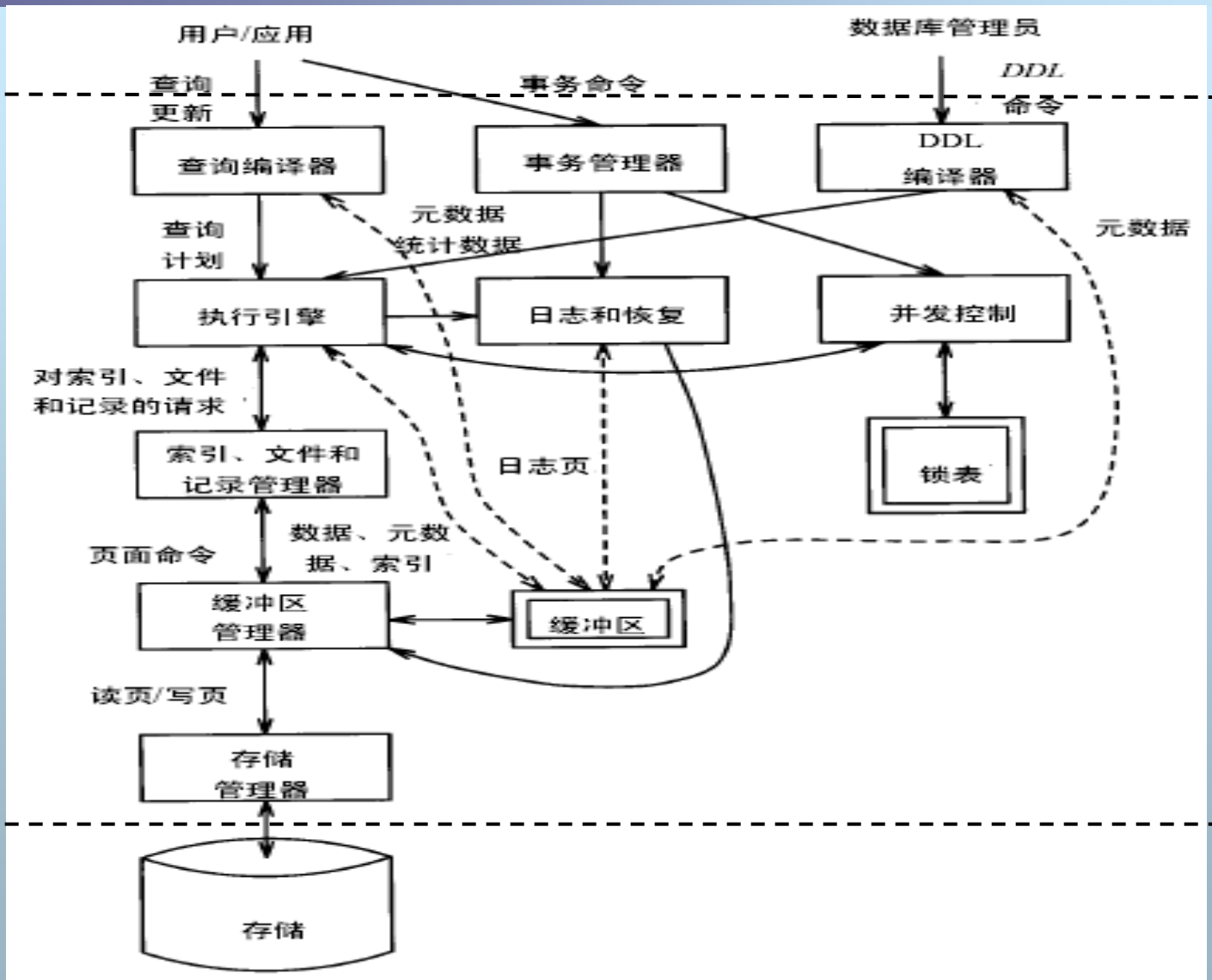
Data Storage

金培权（jpq@ustc.edu.cn）

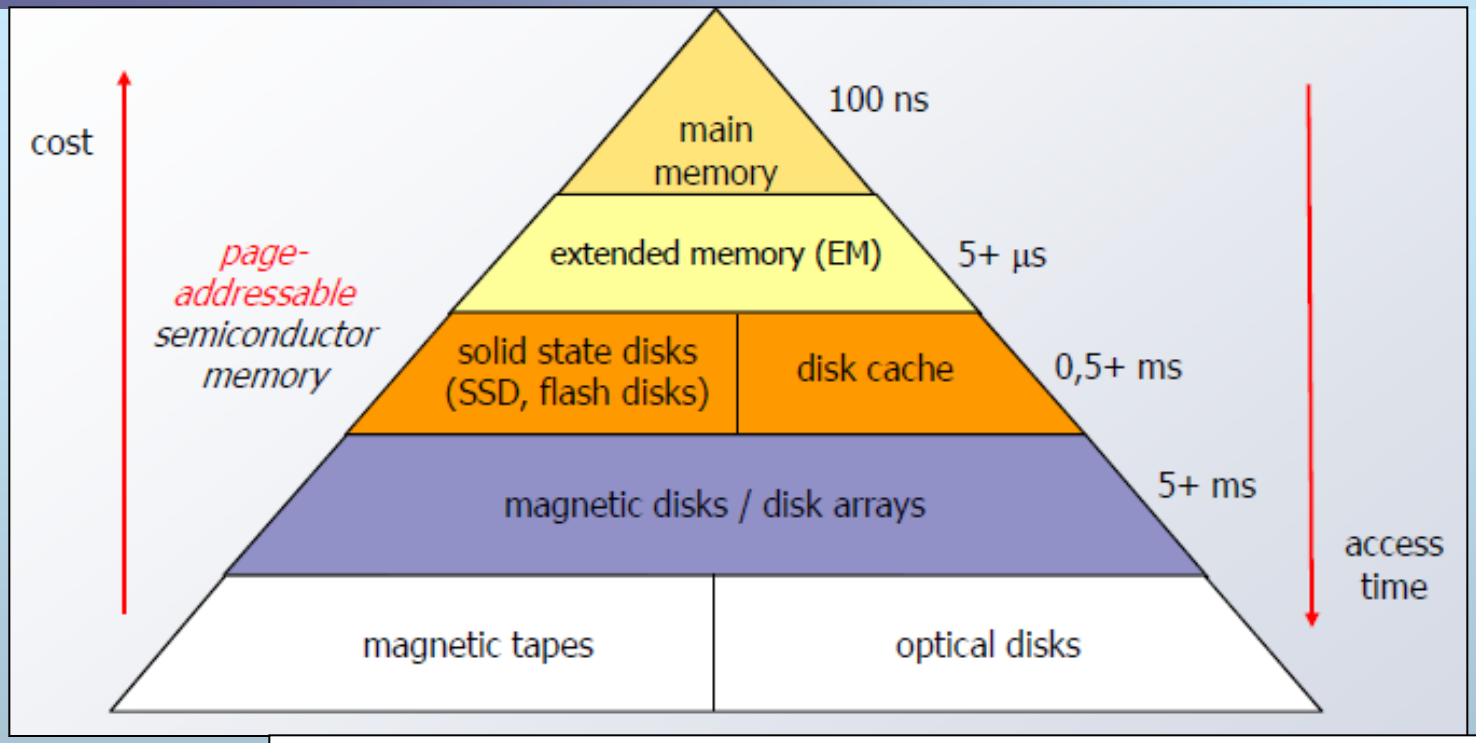
DBMS一般架构

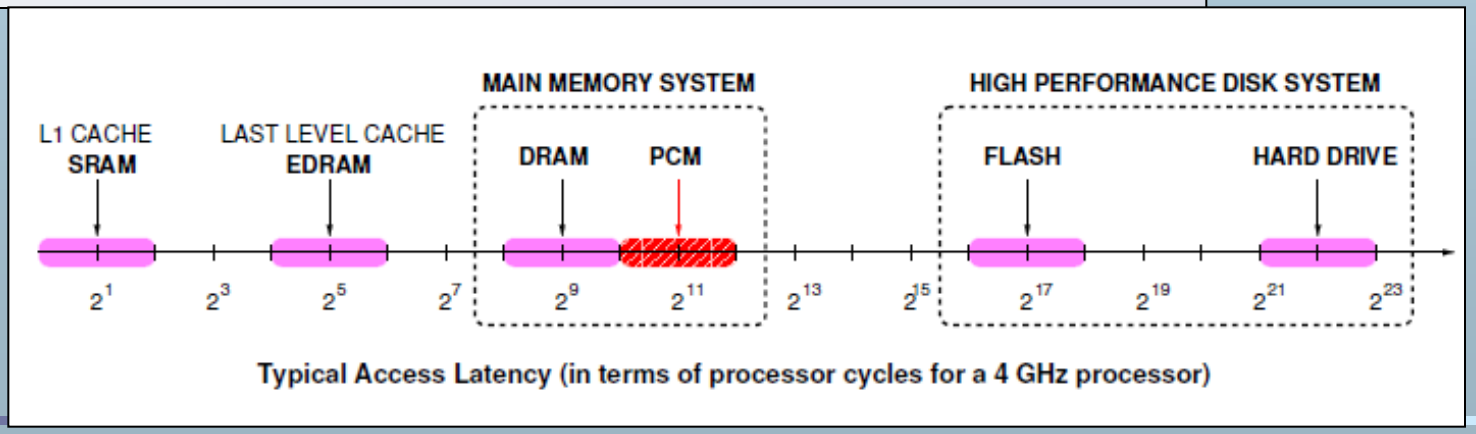

用户/应用
数据库管理员
查询
更新
事务命令
DDL
命令
查询编译器
事务管理器
DDL
编译器
元数据
统计数据
元数据
查询
计划
执行引擎
日志和恢复
并发控制
对索引、文件
和记录的请求
索引、文件和
记录管理器
日志页
锁表
页面命令
数据、元数
据、索引
缓冲区
管理器
缓冲区
读页/写页
存储
管理器
存储
用户
DBMS
数据库

主要内容

