第六章 存储系统与技术

张华平 副教授 博士

Email: kevinzhang@bit.edu.cn

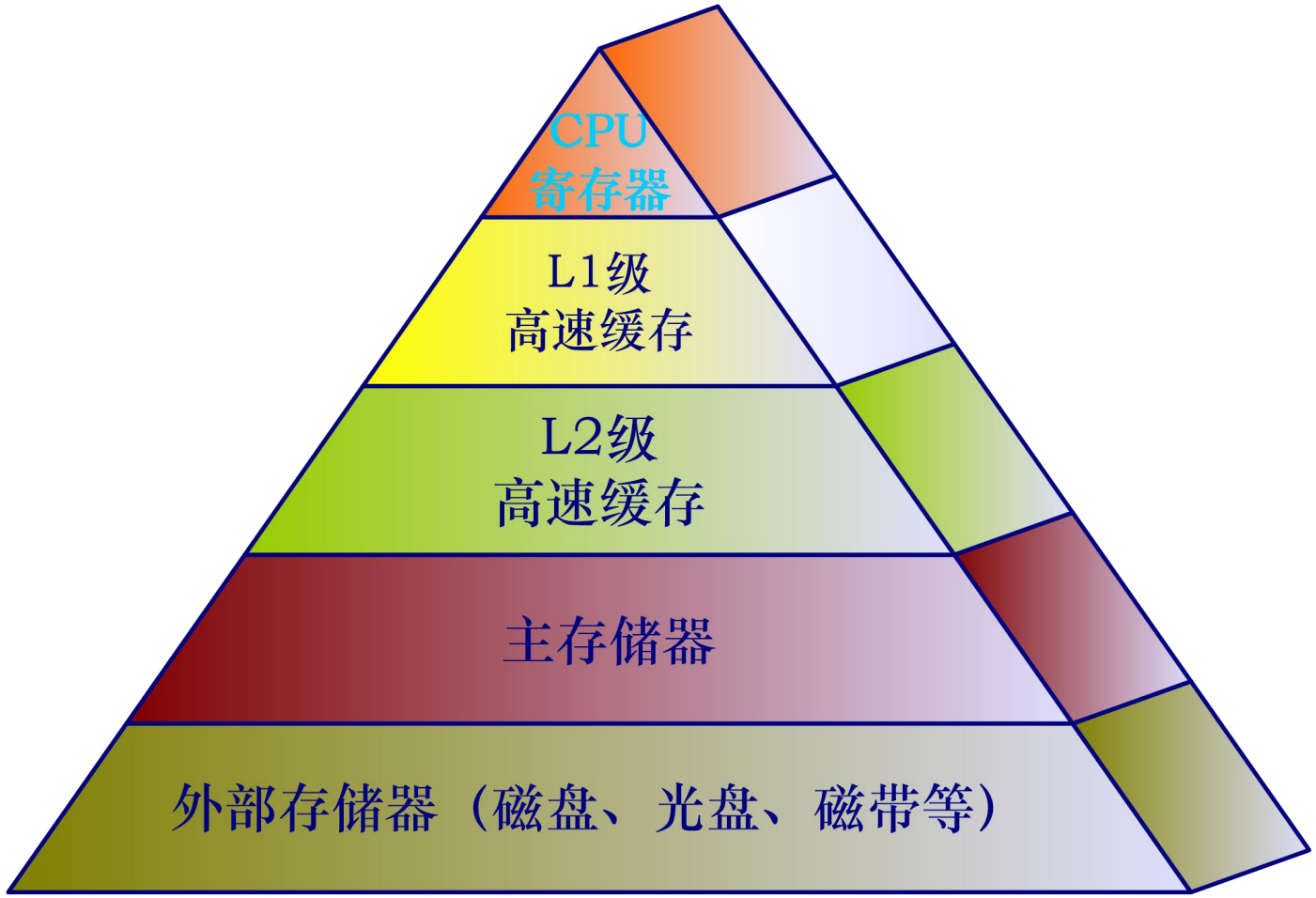
Website: http://www.nlpir.org/

@ICTCLAS张华平博士

大数据搜索挖掘实验室（wSMS@BIT）

（1）【重点讲解】Cache工作原理
（2）【重点讲解】DDR读写时序
（3）【一般性讲解，概念为主】辅助存储器/扇区编址
（4）【简单了解，不作要求】固态硬盘

存储系统层次结构



6.1 高速缓冲存储器

- Cache：SRAM构成
  - SRAM：1～2时钟周期读写一次数据（Cache）
  - DRAM：多个时钟周期读写一次数据（RAM/主存）
- Cache的局部性原理
  - 时间局部性：Cache访问速度
  - 空间局部性：Cache访问容量
- Cache的访问结构
  - 贯通查找式（Look Through）结构
  - 旁路读出式（Look Aside）结构

