# 第2周 主存储器组织

第1讲 存储器基本概念

第2讲 主存的基本结构

第3讲 主存的性能指标

第4讲 半导体存储器组织

第5讲 内存条组织与总线宽度

第6讲 主存模块的连接与读写操作

# 回顾：程序及指令的执行过程

- 在内存存放的指令实际上是机器代码（0/1序列）

```
08048394 <add>:
1  8048394:  55          push   %ebp
2  8048395:  89  e5       mov    %esp, %ebp
3  8048397:  8b  45 0c    mov    0xc(%ebp), %eax
4  804839a:  03  45 08    add    0x8(%ebp), %eax
5  804839d:  5d          pop    %ebp
6  804839e:  c3          ret
```

栈是主存中的一个区域！

- 对于add函数的执行，以下情况下都需要访存：
  - ✓每条指令都需从主存单元取到CPU执行 取指
  - ✓PUSH指令需把寄存器内容压入栈中 存数 POP指令则相反 取数
  - ✓第3条mov指令需要从主存中取数后送到寄存器 取数
  - ✓第4条add指令需要从主存取操作数到ALU中进行运算 取数
  - ✓ret指令需要从栈中取出返回地址，以能正确回到调用程序执行 取数

访存是指令执行过程中一个非常重要的环节！ 取指、取数、存数

# 基本术语

- 记忆单元（存储基元 / 存储元 / 位元）（Cell）
  - 具有两种稳态的能够表示二进制数码0和1的物理器件
- 存储单元 / 编址单位（Addressing Unit）
  - 具有相同地址的位构成一个存储单元，也称为一个编址单位
- 存储体/ 存储矩阵 / 存储阵列（Bank）
  - 所有存储单元构成一个存储阵列
- 编址方式（Addressing Mode）
  - 字节编址、按字编址
- 存储器地址寄存器（Memory Address Register - MAR）
  - 用于存放主存单元地址的寄存器
- 存储器数据寄存器（ Memory Data Register-MDR (或MBR) ）
  - 用于存放主存单元中的数据的寄存器

# 存储器分类

## 依据不同的特性有多种分类方法

### （1）按工作性质/存取方式分类

- 随机存取存储器 Random Access Memory (RAM)
  - 每个单元读写时间一样，且与各单元所在位置无关。如：内存。

    （注：原意主要强调地址译码时间相同。现在的DRAM芯片采用行缓冲，因而可能因为位置不同而使访问时间有所差别。）
- 顺序存取存储器 Sequential Access Memory (SAM)
  - 数据按顺序从存储载体的始端读出或写入，因而存取时间的长短与信息所在位置有关。例如：磁带。
- 直接存取存储器 Direct Access Memory(DAM)
  - 直接定位到读写数据块，在读写数据块时按顺序进行。如磁盘。
- 相联存储器 Associate Memory（AM）

  Content Addressed Memory (CAM)
  - 按内容检索到存储位置进行读写。例如：快表。

# 存储器分类

（2）按存储介质分类

半导体存储器：双极型，静态MOS型，动态MOS型

磁表面存储器：磁盘（Disk）、磁带 （Tape）

光存储器：CD，CD-ROM，DVD

（3）按信息的可更改性分类

读写存储器（Read / Write Memory)：可读可写

只读存储器（Read Only Memory)：只能读不能写

（4）按断电后信息的可保存性分类

非易失（不挥发）性存储器(Nonvolatile Memory)

信息可一直保留， 不需电源维持。

（如 ：ROM、磁表面存储器、光存储器等）

易失（挥发）性存储器(Volatile Memory)

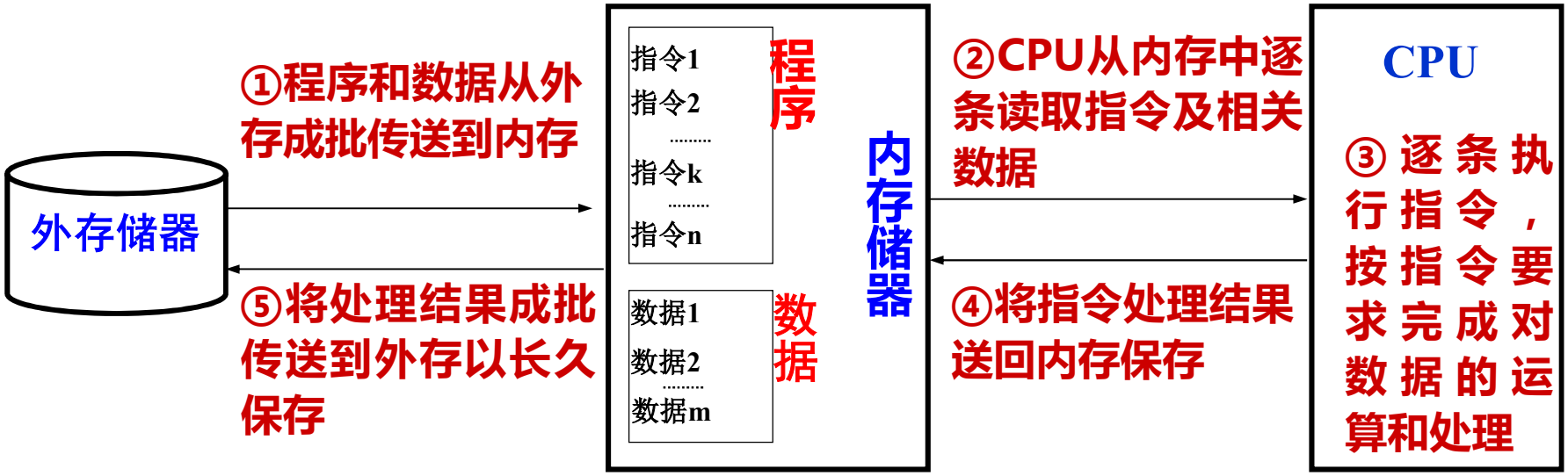
电源关闭时信息自动丢失。（如：RAM、Cache等）

# 存储器分类

（5）按功能/容量/速度/所在位置分类

- 寄存器(Register)
  - 封装在CPU内，用于存放当前正在执行的指令和使用的数据
  - 用触发器实现，速度快，容量小（几~几十个）
- 高速缓存(Cache)
  - 位于CPU内部或附近，用来存放当前要执行的局部程序段和数据
  - 用SRAM实现，速度可与CPU匹配，容量小（几MB）
- 内存储器MM（主存储器Main (Primary) Memory）
  - 位于CPU之外，用来存放已被启动的程序及所用的数据
  - 用DRAM实现，速度较快，容量较大（几GB）
- 外存储器AM (辅助存储器Auxiliary / Secondary Storage)
  - 位于主机之外，用来存放暂不运行的程序、数据或存档文件
  - 用磁盘、SSD等实现，容量大而速度慢

# 内存与外存的关系及比较




✓ 外存储器（简称外存或辅存）

- 存取速度慢
- 成本低、容量很大
- 不与CPU直接连接，先传送到内存，然后才能被CPU使用。
- 属于非易失性存储器，用于长久存放系统中几乎所有的信息

✓ 内存储器（简称内存或主存）

- 存取速度快
- 成本高、容量相对较小
- 直接与CPU连接，CPU对内存中可直接进行读、写操作
- 属于易失性存储器(volatile)，用于临时存放正在运行的程序和数据

# 主存的结构

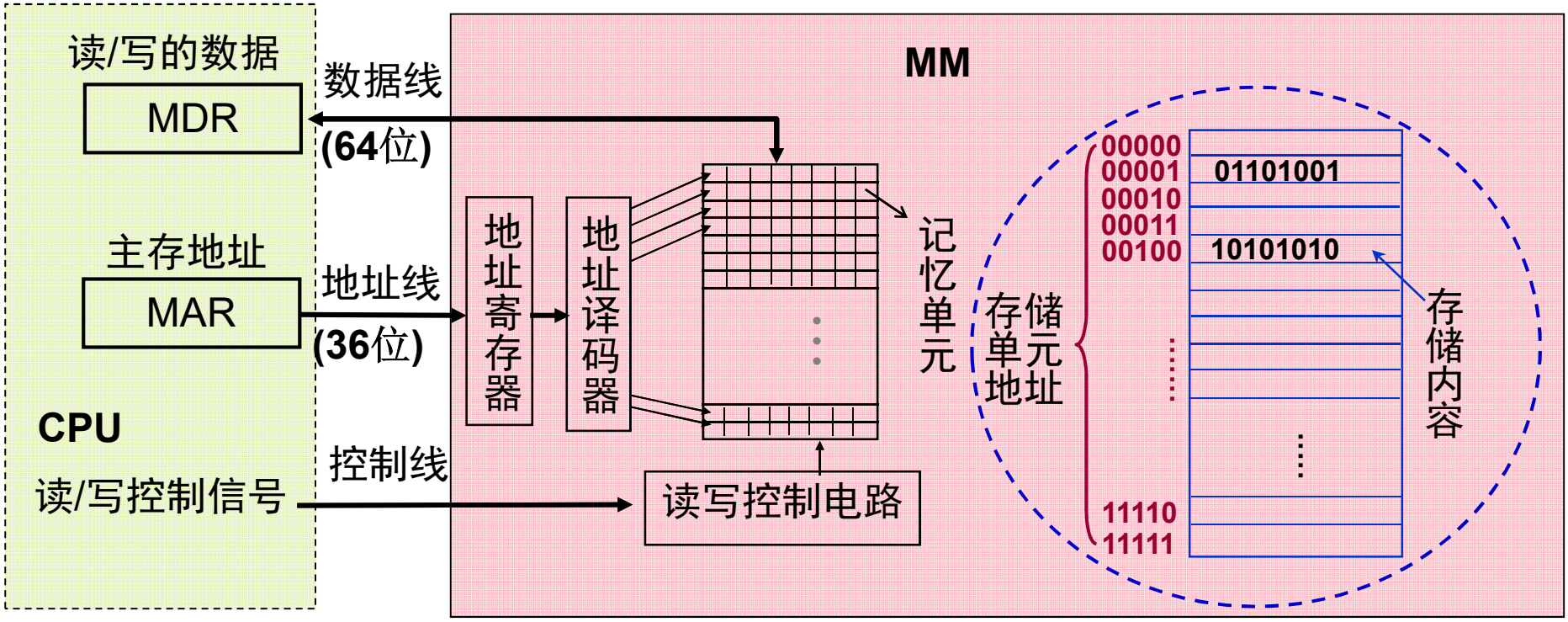
**问题：主存中存放的是什么信息？CPU何时会访问主存？**

**指令及其数据！CPU执行指令时需要取指令、取数据、存数据！**

**问题：地址译码器的输入是什么？输出是什么？可寻址范围多少？**

**输入是地址，输出是地址驱动信号（只有一根地址驱动线被选中）。**

**可寻址范围为0~$2^{36}$-1，即主存地址空间为64GB（按字节编址时）。**




**主存地址空间大小不等于主存容量（实际安装的主存大小）！**

**若是字节编址，则每次最多可读/写8个单元，给出的是首(最小)地址.**

# 主存的主要性能指标

- 性能指标：
  - 按字节连续编址，每个存储单元为1个字节（8个二进位）
  - 存储容量：所包含的存储单元的总数（单位：MB或GB）
  - 存取时间$T_A$：从CPU送出内存单元的地址码开始，到主存读出数据并送到CPU（或者是把CPU数据写入主存）所需要的时间（单位：ns，1 ns = $10^{-9}$ s），分读取时间和写入时间
  - 存储周期$T_{MC}$：连读两次访问存储器所需的最小时间间隔，它应等于存取时间加上下一次存取开始前所要求的附加时间，因此，$T_{MC}$比$T_A$大（ 因为存储器由于读出放大器、驱动电路等都有一段稳定恢复时间，所以读出后不能立即进行下一次访问。 ）