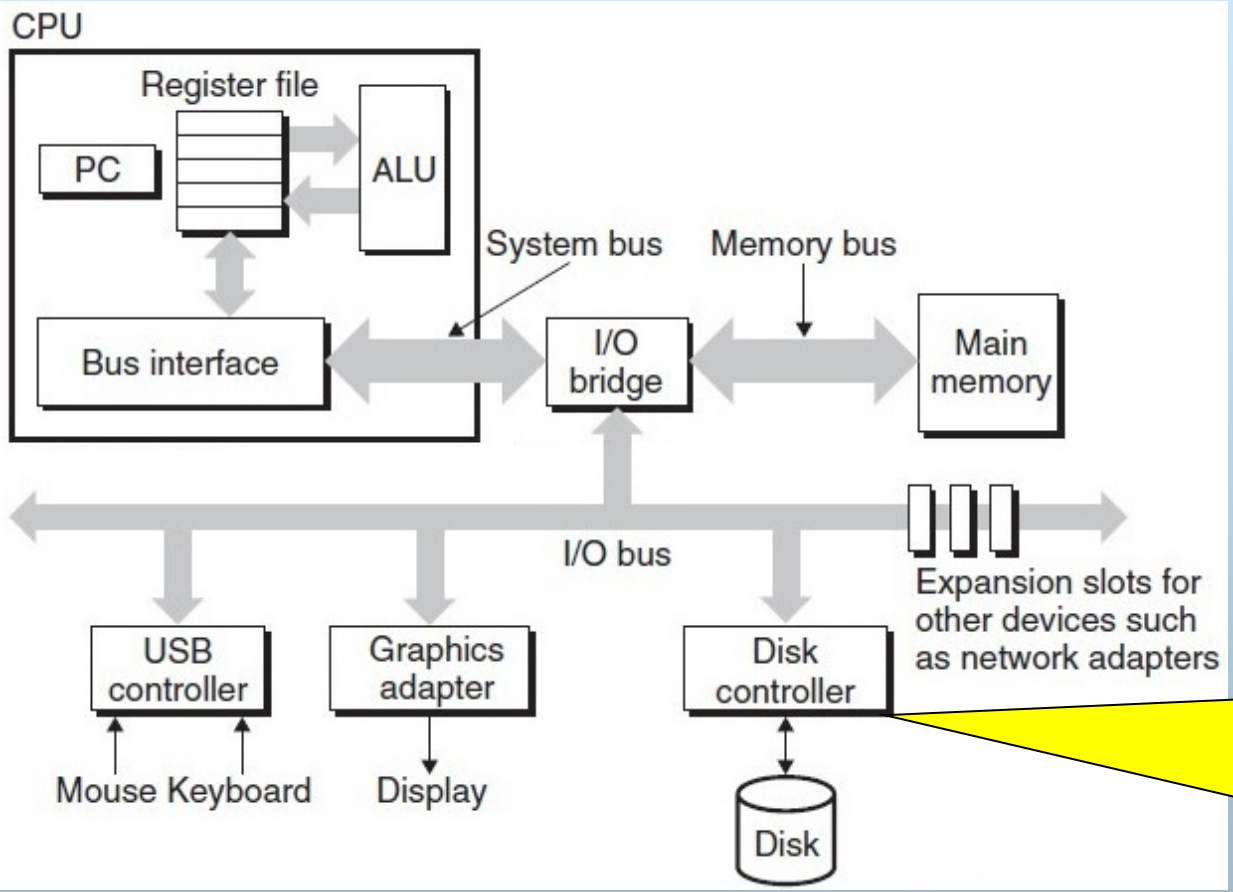
- 存储器结构（Disk Structure）
- 磁盘块存取时间（Block Access Time）
- 磁盘例子: Megatron747
- 磁盘存取优化（Optimization）
- 新型存储

一、存储器结构


cost
page-addressable semiconductor memory
main memory
100 ns
extended memory (EM)
5+ µs
solid state disks (SSD, flash disks)
disk cache
0,5+ ms
magnetic disks / disk arrays
5+ ms
magnetic tapes
optical disks
access time