Cache访问结构

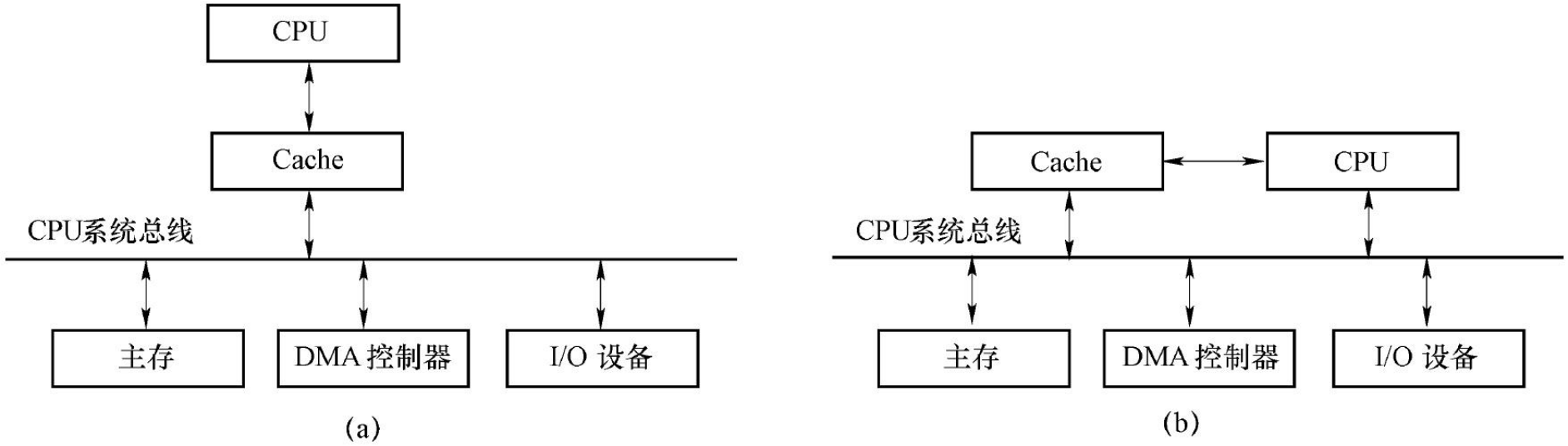


图 6－2　Cache 在数据访问中的位置

（a）Look Through 结构；（b）Look Aside 结构

(a) cache平均访问时间 = cache访问时间+（1−命中率）×未命中时主存访问时间

(b) cache平均访问时间 = 命中率×cache访问时间+（1−命中率）×未命中时主存访问时间

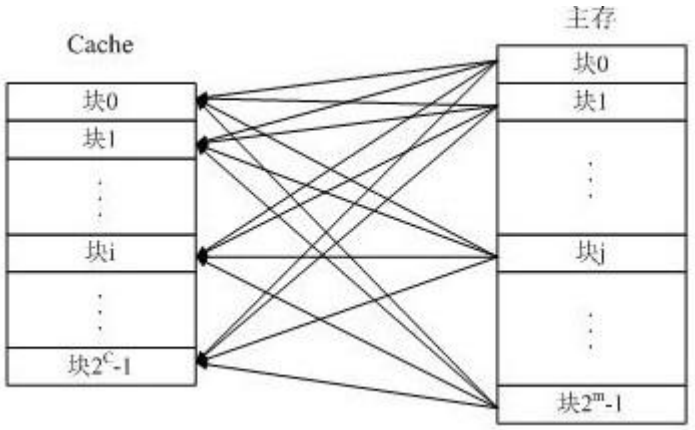
Cache映射

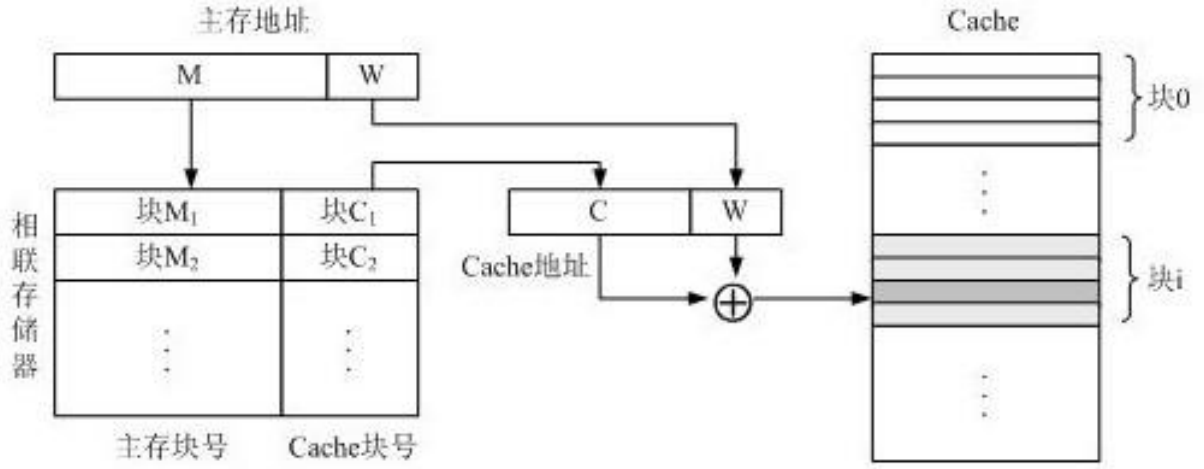
主存和Cache之间一次交换的数据单位是一个数据块；数据块大小固定，由若干个字组成，主存和Cache的数据块大小相同；Cache对程序员透明，CPU的访主存地址需转换成访Cache地址；主存地址与Cache地址之间的转换是与主存块与 Cache块之间的映射关系紧密联系。