（就像一趟火车运行时间和发车周期是两个不同概念一样。）

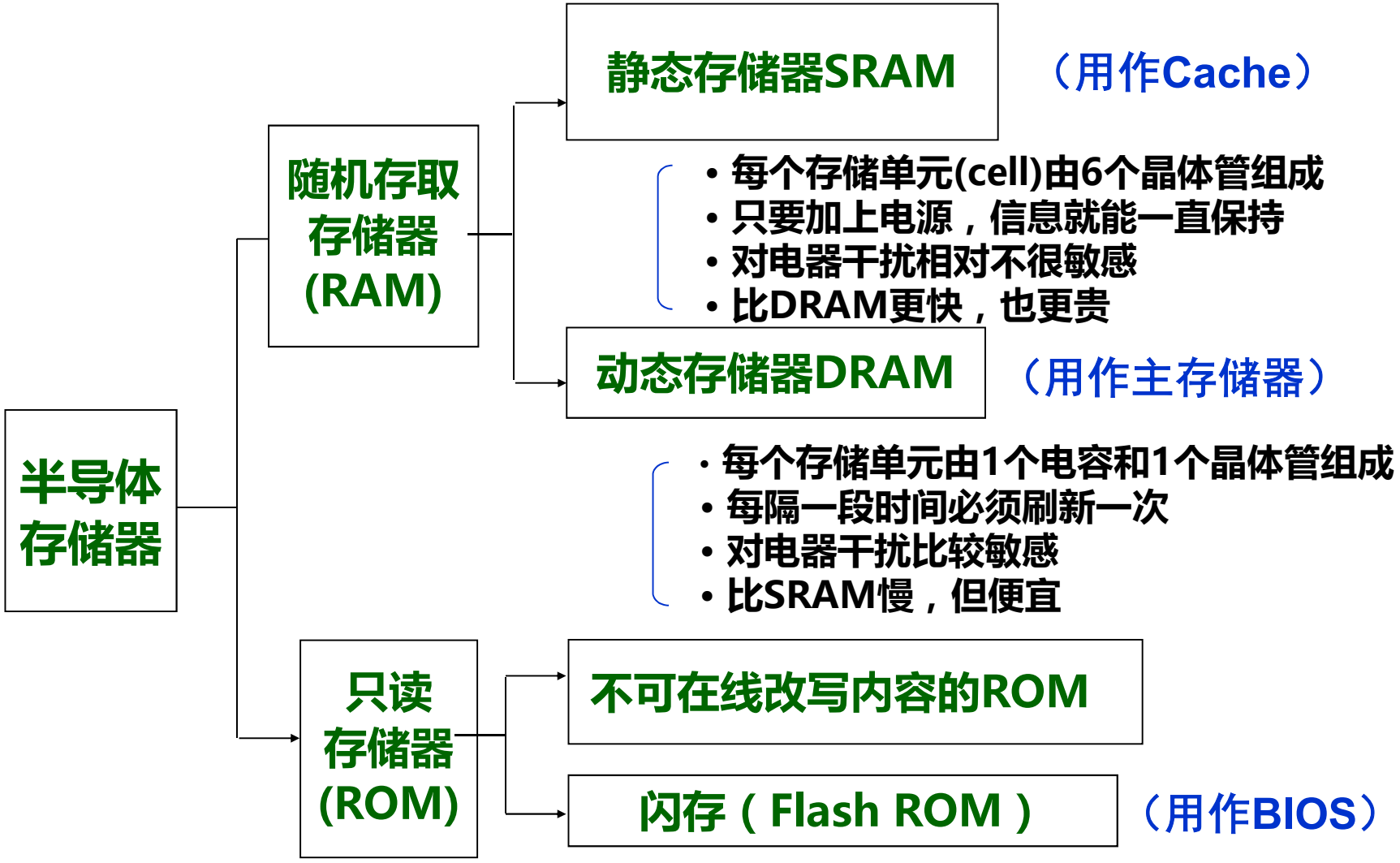
# 时间、存储容量（或带宽）的单位

## Notations and Conventions for Numbers

| Prefix | Abbreviation | Meaning | Numeric Value |
|---|---|---|---|
| mill | m | One thousandth | $10^{-3}$ |
| micro | μ | One millionth | $10^{-6}$ |
| nano | n | One billionth | $10^{-9}$ |
| pico | p | One trillionth | $10^{-12}$ |
| femto | f | One quadrillionth | $10^{-15}$ |
| atta | a | One quintillionth | $10^{-18}$ |
| kilo | K (or k) | Thousand | $10^{3}$ or $2^{10}$ |
| mega | M | Million | $10^{6}$ or $2^{20}$ |
| giga | G | Billion | $10^{9}$ or $2^{30}$ |
| tera | T | Trillion | $10^{12}$ or $2^{40}$ |
| peta | P | Quadrillion | $10^{15}$ or $2^{50}$ |
| exa | E | Quintillion | $10^{18}$ or $2^{60}$ |

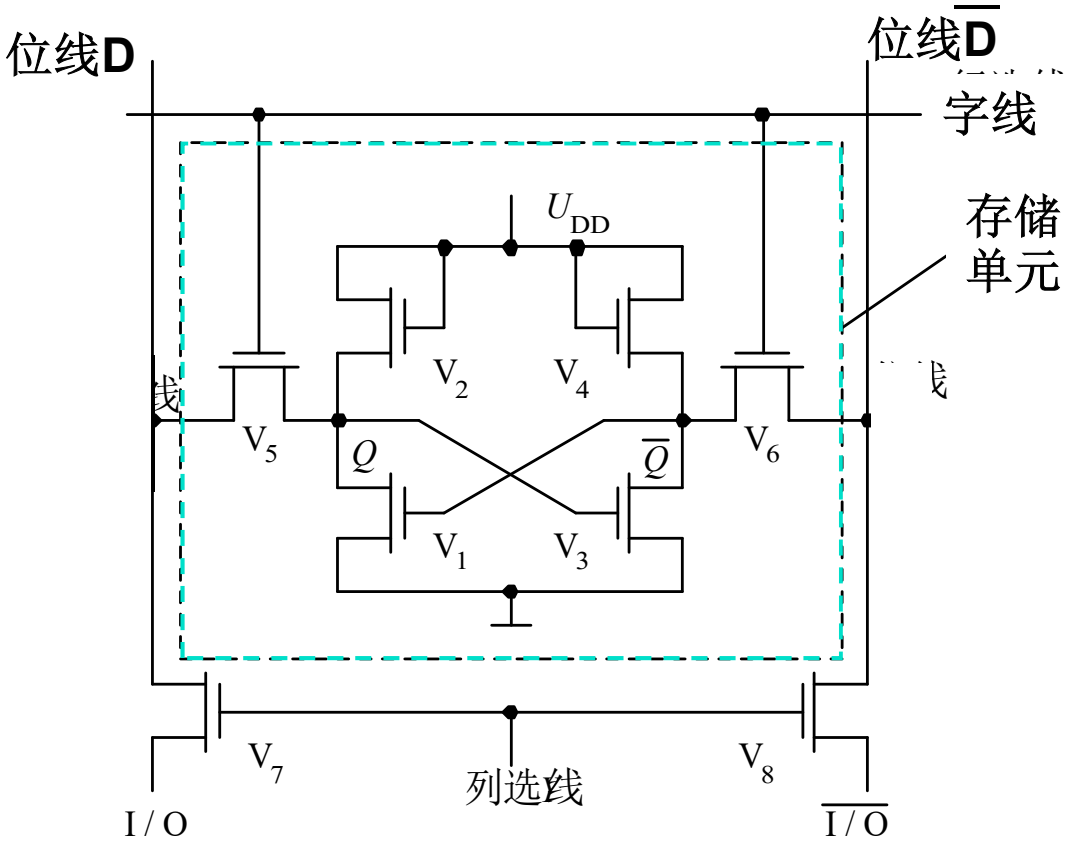
# 内存储器的分类及应用

- 内存由半导体存储器芯片组成，芯片有多种类型：



# 六管静态MOS管电路（不作要求）

## 6管静态NMOS记忆单元




**SRAM中数据保存在一对正负反馈门电路中，只要供电，数据就一直保持，不是破坏性读出，也无需重写，即无需刷新！**

信息存储原理： 看作带时钟的RS触发器

**保持时：**
- 字线为0（低电平）

**写入时：**
- 位线上是被写入的二进位信息0或1
- 置字线为1
- 存储单元(触发器)按位线的状态设置成0或1

**读出时：**
- 置2个位线为高电平
- 置字线为1
- 存储单元状态不同，位线的输出不同

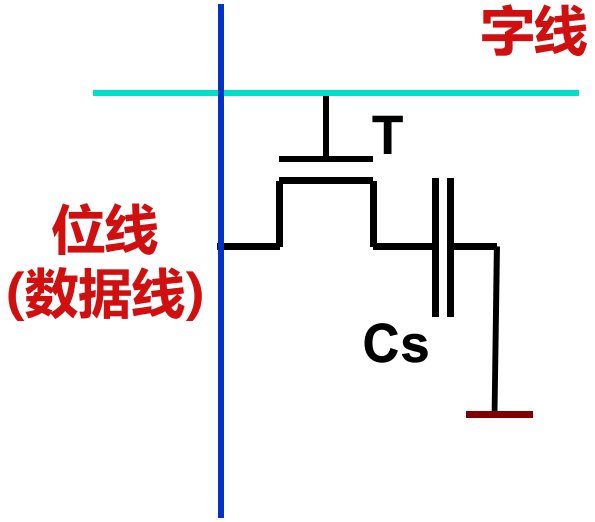
# 动态单管记忆单元电路（不作要求）

**读写原理：字线上加高电平，使T管导通。**

**写“0”时，数据线加低电平，使$C_S$上电荷对数据线放电；**

**写“1”时，数据线加高电平，使数据线对$C_S$充电；**

**读出时，数据线上有一读出电压。它与$C_S$上电荷量成正比。**




**优点：电路元件少，功耗小，集成度高，用于构建主存储器**

**缺点：速度慢、是破坏性读出（需读后再生）、需定时刷新**

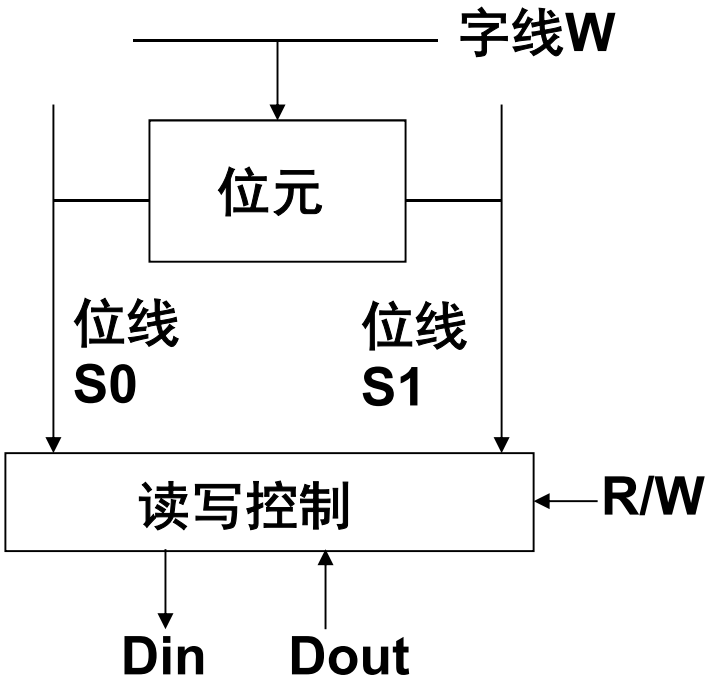
**刷新：DRAM的一个重要特点是，数据以电荷的形式保存在电容中，电容的放电使得电荷通常只能维持几十个毫秒左右，相当于1M个时钟周期左右，因此要定期进行刷新（读出后重新写回），按行进行（所有芯片中的同一行一起进行），刷新操作所需时间通常只占1%~2%左右。**

