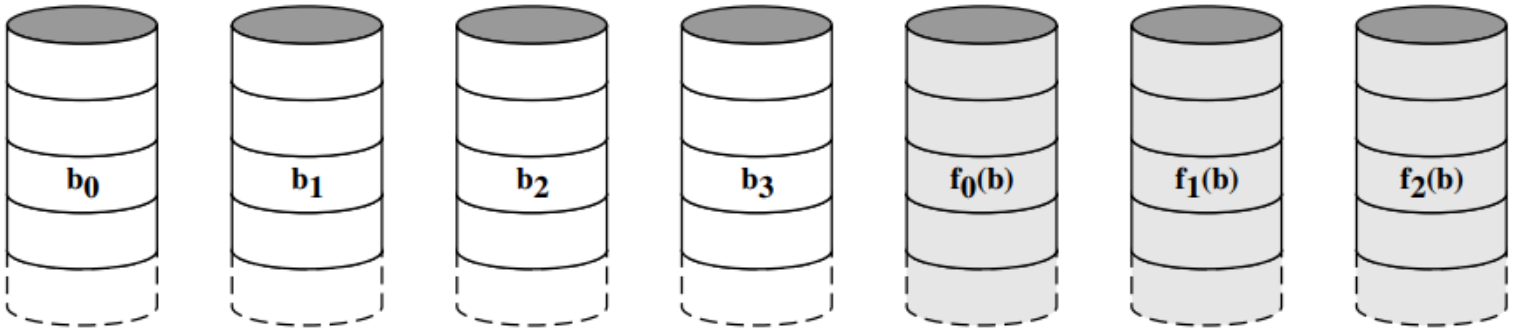
RAID2 -1

- Using the parallel access technology, all disks participate in the execution of each I/O request. All disks are synchronized 采用并行访问技术，所有磁盘都参与I/O操作，磁盘同步运行
- Very small stripes，often single byte/word 小条带，一个字节或一个字
- Calculate error correction code, and then the data and check code are stored on different disks in the disk array 计算纠错码，并将数据和纠错码分别存储在不同的磁盘上
- Hamming code is often used to correct error 汉明码经常用来进行纠错

RAID2 -2

- All disks need to be read at the same time 读取需要所有盘参与
- All disks need to participate in the write, and the error correction code needs to be calculated 写需要所有盘参与，并且需要计算纠错码
- Lots of redundancy 大量的冗余
  - Expensive 昂贵
  - Not used 用的不多

Strip of RAID2 条带分布



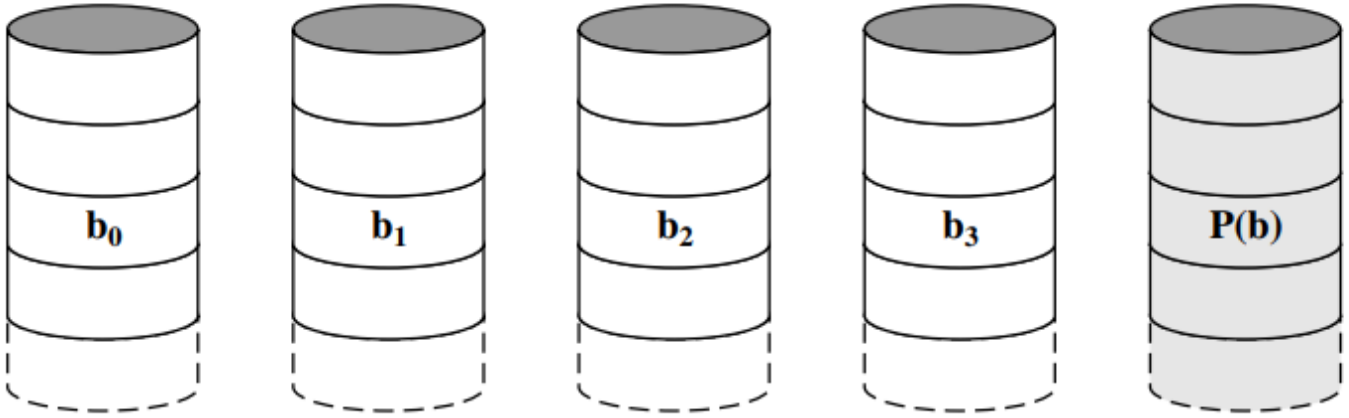
(c) RAID 2 (Redundancy through Hamming code) 通过汉明码冗余

- 4个数据存储盘，3个纠错码存储盘
- 小条带，一个字或一个字节
- 读的时候，b0~b3，f0（b），f1（b），f2（b）都送到阵列控制器，控制器进行纠错计算，并把正确的结果输入。
- 写的时候，需要计算生成纠错码，再写到数据盘和纠错盘中
- 写速度会比较慢

RAID3

- Similar to RAID 2，parallel storage and check bit used 和RAID2类似，使用并行存储和校验位
- Only one redundant disk, no matter how large the array 只有1个冗余盘，无论阵列多大
- Simple parity bit for each set of corresponding bits 简单的在相应的位上做奇偶校验
- Data on failed drive can be reconstructed from surviving data and parity info 损坏磁盘上的数据可以通过其他盘的数据和校验盘的数据进行重构
- Very high transfer rates 高传输速率

Strip of RAID3 条带分布



(d) RAID 3 (Bit-interleaved parity) 位交错奇偶校验

- 小条带，b0~b3 为数据位，P（b）为奇偶校验位
- 数据分布在不同的盘上，可以并行传送，读速度高
- 写操作的时候需要计算校验位，减慢了写操作的性能
- 重构容易
  - X4(i) = X3(i)⊕ X2(i)⊕ X1(i)⊕ X0(i)
  - X1(i) = X4(i)⊕ X3(i)⊕ X2(i)⊕ X0(i)

