数据结构与算法（九）

张铭 主讲

采用教材：张铭，王腾蛟，赵海燕 编写
高等教育出版社，2008. 6（“十一五”国家级规划教材）

http://www.jpk.pku.edu.cn/pkujpk/course/sjjg

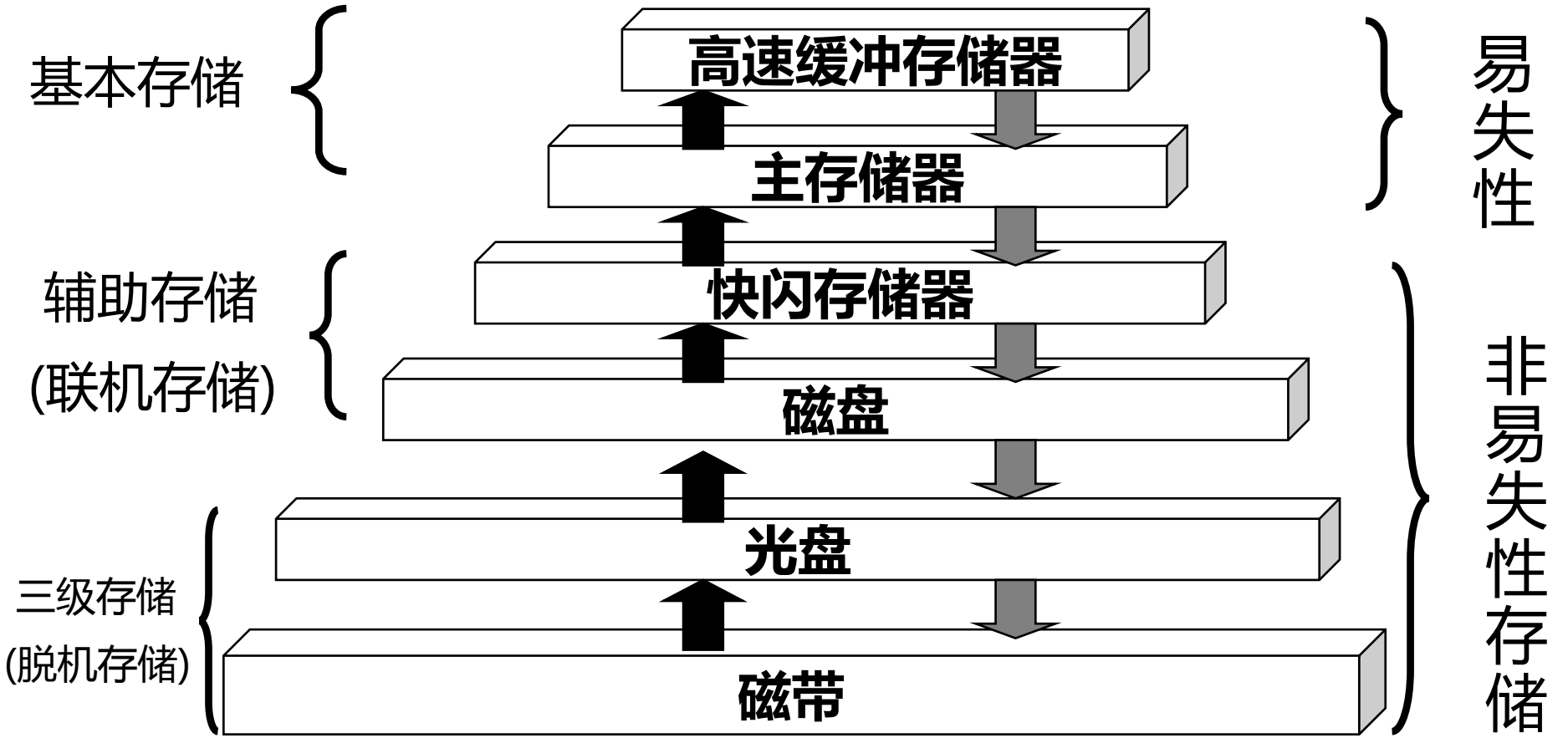
第9章 文件管理和外排序

- 9.1 主存储器和外存储器
- 9.2 文件的组织和管理
  - 9.2.1 文件组织
  - 9.2.2 C++ 的流文件
- 9.3 外排序

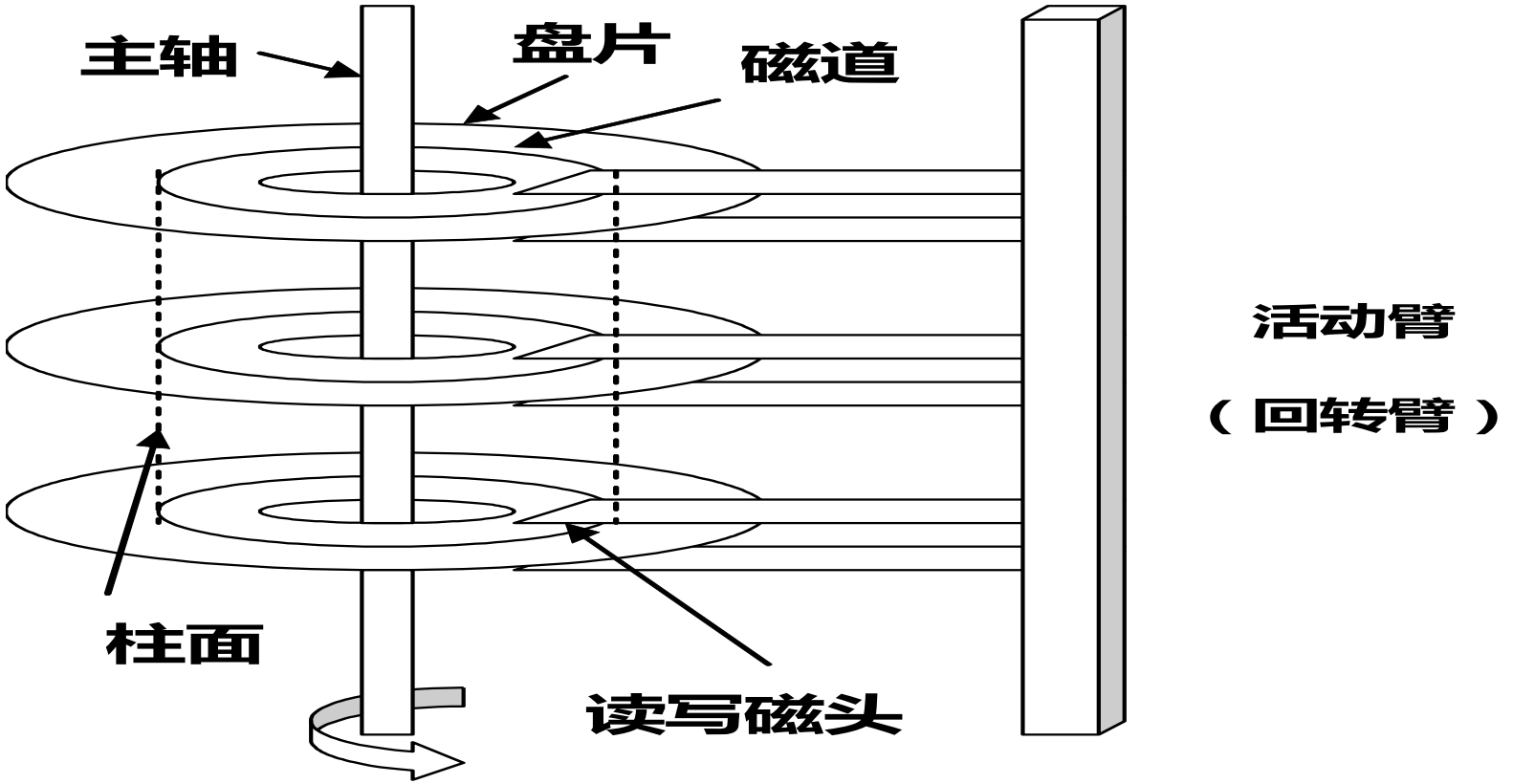
主存储器和外存储器

- 计算机存储器主要有两种：
  - 主存储器（primary memory 或者 main memory ，简称“内存” ，或者“主存”）
    - 随机访问存储器（Random Access Memory, 即 RAM）
    - 高速缓存（cache）
    - 视频存储器（video memory）
  - 外存储器（peripheral storage 或者 secondary storage，简称“外存”）
    - 硬盘 (几百G - 几百T， 10^{12}B)
    - 磁带 (几个P， 10^{15}B)