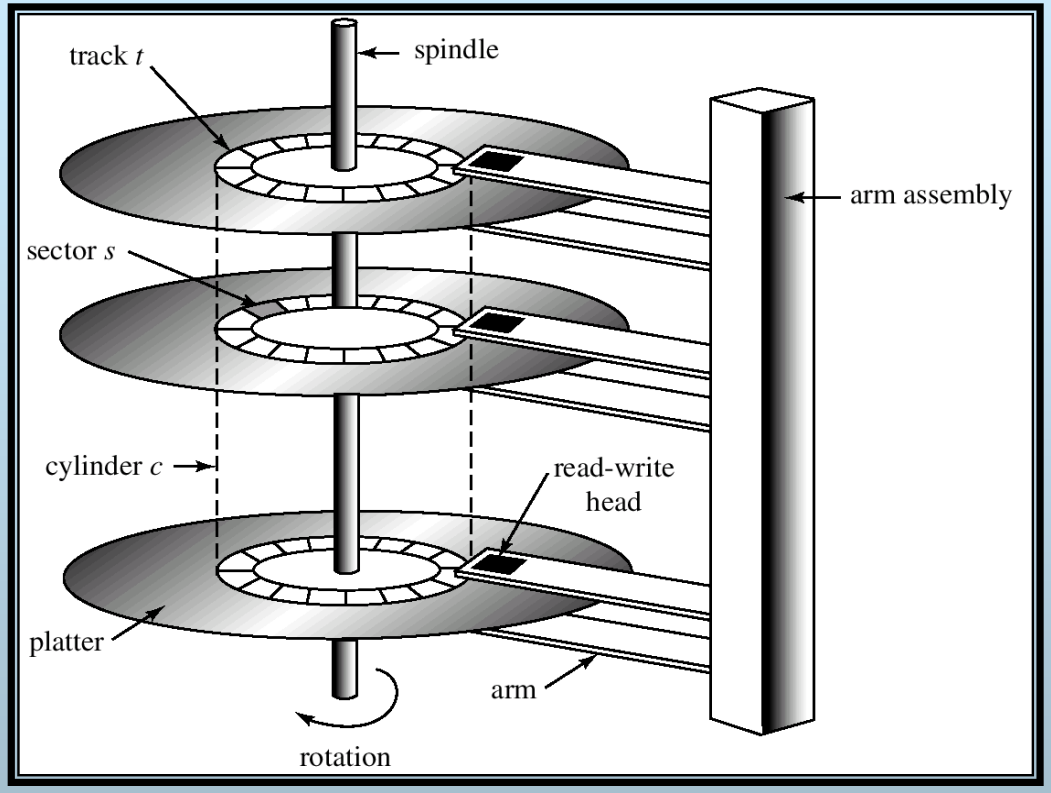

MAIN MEMORY SYSTEM
HIGH PERFORMANCE DISK SYSTEM
L1 CACHE
SRAM
LAST LEVEL CACHE
EDRAM
DRAM
PCM
FLASH
HARD DRIVE
2^1
2^3
2^5
2^7
2^9
2^11
2^13
2^15
2^17
2^19
2^21
2^23
Typical Access Latency (in terms of processor cycles for a 4 GHz processor)