- 全相联映射
- 直接相联映射
- 组相联映射

6.1 高速缓冲存储器

全相联地址映射方法





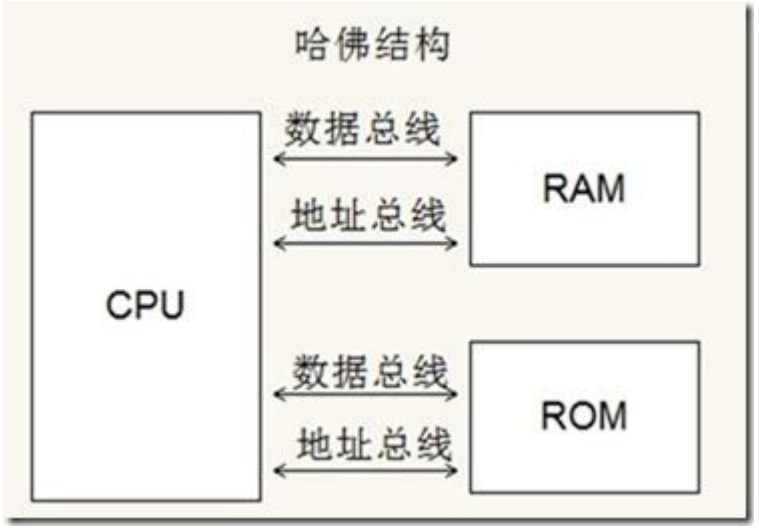
Cache替换策略

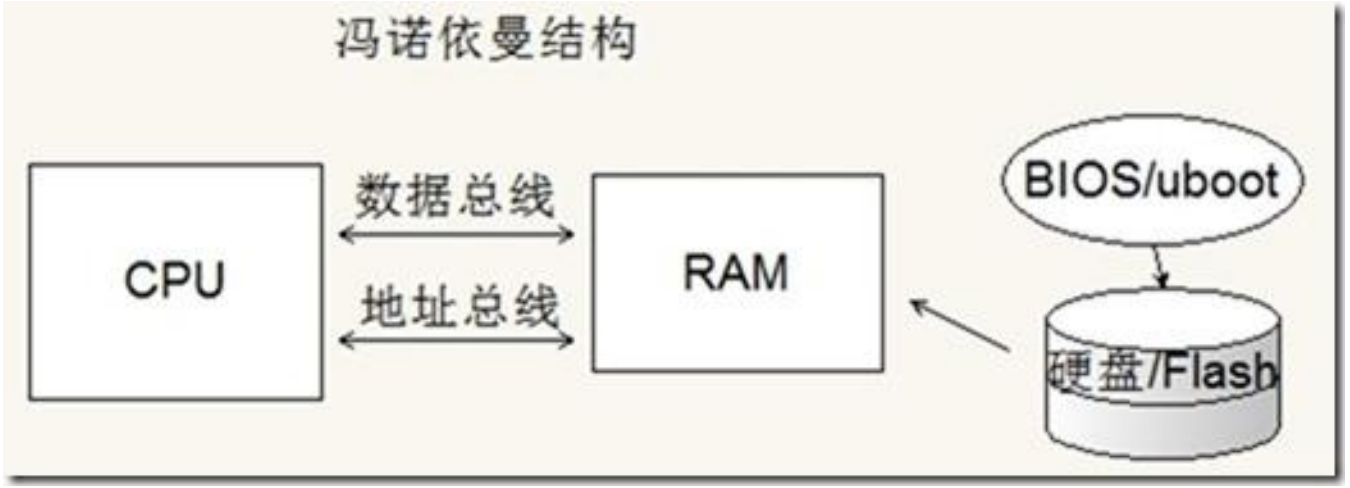
- 主存的一个块要调入Cache存储器时，如果Cache存储器中没有空闲的行，就必须从中选取一行，用新的块覆盖其原有的内容。这种替换应该遵循一定的规则，其目标是选取在下一段时间内被存取的可能性最小的块，替换出Cache。
- 随机算法、先进先出（FIFO）算法和近期最少使用（LRU）算法

微机中的Cache

- 一级缓存（L1 cache）：一级缓存中采哈佛结构，分为数据缓存（Data Cache，D-Cache）和指令缓存（Instruction Cache，I-Cache），分别用来存放数据和指令。
- 二级缓存（L2 Cache）
- 三级缓存（L3 Cache）
- 追踪缓存（Execution Trace Cache，T-Cache或ETC）：P4中替代一级指令缓存，容量为12KμOps（μOps，微指令），能存储12000条解码后的微指令。

哈佛结构与普林斯顿/冯诺依曼结构





Cache一致性协议

写Cache的过程因为涉及到对内容的修改，存在导致Cache内容和对应内存内容不一致的可能性。

单核CPU一致性处理

- 未命中时的Cache写策略：数据直接写入内存。含有写入数据的内存块可以根据需要决定是否随后调入Cache中。
- 命中时的Cache写策略：直写式（Write Through）及回写式（Write Back）。

命中时的Cache写策略

- 直写式：CPU在向Cache写入数据的同时，立即把数据写入内存，以保证Cache和内存中相应单元数据的一致性。直写式策略的特点是简单可靠，但由于CPU每次更新数据时都要对内存写入，写入速度受到影响。
- 回写式：CPU只向Cache写入数据，不立即写入内存。Cache为每一行设置一个标志位（dirty，脏位），为1时表示Cache中的数据尚未更新到内存。要替换这一行时，数据必须先写入内存的块之后，才被其他块所使用。回写式策略的特点是发生命中时CPU更新数据较快，但Cache的结构复杂，而且在回写前会暂时出现Cache中的数据和内存不一致的情况。

多核CPU的MESI协议

- 多核环境，每个核又都有自己的缓存，那么就需要更复杂的协议来保持一致性，通常利用MESI及其衍生协议（比如MESIF协议和MOESI协议等）来达到目的。
- 修改（Modified）缓存段
- 独占（Exclusive）缓存段
- 共享（Shared）缓存段
- 无效（Invalid）缓存段

MESI协议的一致性处理

- 对于无效缓存段（I状态），相当于未加载进Cache。进行读写操作，则首先需要将对应的内存块调入。此时Cache和内存对应的块内容是一致。
- 对于共享缓存段（S状态），可以在多个处理器中存在相同的拷贝，但因为只能读不能写，所以也不存在不一致的可能性。
- 对于独占缓存段（E状态），表示当前Cache行中包含的数据有效，并且该数据仅在当前处理器的Cache中有效，而不在其他处理器的Cache中存在拷贝。在该Cache行中的数据是当前处理器系统中最新的数据拷贝，而且与存储器中的数据一致。
- 对于修改缓存段（M状态），表示当前Cache行中包含的数据与存储器中的数据不一致，而且它仅在本处理器的Cache中有效，不在其他处理器的Cache中存在拷贝，因此其他处理器不会读出无效的、过期数据。当处理器对这个Cache行执行替换操作时，会触发系统总线的写周期，将Cache行中被修改过的数据（脏数据）与内存中的数据进行同步，从而保持一致性。

- 只有当缓存段处于E或M状态时，处理器才能执行写操作，也就是说，只有这两种状态下，处理器是独占这个缓存段的，而对应的内容在其他Cache区域没有复制。
- E状态和M状态的差别在于，E状态的缓存段内容和对应的内存块一致，因此，当退出E状态时，可以转入S状态。而M状态的缓存段内容和对应的内存块不一致，因此，当退出M状态时，需先进行写内存操作。

