



南京大學

NANJING UNIVERSITY

存储器层次结构

殷亚凤

智能软件与工程学院

苏州校区南雍楼东区225

yafeng@nju.edu.cn , <https://yafengnju.github.io/>



存储器层次结构

- 存储器概述
- 半导体随机存取存储器
- 外部辅助存储器
- 存储器的数据校验
- 高速缓冲存储器
- 虚拟存储器





存储器概述——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）**
 - 按字节编址、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数据的寄存器





存储器概述——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)
 - 按地址访问，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。





存储器概述——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)





存储器概述——存储器分类

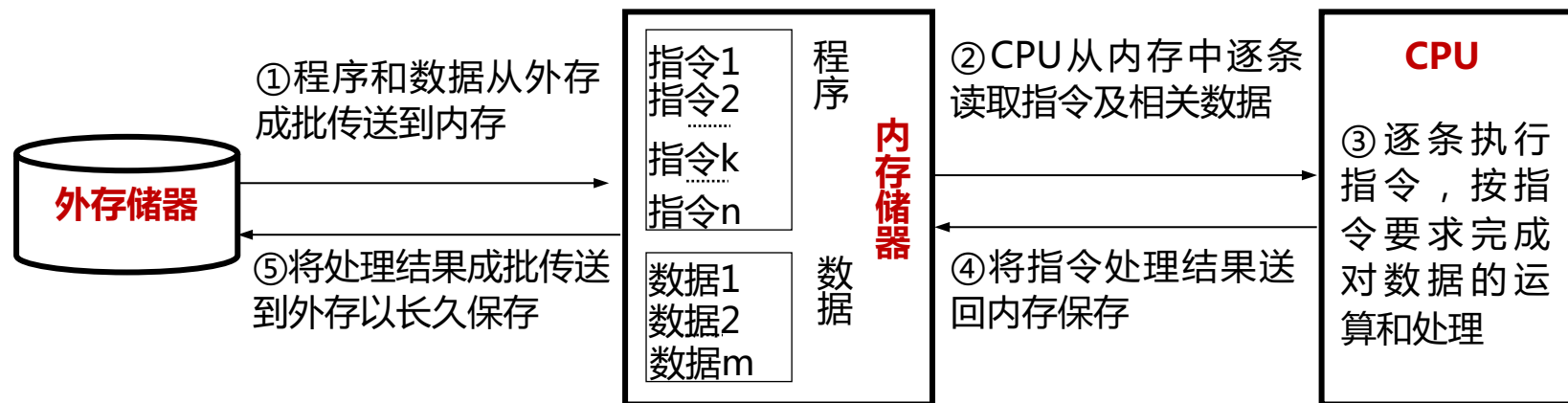
(5) 按功能/容量/速度/所在位置分类

- 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）
- 主存储器MM (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
 - 用磁表面或光存储器实现，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

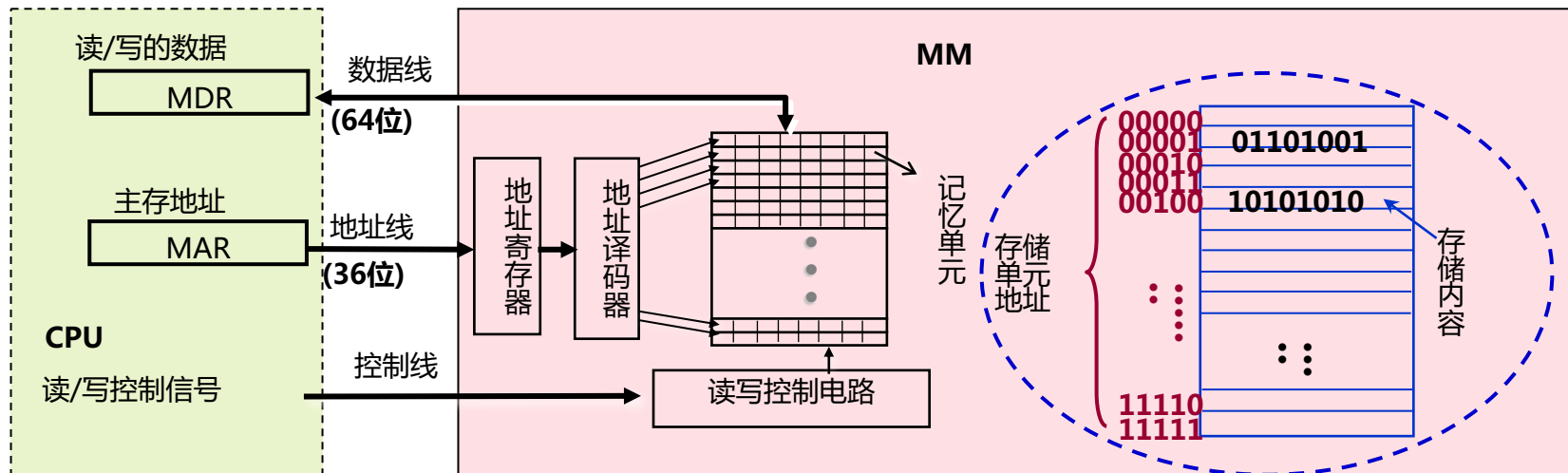
✓ 内存（简称内存或主存）

- 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 $0 \sim 2^{36}-1$ ，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， $1\text{ ns} = 10^{-9}\text{ s}$ ），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。)



存储器的层次化结构

为了缩小存储器和处理器两者之间在性能方面的差距，通常在计算机内部采用**层次化的存储器体系结构**。

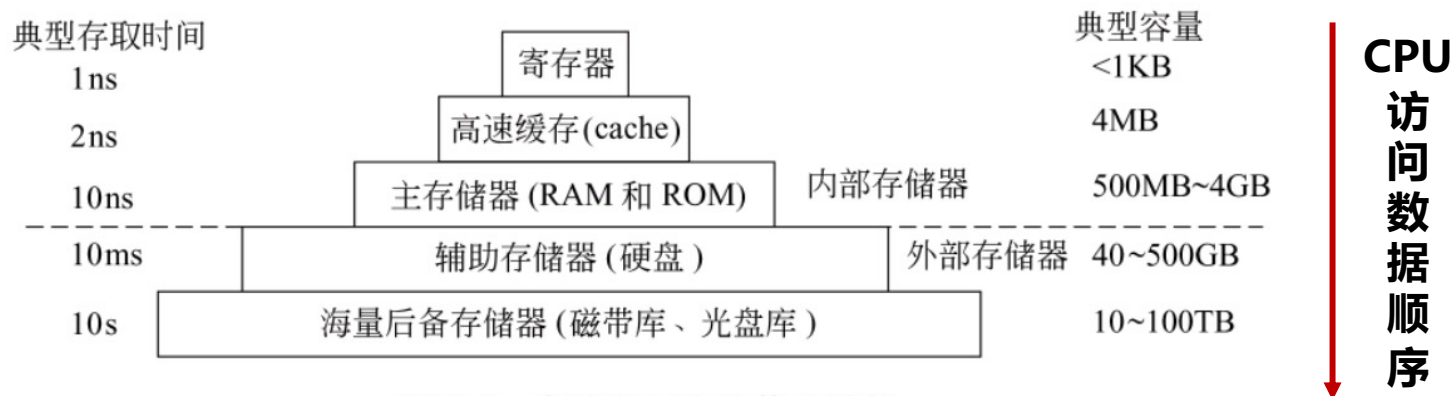


图 7.2 存储器层次化体系结构

- 速度越快，容量越小，越靠近CPU。
- CPU可以直接访问内部存储器；而外部存储器信息要先被取到主存，再被CPU访问。
- 数据一般只在相邻层之间复制传输，而且总是从慢速存储器复制到快速存储器。