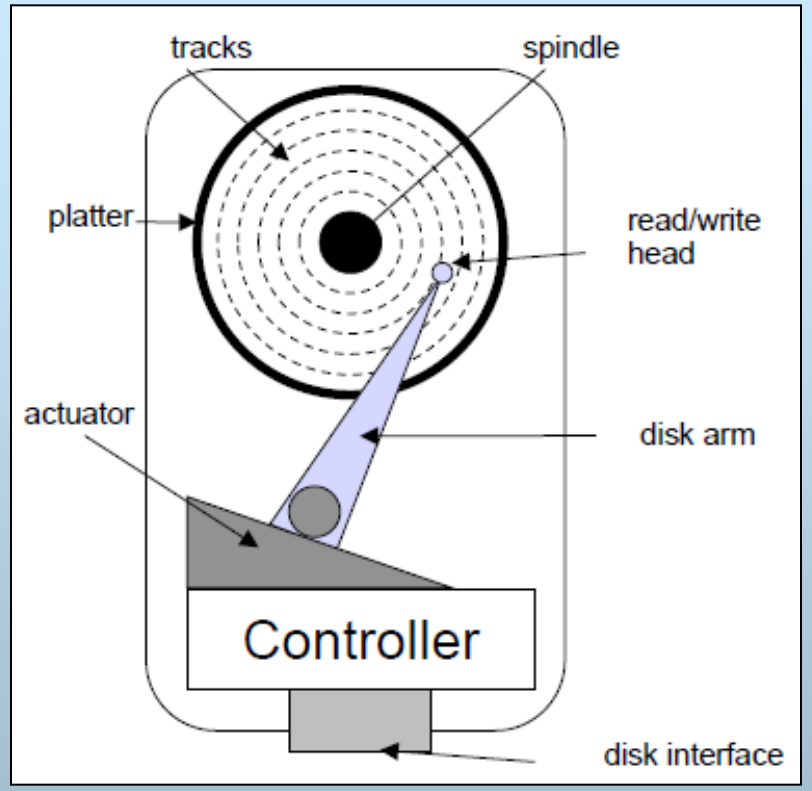
1、计算机系统结构



磁盘驱动器与计算机的接口。控制磁盘臂实现对磁盘（扇区）的读写

2、典型磁盘结构

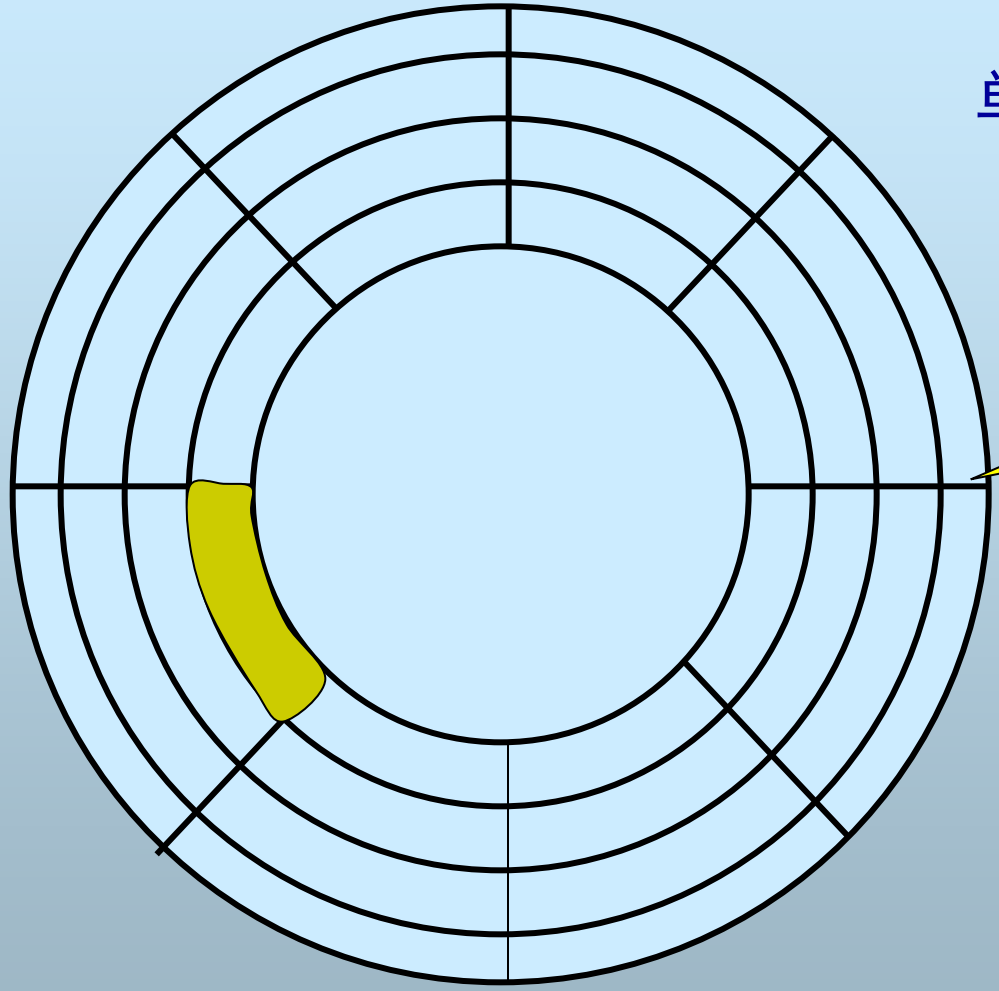




几个术语

- 盘片platter, 盘面 surface, 磁头 R/W head, 磁道 track, 柱面 cylinder, 扇区 sector

2、典型磁盘结构



单个盘片俯视图

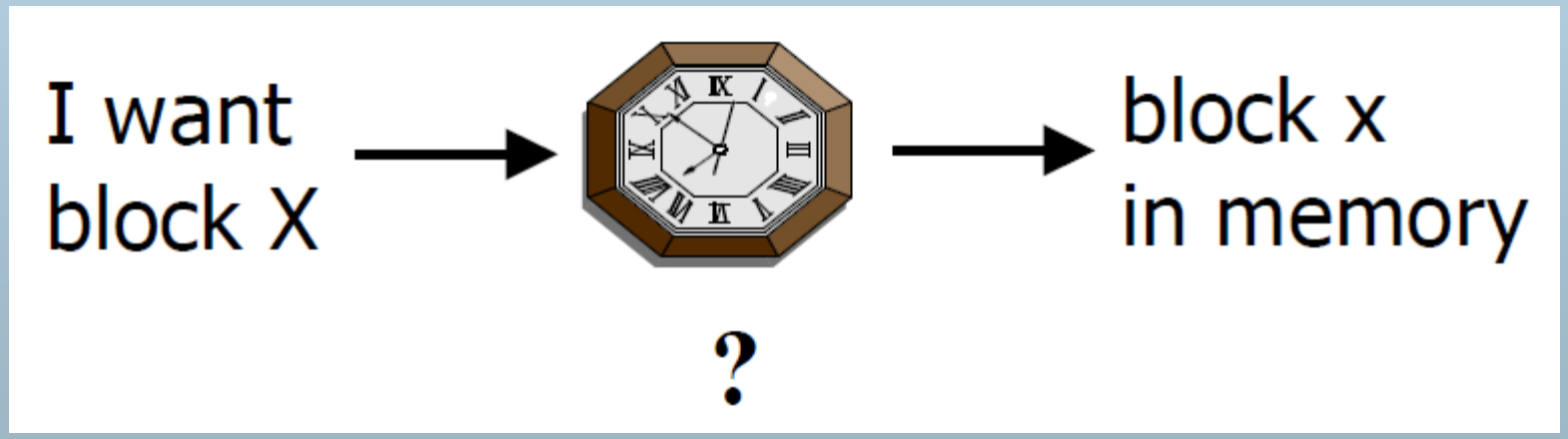
扇区间隙

二、磁盘块存取时间

- 块（Block）
  - OS或DBMS进行磁盘数据存取的最小逻辑单元，由若干连续扇区构成
  - 块是DBMS中数据存取的最小单元
  - 扇区是磁盘中数据存储的最小单元

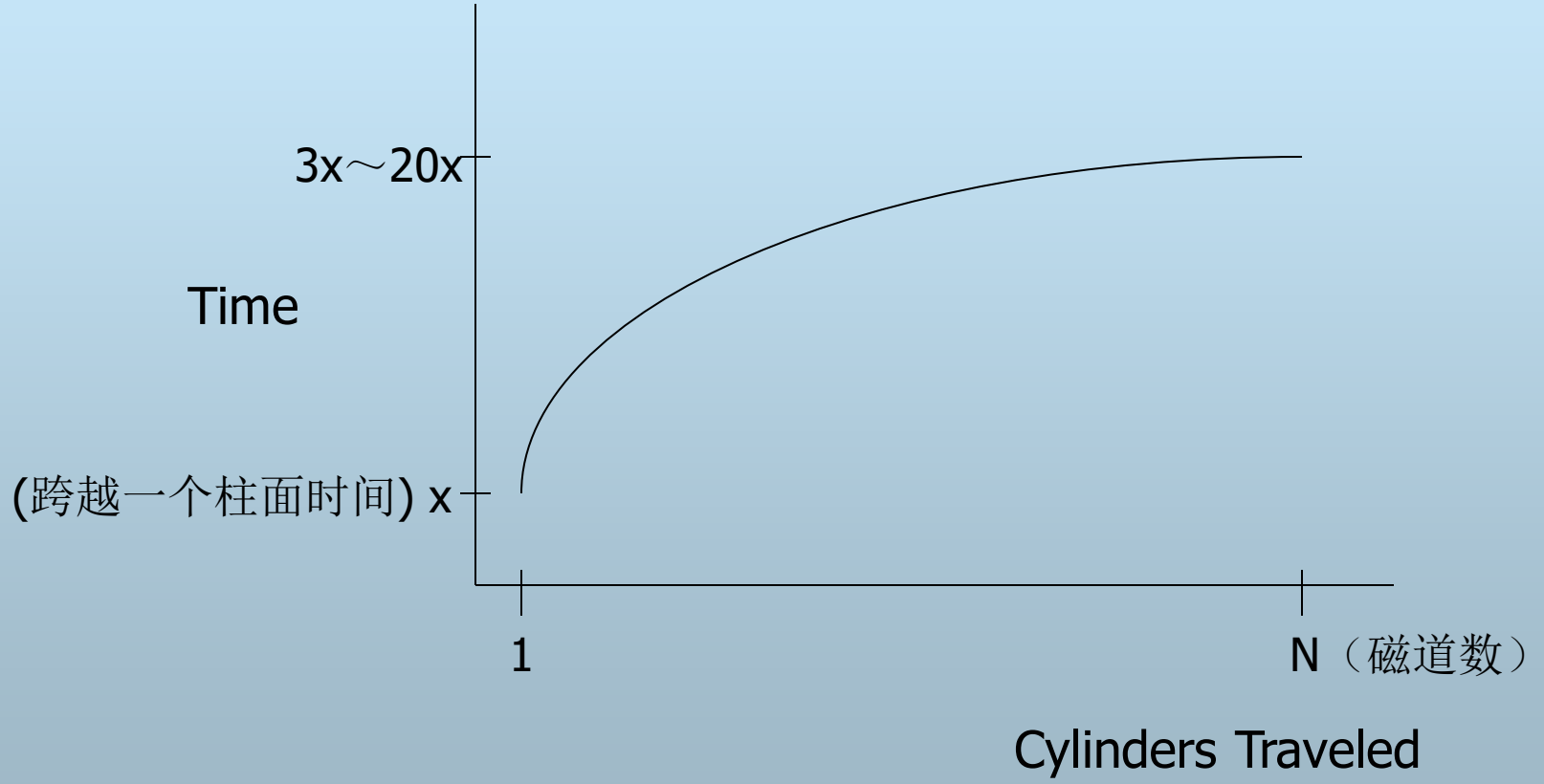
1、读块时间

- 磁盘块读取时间
  - 从“发出块存取请求”到“块位于主存”的时间
  - ＝寻道时间S＋旋转延迟R＋传输时间T＋其它延迟



2、寻道时间（Seek Time）

- 磁头定位到所要的柱面所花费的时间



3、平均寻道时间

$$S = \frac{\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \text{SEEKTIME}\ (i \rightarrow j)}{N(N-1)}$$