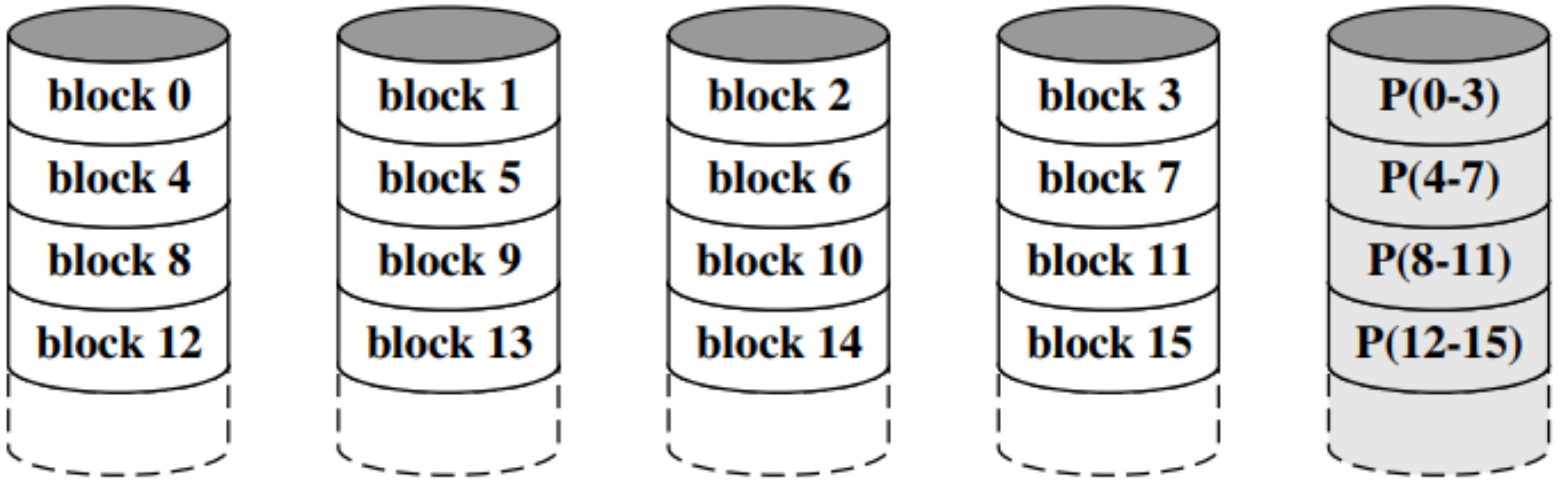
RAID4

- Each disk operates independently 每个盘独立操作
- Good for high I/O request rate 适合高IO请求
- Large stripes 大的条带
- Bit by bit parity calculated across stripes on each disk 跨条带

  逐位计算奇偶校验位

  – X5（i）=X1（i）⊕X2（i）⊕X3（i）⊕X4（i）
- Parity stored on parity disk 校验位保存在校验盘上
- Update parity when writing data 写入时需要重新计算校验码

Strip of RAID4 条带分布



(e) RAID 4 (Block-level parity)

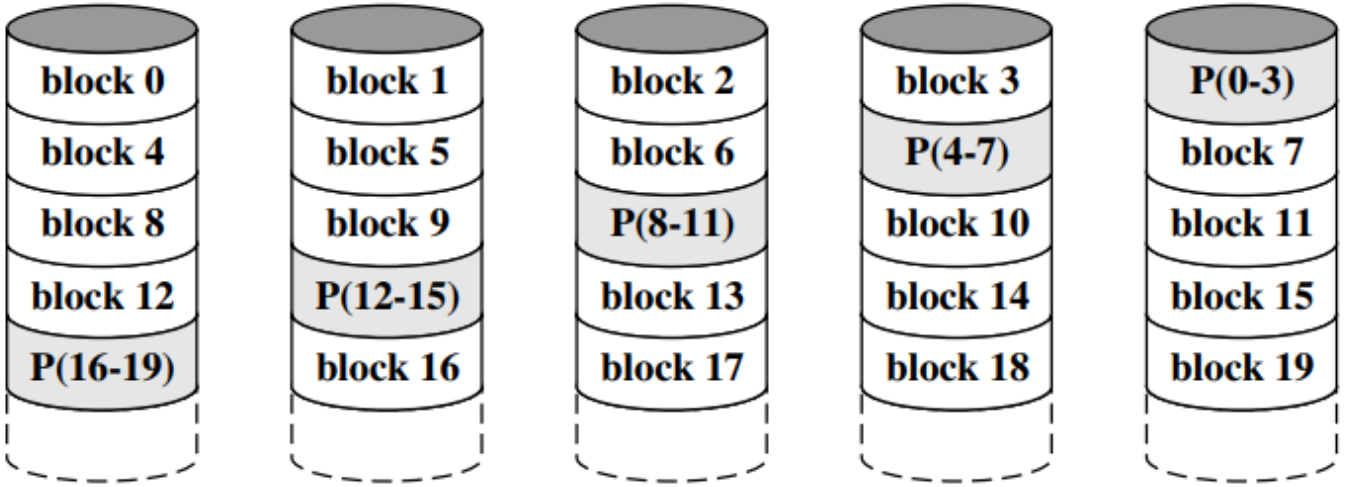
块级别校验

- 1~4号盘为数据盘，5号盘为校验盘
- 大条带设计
- 并行读取满足独立IO请求，适合高IO请求的场景
- 每次写操作都会要更新校验盘，校验盘的性能影响整体性能

RAID5

- Like RAID 4 和RAID4类似
- Parity striped across all disks 校验条带在各个盘上
- Round robin allocation for parity stripe 校验条带轮流在各个盘上
- Avoids RAID 4 bottleneck at parity disk 避免RAID4的校验盘的瓶颈
- Commonly used in network servers 通常用在网络服务器上

Strip of RAID5 条带分布



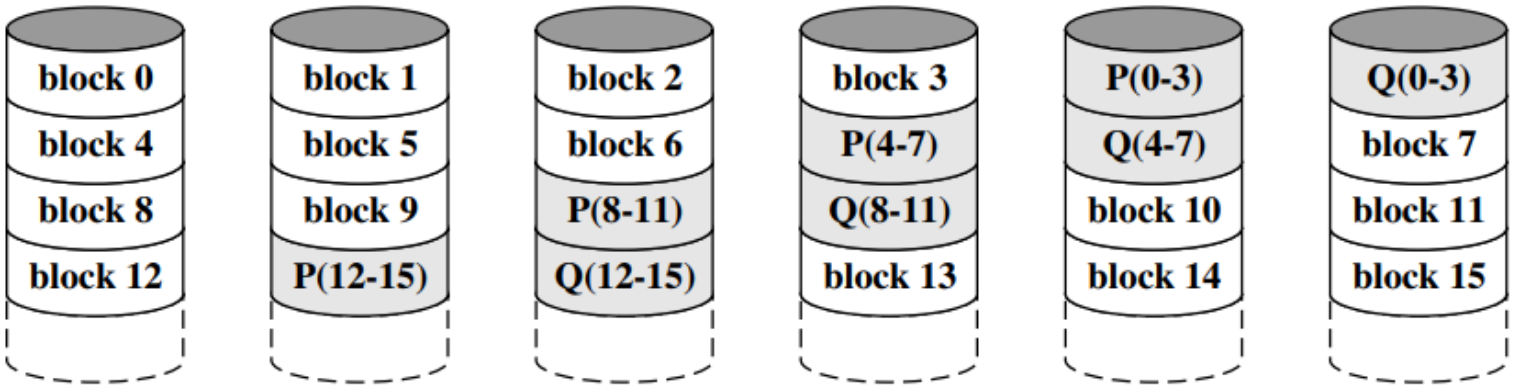
(f) RAID 5 (Block-level distributed parity) 块级别分布式校验