存储器层次结构

- 存储器概述
- **半导体随机存取存储器**
- 外部辅助存储器
- 存储器的数据校验
- 高速缓冲存储器
- 虚拟存储器





半导体随机存取存储器——基本存储元件

六管静态MOS管存储元件

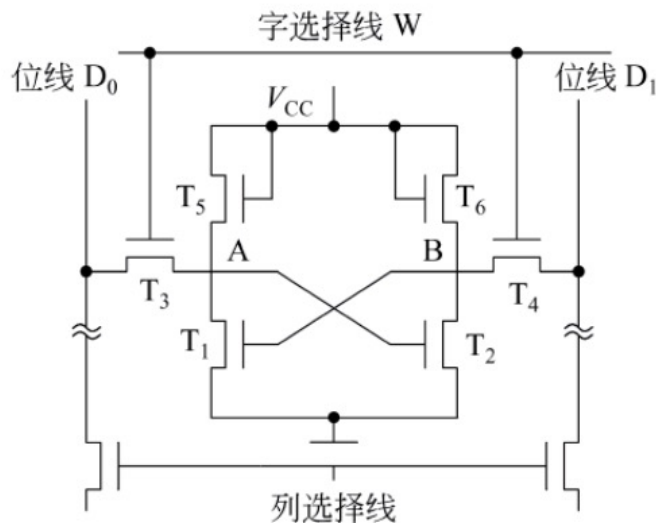


图 7.3 六管静态存储元件

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



半导体随机存取存储器——基本存储元件

• 单管动态MOS管存储元件

■ 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。

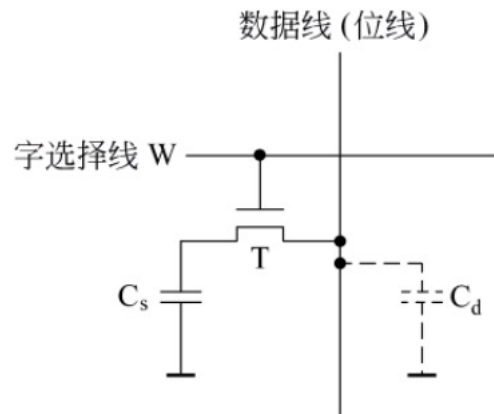


图 7.4 单管动态存储元件

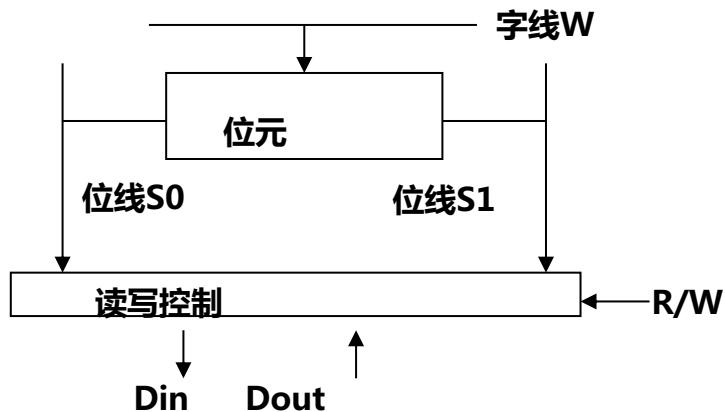




静态存储元件和动态存储元件的比较

静态存储元件：

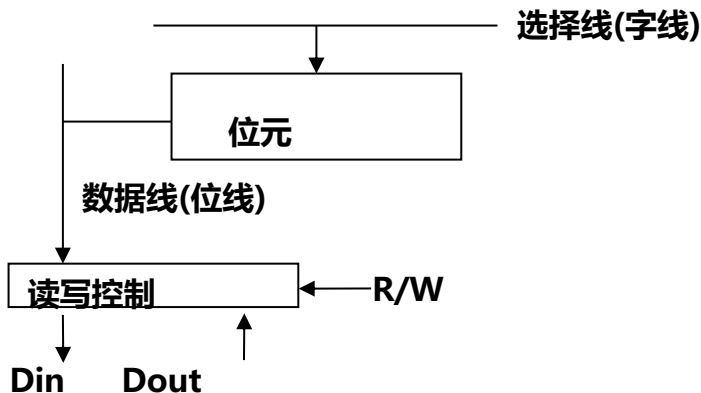
- MOS管多，功耗大，集成度低；
- 可保持记忆状态，无须刷新；
- 读写速度快；
- 价格昂贵。



SRAM(适合做Cache)

动态存储元件：

- MOS管少，功耗小，集成度高；
- 必须定时刷新；
- 读写速度慢；
- 价格较低。



DRAM(适合做主存)



SRAM芯片和DRAM芯片

存储器芯片由存储体、I/O读写电路、地址译码和控制电路等组成。

- **存储体/存储矩阵**：存储单元的集合；
- **地址译码器**：将地址转换为译码输出线上的高电平；
- **驱动器**：X选择线负载大，所以加驱动器；
- **I/O控制电路**：控制被选中单元的读/写，有放大信息的作用；
- **片选信号**：选中某个芯片
- **读/写控制信号**：控制被选中单元进行读或写。