S一般在10 ms～40 ms之间

4、旋转延迟(Rotation Latency)

- 磁盘转动到块的第一个扇区到达磁头所需的时间
- 平均时间为旋转1/2周所费的时间

一个7200RPM的磁盘

平均旋转延迟 R≈4.17 ms

5、传输延迟(Transfer Time)

- 块的扇区及其间隙旋转通过磁头所需的时间
- 如果磁道大约有100 000字节，约10ms转一周，则每秒可从磁盘读取约10M字节，一个4K字节的块传输时间小于0.5ms

6、其它延迟

- CPU请求I/O的时间 (CPU time to issue I/O)
- 争用磁盘控制器时间 (Contention for controller)
- 争用总线和主存的时间 (Contention for bus, memory)

典型情况：0

7、如何读下一块？

- **CASE 1：下一块在同一柱面上**
  - 旋转延迟＋传输时间＋其它（忽略）
- **CASE 2：不在一个柱面上**
  - 寻道＋旋转＋传输＋其它

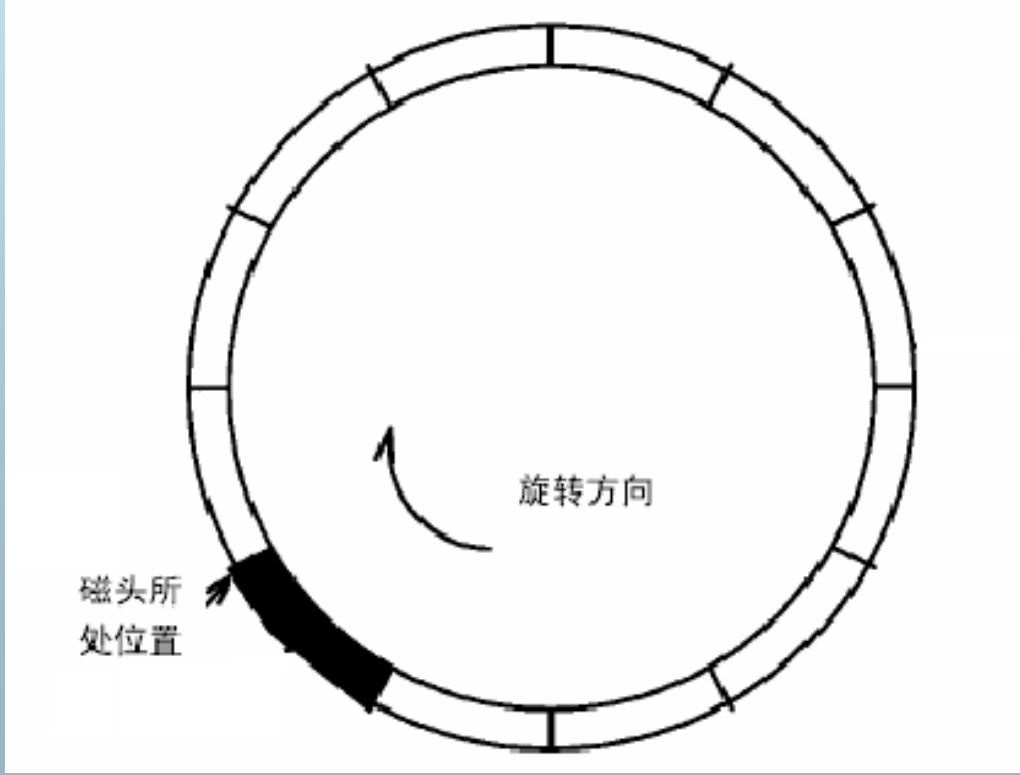
Random I/O

Sequential I/O

For a 4KB block
Random I/O ≈ 20ms
Sequential I/O ≈ 1 ms

8、写块

- 与读块类似
- 如果需要校验块是否正确写入，则需要加上一次旋转时间和一次块传输时间



9、块修改

- 将块读入主存
- 在主存中完成修改
- 将块重新写入磁盘

10、块地址

- 物理设备号
- 柱面号
- 盘面号（或磁头号）
- 扇区号

三、磁盘例子: Megatron747

- 参数
  - 3.5 inch
  - 3840 RPM
  - 8 surfaces
  - 8192 tracks/surface
  - 256 sectors/track
  - 512 bytes/sector
- Megatron 747大小
  - $8*8192*256*512 = 2^{33} = 8$ GB

1、Megatron 747参数

- 寻道时间 (最大)：17.4 ms
- 磁头启动停止1 ms，每移动500个柱面需1ms
- 1 block = 4 KB = 8 sectors
- 块之间的间隙占块的10%大小
- 每磁道大小=(256/8)*4 KB=128 KB=32块
- 每柱面大小=8*128KB=1 MB

2、Megatron747存取时间

- 3840 RPM → 1/64 秒/转 = 15.625 ms
- 读取一个磁道时间=15.625 ms, 其中
  - 用于磁道数据的时间=15.625 * 0.9=14.0625 ms
  - 用于扇区间隙的时间=15.625*0.1=1.5625 ms
- 读取一个块的时间=15.625/32－1.5625/256 ≈ 0.482 ms
  - 读取数据的时间=15.625/32 * 0.9 ≈ 0.439 ms

2、Megatron747存取时间

- OS或DBMS随机读取一块的最大时间
  - T=S+R+T
    =17.4 + 15.625 + 0.482 ≈ 33. 507 ms
- 最小时间：0.482 ms
- 平均时间
  - T=S+R+T
    =6.5 + 7.8125 + 0.482 ≈ 14.8 ms