6.2 内部存储器

内存分类

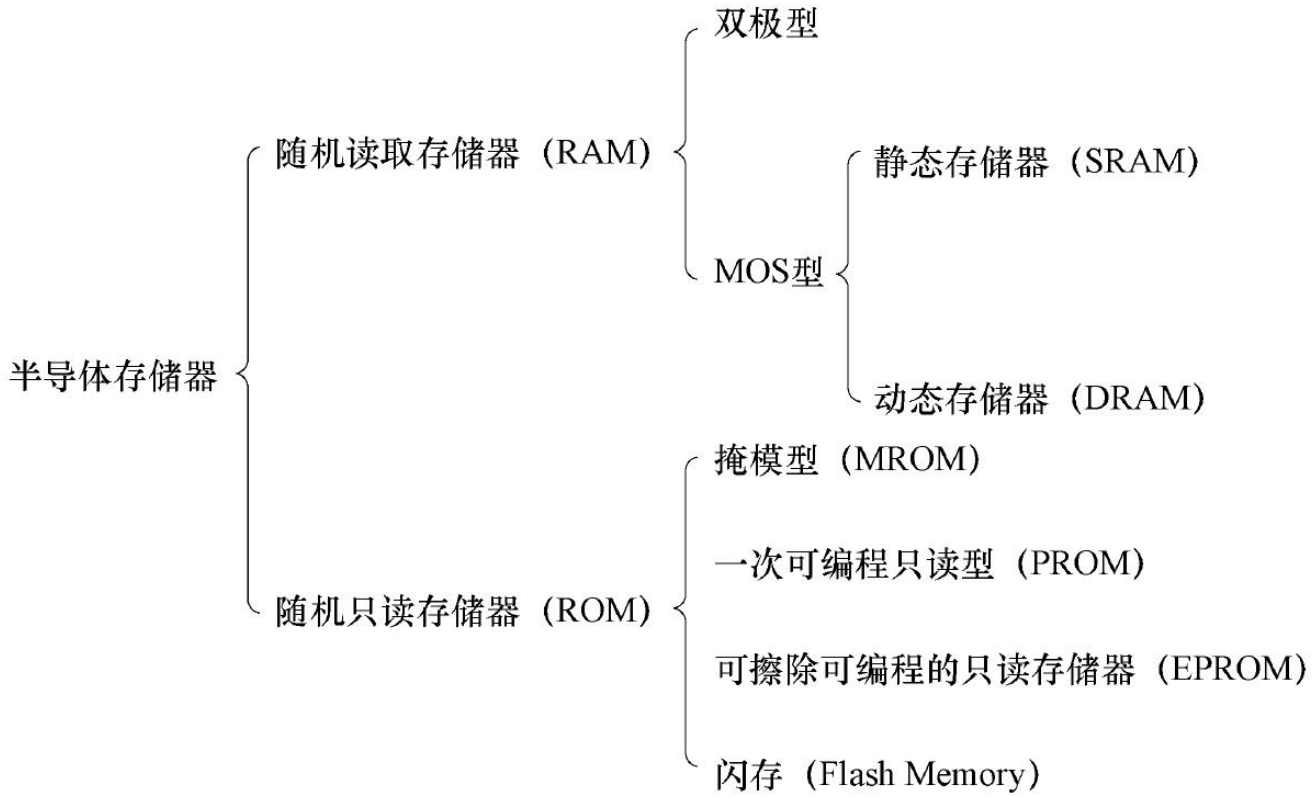


图 6－4　半导体存储器的分类

6.2 内部存储器

- RAM
  - 双极型（Bipolar）：存取速度高，成本高，用作Cache
  - MOS型
    - ⑩ 静态RAM（Static RAM，SRAM）：不需要刷新
    - ⑩ 动态RAM（Dynamic RAM，DRAM）：定期刷新
- ROM
  - 掩模型MROM（Mask ROM）：制造厂家写入数据
  - 可编程只读存储器PROM（Programmable ROM）：1次编程
  - 可擦除可编程的只读存储器EPROM（Erasable PROM）
    - ⑩ 多次编程
  - 闪存（Flash Memory）

主要技术指标和参数

- 存储容量

存储单元数量=行数×列数×数据深度×L-Bank的数量

表 6-1　128M 位内存芯片的布局

布局	存储单元数	位宽	Bank 数	行地址	列地址
8M×4×4	8M	4	4	A0～11	A0～9、A11
4M×8×4	4M	8	4	A0～11	A0～9
2M×16×4	2M	16	4	A0～11	A0～8

每个Bank中8M个存储单元（2^{12}行和2^{11}列），总Bank数为4，数据深度为4，总容量128M bit

内存带宽

- 内存的数据传输速度，是衡量内存的重要指标

带宽=总线宽度×总线频率×一个时钟周期内交换的数据包个数

- 例6.1 已知总线频率，试计算如下内存带宽。

PC100 SDRAM 外频100MHz时，带宽=64×100/8=800（MB/s）

PC133 SDRAM 外频133MHz时，带宽=64×133/8=1 064（MB/s）

DDR DRAM 外频100MHz时，带宽=64×100×2/8=1.6（GB/s）

- **存储器访存速度**
  - 存储周期（MC）/访存时间（AC）
- **错误校验**
  - 奇偶校验（Parity）/ECC校验

内存模组

- 为了节省主板空间和增强配置的灵活性，现在的主板多采用内存条结构。将存储器芯片、电容、电阻等元件焊接在一条PCB（印制电路板）上组装起来合成一个内存模组（RAM Module），俗称内存条，由内存控制器管理。
- SIMM（30/72）、DIMM（168/184）、RIMM（184）

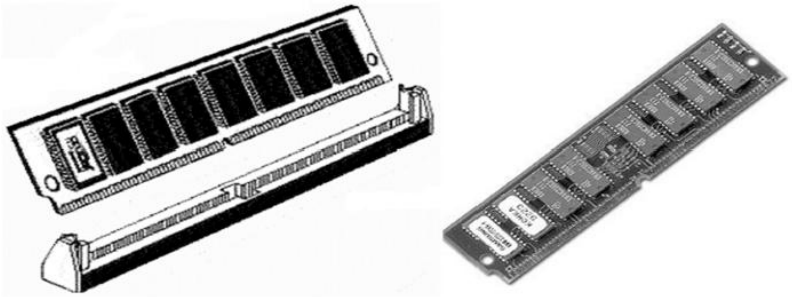


图 6-6 SIMM

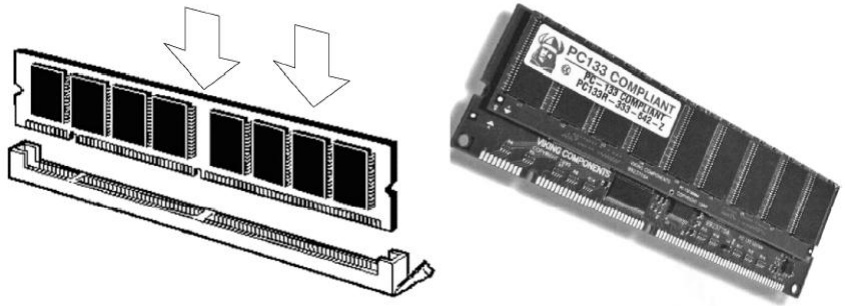


图 6-7 DIMM

内存颗粒

- SDRAM：同步动态随机存储器（Synchronous DRAM，SDRAM），内存与系统总线速度同步。
- CL-t_{RCD}-t_{RP}-t_{RAS}