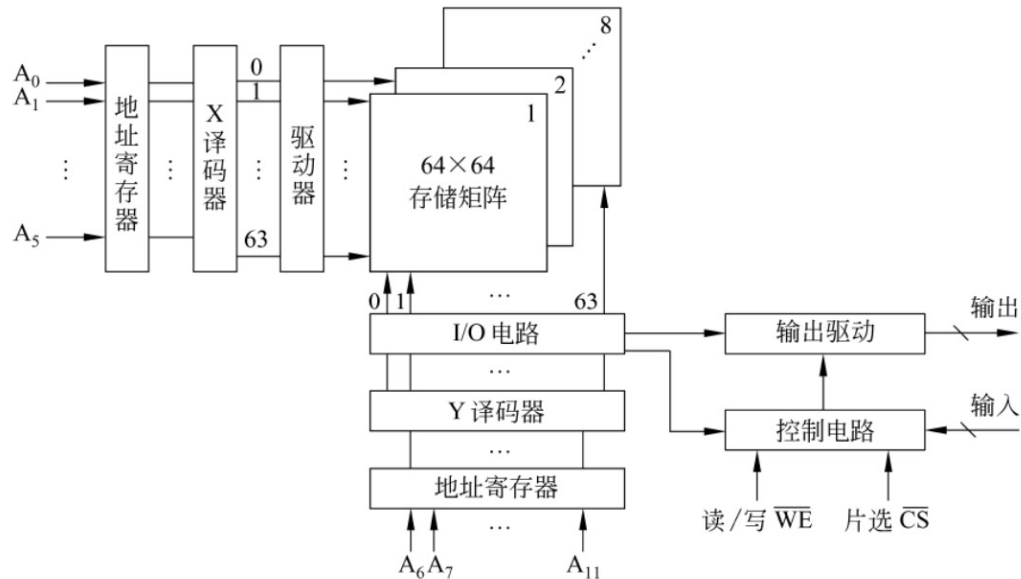
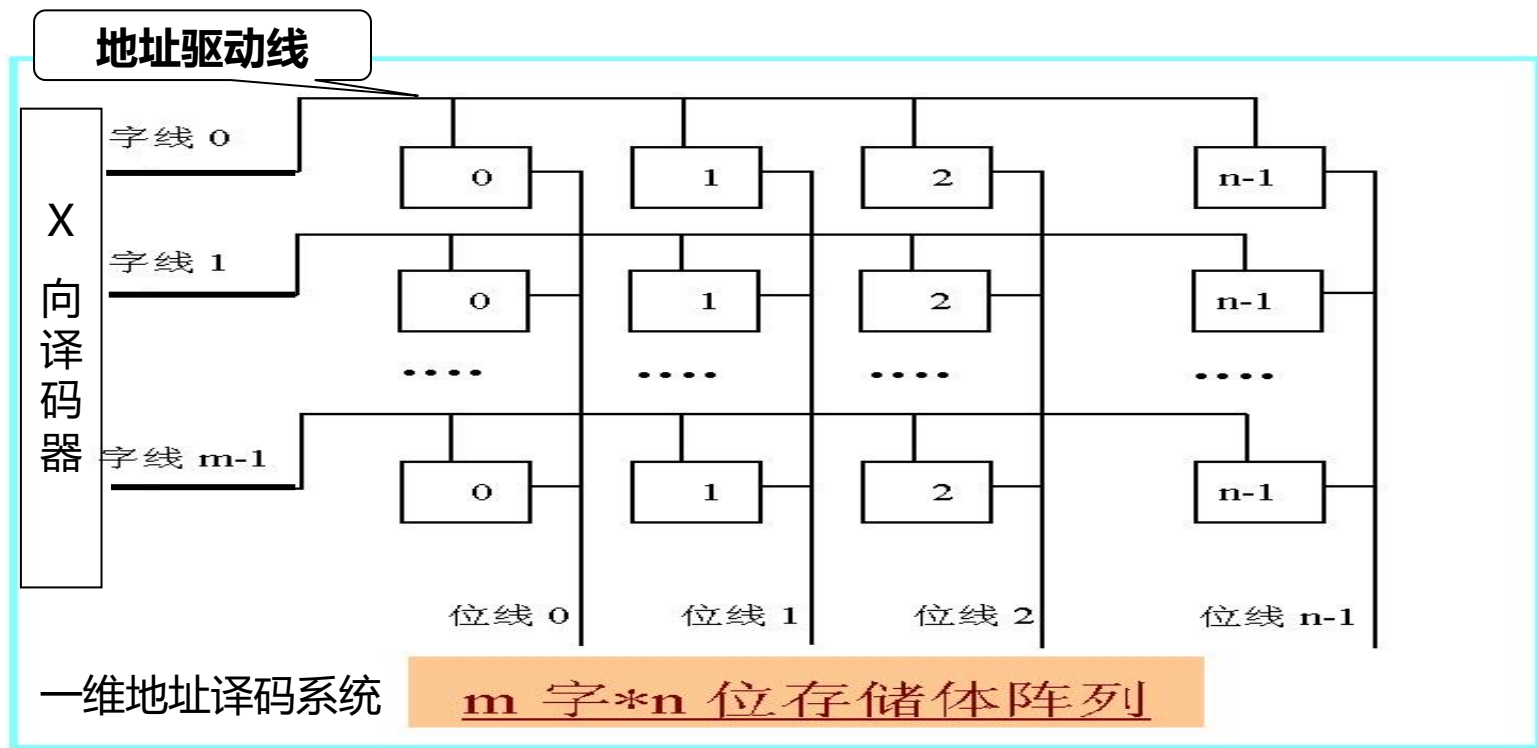