- 大条带设计
- 没有固定的校验盘，校验条带依次分布在各个盘上
- 避免了校验盘的性能瓶颈

RAID6

- For higher fault tolerance 为了更好的容错能力
- Two parity calculations 2个奇偶校验计算
- Stored in separate blocks on different disks 存放在不同磁盘的不同块上
- User requirement of N disks needs N+2 N个数据磁盘总共需要N+2个磁盘
- High data availability 数据高可靠性
  - Three disks need to fail for data loss 数据丢失需要至少坏3块盘
- Significant write penalty 写惩罚很明显
  - Two parity disks need to be updated every time data is written 每次写入需要更新2个校验盘

Strip of RAID6 条带分布



(g) RAID 6 (Dual redundancy) 双冗余

- 2个不同算法得到的校验位，分别存储在不同的磁盘上
- 校验位不固定，轮流存储，避免校验盘的I/O瓶颈
- 提供很高的可用性
- 应用场景：文件和应用服务器，数据库服务器

RAID comparison RAID比较

冗余性 可靠性 性能 价格

RAID	Redundancy	Reliability	Performance	Price
RAID0	N+0	Low	Middle	Low
RAID1	N+N	Highest	Lowest	Highest
RAID2	N+M	High	Middle	High
RAID3	N+1	Middle	Fast	Middle
RAID4	N+1	Middle	Fast	Middle
RAID5	N+1	Middle	Fastest	Middle
RAID6	N+2	High	Middle	High

Outline

- Magnetic Disk 磁盘
- RAID RAID技术
- Solid State Drives 固态硬盘
- Optical Memory 光盘
- Magnetic Tape 磁带

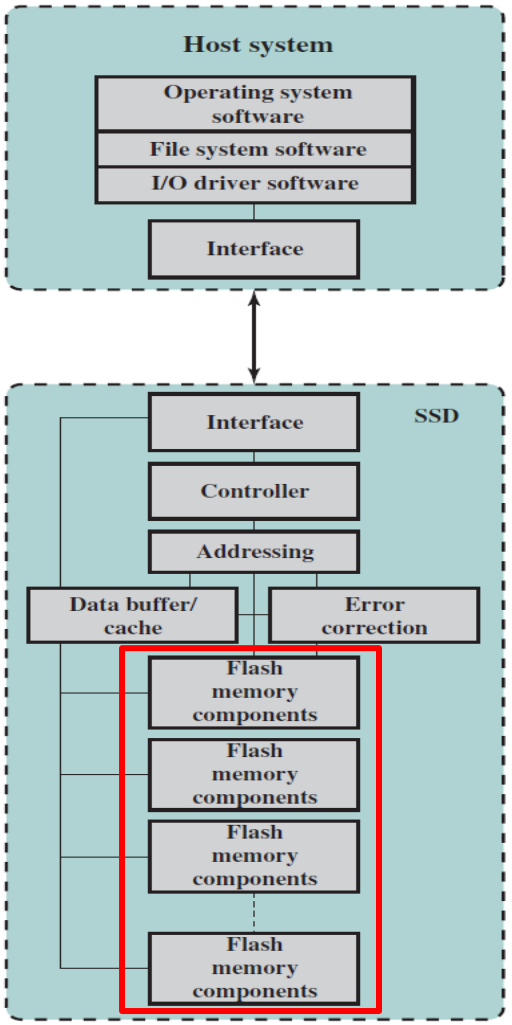
Comparison between SSD and HD SSD和HD比较

	NAND Flash Drives	Seagate Laptop Internal HDD
File copy/write speed	200–550 Mbps	50–120 Mbps
Power draw/battery life	Less power draw, averages 2–3 watts, resulting in 30+ minute battery boost	More power draw, averages 6–7 watts and therefore uses more battery
Storage capacity	Typically not larger than 512 GB for notebook size drives; 1 TB max for desktops	Typically around 500 GB and 2 TB max for notebook size drives; 4 TB max for desktops
Cost	Approx. $0.50 per GB for a 1-TB drive	Approx. $0.15 per GB for a 4-TB drive

项目	NAND 闪存驱动器	希捷笔记本内部 HDD
文件复制 / 写速度	200 ～ 550 Mbps	50 ～ 120 Mbps
电源 / 电池寿命	功耗更少，平均 2 ～ 3W，电池续航时间为 30 分钟	功耗更多，平均 6 ～ 7W，因此要使用更多电池
存储容量	对笔记本大小的驱动器，一般不超过 512 GB；台式机最大为 1 TB	对笔记本大小的驱动器，一般约为 500 GB，最大为 2 TB；台式机最大为 4 TB
成本	对 1 TB 驱动器来说，每 GB 约为 0.50 美元	对 4 TB 驱动器来说，每 GB 约为 0.15 美元

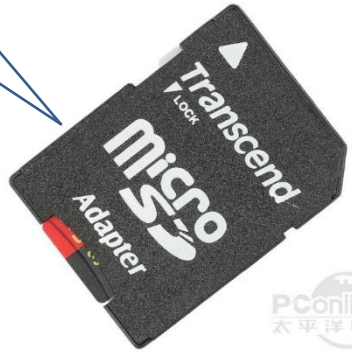
Solid state drives 固态硬盘

- **A solid state drive** is a memory device made with solid state components that can be used as a replacement to a hard disk drive 是一种由固态存储芯片制成的存储器，替代硬盘
- A type of EEPROM 属于一种电可擦除可编程ROM
- Include interface module, controller, address module, data buffer, error correction module and storage module 包括接口模块、控制器、地址模块、数据缓冲、错误校验模块和存储模块



Example of SSD　SSD的几个举例

CF卡、SD卡、MiniSD、MicroSD



固态移动硬盘

U盘

电脑中的固态硬盘

Types of SSD -1 SSD的类型1

- Flash Based SSD 基于Flash的SSD
  - Flash chip as storage medium, often called SSD Flash芯片为存储介质，通常称为SSD
  - Movable 可移动
  - No need power supply 不需要供电
  - Commonly used in notebook, micro hard disk, memory card, U disk, etc. 常用于笔记本，硬盘，记忆卡，U盘等
  - Long life and high reliability 长寿命，高可靠
    - SLC P/E more than 10000 times SLC的PE次数可达10000以上
    - MLC can reach more than 3000 times MLC寿命可达3000次以上
    - TLC can also reach about 1000 times TLC寿命可达1000次以上
    - QLC can also ensure 300 times QLC能达到300次

Types of SSD -2 SSD的类型2

- DRAM Based SSD 基于DRAM的SSD
  - DRAM as storage medium, application range is narrow 使用DRAM作为存储介质，应用受限
  - High performance, theoretically unlimited write 高性能，无限写
  - Independent power supply is needed 需要独立电源供电来保护数据
- 3D Xpoint SSD 基于3D Xpoint的SSD
  - Based on 3D xpoint, close to DRAM 基于3D Xpoint，和DRAM类似
  - Nonvolatile storage 非易失性的
  - Low read delay ,% of the existing SSD 读延迟很小，是SSD的1%
  - unlimited storage life 寿命长
  - Density is relatively low and cost is very high 存储密度低，成本高