# 半导体RAM的组织

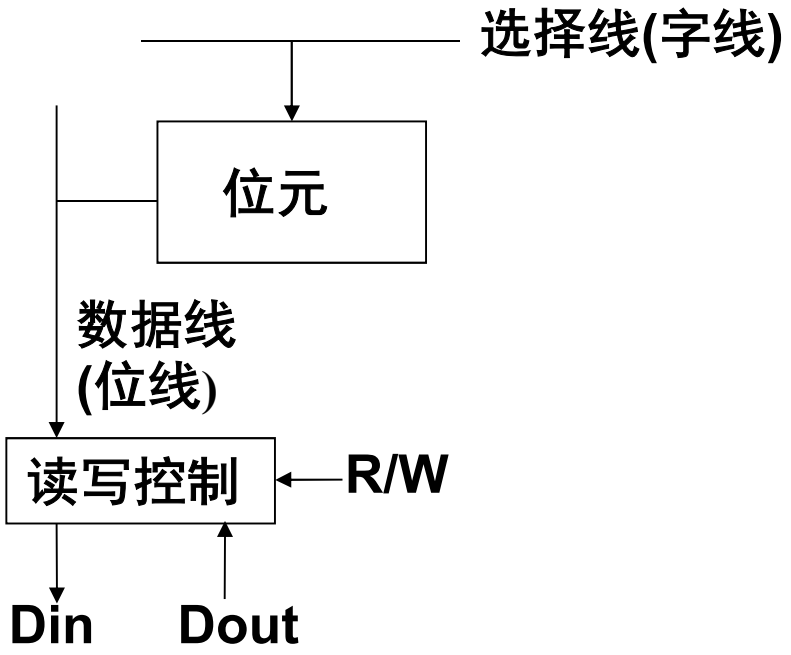
记忆单元(Cell)→存储器芯片(Chip)→ 内存条（存储器模块）

存储体(Memory Bank)： 由记忆单元(位元)构成的存储阵列

记忆单元的组织：

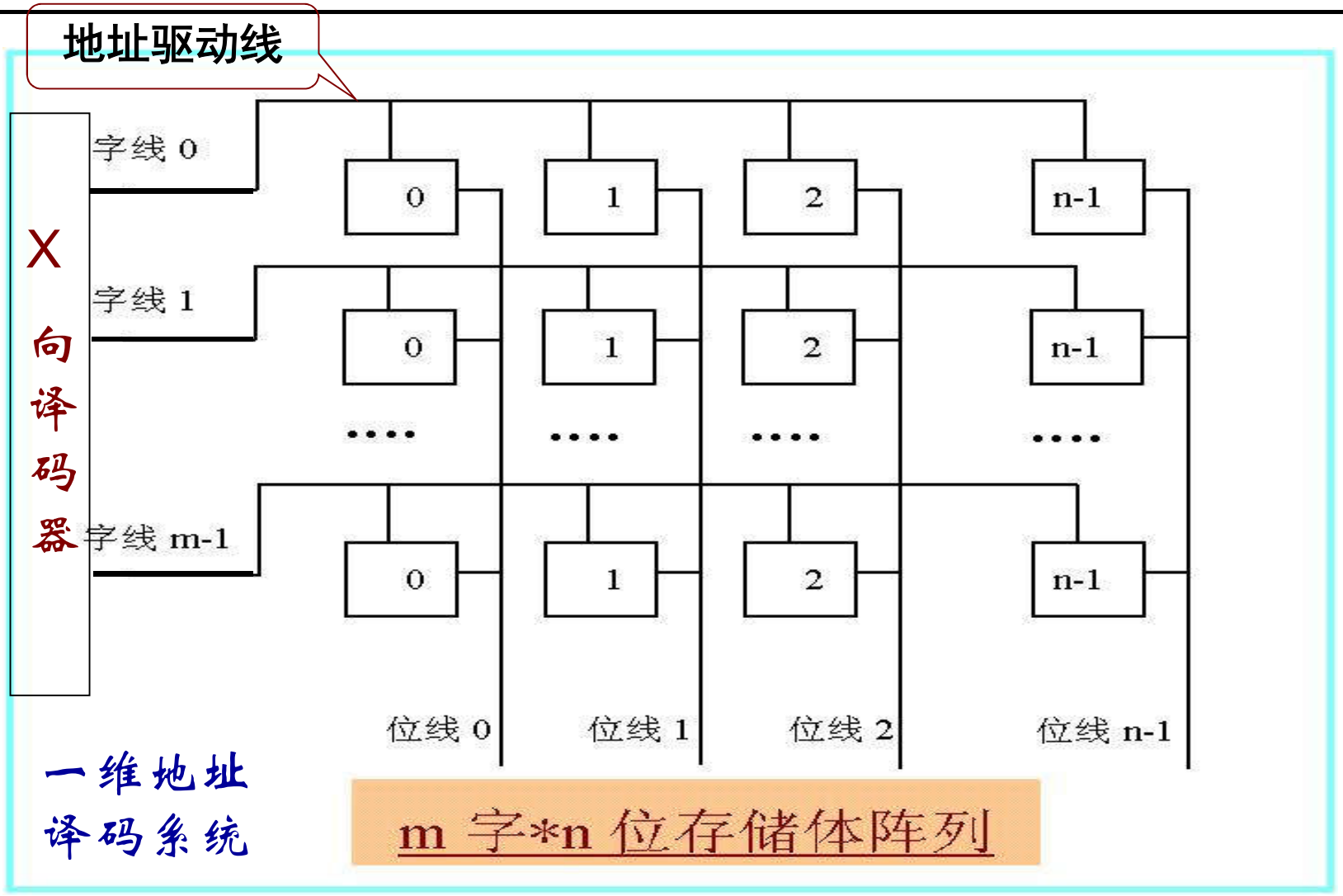



SRAM



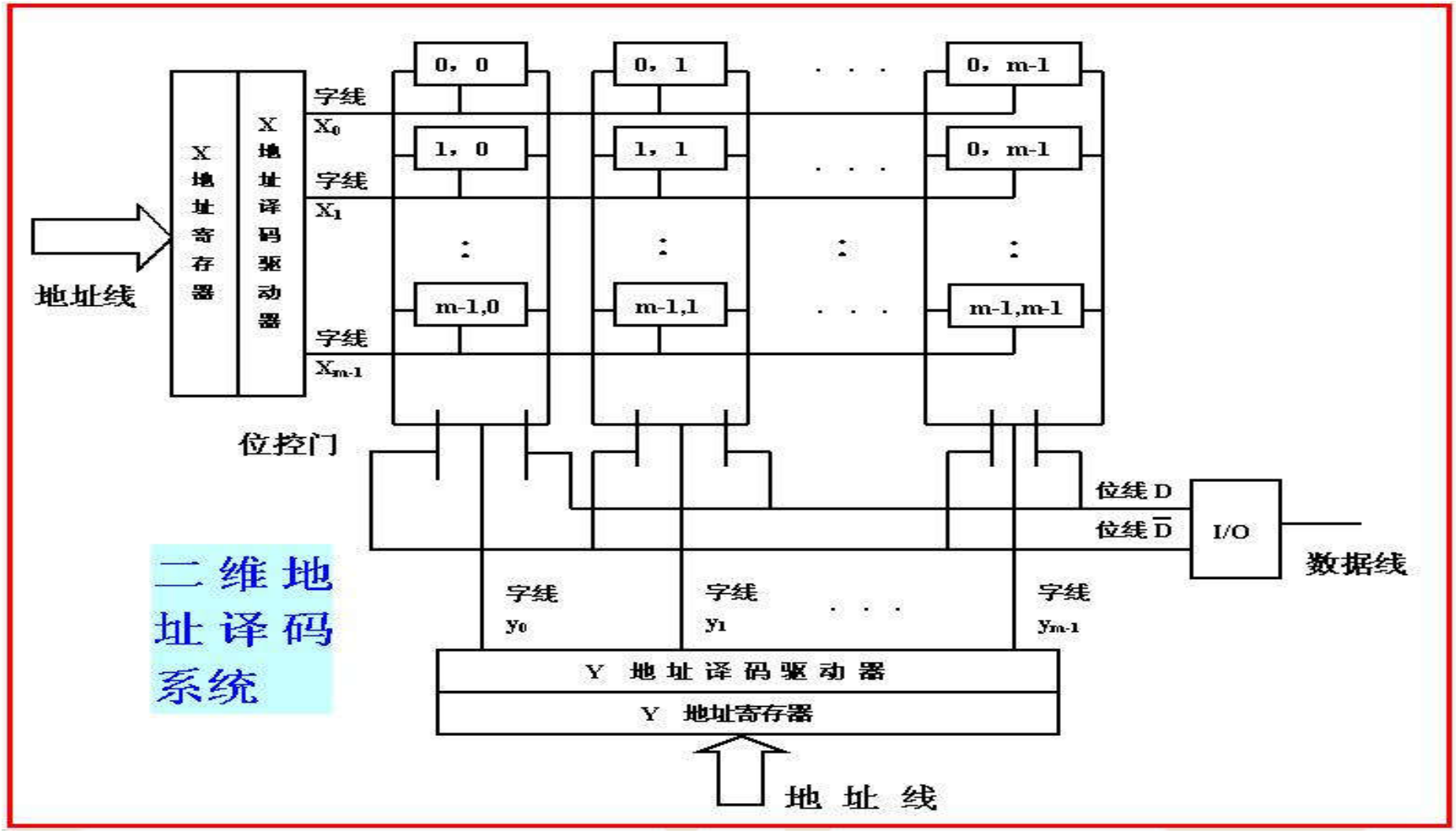

DRAM

# 字片式存储体阵列组织（不作要求）




**一般SRAM为字片式芯片，只在x向上译码，同时读出字线上所有位！**

# 位片式存储体阵列组织（不作要求）




**位片式在字方向和位方向扩充，需要有片选信号**
**DRAM芯片都是位片式**

# 举例：典型的16M位DRAM（4Mx4）

16M位 = 4Mbx4 = 2048x2048x4 = $2^{11}$x$2^{11}$x4

(1) 地址线：11根线分时复用，由RAS和CAS提供控制时序。

(2) 需4个位平面，对相同行、列交叉点的4位一起读/写

(3) 内部结构框图

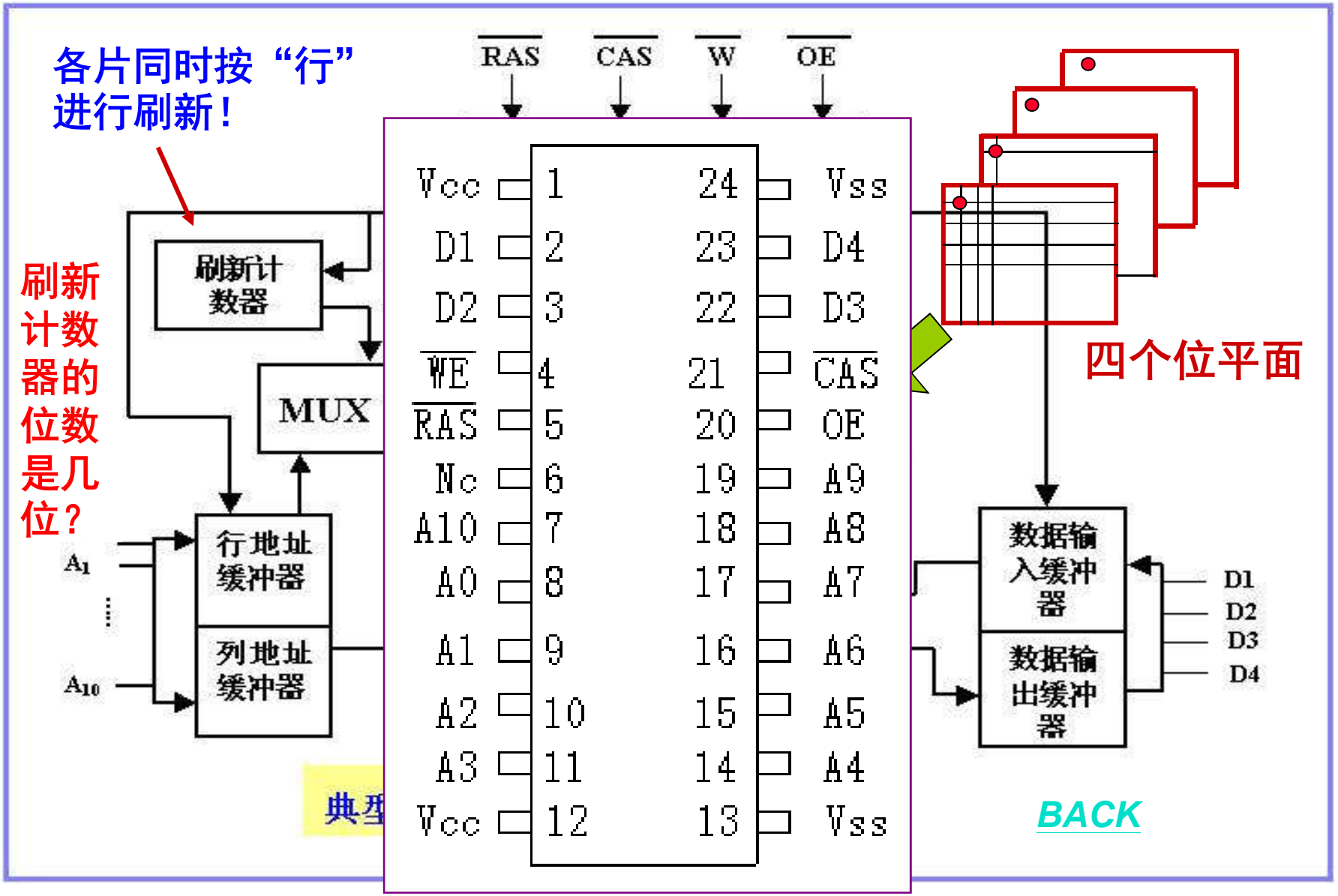
**问题：**

**为什么每出现新一代DRAM芯片，容量至少提高到4倍？**

行地址和列地址分时复用，每出现新一代DRAM芯片，至少要增加一根地址线。每加一根地址线，则行地址和列地址各增加一位，所以行数和列数各增加一倍。因而容量至少提高到4倍。