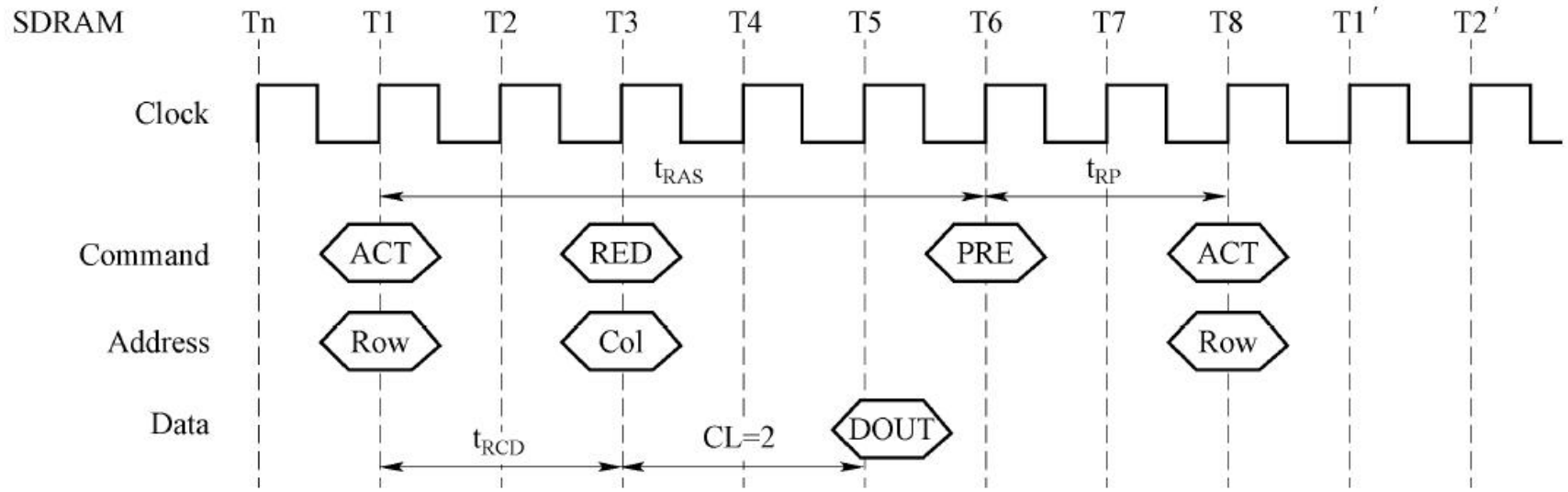


图 6-8 SDRAM 存取时序（BL=1）

DDR

- SDRAM 在时钟的上升沿进行数据传输，一个时钟周期内只传输一次数据；而DDR 内存则是一个时钟周期内传输两次数据，它能够在时钟的上升沿和下降沿各传输一次数据，因此称为双倍速率同步动态随机存储器（Double Date Rate SDRAM，DDR）
- DDR2/DDR3/DDR4
  - ⑩DDR2内存的DQS采用差分信号/内存采用1.8V电压/TSOP
  - ⑩DDR3采用1.5V电压/8位预读/FBGA
  - ⑩DDR4采用16位预读/1.2V电压/DBI及CRC功能/更高的频率

SPD芯片

- 记录内存参数信息

内存比较

表 6－3　SDRAM、DDR、DDR2 和 DDR3 比较

SDRAM 类型	DDR3 SDRAM	DDR2 SDRAM	DDR SDRAM	SDRAM
时钟频率/MHz	400/533/667	200/266/333/400	100/133/166/200	100/133/166
数据传输速率/（Mb・s^{-1}）	800/1066/1 333	400/533/667/800	200/266/333/400	100/133/166
位宽	x4/x8/x16	x4/x8/x16	x4/x8/x16/x32	x16/x32
预读宽度/位	8	8	8	1
时钟输入	差分（CK，$\overline{CK}$）	差分（CK，$\overline{CK}$）	差分（CK，$\overline{CK}$）	Single clock
BL（突发长度）	4　（Burst chop），8	4，8	2，4，8	1，2，4，8，full page
数据选通信号	差分（DQS）	差分（DQS）	无	无
电压/V	1.5	1.8	2.5	3.3/2.5
标准	SSTL_15	SSTL_18	SSTL_2	LVTTL
CL 范围	5，6，7，8，9，10	3，4，5	2，2.5，3	2，3
封装形式	FBGA	FBGA	TSOP（II）/FBGA/LQFP	TSOP（II）/FBGA