Development history -1 SSD的发展历史1

- In 1970, sun Storage Tek developed the first solid state hard disk drive. 1970年，SUN storage tek开发了首个固态硬盘驱动器
- In 1984, Toshiba invented flash memory. 1984年，东芝发明了闪存盘
- In 1989, The world ‘s first solid state drive appeared. 1989年，首个固态硬盘出现
- In March 2006, Samsung first released a 32GB SSD laptop 2006年，三星发布了第一个32GB的SSD笔记本
- In 2007, SanDisk released 32GB SSD 2007年，SanDisk发布了32GB的SSD
- In June 2007, Toshiba launched its first 120GB SSD laptop 2007年6月，东芝发布了第一款120GB的SSD笔记本

Development history -2 SSD的发展历史2

- In September 2008, official release of memoRight SSD， Chinese enter the SSD industry 2008年9月，忆正发布MemoRight，中国公司进入该领域
- In 2009, SSD blowout development, storage virtualization officially entered a new stage. 2009年，SSD爆发式发展，存储虚拟化进入新阶段
- In February 2010, MgO released the world ‘s first SATA 6gbps interface solid state disk 2010年2月，美光发布6Gbps的SSD
- At the end of 2010, Renice launched the world ‘s first high - performance SATA SSD 2010年底，Renice公司发布第一款高性能SSD
- In 2013, Samsung launched VNand 3D flash memory 2013年，三星发布Vnand 3D闪存

Compare of NOR and NAND -1 NOR和NAND的比较1

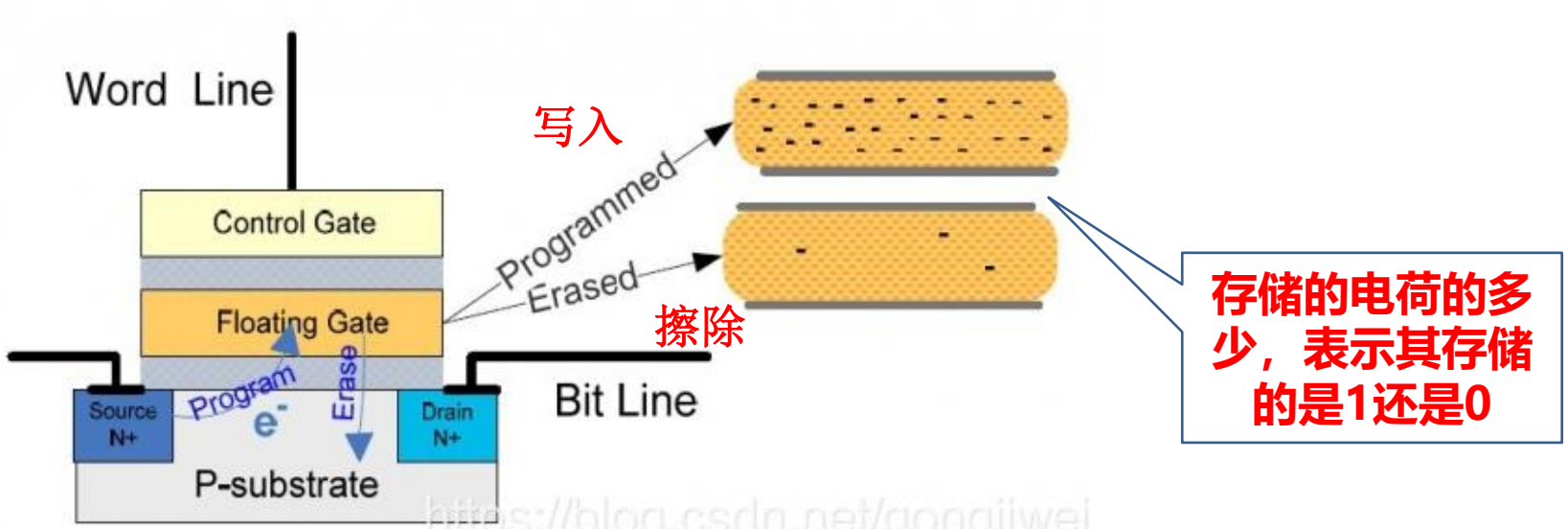
- Flash memory include NOR Flash and NAND FLASH 包括NOR Flash和NAND FLASH两类
- NOR
  - Developed by Intel in 1988 Intel公司1988年开发
  - With dedicated address and data line (similar to SRAM), read and write in byte mode 专用地址和数据线，跟SRAM类似，读写采用字节模式
  - Fast access speed and small storage capacity 高速访问，容量小
  - Data can be read randomly by byte 可以按字节随机读取
  - Program can be executed in the NOR 程序可以在NOR内执行
  - Speed of writing and erasing is not fast, which affects its performance 写入和擦除的速度都不快，影响了它的性能
  - Suitable for program storage, such as BIOS 适合程序存储，比如BIOS

Compare of NOR and NAND -2 NOR和NAND的比较2

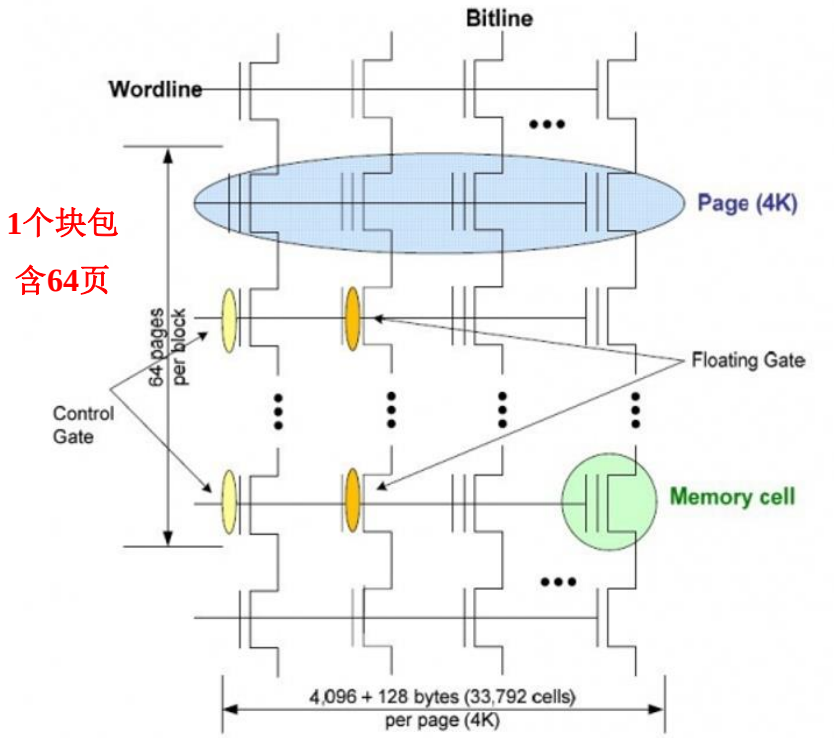
- NAND
  - Reading and writing in blocks 按块读写
  - Slower in reading, but much faster in writing and erasing than NOR
    比NOR读要慢，但写和擦除要快很多
  - Smaller volume and higher storage density than NOR flash memory
    体积小，存储密度大
  - Suitable for storing large amount of data 适合存储大量的数据
  - Including four types: SLC（Single-Level Cell）,MLC（Multi-Level Cell）,TLC（Trinary-Level Cell）,QLC（Quad-Level Cell）包括四种类型：SLC、MLC、TLC和QLC