SKIP

# 举例：典型的16M位DRAM（4Mx4）

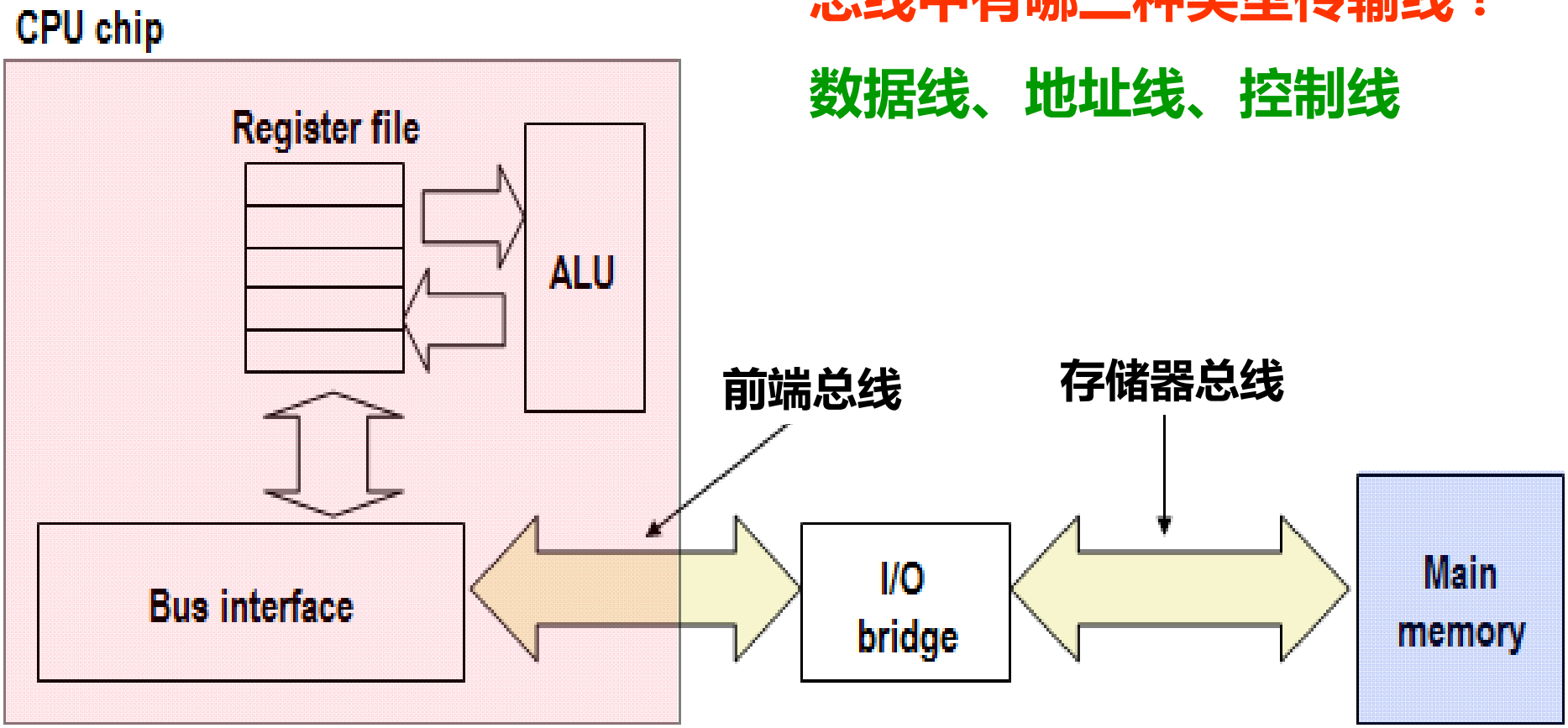


# 主存模块的连接和读写操作

- 主存与CPU的连接

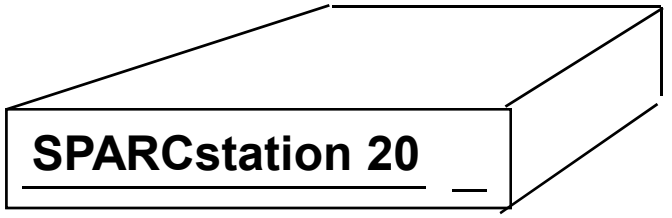
总线中有哪三种类型传输线？

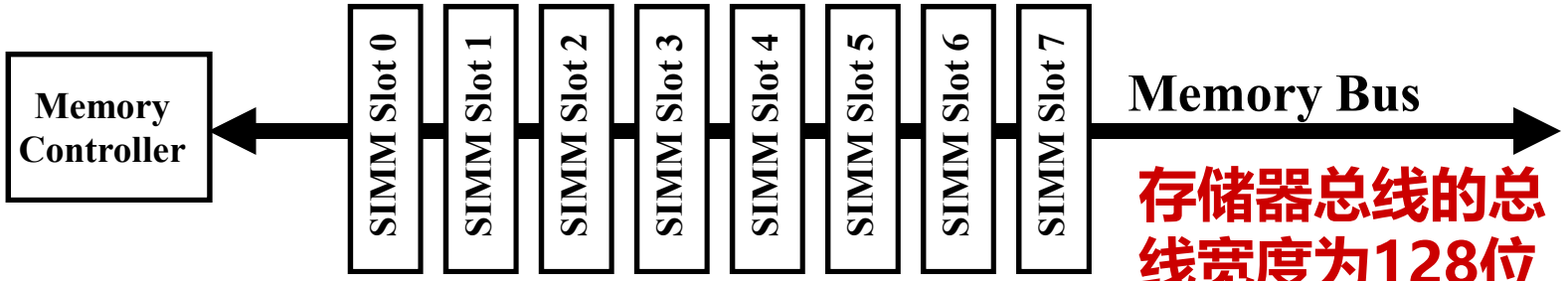
数据线、地址线、控制线



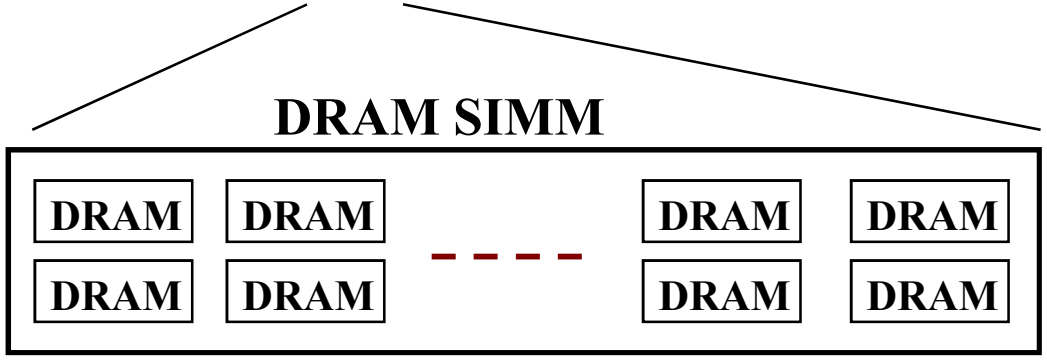
# 举例：SPARCstation 20's Memory Module

**总线宽度**是指总线中**数据线**的条数



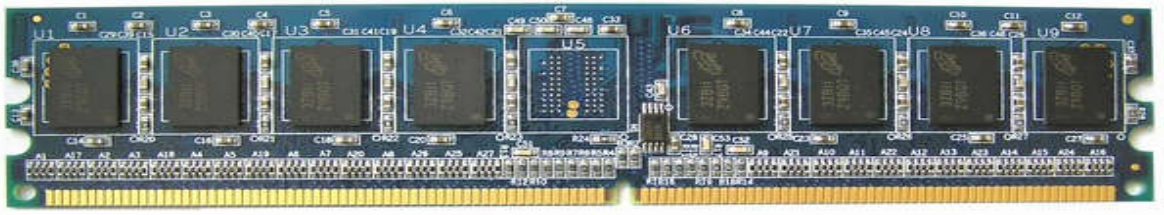

**存储器总线的总线宽度为128位**




**每个内存条最多能同时读出128位数据**

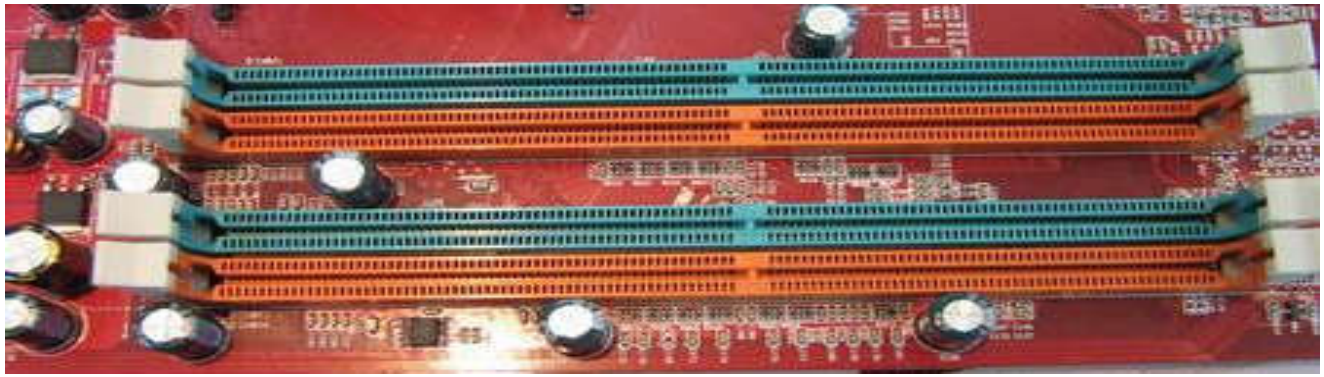
**每次访存操作总是在某一个内存条内进行！**

# PC机主存储器的物理结构



- 由若干内存条组成
- 内存条的组成：

  把若干片DRAM芯片焊装在一小条印制电路板上制成
- 内存条必须插在主板上的内存条插槽中才能使用



目前流行的是DDR2、DDR3内存条：

- 采用双列直插式，其触点分布在内存条的两面
- DDR条有184个引脚，DDR2有240个引脚
- PC机主板中一般都配备有2个或4个DIMM插槽

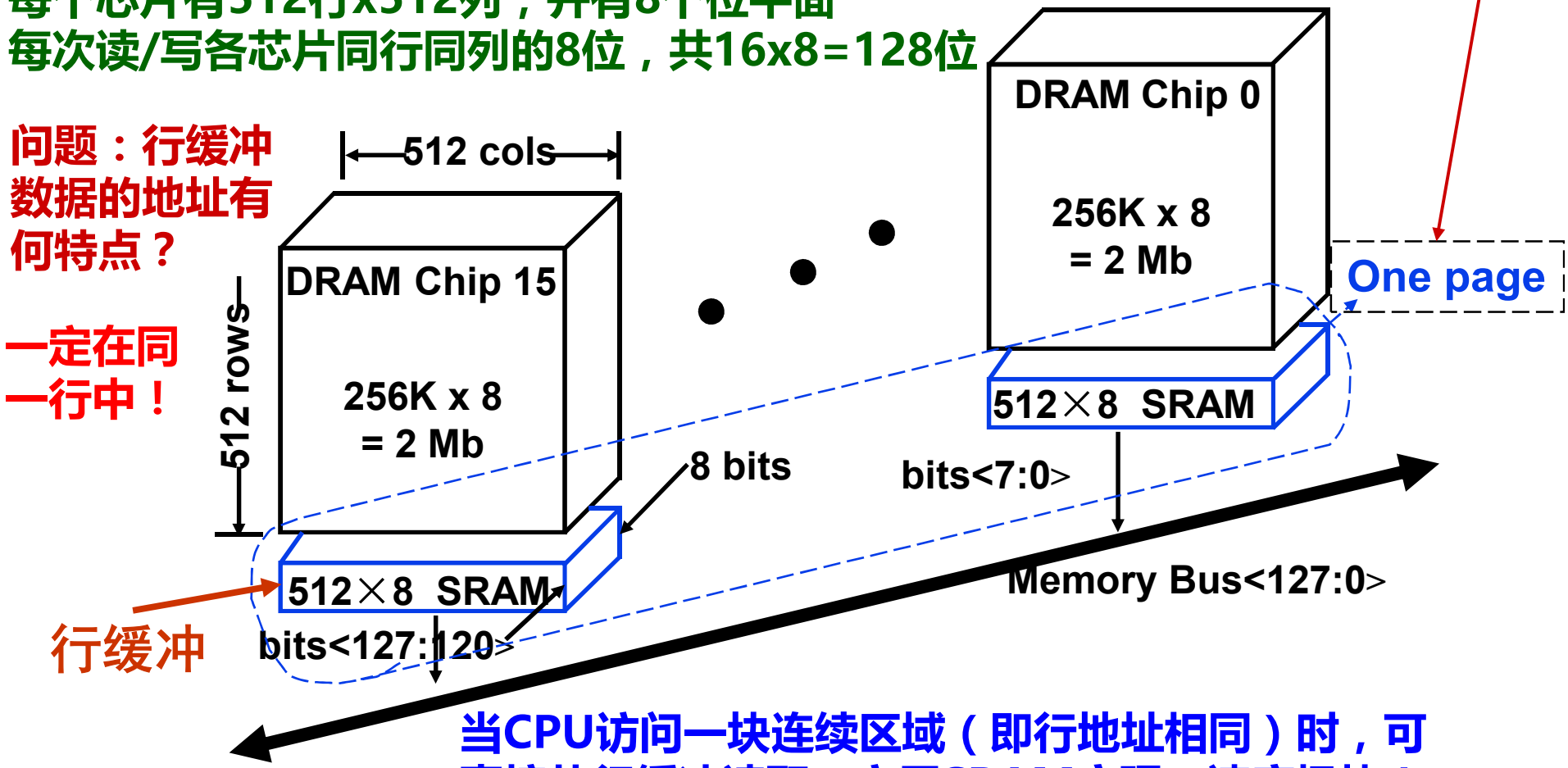
# 举例：SPARCstation 20’s内存条(模块)

- one memory module（内存条）
  - Smallest: 4 MB = 16x 2Mb DRAM chips, 8 KB of Page SRAM
  - Biggest: 64 MB = 32x 16Mb chips, 16 KB of Page SRAM

每个芯片有512行x512列，并有8个位平面
每次读/写各芯片同行同列的8位，共16x8=128位

问题：行缓冲数据的地址有何特点？

一定在同一行中！

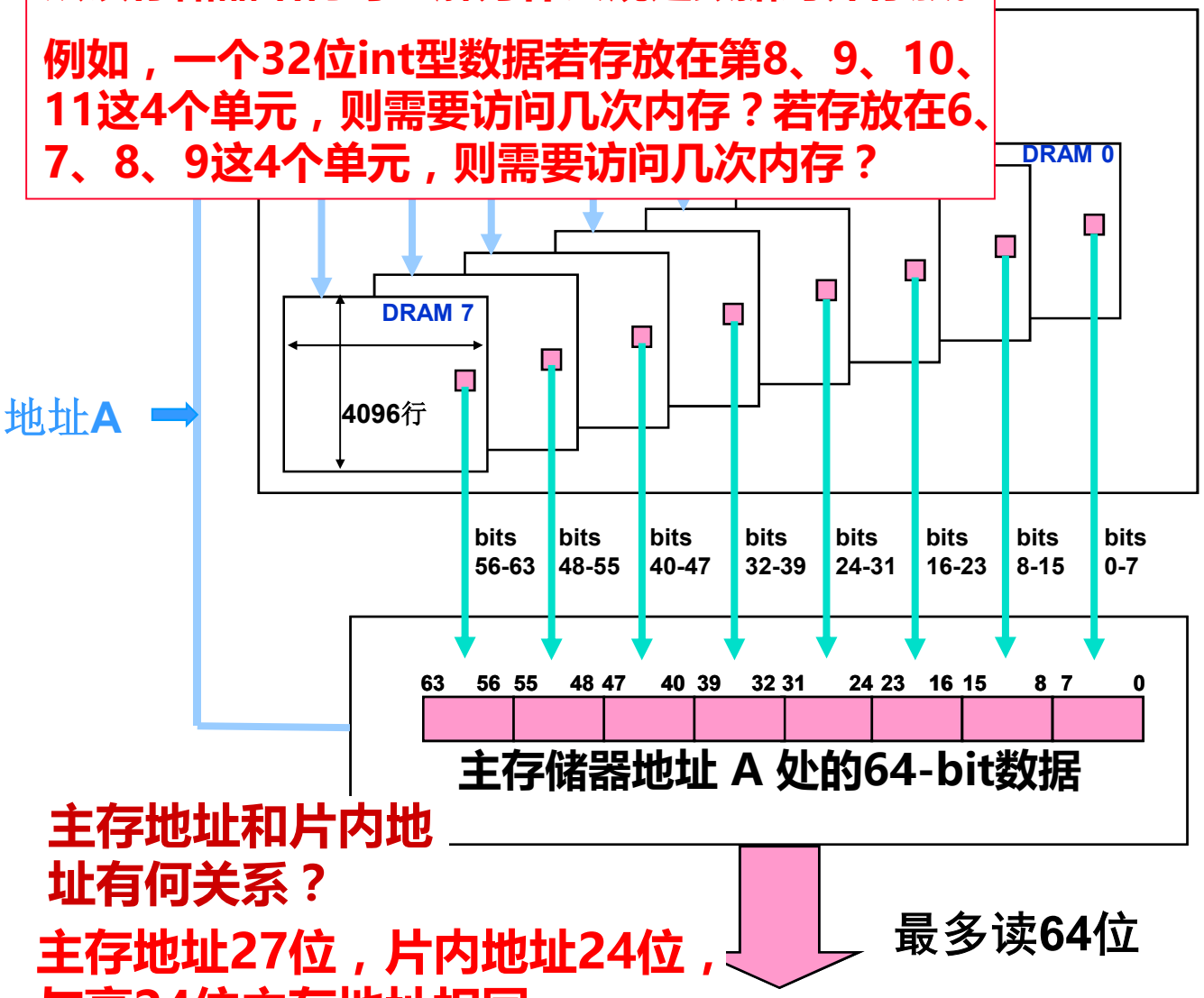



当CPU访问一块连续区域（即行地址相同）时，可直接从行缓冲读取，它用SRAM实现，速度极快！

# 举例：128MB的DRAM存储器

从该存储器结构可理解为什么规定数据对齐存放。

例如，一个32位int型数据若存放在第8、9、10、11这4个单元，则需要访问几次内存？若存放在6、7、8、9这4个单元，则需要访问几次内存？




分别访问1次和2次

- 由8片DRAM芯片构成
- 每片 16Mx8 bits
- 行地址、列地址各12位
- 每行共4096列(8位/列)
- 选中某一行并读出之后再由列地址选择其中的一列(8个二进位) 送出