平均寻道数＝8192/3＝2730（see Fig. 13. 9）

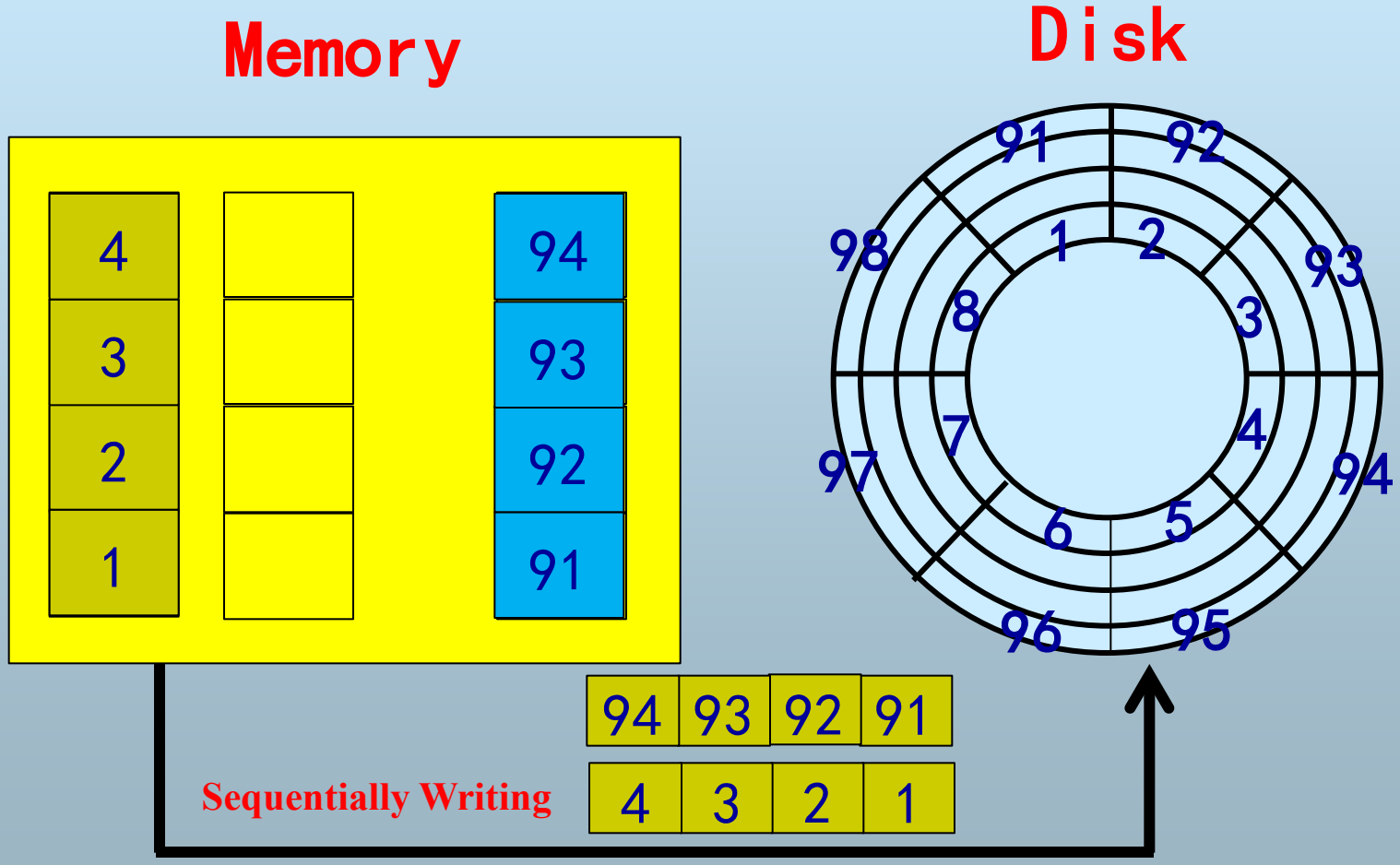
1＋2730/500 ＝ 6. 5

四、磁盘存取优化

- 按柱面组织数据
  - 减少平均寻道时间
- 磁盘调度算法
  - 如电梯算法 (Elevator Algorithm)
- 磁盘阵列(Disk Arrays)
- 磁盘镜像(Disk Mirrors)
- Random IO to Sequential IO
- 预取(Pre-fetch)和缓冲(Buffering)

1、Random IO to Sequential IO

Page requests: 1—91—2—92—3—93—4—94……



优化前：
8次随机IO
~160ms

优化后：
2次随机IO
6次顺序IO
~46ms

1、预取/缓冲

- 双缓冲(Double Buffering)
- 单缓冲(Single Buffering)

例：一个文件由一系列块构成：B1，B2，…

设有一程序，按下面顺序处理数据：

1、处理B1

2、处理B2

3、处理B3

……

2、单缓冲处理策略

(1) 将B1读入缓冲区

(2) 在缓冲区中处理B1中的数据

(3) 将B2读入缓冲区

(4) 处理缓冲区中的B2数据

...

3、单缓冲处理分析

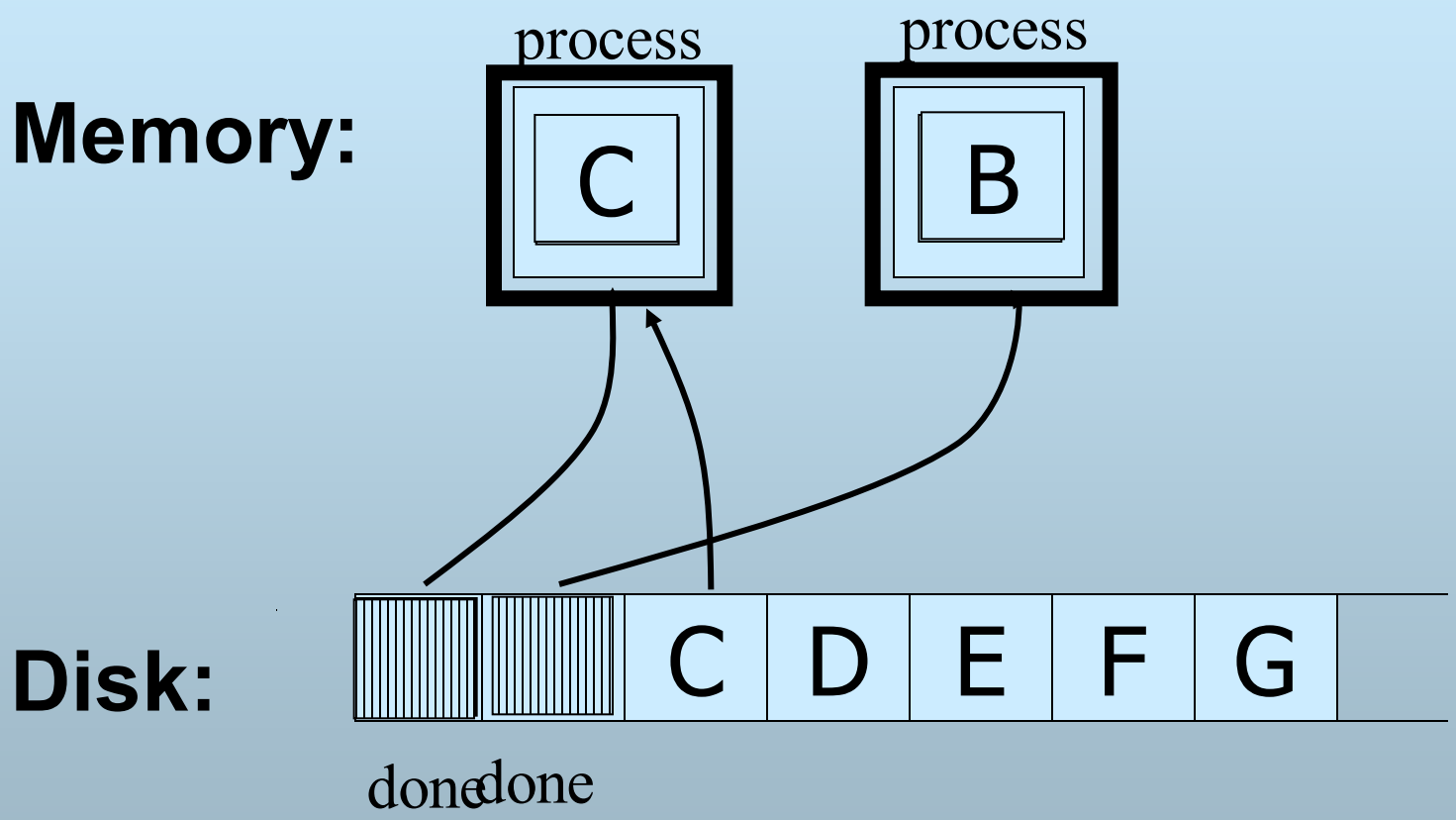
设　P = 在缓冲区中处理一块的时间

R = 将一块读入缓冲区的时间

n = 块数

单缓冲处理时间 = n(P+R)

4、双缓冲



4、双缓冲分析

P = Processing time/block
R = IO time/block
n = # blocks

- 双缓冲处理时间＝R＋nP（P>=R）
  ＝nR＋P（R>=P）
- 单缓冲处理时间＝n(R + P)