内存组织示例

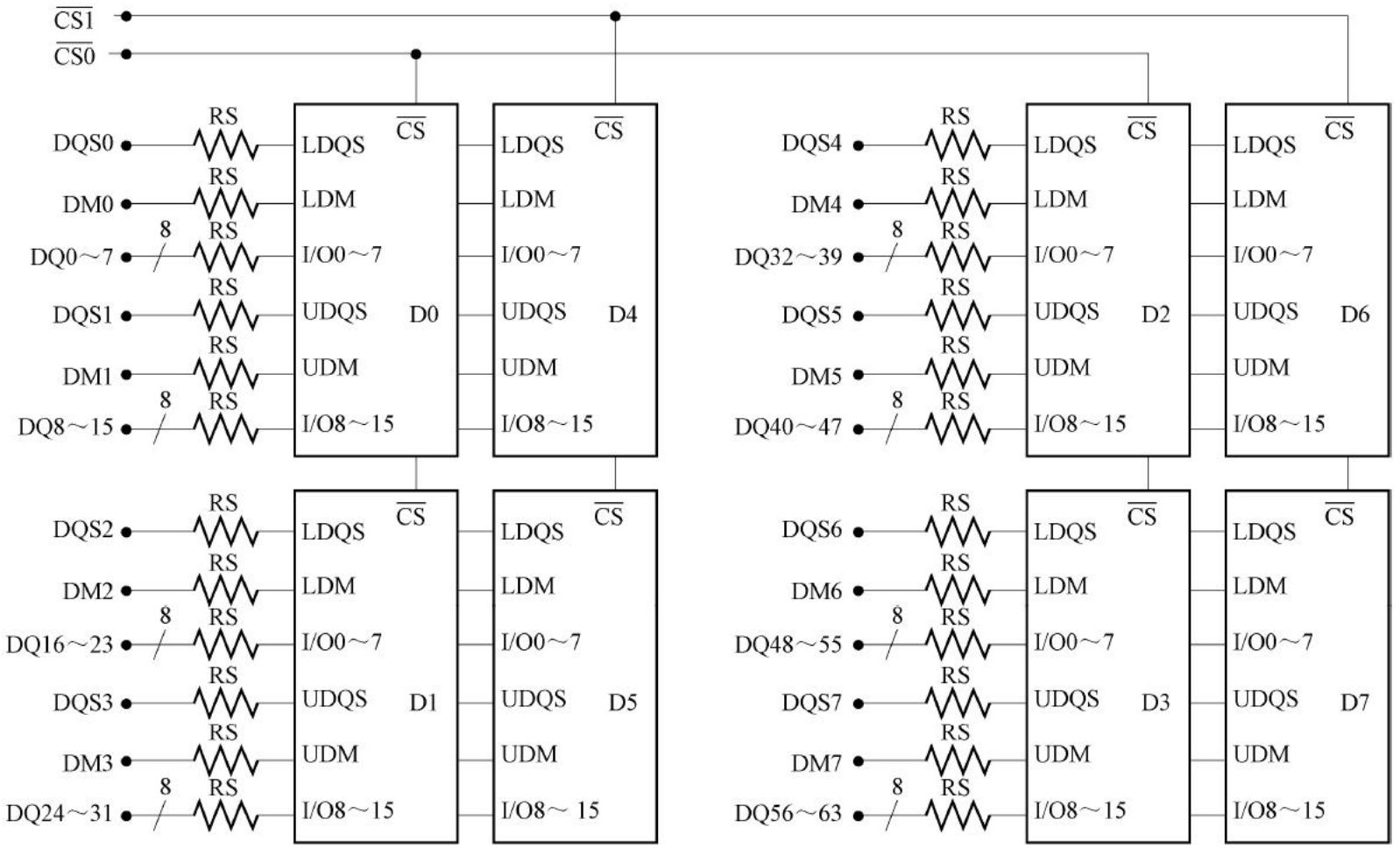
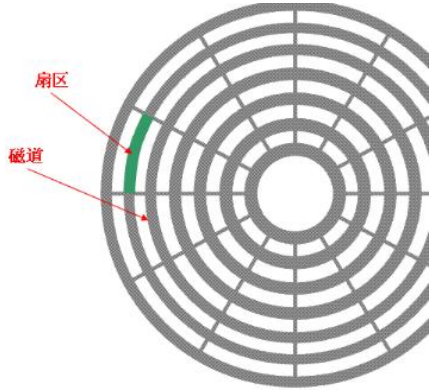


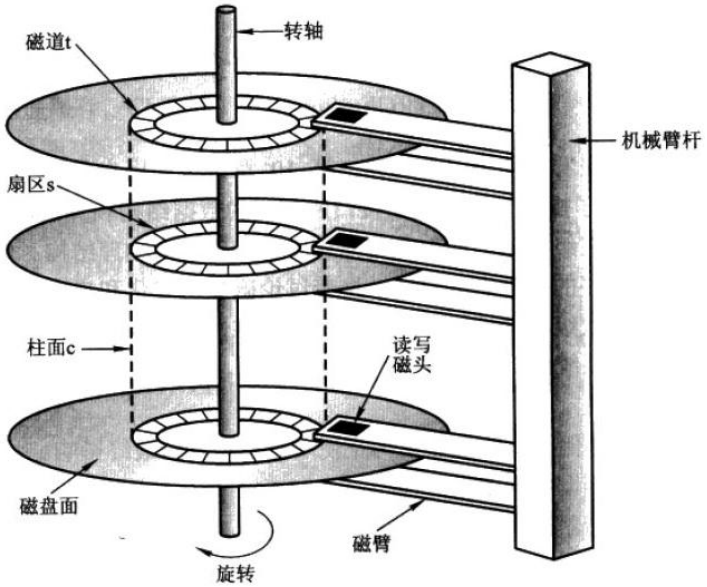
图 6－14　DDR 内存组织

6.3 辅助存储器

硬盘

- ATA、IDE、SATA
- 固态硬盘/Flash





- **工作原理**
  - 磁盘片/2个盘面
  - 磁道：0磁道是硬盘上非常重要的位置。硬盘的主引导记录区（Main Boot Record，MBR）就保存在0磁头0柱面1扇区。
  - 扇区：扇区是最小的读写单位，每个扇区中的数据作为一个单元同时读出或写入，扇区从1开始编号。

HDD主要技术指标

- 容量：可达15TB/WD
- 转速：7200转、5400转/分钟，可达15000/用于服务器
- 缓存：64MB、256MB等
- 时间相关参数
  - ❿ 平均寻道时间：磁头移动至指定磁道所用的时间
  - ❿ 平均潜伏时间：磁头移动到数据所在的磁道后，等待指定数据块继续转动到磁头下的时间
  - ❿ 平均访问时间：磁头开始移动到找到指定数据的平均时间
  - ❿ 平均访问时间＝平均寻道时间+平均潜伏时间+数据读取时间
- 数据传输率：突发数据传输率、持续传输率
- 接口类型：IDE、SCSI、SATA

ATAPI标准

- ATA总线：40 针或者80针连接器，针脚间距0.1in（2.54mm），通常有“键控”，以防止安装时颠倒方向。IDE 电缆的长度不能超过0.46m（18in）。TTL电平。