物理存储介质概览

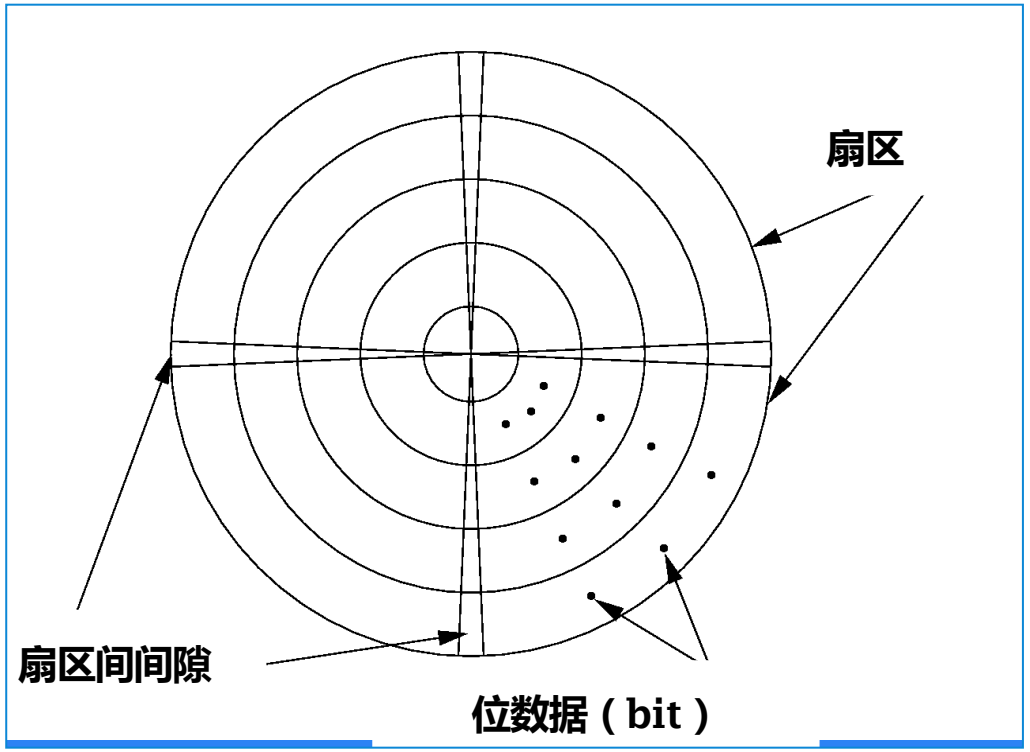

基本存储
辅助存储
(联机存储)
三级存储
(脱机存储)
高速缓冲存储器
主存储器
快闪存储器
磁盘
光盘
磁带
易失性
非易失性存储

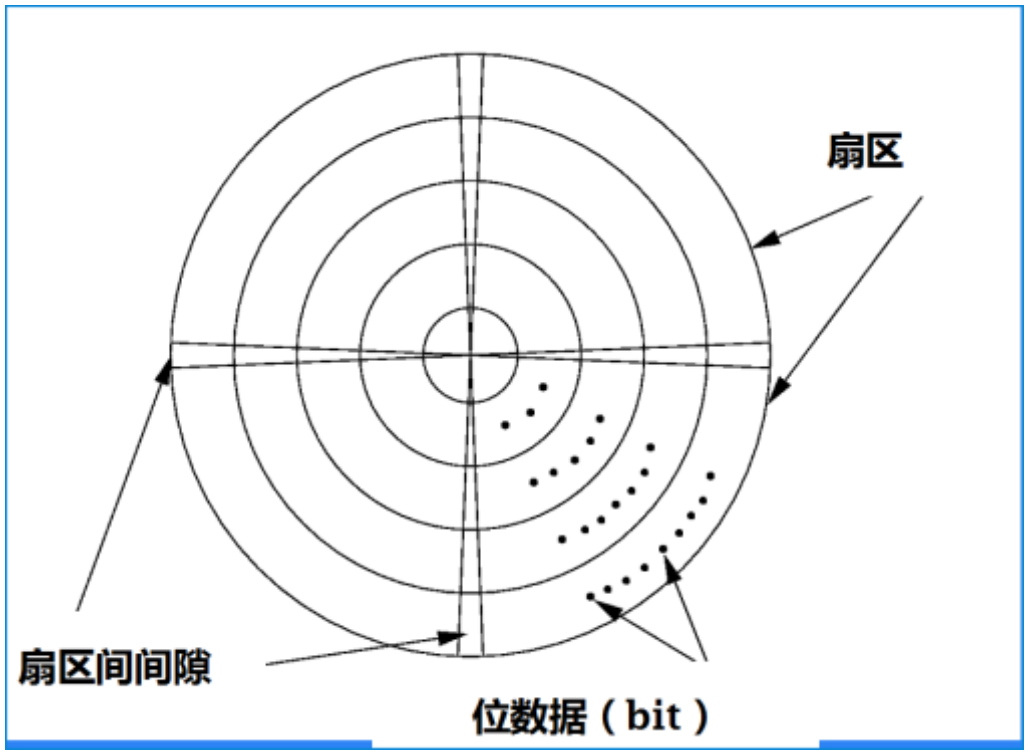
磁盘的物理结构



9.1 主存储器和外存储器

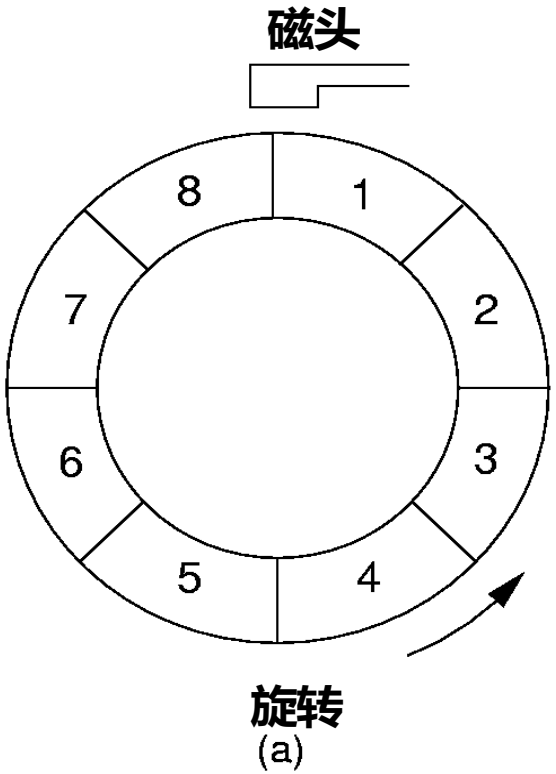
磁盘盘片的组织

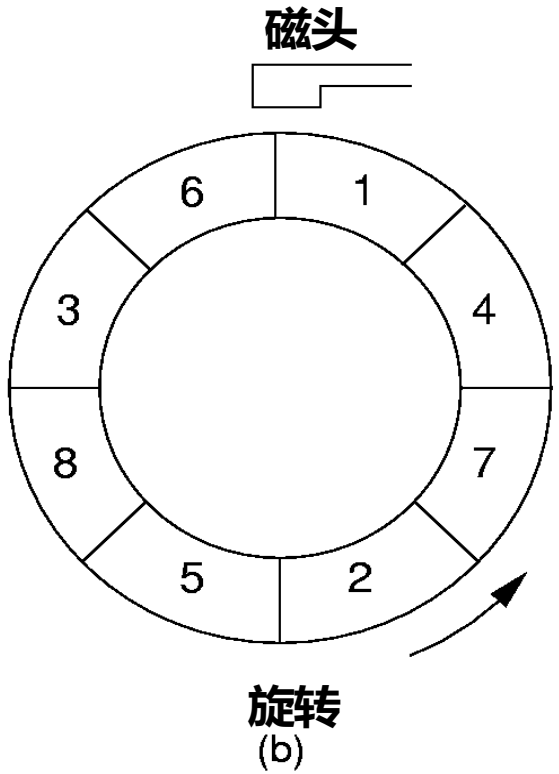




磁盘磁道的组织（交错法）

- 每页 512 字节 或 1024 字节





（a）没有扇区交错；　　（b）以 3 为交错因子

内存的优缺点

- 优点：访问速度快
- 缺点：造价高，存储容量小，断电丢数据
- CPU 直接与主存沟通，对存储在内存地址的数据进行访问时，所需要的时间可以看作是一个很小的常数

外存的优缺点

- 优点：价格低、信息不易失 、便携性
- 缺点：存取速度慢
  - 一般的内存访问存取时间的单位是 **纳秒**（1 纳秒 = 10^{-9} 秒）
  - 外存一次访问时间则以 **毫秒**（1 毫秒 = 10^{-3} 秒）或秒为数量级
- 牵扯到外存的计算机程序应当尽量 **减少外存的访问次数**， 从而减少程序执行的时间