5、缓冲的缺点

- 主存代价
- 缓冲区管理
- 一致性维护

6、块大小选择

- 大块
  - I/O次数↓
  - 可能读入大量无用数据
  - 每次I/O要花费更多时间
- 趋势
  - 大块

五、新型存储

- 计算机系统性能依赖于
  - 处理器的数据计算能力
  - 存储层次向处理器传输数据的能力
- 随着多\众核、多线程技术的发展，传统存储器件构成的存储层次面临的存储墙问题愈发严重
  - 处理单元（核）数的增长与存储数据供应能力（容量）不匹配
  - SRAM\DRAM的功耗过高
- 新型存储器件包括：闪存、相变存储器、磁阻式存储、电阻式存储器、忆阻器等等。具备一个共同特点：非易失性
  - 优点：高存储密度、低功耗、无机械延迟、存取速度快、便携、抗震、低噪音等
  - 缺点：读写性能不对称、读写次数有限、可靠性不高等

1、闪存 Flash Memory

Tape is Dead
Disk is Tape
Flash is Disk

Jim Gray
1998年图灵奖获奖演说

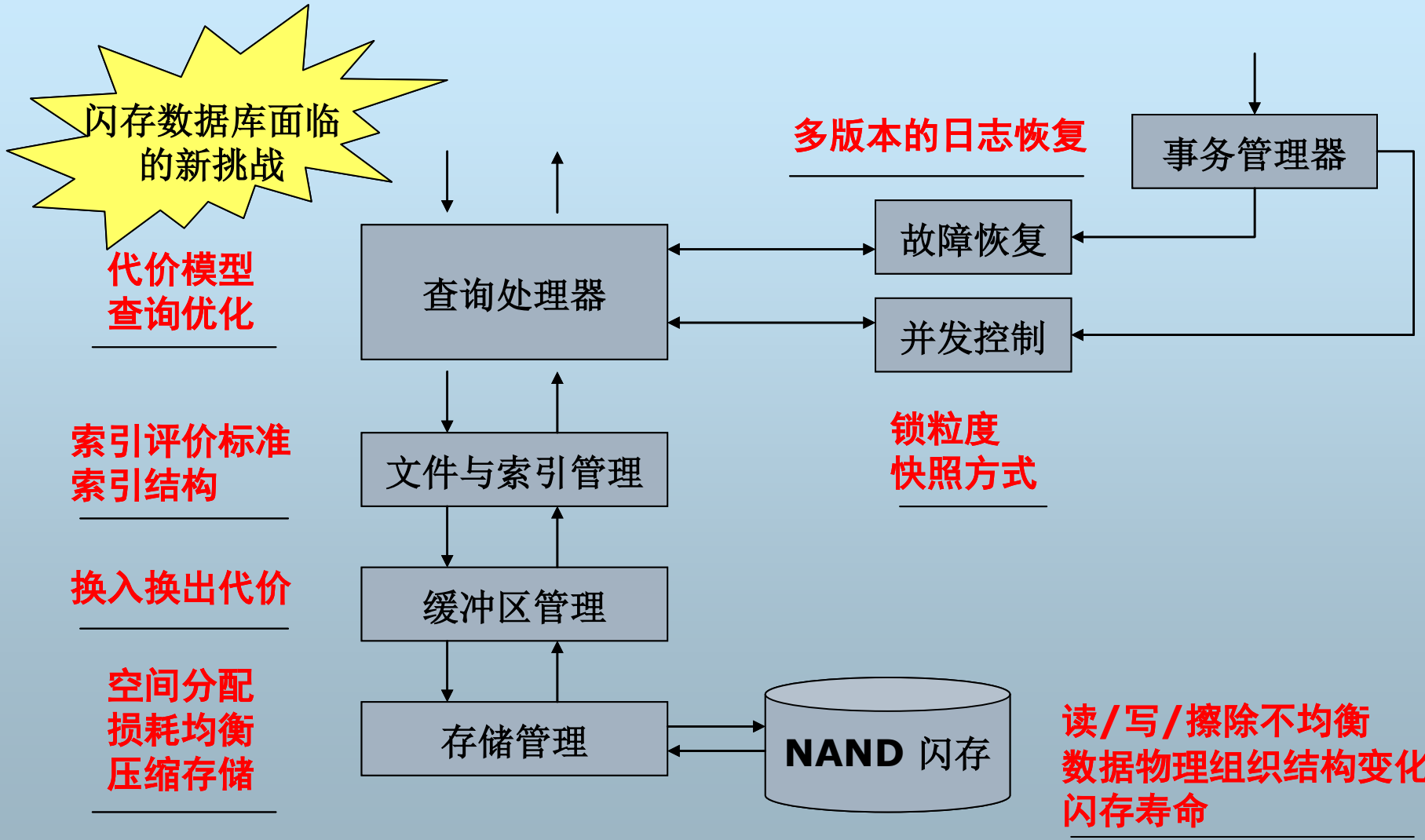
1、闪存 Flash Memory

- 闪存的工业化程度最高
  - SSD（solid state drive)
  - 闪存芯片+控制器+FTL (WL, LBA-PBA, GC)

- （NAND）闪存的特点
  - 读写不对称
    - 写慢读快
  - 写前擦除：异位更新、块擦除操作
  - 寿命有限：块擦除次数有限
    - SLC （约10万次擦写)
    - MLC（小于1万次)
    - TLC（小于1000次)
  - 按页读写
    - E.g., 1 page =2 KB
  - 按块擦除
    - E.g., 1 block = 64 pages

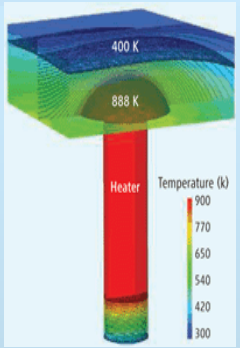
Media	Access time		
	Read	Write	Erase
Magnetic† Disk	12.7 ms (2 KB)	13.7 ms (2 KB)	N/A
NAND Flash‡	80 μs (2 KB)	200 μs (2 KB)	1.5 ms (128 KB)