图 7.5 存储器芯片结构图





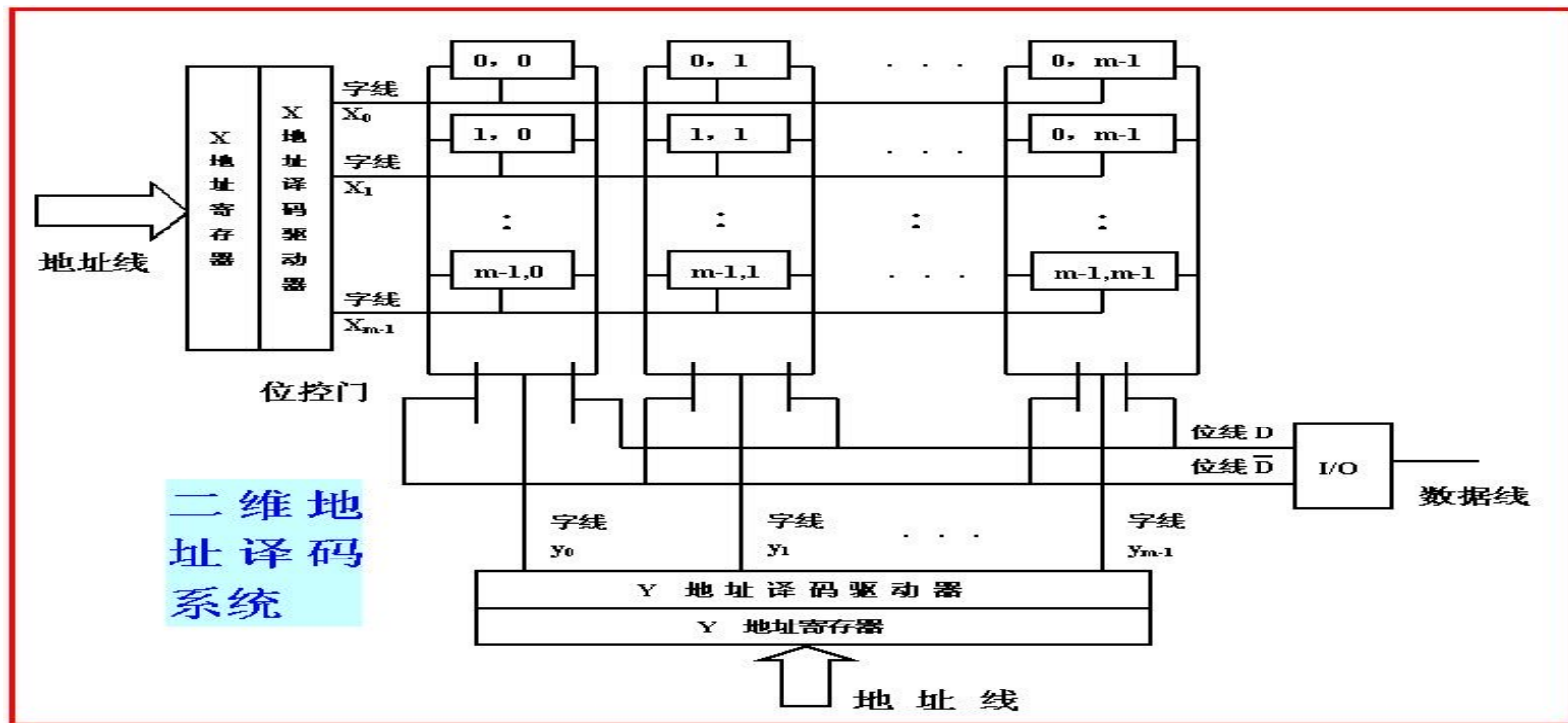
字片式存储体阵列组织（不作要求）



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





CPU与存储器之间的通信方式

- CPU和主存之间有同步和异步两种通信方式

- 异步方式（读操作）过程（需握手信号）

- CPU送地址到地址线，主存进行地址译码
 - CPU发读命令，然后等待存储器发回“完成”信号
 - 主存收到读命令后开始读数，完成后发“完成”信号给CPU
 - CPU接收到“完成”信号，从数据线取数

(写操作过程类似)

- 同步方式的特点

- CPU和主存由统一时钟信号控制，无需应答信号（如“完成”）
 - 主存总是在确定的时间内准备好数据
 - CPU送出地址和读命令后，总是在确定的时间取数据
 - 存储器芯片必须支持同步方式





SDRAM芯片技术

- SDRAM是同步存储芯片

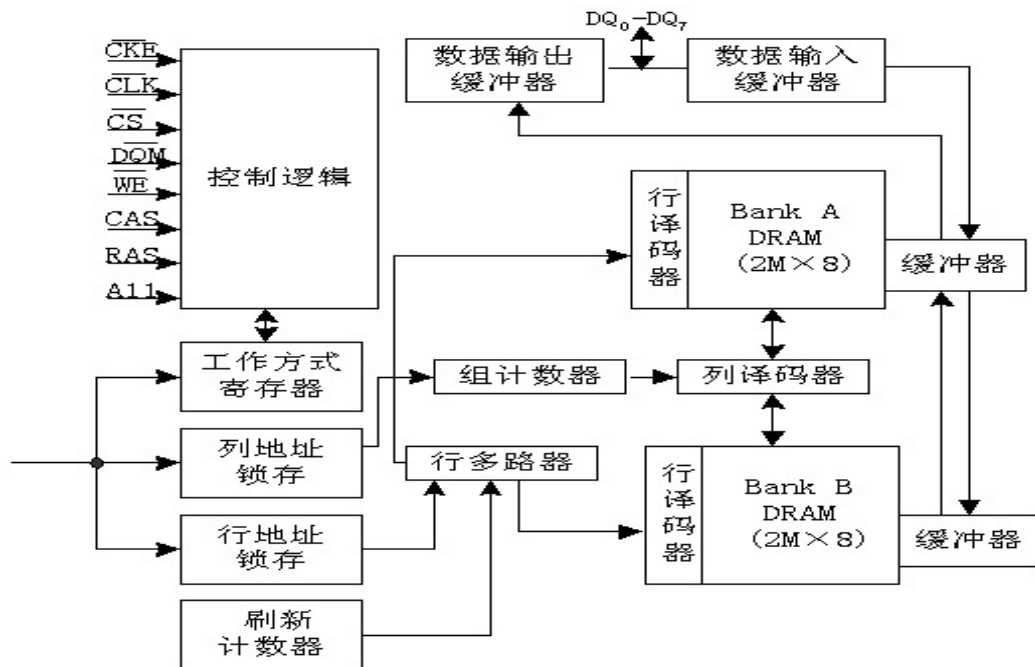
- 每步操作都在系统时钟控制下进行
- 有确定的等待时间（读命令开始到数据线有效的时间, 称为CAS潜伏期）
CL，例如 CL=2 clks
- 连续传送（Burst）数据个数 BL=1 / 2 / 4 / 8
- 多体(缓冲器)交叉存取
- 利用总线时钟上升沿与下降沿同步传送



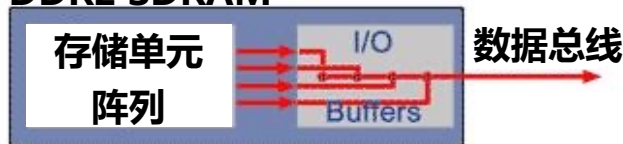


SDRAM芯片技术

- SDRAM芯片的内部结构
- 同步方式
- DDR SDRAM技术：
每个时钟内传送两个数据
- DDR2 SDRAM技术：
每个时钟内传送4个数据
- DDR3 SDRAM：
每个时钟内传送8个数据



DDR2 SDRAM

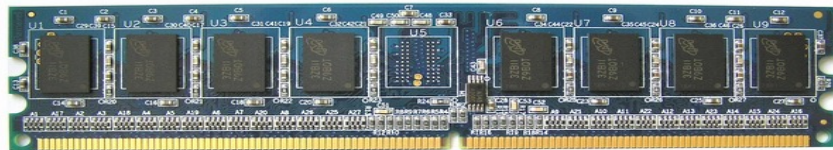




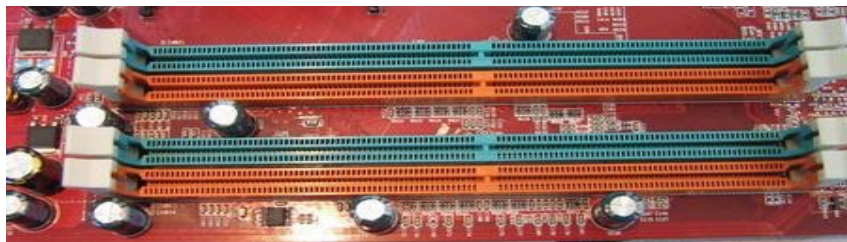
内存条和内存条插槽

- **内存条**：把若干片DRAM芯片焊装在一小条印制电路板上制成
- **内存条插槽**：存储器总线
- 内存条必须插在主板上的内存条插槽中才能使用（相同颜色插槽可以并行传输）