NAND Flash Storage unit NAMD flash存储单元

Flash的内部存储是金属-氧化层-半导体-场效晶体管 MOSFET，里面有个悬浮门 Floating Gate，是真正存储数据的单元。



NAND Flash Diagram -1 NAND Flash框图1

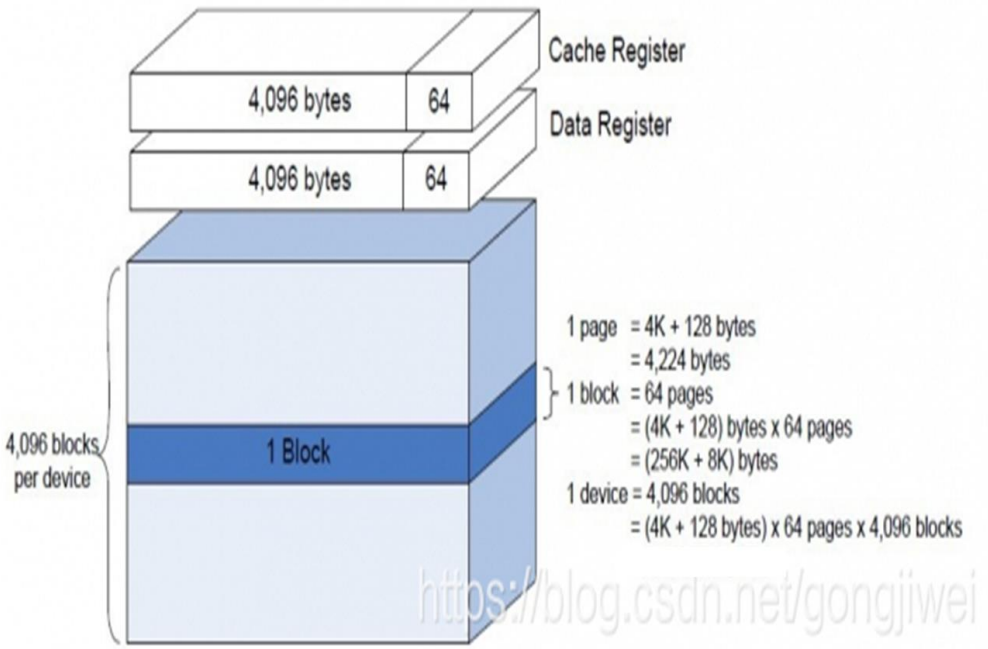


1个块包含64页

1页包含33792个单元

- NAND flash的单层单元，由块、页和存储单元组成
- 页是最小的读/写单位。每页有33,792个存储单元，每页4096Byte + 128Byte
- 块是最小的可擦除单位，1个块由64页组成，256kB+8kB
- 写数据之前需要先擦除然后才能写入

NAND Flash Diagram -2 NAND Flash框图2



- 一个NAND闪存由若干个块堆叠而成
- 4096个块，每个块包含64页，每个页包含4k 字节的可用存储空间，以及128字节的冗余空间。
- SSD的容量为4096*64*4K bytes=1GB。

Advantages of SSD SSD的优点

- Read and write fast. Continuous R/W over 500MB / s, access time less than 0.1ms (mechanical hard disk generally in 12 ~ 14ms) 读写速度快。连续读写超过500MB/s, 访问时间低于0.1ms，机械硬盘一般是12~14ms
- Shockproof and fall Resistant 防震抗摔
- Low power consumption 低功耗
- No noise 没有噪音
- Wide working temperature range, -10 ~ 70 ℃ (mechanical hard disk generally in 5 ~ 55 ℃) 工作温度在-10~70 ℃，机械硬盘一般是5~55 ℃
- Light 重量轻

Disadvantages of SSD SSD的缺点

- Smaller capacity than mechanical hard disk 容量小于机械硬盘
- Life limit 寿命限制
  - Limit the number of erasures. SLC about 10000 times, MLC about 3000 times, TLC about 1000 times, QLC only 300 times 擦写次数有限，SLC只有10000次，MLC 3000次，TLC1000次，QLC只有300次
  - Through the balanced algorithm to manage the storage unit, reduce the unnecessary amount of writing, improve the service life 需要通过平衡算法管理存储单元，减少写的次数，提高寿命
- Expensive than mechanical hard disk 比机械硬盘贵
- Data storage time is limited. SLC is about 10 years, MLC is shorter 数据存储时间有限。SLC大概10年，MLC更短

Outline

- Magnetic Disk 磁盘
- RAID RAID技术
- Solid State Drives 固态硬盘
- Optical Memory 光盘
- Magnetic Tape 磁带

Optical storage CD-ROM 光盘存储CD-ROM

- CD—Compact Disk 压缩光盘
- Invented in 1983 1983年发明
- Originally for audio，later used to store data 最早是用于音频
- 650Mbytes giving over 70 minutes audio，about 15 songs
  650MB的容量，70分钟的音频，大概15首歌
- Polycarbonate coated with highly reflective coat, usually aluminium 聚碳酸酯上涂有高反射膜，如铝
- Data stored as pits 数据用坑来表示
- Read by reflecting laser 通过反射激光读取