芯片内地址是否连续？

不连续，交叉编址，可同时读写所有芯片。

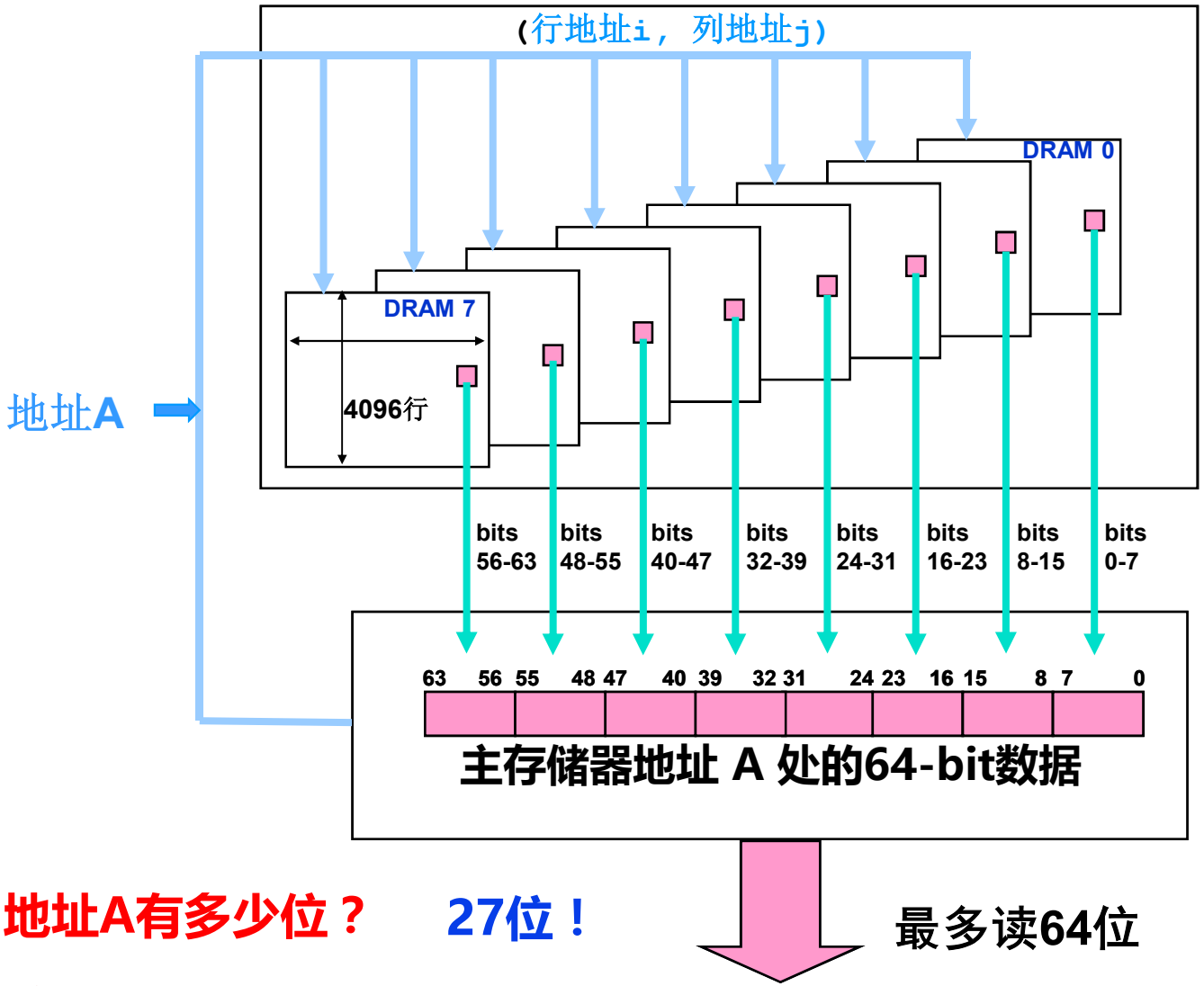
- 行、列地址为(i,j)的8个单元

主存地址和片内地址有何关系？

主存地址27位，片内地址24位，与高24位主存地址相同。

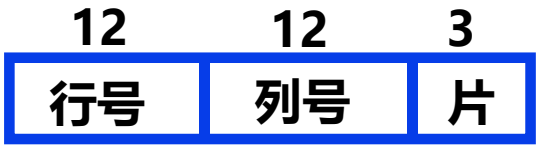
主存低3位地址的作用是什么？ 确定8个字节中的哪个，即用来选片。

# 复习：128MB的DRAM存储器




- 行、列地址为(i,j)的8个单元

地址A如何划分？




低3位用来选片

地址A有多少位？ 27位！

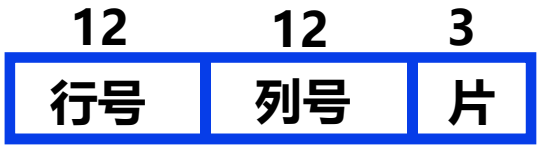
假定首地址为i，则地址分布如下：

在DRAM行缓冲中数据的地址有何特点？

同一行地址连续，共8*4096B=$2^{15}$B=32768个单元

# 复习：128MB的DRAM存储器

地址A如何划分？　低3位用来选片




在DRAM行缓冲中数据的地址有何特点？

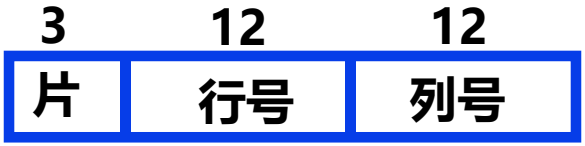
假定首地址为i（i=32768*k，k为行号 ），则地址分布如下：

| | Chip0 | Chip1 | … | Chip7 |
|---|---|---|---|---|
| 第0列 | i | i+1 | | i+7 |
| 第1列 | i+8 | i+9 | | i+15 |
| … | | | | |
| 第4095列 | i+8*4095 | i+1+8*4095 | | i+7+8*4095 |