9.1 主存储器和外存储器

- KB (kilo byte)　　10^3B (页块)
- MB (mega byte)　　10^6B (高速缓存)
- GB (giga)　　10^9B (内存、硬盘)
- TB (tera)　　10^{12}B (磁盘阵列)
- PB (peta)　　10^{15}B (磁带库)
- EB = 10^{18}B ; ZB = 10^{21}B ; YB = 10^{24}B
- Googol 是 10 的 100 次方

文件的逻辑结构

- 文件是记录的汇集
  - 一个文件的各个记录按照某种次序排列起来，各纪录间就自然地形成了一种线性关系
- 因而，文件可看成是一种线性结构

文件的组织和管理

- 逻辑文件(logical file)
  - 对高级程序语言的编程人员而言
  - 连续的字节构成记录，记录构成逻辑文件
- 物理文件(physical file)
  - 成块地分布在整个磁盘中
- 文件管理器
  - 操作系统或数据库系统的一部分
    - 文件的记录无结构，数据库文件是结构型记录
  - 把逻辑位置映射为磁盘中具体的物理位置

文件组织

- 文件逻辑组织有三种形式：
  - 顺序结构的定长记录
  - 顺序结构的变长记录
  - 按关键码存取的记录
- 常见的物理组织结构：
  - 顺序结构——顺序文件
  - 计算寻址结构——散列文件
  - 带索引的结构——带索引文件
    - 倒排是一种特殊的索引

文件上的操作

- 检索：在文件中寻找满足一定条件的记录
- 修改：是指对记录中某些数据值进行修改。若对关键码值进行修改，这相当于删除加插入
- 插入：向文件中增加一个新记录
- 删除：从文件中删去一个记录
- 排序：对指定好的数据项，按其值的大小把文件中的记录排成序列，较常用的是按关键码值的排序

C++ 的标准输入输出流类

- 标准输入输出流类
  - istream 是通用输入流和其它输入流的基类，支持输入
  - ostream 是通用输出流和其它输出流的基类，支持输出
  - iostream 是通用输入输出流和其它输入输出流的基类，支持输入输出
- 3个用于文件操作的文件类
  - ifstream 类，从 istream 类派生，支持从磁盘文件的输入
  - ofstream 类，从 ostream 类派生，支持向磁盘文件的输出
  - fstream 类，从 iostream 类派生，支持对磁盘文件的输入和输出

fstream类的主要成员函数

文件指针 **定位**；在当前文件指针位置 **读取**；向当前文件指针位置 **写入**

```
#include <fstream.h>                          // fstream = ifstream + ofstream
void fstream::open(char*name, openmode mode);
                                              // 打开文件
fstream::read(char*ptr, int numbytes);  // 从文件当前位置读入字节
fstream::write(char*ptr, int numbtyes); // 向文件当前位置写入字节
                              // seekg和seekp：在文件中移动当前位置
                              // 以便在文件中的任何位置读出或写入字节
fstream::seekg(int pos);                      // 输入时用于设置读取位置
fstream::seekg(int pos, ios::curr);
fstream::seekp(int pos);                      // 设置输出时的写入位置
fstream::seekp(int pos, ios::end);
void fstream::close();                        // 处理结束后关闭文件
```

缓冲区和缓冲池

- 目的：减少磁盘访问次数的
- 方法：缓冲（buffering）或缓存（caching）
  - 在内存中保留尽可能多的块
  - 可以增加待访问的块已经在内存中的机会
- 存储在一个缓冲区中的信息经常称为一页（page），往往是一次 I/O 的量
- 缓冲区合起来称为缓冲池（buffer pool）

替换缓冲区块的策略

- 新的页块申请缓冲区时，把最近最不可能被再次引用的缓冲区释放来存放新页
  - “先进先出”（FIFO）
  - “最不频繁使用”（LFU）
  - “最近最少使用”（LRU）

思考

1. 查询内存、硬盘、磁带、高速缓存等设备每字节的价格
2. 查询当前主流硬盘的性能指标
   - 容量（G）
   - 磁盘旋转速度（rpm）
   - 交错因子
   - 寻道时间
   - 旋转延迟时间

数据结构与算法

谢谢聆听

国家精品课“数据结构与算法”
http://www.jpk.pku.edu.cn/pkujpk/course/sjjg/

张铭，王腾蛟，赵海燕
高等教育出版社，2008. 6。“十一五”国家级规划教材