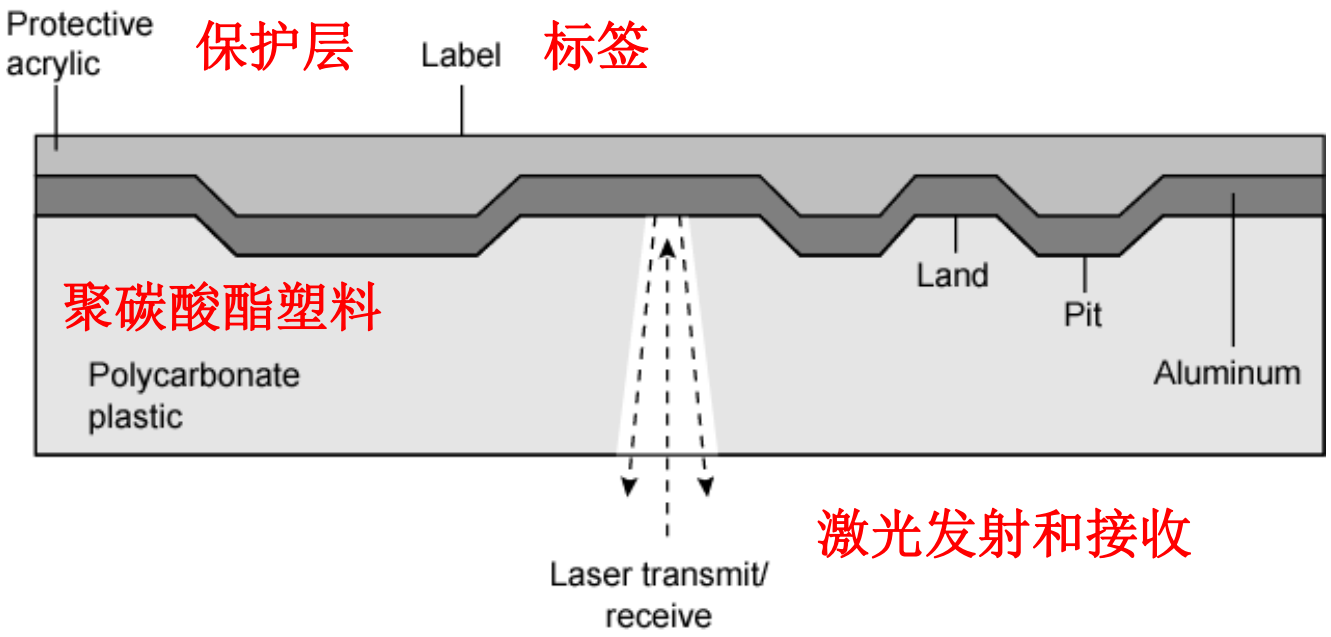
Optical storage CD-ROM 光盘存储CD-ROM

- Surface area of the optical disc is large. In order to store more data, hard disk storage cannot be used 光盘的表面积大，为了存储更多的数据，不能采用硬盘的存储方式
- A spiral that rotates outward from the center to the outermost edge 一条螺旋线，从中心开始往外旋转，一直到最外边
- Length of the innermost and outermost sectors is the same 最内和最外的扇区的长度一样
- Constant packing density，no loss of capacity，恒定数据密度，不损失容量
- Constant linear velocity，variable speed rotation 恒定线速度，变速旋转

CD and its driver CD和CD驱动器

- CD最开始就是音乐光盘的代名词
- 后来在逐渐用到数据存储领域

CD operation



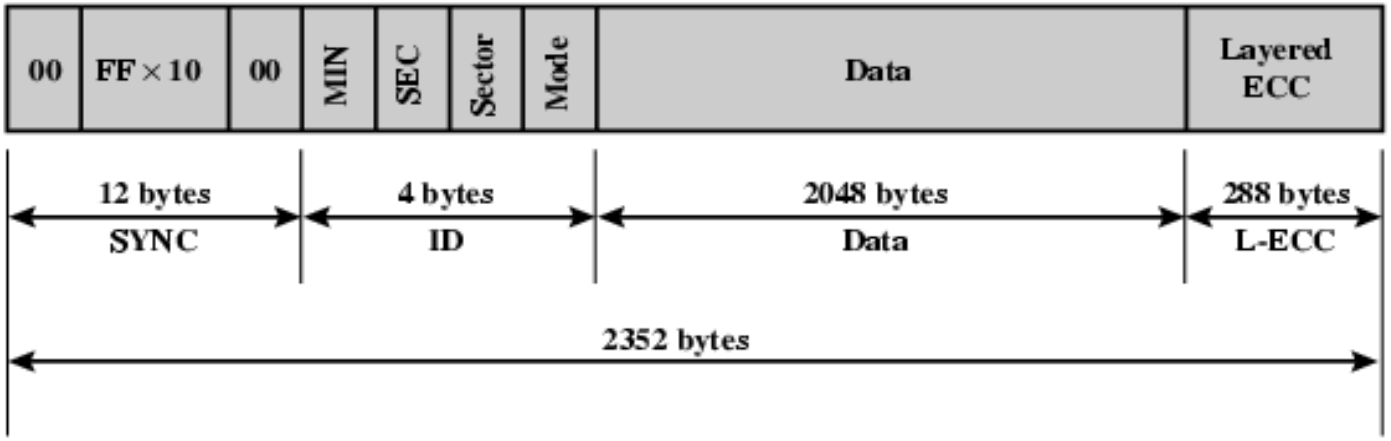
- CD盘的背面是保护层，标签就贴在保护层上
- 正面是聚碳酸材料，其上通过母盘，将有数据的地方压成一个小坑
- 读的时候，激光从下面往上照射，遇到坑的时候，激光反射就不均匀；而遇到没有坑的台，激光反射很均匀
- 反射的光线经过分析后，就可以得到在相应位置上保存的0或者1

CD-ROM Drive Speeds CD-ROM驱动器速度

- Constant linier velocity, not like hard disk 恒定线速度
- Audio is single speed 音频是单速
  - 1.2 m/s
  - Track (spiral) is 5.27km long 磁道总共是5.27公里长
  - Gives 4391 seconds = 73.2 minutes 大概是73分钟
- For data CD, can be rotated in high multiples for faster data reading 对于数据CD，能够以高倍速旋转以提高读取速度
  - e.g. 24x
  - Quoted figure is maximum drive can achieve 标注的数字是驱动器最大能达到的速度

CD-ROM format CD-ROM数据格式

扇区 模式



- 数据按照块（扇区）的方式组织在光盘上
- 块的开头是12个字节的同步标识，采用了特殊的字符，开头和结尾的1个字节全为0，中间10个字节全为1
- 同步标识后面是4个字节的块头，包含块（扇区）地址和模式。模式0表示是空的，没有数据。模式1表示后面是2048个字节的数据和288位的纠错码。模式2表示后面是2336个字节的数据，没有纠错码
- 数据域，是用户的数据。

Random Access on CD-ROM CD-ROM的随机访问

- Difficult 困难
- Move head to rough position 移动光驱头到指定的位置
- Set correct speed 设置合适的速度
- Read address 读取地址
- Adjust to required location 微调光驱头到扇区

Advantages & Disadvantages of CD-ROM 优缺点

- Large capacity（once） 大容量
- Easy to mass produce 容易大批量生产
- Removable 可移动
- Robust 健壮
- Expensive for small runs 小规模时价格贵
- Read only 只读
- Slow 慢

Other Optical Storage 其他光学存储

- CD-Recordable (CD-R) 可刻录CD
  - Disc is coated with a dye layer that can be activated by laser光盘上涂了一层染色层，可以通过激光激活
  - Write only once 只能写一次
  - Now affordable 现在不贵
  - Applicable to document archiving 适用于文档的归档
  - Compatible with CD-ROM drives 和CD-ROM兼容
- CD-RW 可重复写CD
  - Erasable 可擦写
  - Getting cheaper 便宜
  - Can be used as a secondary storage device 可以作为辅助存储设备
  - Mostly CD-ROM drive compatible 和大多数CD-ROM兼容