- **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



内存条



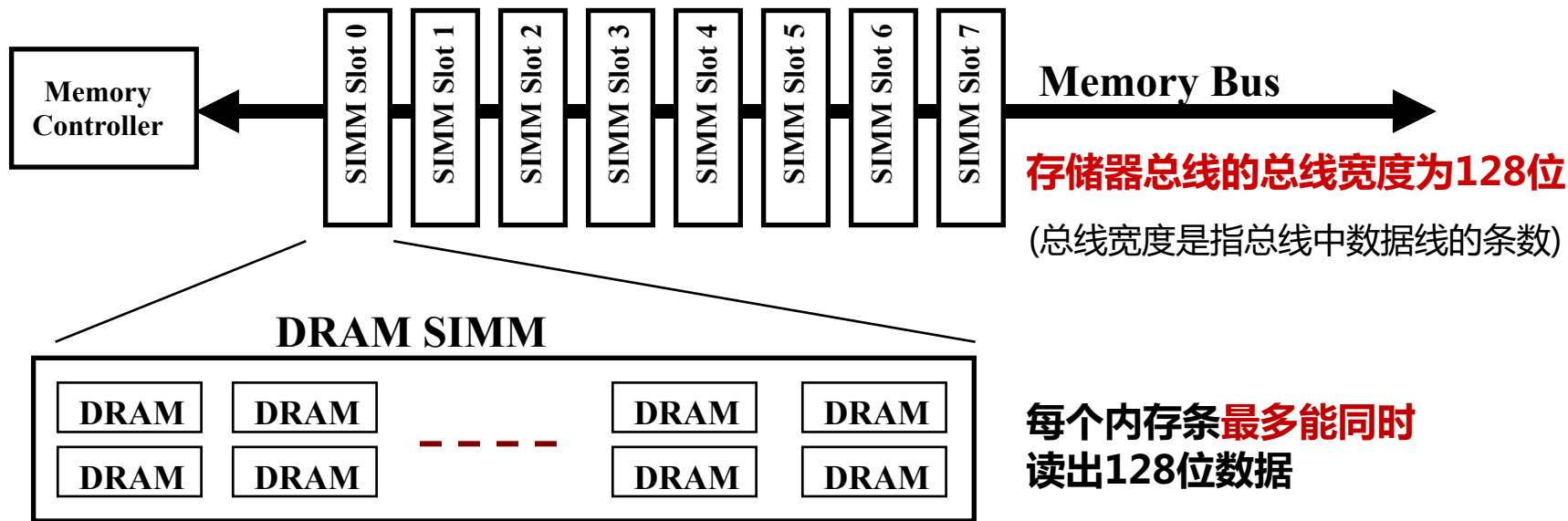
内存条插槽





内存条和内存条插槽

存储器控制器、存储器总线、内存条、DRAM芯片之间的连接



每次访存操作总是在某一个内存条内进行！





存储器芯片的扩展

- **字扩展（位数不变、扩充容量）**

用16K×8位芯片扩成64K×8位存储器需几个芯片？地址范围各为什么？

- 字方向扩展4倍，即4个芯片。0000-3FFFH，4000-7FFFH，8000-BFFFH，C000-FFFFH，地址共16位，高两位由外部译码器译码生成4个输出，分别连到4个片选信号，片内地址有14位
- 地址线、读/写控制线等对应相接，片选信号连译码输出

- **位扩展（字数不变，位数扩展）**

用4096×1位芯片构成4K×8位存储器需几个芯片？地址范围各是多少？

- 位方向扩展8倍，字方向无需扩展。即8个芯片，地址范围都一样：000-FFFH，地址共12位，全部作为片内地址
- 芯片的地址线及读/写控制线对应相接，而数据线单独引出

- **字位同时扩展（字和位同时扩展）**

用16K×4位芯片构成64K×8位存储器需几个芯片，地址范围各是多少？

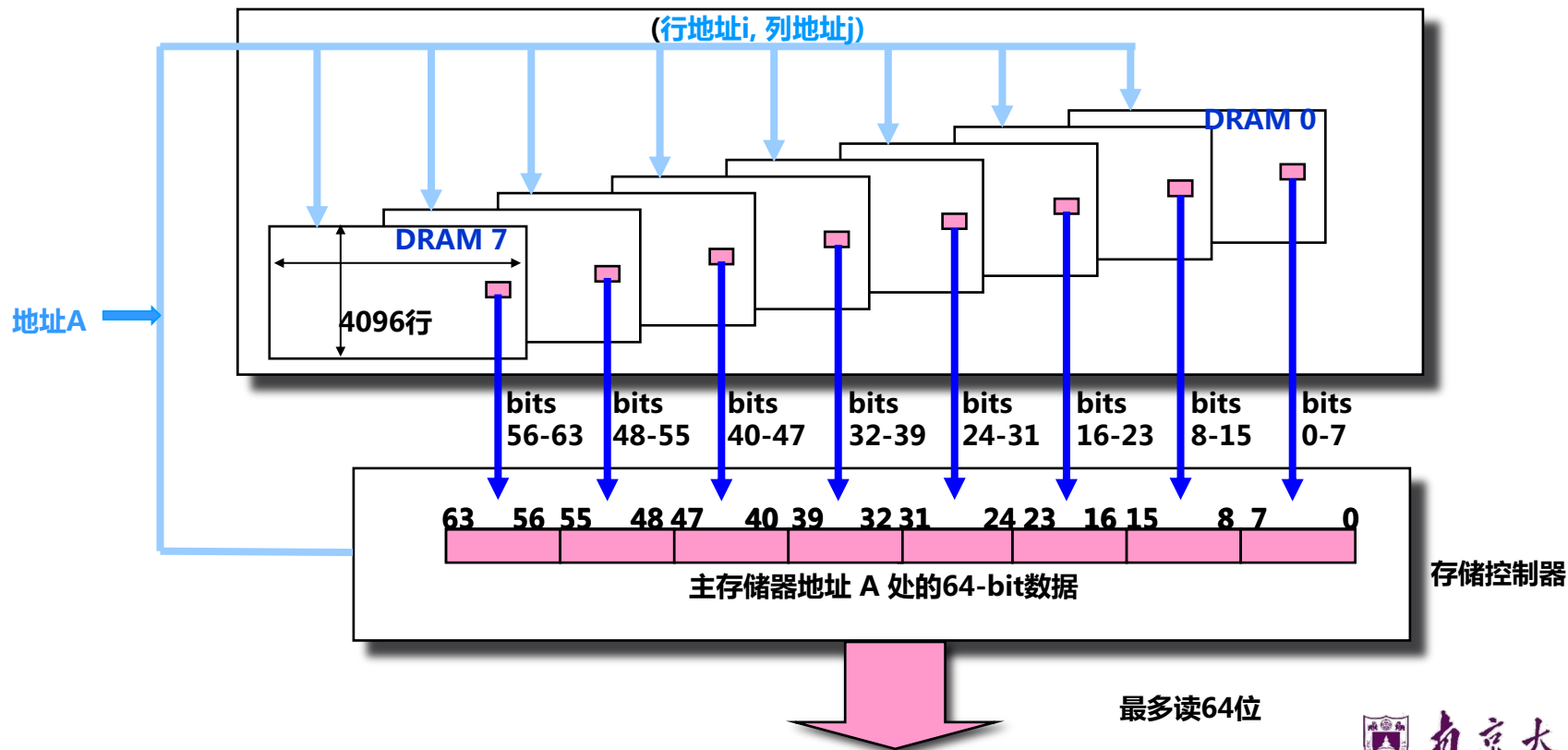
- 字向4倍、位向2倍，8个芯片。0000-3FFFH，4000-7FFFH，8000-BFFFH，C000-FFFFH
- 地址线、读/写控制线等对应相接，片选信号则分别与外部译码器各个译码输出端相连