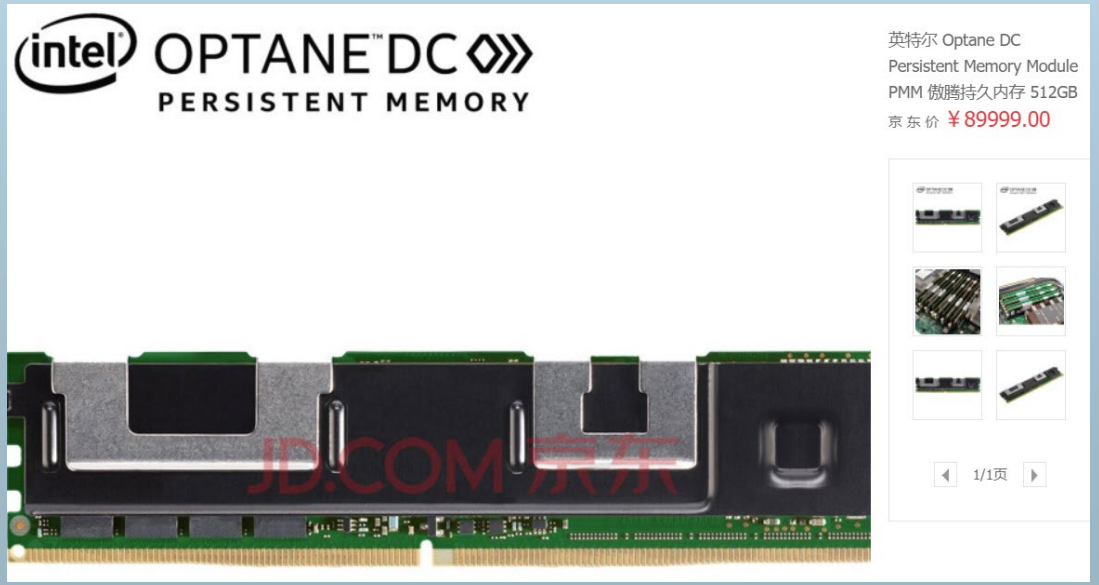
1、闪存 Flash Memory

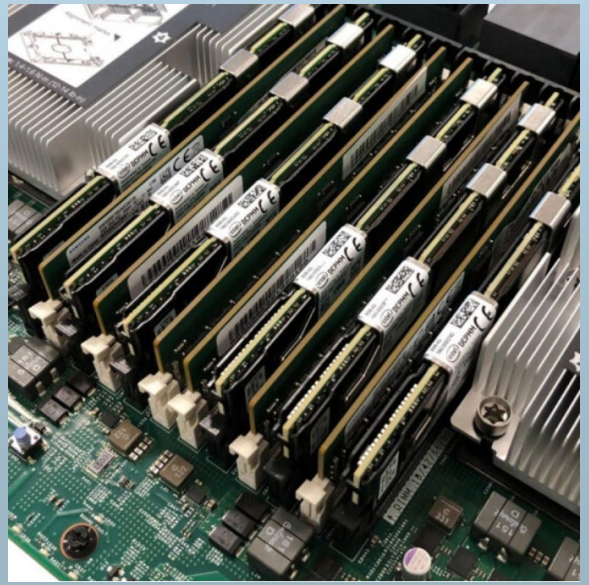


2、相变存储器 Phase Change Memory

- **PCM**
  - 起源于20世纪60年代
  - 电阻式非易失性半导体存储器
  - 以硫族化物材料作为存储介质，利用相变材料在不同结晶状态时呈现出显著的电阻值差异性来实现数据存储







2、相变存储器

- **PCM vs. Flash**

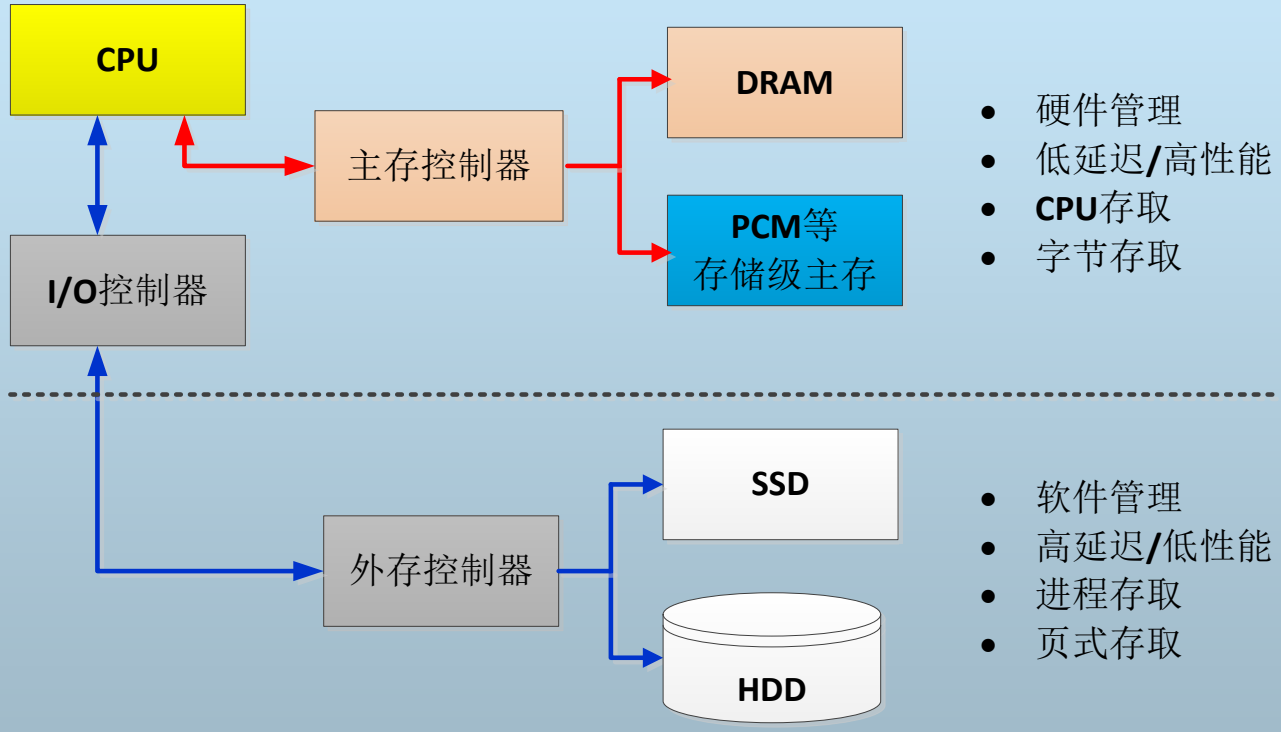
	DRAM	PCM	NAND Flash
Page size	64B	64B	4KB
Page read latency	20-50ns	~ 50ns	~ 25 µs
Page write latency	20-50ns	~ 1 µs	~ 500 µs
Write bandwidth	~GB/s per die	50-100 MB/s per die	5-40 MB/s per die
Erase latency	N/A	N/A	~ 2 ms
Endurance	∞	$10^6 - 10^8$	$10^4 - 10^5$
Read energy	0.8 J/GB	1 J/GB	1.5 J/GB [28]
Write energy	1.2 J/GB	6 J/GB	17.5 J/GB [28]
Idle power	~100 mW/GB	~1 mW/GB	1–10 mW/GB
Density	1×	2 – 4×	4×

2、相变存储器

- **PCM vs. DRAM**

	DRAM	PCM	NAND Flash
Page size	64B	64B	4KB
Page read latency	20-50ns	~ 50ns	~ 25 µs
Page write latency	20-50ns	~ 1 µs	~ 500 µs
Write bandwidth	~GB/s per die	50-100 MB/s per die	5-40 MB/s per die
Erase latency	N/A	N/A	~ 2 ms
Endurance	∞	$10^6 - 10^8$	$10^4 - 10^5$
Read energy	0.8 J/GB	1 J/GB	1.5 J/GB [28]
Write energy	1.2 J/GB	6 J/GB	17.5 J/GB [28]
Idle power	~100 mW/GB	~1 mW/GB	1–10 mW/GB
Density	1×	2 – 4×	4×

3、基于新型存储的计算机架构



- 硬件管理
- 低延迟/高性能
- CPU存取
- 字节存取

- 软件管理
- 高延迟/低性能
- 进程存取
- 页式存取

小 结

- 存储器结构（Disk Structure）
  - 柱面、磁道、扇区、块
- 磁盘块存取时间（Block Access Time）
  - 寻道时间、旋转延迟、传输时间
  - 块存取时间分析
- 磁盘例子: Megatron747
- 磁盘存取优化（Optimization）
- 新型存储