DVD

- DVD invented to increase storage capacity 发明DVD用于提高容量
- Digital Video Disk 数字视频光盘
  - Used to indicate a player for movies 用于表示电影播放器
  - Only plays video disks 只能播放视频光盘
- Digital Versatile Disk 数字多功能光盘
  - Used to indicate a computer driver 用于表示计算机驱动器
  - Will read computer disks and play video disks 能够读计算机光盘，也能播放视频光盘

DVD technology

- Very high capacity (4.7G per layer) 高容量，4.7G每层
- Full length movie on single disk 单张盘上可以放完整的电影
  - Using MPEG compression 使用MPEG压缩
  - Finally standardized (honest!) 标准化
  - Movies carry regional coding 电影包含地区代码
  - Players only play correct region films 只能播放对应区域的电影
- Multi-layer，double-side 多层，双面
- Content can be fixed 内容可以固定

DVD – writable 可写DVD

- Loads of trouble with standards 缺乏标准化
- First generation DVD drivers may not read first generation DVD-W disks 第一代DVD驱动器不能读第一代DVD-W盘
- First generation DVD drivers may not read CD-RW disks第一代DVD驱动器不能读CD-RW盘
- Wait for it to settle down before buying! 购买之前确定兼容性

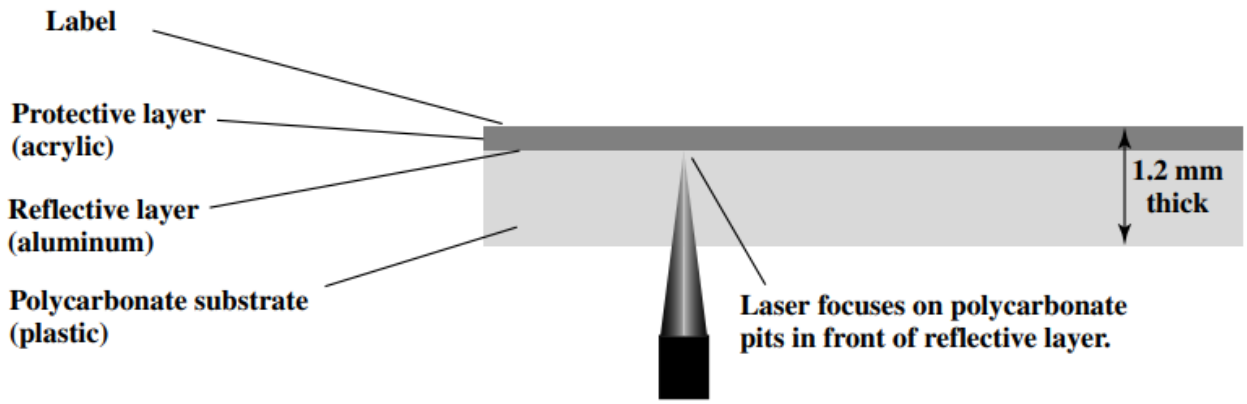
CD and DVD CD和DVD的盘片结构

标签

保护层

发射层

聚碳酸酯衬底



(a) CD-ROM–Capacity 682 MB

聚碳酸酯衬底2

半反射层2

聚碳酸酯层2

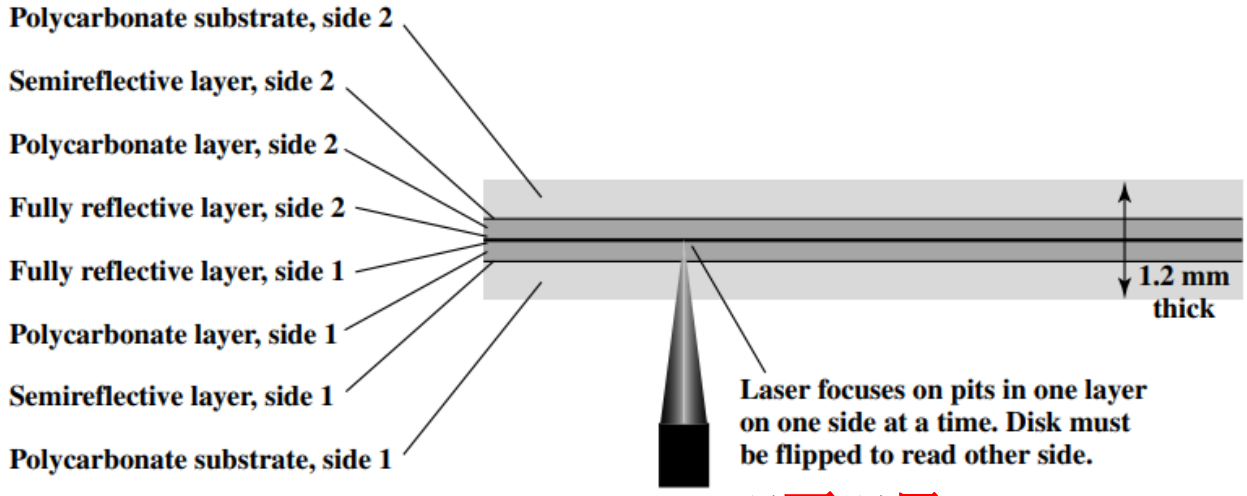
全反射层2

全反射层1

聚碳酸酯层1

半反射层1

聚碳酸酯衬底1



双面双层DVD，17GB

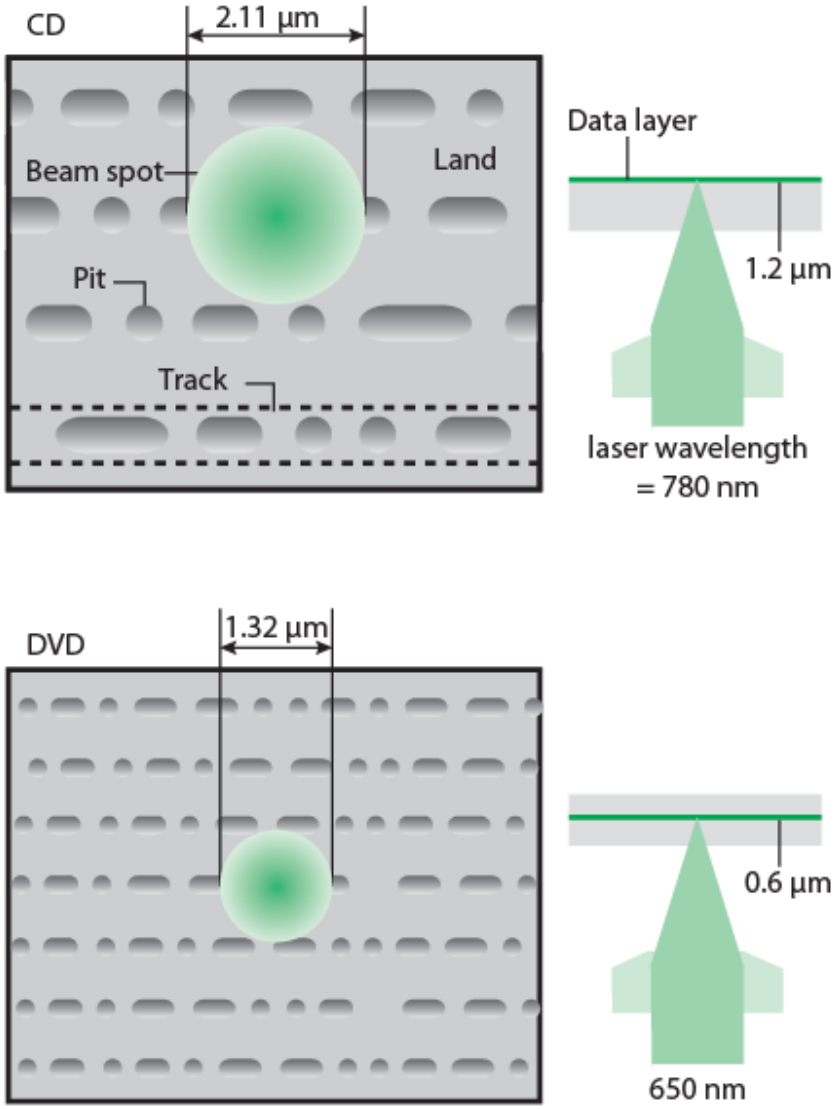
(b) DVD-ROM, double-sided, dual-layer–Capacity 17 GB

读另一面，需要盘片翻面

High Definition Optical Disks 高清晰度光盘

- Designed for high definition videos 为高清视频设计
- Much higher capacity than DVD 比DVD的容量高
  - Shorter wavelength laser , Blue-violet range 更短的波长激光，蓝-紫之间
  - Smaller pits 更小的坑
- HD-DVD
  - 15GB single side single layer 单面单层15GB
- Blue-ray 蓝光
  - Data layer closer to laser, tighter focus, less distortion, smaller pits 数据层和激光头更近，聚焦更好，减少失真，坑更小
  - 25GB on single layer 单层25GB
  - Available read only (BD-ROM), Recordable once (BD-R) and re-recordable (BD-RE) 包括只读，可刻录，可重复刻录三种

Optical Memory Characteristics 光存储器特性

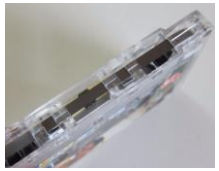