有两种容量扩展方式：交叉编址和连续编址。

上述例子都是何种编址方式？ 连续编址！



DRAM芯片的扩展





DRAM芯片的扩展

- 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？

- 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。

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个二进制位) 送出

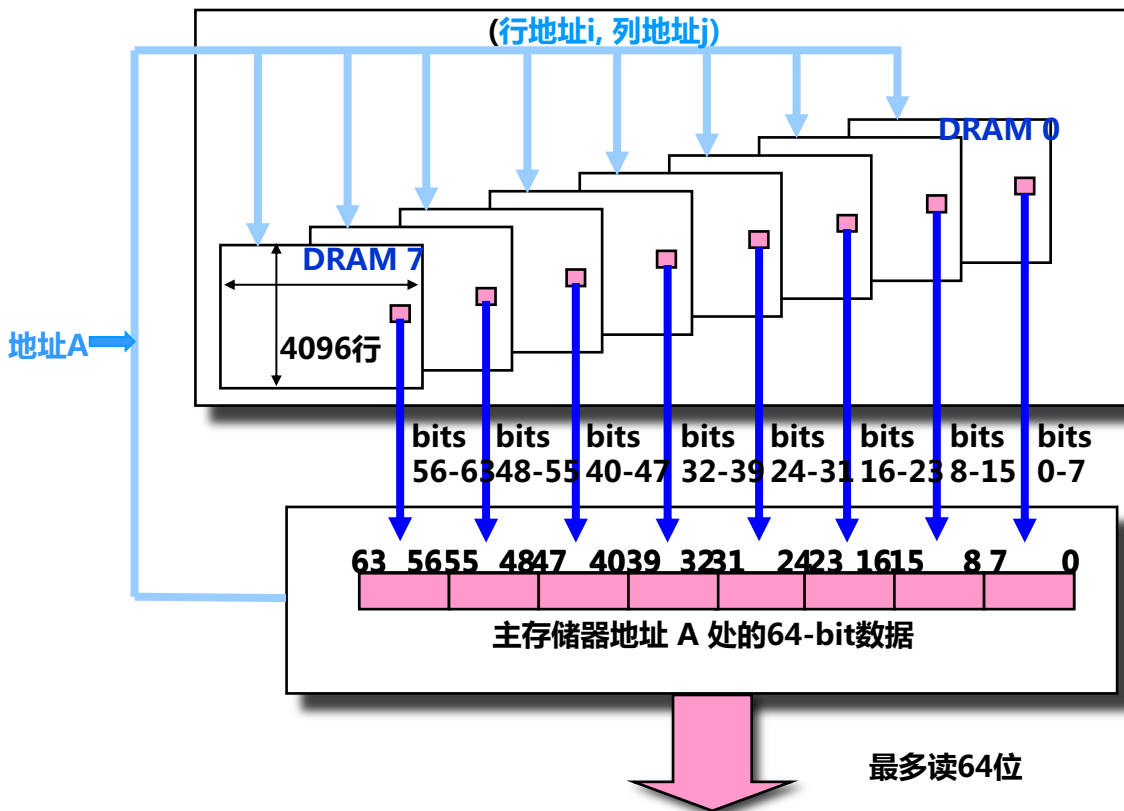
- 芯片内地址是否连续？

不连续，交叉编址，可同时读写所有芯片。





DRAM芯片的扩展



□ : 行、列地址为(i,j)的8个单元

地址A如何划分？

12	12	3
行号	列号	片

低3位用来选片

地址连续，共
 $8 \times 4096 = 2^{15} = 32768$ 个单元

存储控制器

地址A有多少位？ 27位！

在DRAM行缓冲中数据的地址有何特点？



DRAM芯片的扩展

12	12	3
行号	列号	片

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，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 $8*4096=2^{15}=32768$ 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？