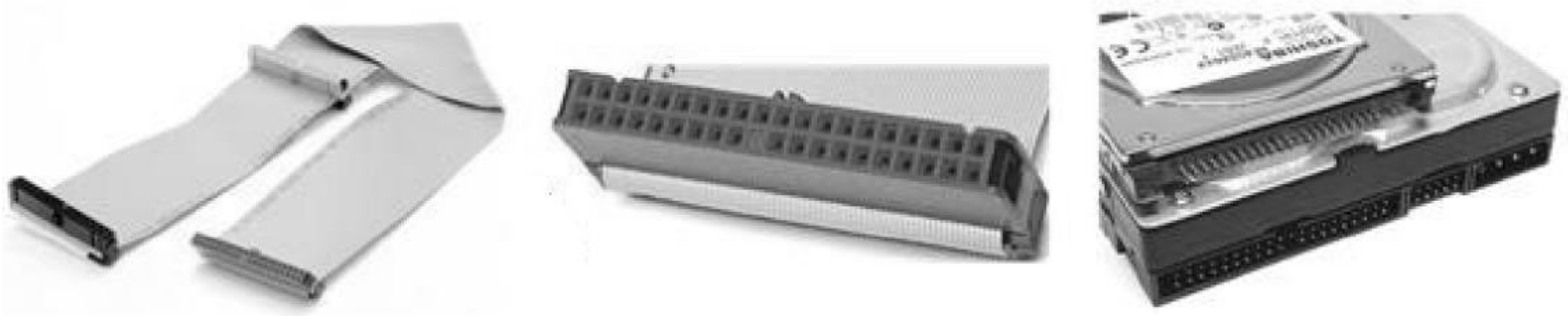


图 6-15　ATA 电缆和硬盘接口

ATA接口及其发展

- ATA-1至ATA-6

6.3 辅助存储器

- **ATA接口的编程模型：**扇区编制模式CHS
  - N个盘面对应2N个磁头，磁头：0.1.2…nH
  - N个磁道对应N个柱面，柱面：0.1.2…nC
  - 每个磁道又被分为多个扇区，扇区：1.2.3…nS
  - 总nC×nH×nS个扇区
  - 0柱面0磁头1扇区是整个硬盘的第1个扇区
  - 每个扇区512字节

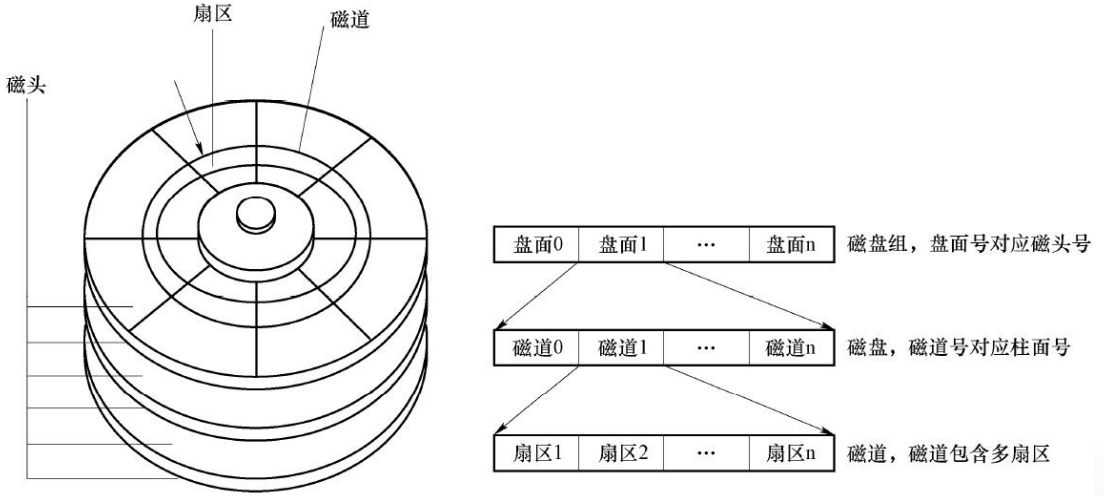


图 6－17　硬盘结构

CHS所带来的容量限制

表 6-5　不同的 CHS 标准及硬盘容量限制

标准	柱面	磁头	扇区	总位数	硬盘上限
IDE/ATA	16	4	8	28	137 GB
BIOS Int 13H	10	8	6	24	8.06 GB
最小定义	10	4	6	20	504 MB

- 28位的LBA可支持：$2^{28}\times512=137438953472$（B）=128GB。当1GB=$10^9$B 时，也被称为137GB限制
- 48位的LBA理论上可支持的硬盘容量就达到了$2^{48}\times512=2^{57}$（B），大致相当于128PB（1PB=2^{50}B）

LBA编址模式与CHS编制模式互换

- 设一个扇区在LBA编址模式中的地址为L，在CHS编址模式的地址为<C，H，S>，0≤C≤nC-1，0≤H≤nH-1，1≤S≤nS，则L=[(C×nH + H)×nS]+S－1。
- 根据L计算<C，H，S>：

$$S = (L \% nS) + 1$$

$$H = (L \div nS) \% nH$$

$$C = (L \div nS) \div nH$$

PATA接口的传输模式

- PIO模式：CPU执行IN/OUT 指令访问ATA控制器的数据端口，将数据从硬盘读出或者写入硬盘。数据传送率不高，数据传送过程中CPU被占用。
- DMA模式：数据的传送在ATA控制器的端口和内存之间直接进行，不需要通过CPU。

- ATA设备寄存器（略）
  - 可编程ATA寄存器地址
  - 可编程ATA寄存器定义

- PIO方式读写硬盘—以读为例
  - 复位硬盘。
  - 读状态寄存器，直到检测到BSY=0，DRQ=0。超时错则退出。
  - 完成读命令的参数设置，并写入读命令。
  - 查询方式下，读取状态寄存器。正确则继续执行，否则转出错处理。
  - 从数据寄存器读取扇区内容。
  - 当所有的请求扇区的数据被读取后，命令执行结束。
- DMA方式读取硬盘（略）

串行ATA/SATA

- 取代了传统的ATA，区别主要在于连接电缆和数据传输方式上。数据线7芯，电源线15芯。

表 6－14　SATA 接口引脚及功能

SATA 引脚	引脚功能
1	地
2	A+，发送差分对信号
3	A-，发送差分对信号
4	地
5	B+，接收差分对信号
6	B-，接收差分对信号
7	地
-	槽口