凹坑的距离越来越小，磁道越来越密，激光的波长也越来越短。存储容量也越来越大

Outline

- Magnetic Disk 磁盘
- RAID RAID技术
- Solid State Drives 固态硬盘
- Optical Memory 光盘
- Magnetic Tape 磁带

Magnetic Tape 磁带

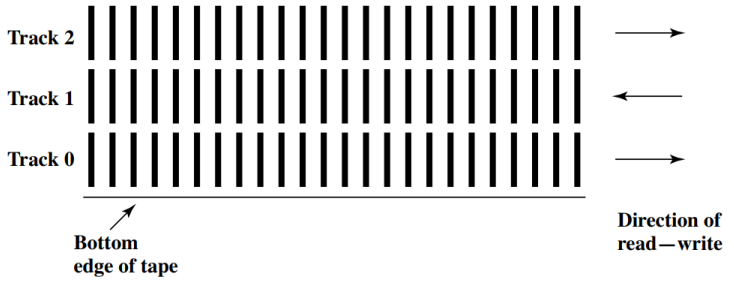





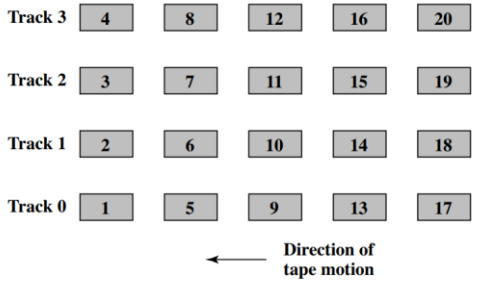
- 存储介质是聚酯薄膜带，上面涂了磁性材料
- 马达带动磁带转动，磁头接触磁带，读取磁带上记录的数据，并转换成电信号
- 磁带在计算机领域也有很广泛的应用，主要用于大批量数据的备份和恢复

Types of magnetic tape 磁带类型

- Two types of recording: 两种类型的磁带
  - Parallel recording 并行记录
  - Serial recording 串行记录
- Earlier tapes used nine tracks – parallel recording 早期磁带使用9个磁道，并行记录
- Modern systems use serial recording, referred to as serpentine recording 现在系统使用串行记录，类似蛇形



(a) Serpentine reading and writing



(b) Block layout for system that reads—writes four tracks simultaneously

Characteristics of tape 磁带的特点

- Serial access 顺序访问
- Slow 比较慢
- Very cheap 便宜
- Backup and archive 备份和恢复
- Linear Tape-Open (LTO) Tape Drives LTO 磁带驱动器
  - Developed late 1990s 1990年发布
  - Open source alternative to proprietary tape systems 开放资源

Linear Tape-Open (LTO) Tape Drives

	LTO-1	LTO-2	LTO-3	LTO-4	LTO-5	LTO-6
Release date	2000	2003	2005	2007	TBA	TBA
Compressed capacity	200 GB	400 GB	800 GB	1600 GB	3.2 TB	6.4 TB
Compressed transfer rate (MB/s)	40	80	160	240	360	540
Linear density (bits/mm)	4880	7398	9638	13300		
Tape tracks	384	512	704	896		
Tape length	609 m	609 m	680 m	820 m		
Tape width (cm)	1.27	1.27	1.27	1.27		
Write elements	8	8	16	16		

Key Terms

Access time	CLV	Floppy disk	Multiple zoned recording	Rotational delay
CD	cylinder	gap	Nonremovable disk	Sector
CD-ROM	DVD	head	Optical memory	Seek time
CD-R	DVD-R	Magnetic disk	platter	Striped data
CD-RW	DVD-RW	Magnetic type	RAID	track
CAV	Fixed-head disk	Movable-head disk	Removable disk	Transfer time

Summary and Question

- 小结
  - 这节课我们对外部存储中的磁盘、RAID等做了详细的分析，并对固态硬盘、光盘等做了简要描述。
- 问题
  - 问题1：磁盘的转动采用什么方式？
  - 问题2：采用RAID 的目的是什么？

Assignments

- Review questions
  - 6.5，6.7，6.11
- Problems
  - 6.2，6.3，6.4

谢谢大家！