3	12	12
片	行号	列号

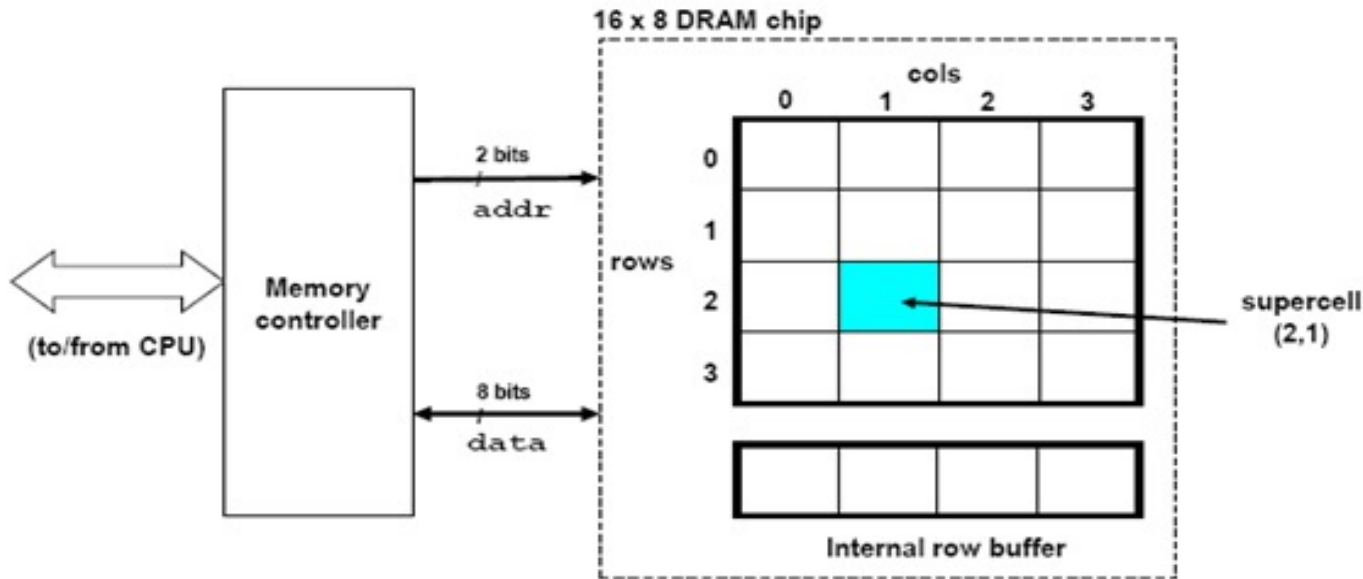




DRAM芯片内部结构

- DRAM芯片内部结构示意图

同时有多个芯片进行读写

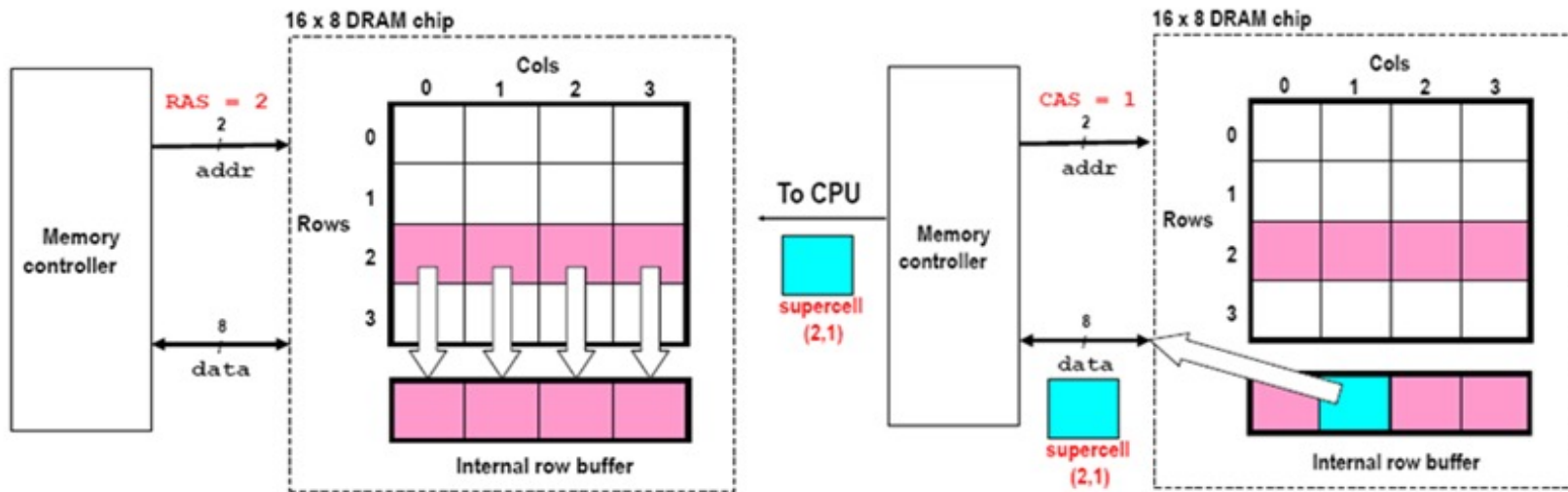


图中芯片容量为16×8位，存储阵列4行×4列，地址引脚采用复用方式，因而仅需2根地址引脚，每个超元（supercell）有8位，需8根数据引脚，有一个内部的行缓冲（row buffer），通常用SRAM元件实现。



DRAM芯片内部结构

- DRAM芯片读写原理示意图

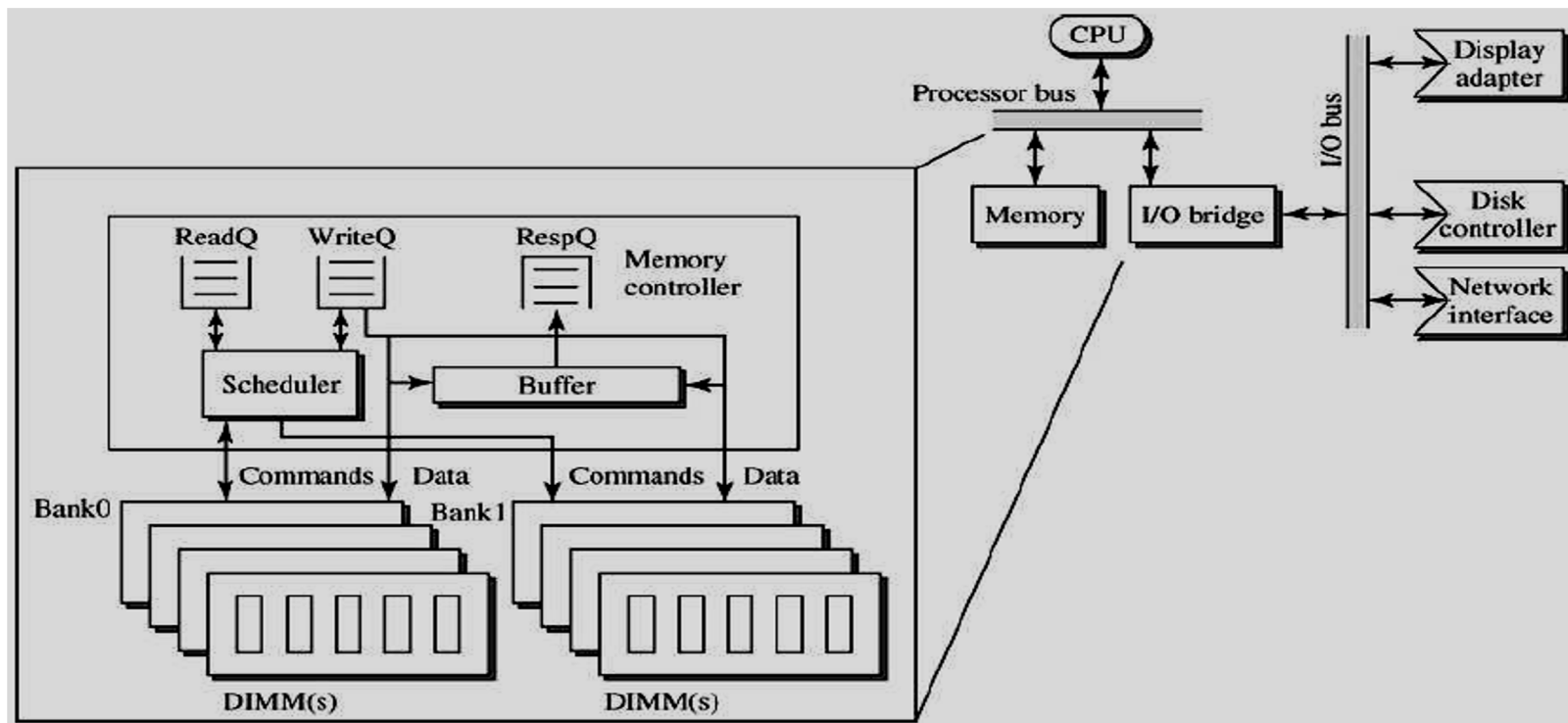


首先，存储控制器将行地址“2”送行译码器，选中第“2”行，此时，整个一行数据被送行缓冲。然后，存储控制器将列地址“1”送列译码器，选中第“1”列，此时，将行缓冲第“1”列的8位数据supercell(2,1)读到数据线，并继续送往CPU。