地址连续，共8*4096B=$2^{15}$B=32768个单元

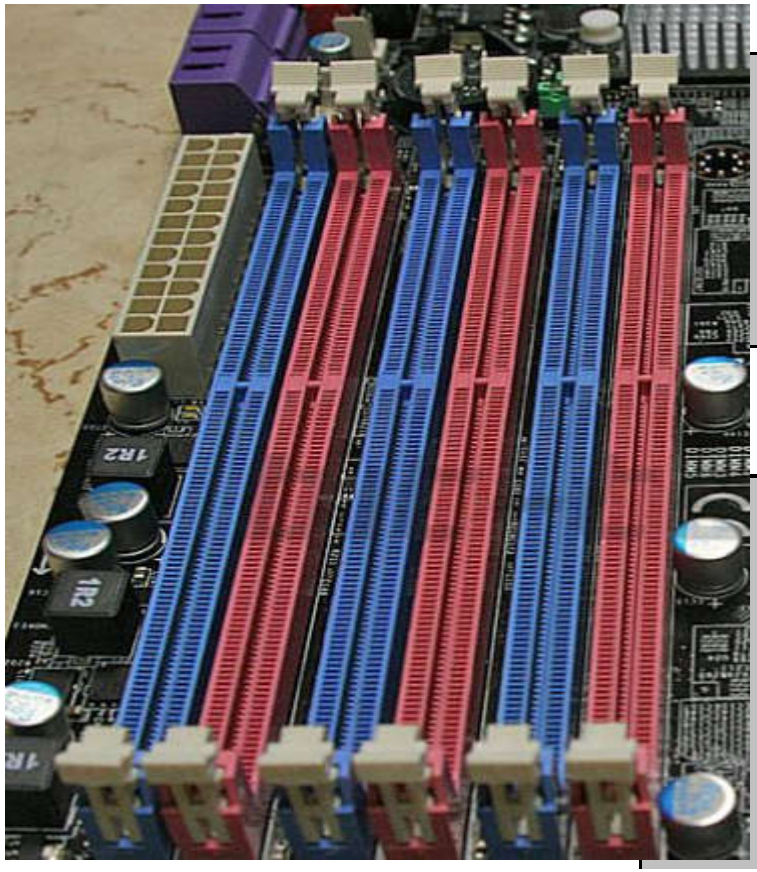
通常，一个主存块包含在行缓冲中
可降低Cache缺失损失

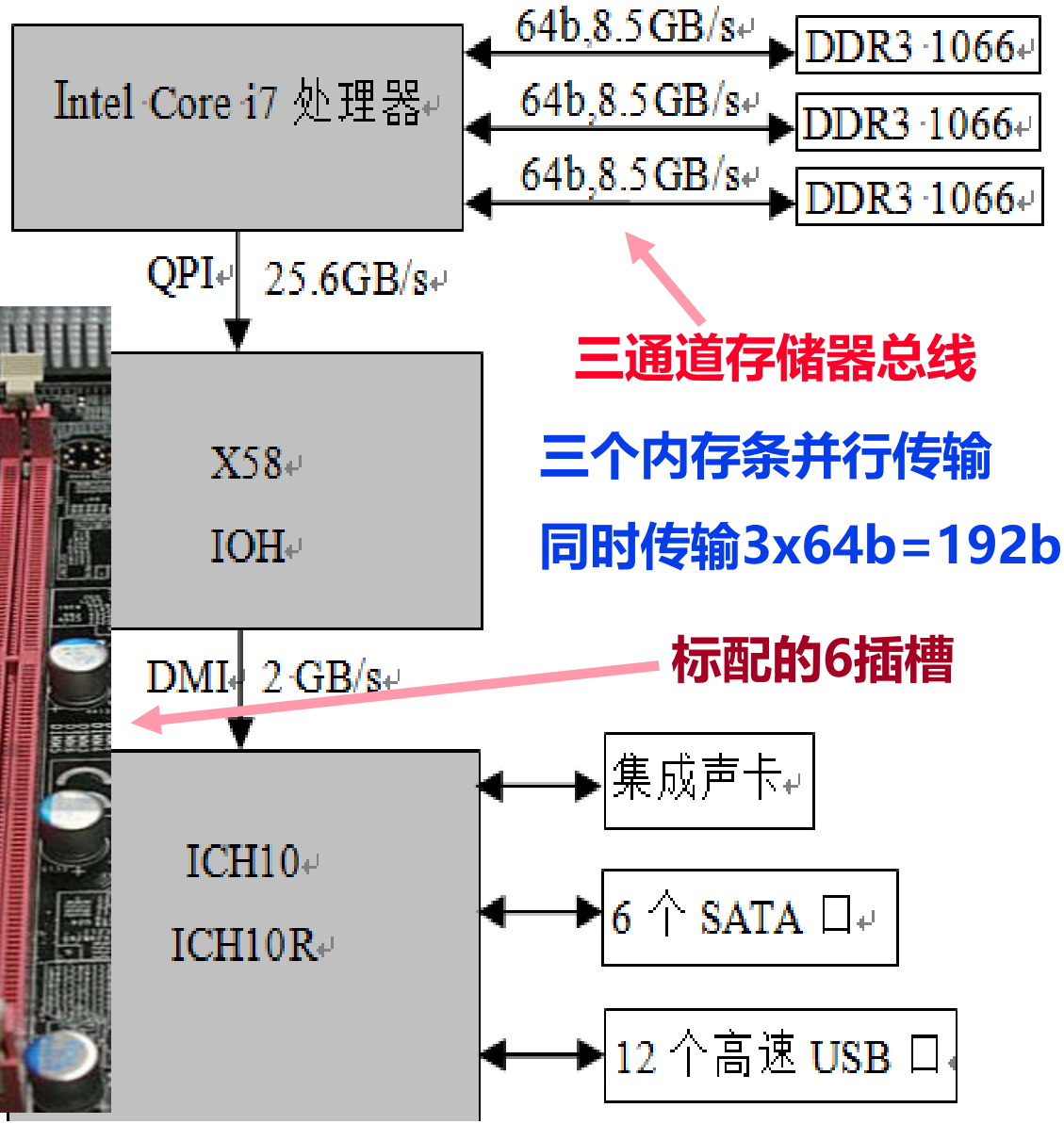
如果片内地址连续，则地址A如何划分？



# 计算机系统互连

只要将同色的三个内存插槽插上内存条，系统便会自动识别进入三通道模式






三通道存储器总线

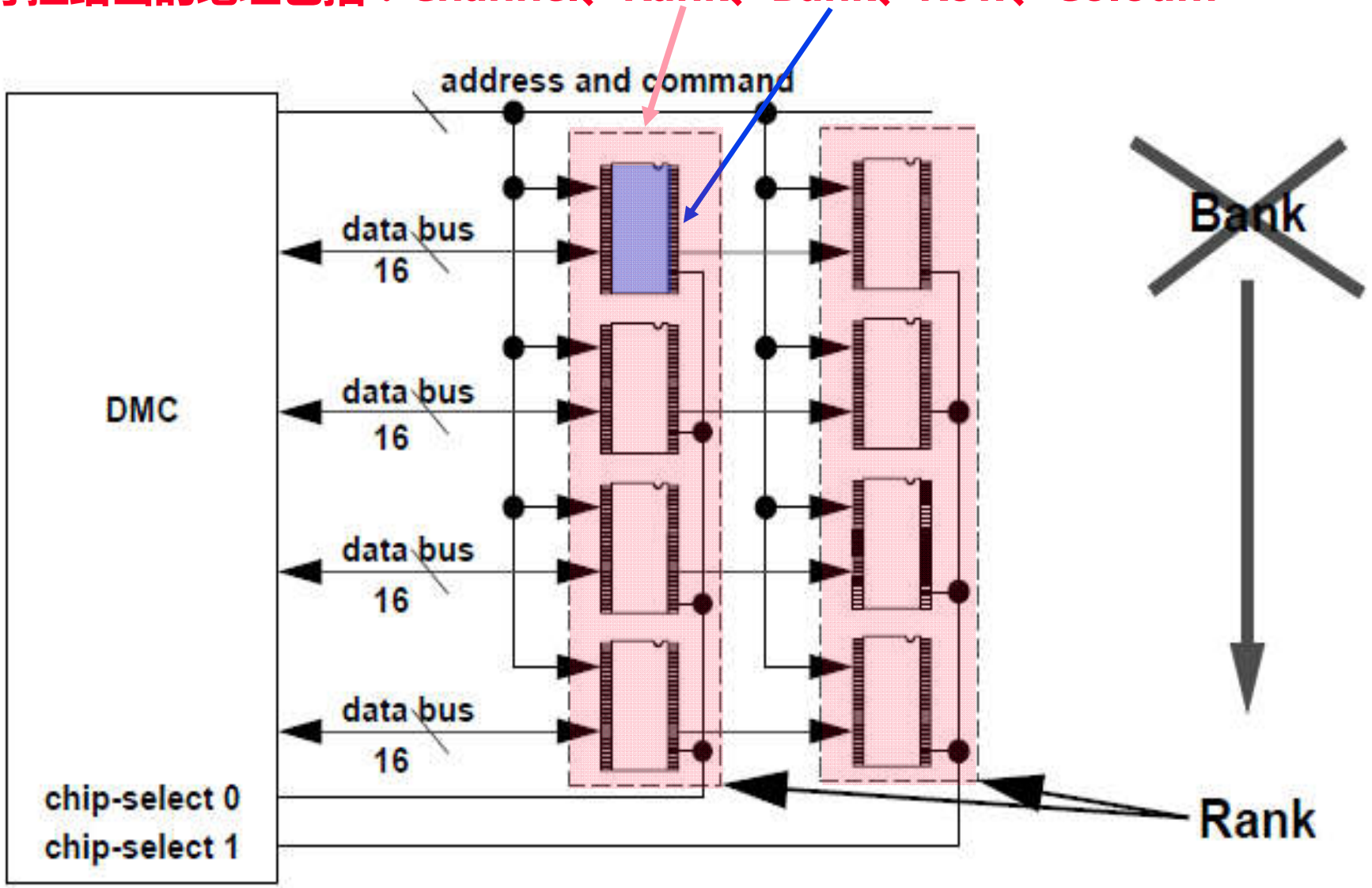
三个内存条并行传输

同时传输3x64b=192b

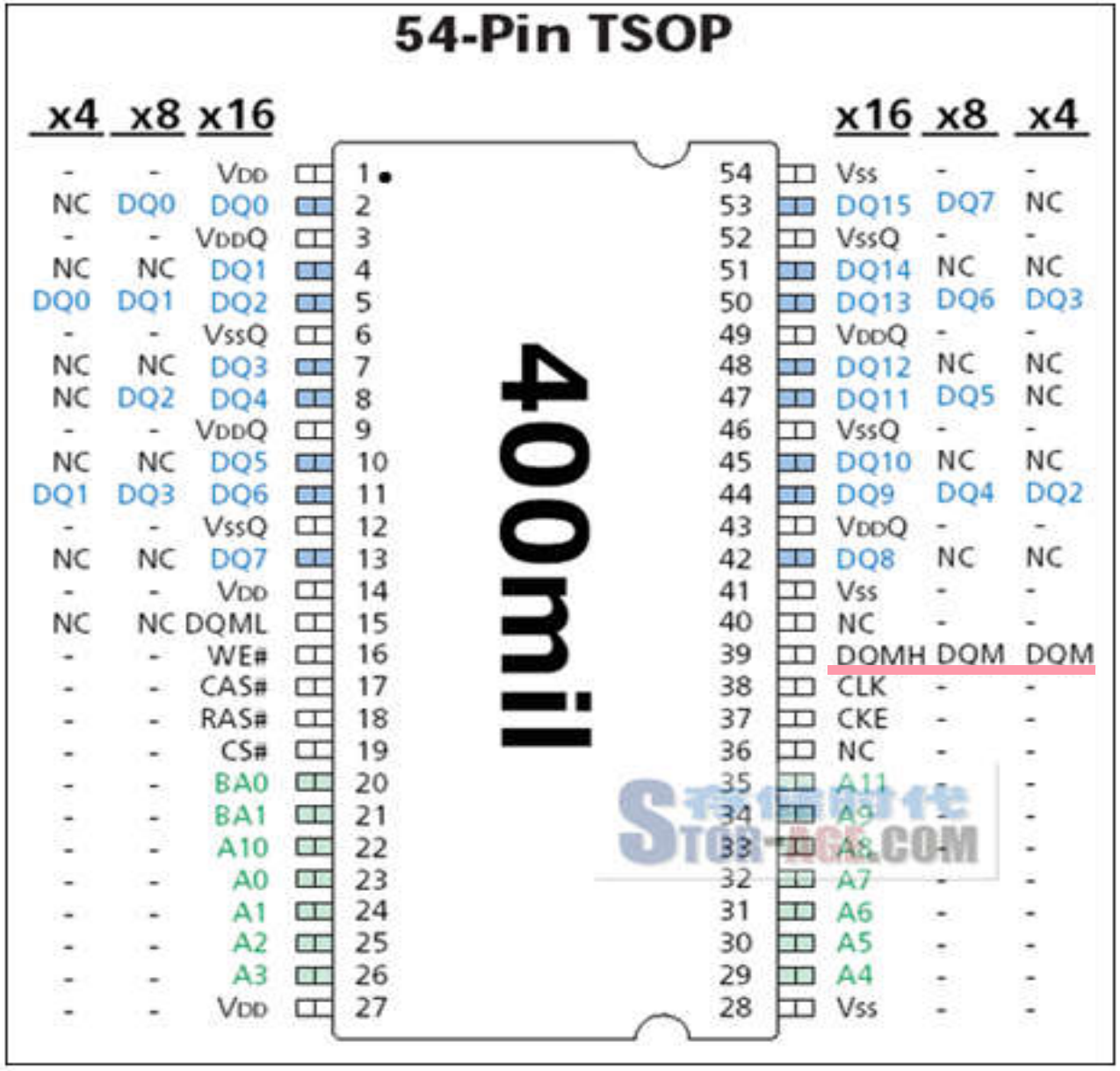
标配的6插槽

# DRAM内存条结构

存控给出的地址包括：Channel、Rank、Bank、Row、Coloum



# SDRAM芯片的引脚




DQM（数据掩码信号）：用于选择Burst传输中的哪个数据，比如，burst长度是4时，则表示需要传输4个64-bit，此时DQM选择需要传输哪几个64bit。