表 6－15　SATA 技术指标

版本	带宽/（Gb・s^{-1}）	实际速度/（MB・s^{-1}）	线缆最大长度/m
SATA 3.0	6	600	2
SATA 2.0	3	300	1.5
SATA 1.0	1.5	150	1

AHCI技术

- SATA模式，区别于传统的ATA模式，没有开启AHCI功能，启动默认为ATA模式。如NCQ技术及热插拔功能等。

NCQ技术

- 支持NCQ 技术的硬盘在接到读写指令后，根据指令对访问地址进行重新排序，减少读取时间，使数据传输更为高效，同时也能有效延长硬盘的使用寿命。

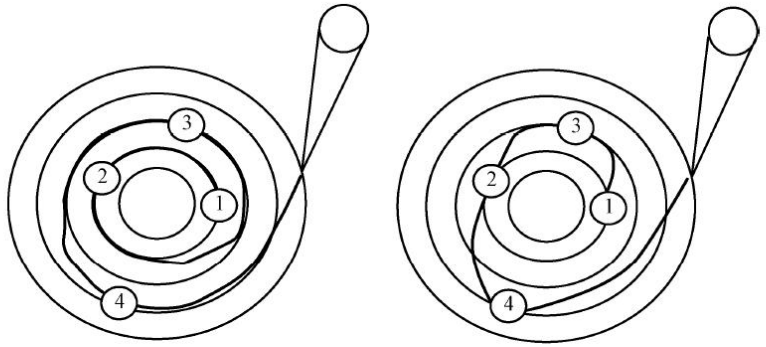


图 6-31　不支持 NCQ 和支持 NCQ 硬盘访问的磁头轨迹示意

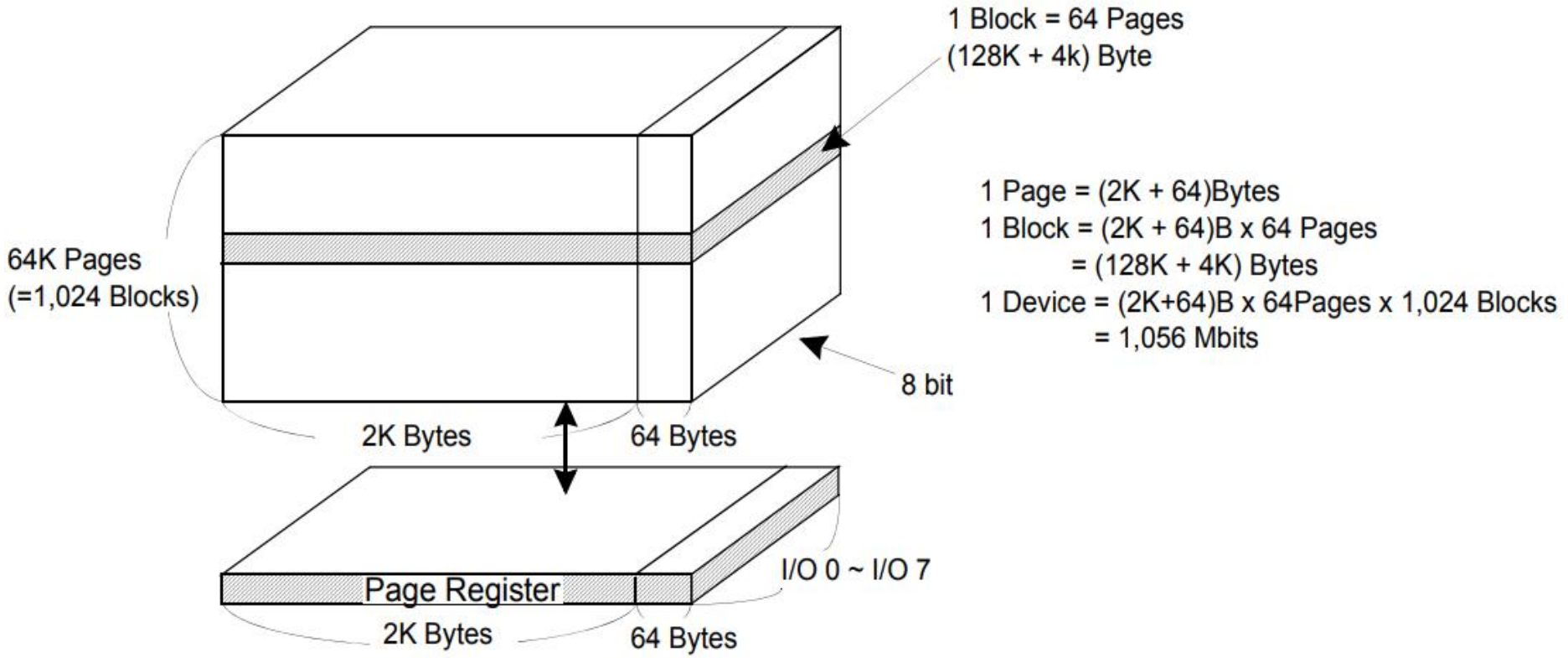
（a）不支持 NCQ；（b）支持 NCQ

- 固态硬盘
  - FLASH介质与DRAM介质
  - 接口规范与非固态硬盘一致
- 优点
  - 速度快；读取时间相对固定，固态硬盘寻址时间与数据存储位置无关；固态硬盘内部不存在任何机械活动部件；工作温度范围更大。
- 缺点
  - 成本高，每单位容量价格是传统硬盘的5～10倍（NAND）；容量低；写入寿命有限；数据损坏后难以恢复；低容量能耗不占优势。

FLASH特点

- 先擦后写。由于FLASH的写操作只能将数据位从1写成0，不能从0写成1，所以在对存储器写入之前，必须先执行擦除操作。
- 读写及擦除接口：对NAND FALSH以页Page为单位读（read）和写（program）数据，以块Block为单位进行擦除（erase）。
- 不能直接对目标地址进行总线操作，通过命令时序进行。
- 存在位翻转可能，通过校验进行处理。
- 坏块无法修复，只能标识。
- 有擦写次数限制。

Samsung 1G bit



[Figure 2] K9F1G08U0E Array Organization

感谢关注聆听！

张华平
Email: kevinzhang@bit.edu.cn
微博：@ICTCLAS张华平博士
实验室官网：
http://www.nlpir.org

大数据千人会