多模块存储器

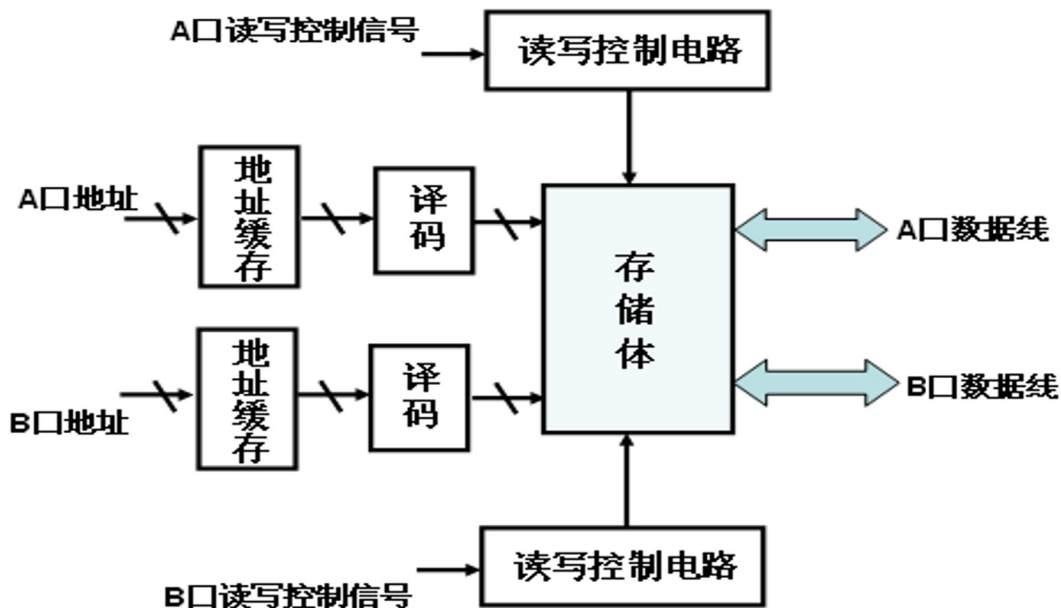
- 多模块技术：2个、4个或多个存储器同时工作





多模块存储器

- 双口存储器（能同时进行两个数据的读/写）
- 两套独立的读/写控制电路、地址缓存、地址译码及地址线和数据线，通常作为双口RAM或指令预取部件





多模块存储器

- **多模块存储器（能提高数据访问速度）**
 - 包含多个小体；
 - 每个体有其自己的MAR、MDR和读写电路；
 - 可独立组成一个存储模块；
 - 可同时对多个模块进行访问。
- **根据不同的编址方式可分为：连续编址、交叉编址**





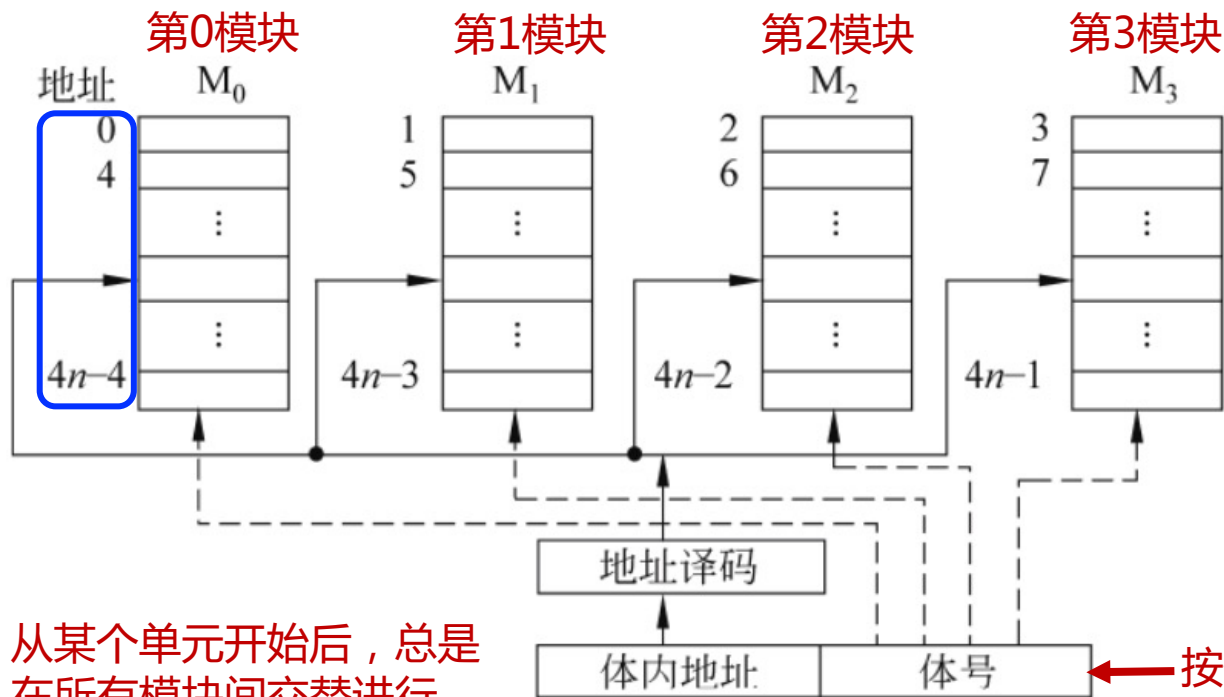
从某单元开始，连续在同一模块访问，然后才跳到下一个模块。

图 7.12 连续编址的多模块存储器

不能提高存储器的吞吐率！



交叉编址方式



从某个单元开始后，总是在所有模块间交替进行。

图 7.13 交叉编址的多模块存储器

为什么能提高吞吐量？多个模块交叉存取！





交叉编址方式：轮流启动、同时启动

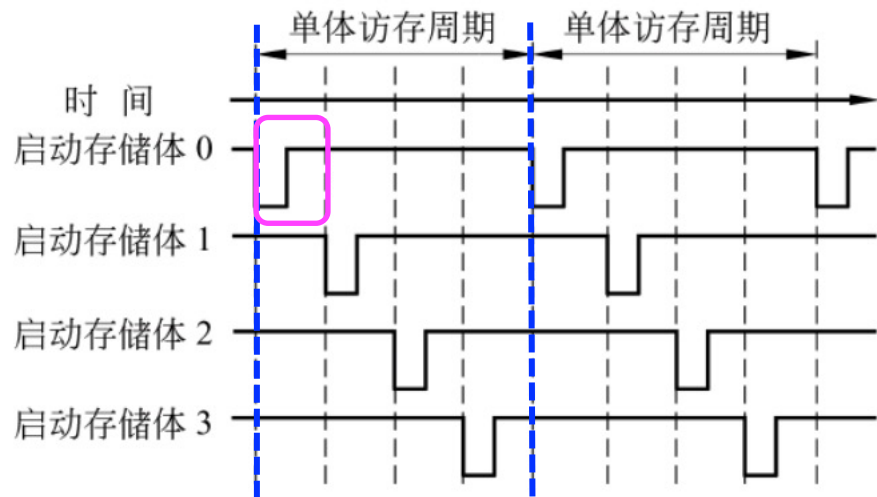


图 7.14 4 体交叉轮流访问方式

如果**所有存储模块**一次并行读写的总位数正好等于存储器总线中的数据位数，则可以采用**同时启动**方式。

- **每个存储模块**一次读写的位数（即存储字）正好等于存储器总线中的数据位数（即总线传输单位），则采用**轮流启动**方式；
- 具有 m 个体的多模块存储器，每隔 $1/m$ 个存储周期启动一个体；
- 存取速度提高 m 倍。





提问

Q & A

殷亚凤

智能软件与工程学院

苏州校区南雍楼东区225

yafeng@nju.edu.cn , <https://yafengnju.github.io/>



南京大學
NANJING UNIVERSITY