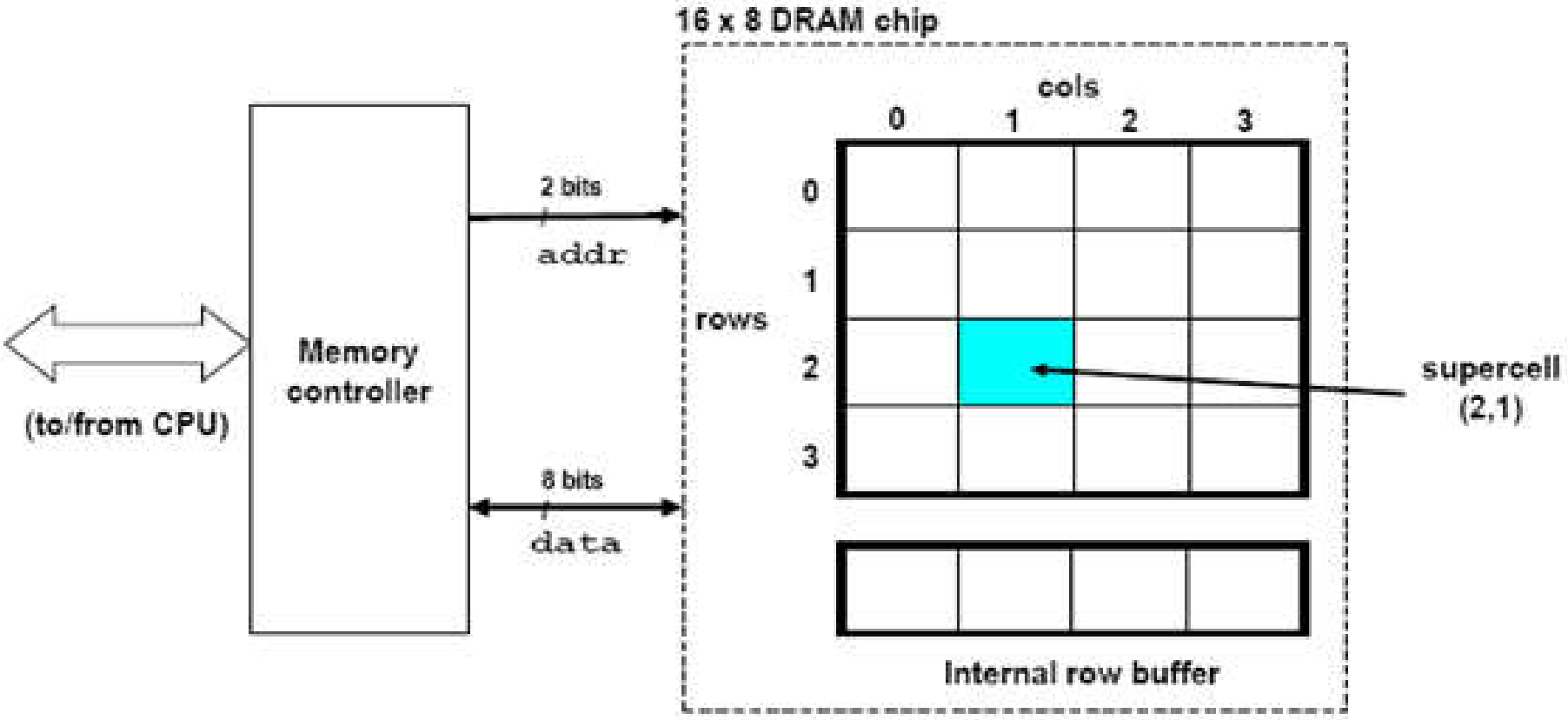
# DRAM芯片的规格

- 若一个$2^n \times b$位DRAM芯片的存储阵列是r行×c列，则该芯片容量为$2^n \times b$位且$2^n = r \times c$。如：16K×8位DRAM，则r=c=128。
- 芯片内的地址位数为n，其中行地址位数为$\log_2 r$，列地址位数为$\log_2 c$。如：16K×8位DRAM，则n=14，行、列地址各占7位。
- n位地址中高位部分为行地址，低位部分为列地址
- 为提高DRAM芯片的性价比，通常设置的r和c满足r≤c且|r-c|最小。
  - 例如，对于8K×8位DRAM芯片，其存储阵列设置为$2^6$行×$2^7$列，因此行地址和列地址的位数分别为6位和7位，13位芯片内地址$A_{12}\ A_{11}...A_1A_0$中，行地址为$A_{12}\ A_{11}...A_7$，列地址为$A_6...A_1A_0$。因按行刷新，为尽量减少刷新次数，故行数越少越好，但是，为了减少地址引脚，应尽量使行、列地址位数一致

# 主存模块的连接和读写操作

° DRAM芯片内部结构示意图

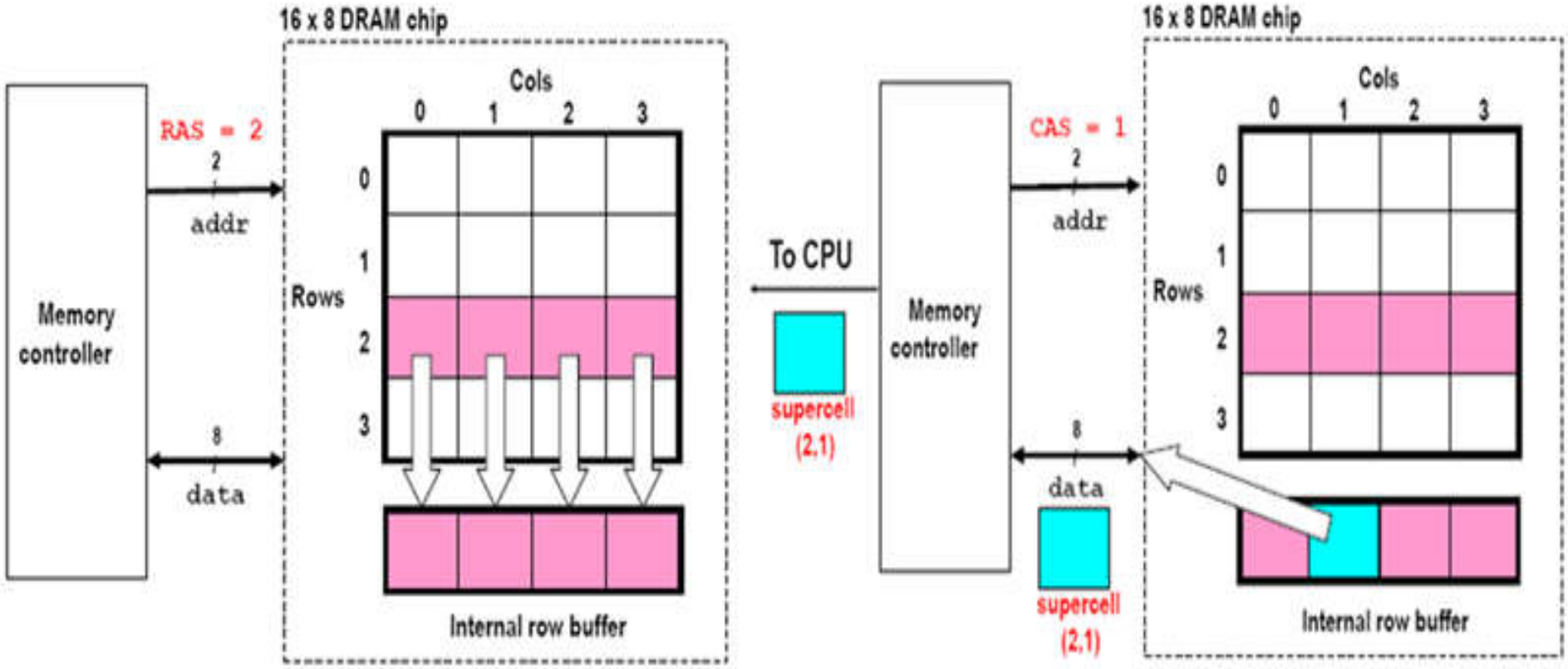
同时有多个芯片进行读写




图中芯片容量为16×8位，存储阵列为4行×4列，地址引脚采用复用方式，因而仅需2根地址引脚，每个超元（supercell）有8位，需8根数据引脚，有一个内部的行缓冲（row buffer），通常用SRAM元件实现。

# 主存模块的连接和读写操作

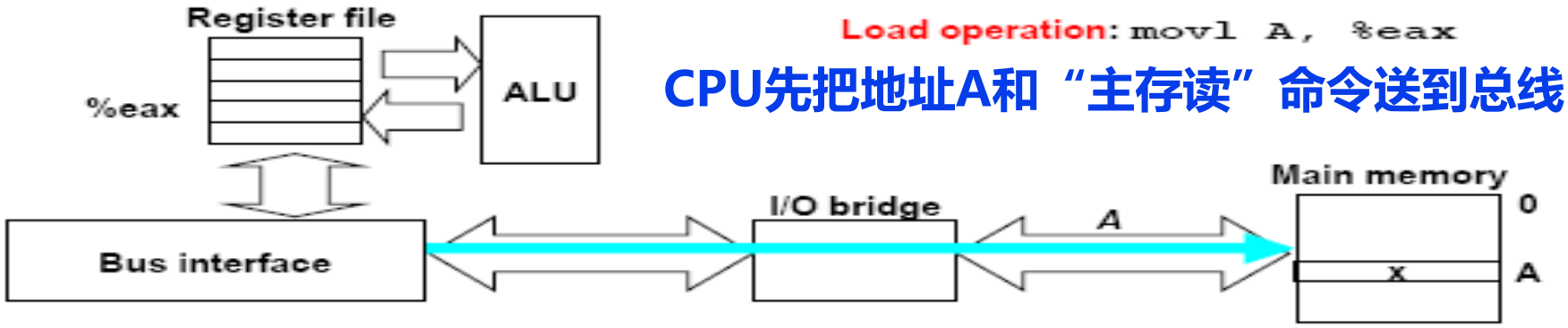
° DRAM芯片读写原理示意图

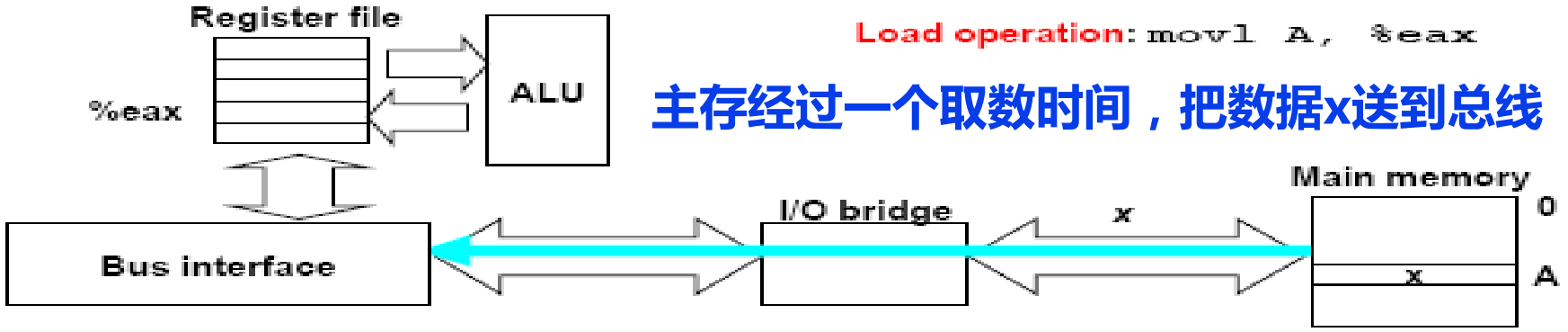


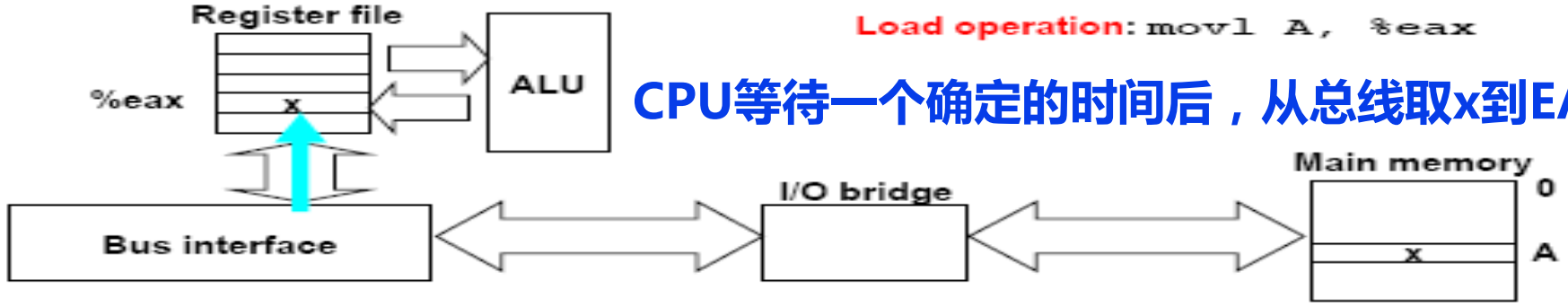

首先，存储控制器将行地址“2”送行译码器，选中第“2”行，此时，整个一行数据被送行缓冲。然后，存储控制器将列地址“1”送列译码器，选中第“1”列，此时，将行缓冲第“1”列的8位数据supercell(2,1)读到数据线，并继续送往CPU。

# 指令“movl 8(%ebp), %eax”操作过程

**由8(%ebp)得到地址A的过程较复杂，涉及MMU、TLB、页表等重要概念！**

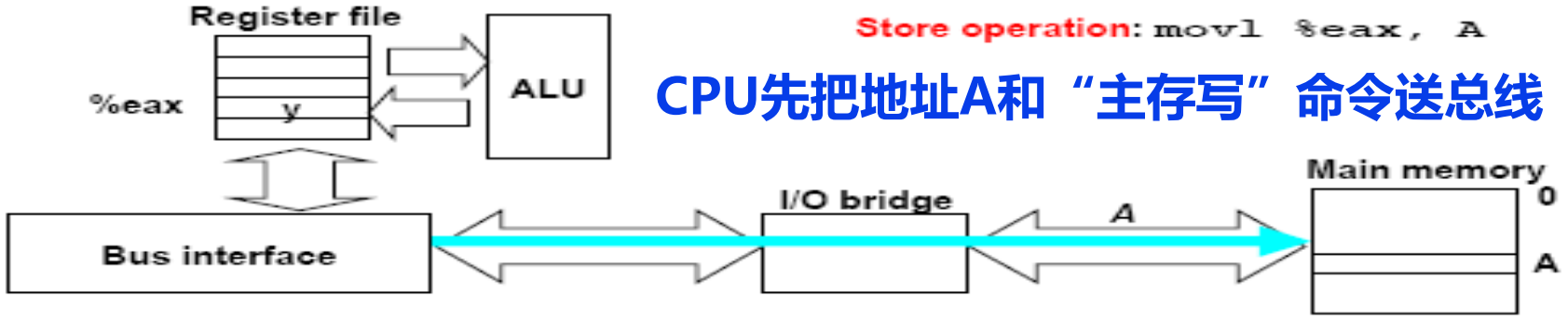


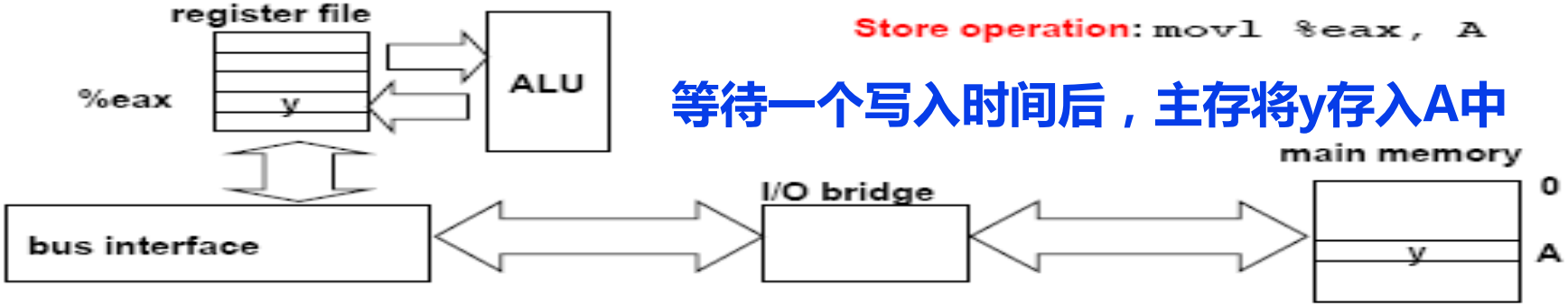



# 指令“movl %eax,8(%ebp)”操作过程

**由8(%ebp)得到地址A的过程较复杂，涉及MMU、TLB、页表等重要概念！**







# 本周小结

- 按介质分半导体、磁表面和光盘存储器；按存取方式分随机访问、直接访问、顺序访问和按内容访问存储器；按信息可更改性分只读、可读可写存储器；按断电后的可保存性分易失性、非易失性存储器。
- 主存储器是易失性的、可读可写的、半导体随机访问存储器。
- 主存储器芯片中存放信息的称为存储阵列；每个存储阵列包含若干个存储单元，每个存储单元由若干个记忆单元（cell）构成，每个记忆单元存放一位信息（0或1）。
- 记忆单元有静态（六管）和动态（单管）两种，前者为SRAM，后者为DRAM。内存条中芯片为DRAM芯片，每个芯片有一个行缓存（用SRAM实现）。
- 主存和CPU之间通过存储器总线（内存条插槽）相连。主存地址由CPU送出；数据信息可以是CPU到主存，或相反。