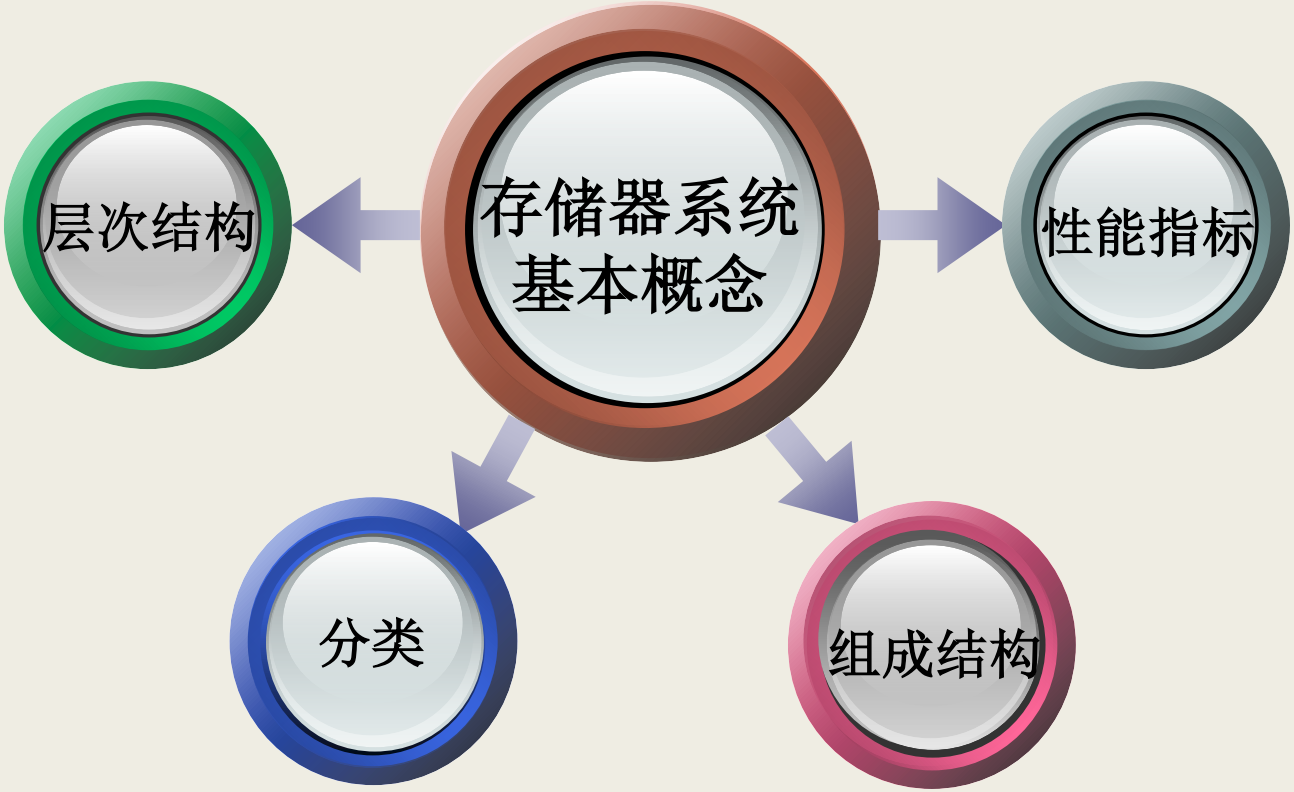
第5章
存储器及存储器子系统(I)

5.1 存储器概述

存储器(Memory)是计算机的主要组成部分之一，是用来存放程序和数据的部件。存储器表征了计算机的“记忆”功能。



5.1.1 存储器的分类（耦合程度）

存储器
- 主存储器（内存储器）
- 辅助存储器（外存储器）
- 高速缓冲存储器（*Cache*）

功能：存放当前正在使用的或经常使用的程序和数据。
特点：通常用半导体存储器作为内存储器。快速存取、容量较小，CPU直接访问
容量：受到地址总线位数的限制。

功能：用来存放相对来说不经常使用的程序或者数据或者需要长期保存的信息。
特点：存取速度慢、容量大，可以保存和修改存储信息，CPU不直接对它进行访问，有专用的设备来管理。

功能：为提高计算机的处理速度，利用Cache来暂存CPU正在使用的指令和数据，可以加快信息传递的速度。
特点：计算题系统中的一个高速小容量存储器，位于CPU和内存之间。主要由高速静态RAM组成。

5.1.2 常用的性能指标

1. 存储容量---是指它可存储的信息的字节数或比特数，决定计算机的速度和处理能力

1KB=2^{10}B=1024B　　1MB=2^{10}KB=1024KB

1GB=2^{10}MB=1024MB　1TB=2^{10}GB=1024GB

半导体存储器芯片容量取决于存储单元的个数和每个单元包含的位数:

存储器容量(S) = 存储单元数(p) × 数据位数(i)

存储单元个数(p)与存储器芯片的地址线条数(k)有密切关系：$p = 2^k$，或 $k = \log_2(p)$。

数据位数i一般等于芯片数据线的根数。

因此存储芯片的容量(S)与地址线条数(k)、数据线的位数(i)之间的关系可表示为：

$$S = 2^k \times i$$

【例】一个存储芯片容量为2048×8，说明它有8条数据线，2048个单元，地址线的条数为：

【例】一个存储芯片有20条地址线和4条数据线，则：

2. 存取速度---决定计算机的运算速度，是指从CPU给出有效的存储器地址到存储器输出有效数据所需的时间，一般以*ns*为单位。访问时间(存取时间)T_A、存取周期T_M、数据传送速率(频宽)B_M
3. 体积和功耗---使用低功耗存储器芯片构成存储系统
4. 可靠性---取决于构成存储器的芯片、配件质量及组装技术；是存储器系统的重要性能指标，通常用平均故障间隔时间(MTBF)来衡量。
5. 性价比---衡量存储器的经济性能

5.1.3 存储系统的层次结构

1、程序的局部性原理

在某一段时间内，CPU频繁访问某一局部的存储器区域，而对此范围外的地址则较少访问的现象。

时间局部性：

最近访问过的代码是不久访问的代码；

空间局部性：

地址相近的代码可能会被一起访问。

层次结构是基于程序的局部性原理的。对大量典型程序运行情况的统计分析得出的结论是：CPU对某些地址的访问在短时间间隔内出现集中分布的倾向。这有利于对存储器实现层次结构。

2、多级存储体系的组成

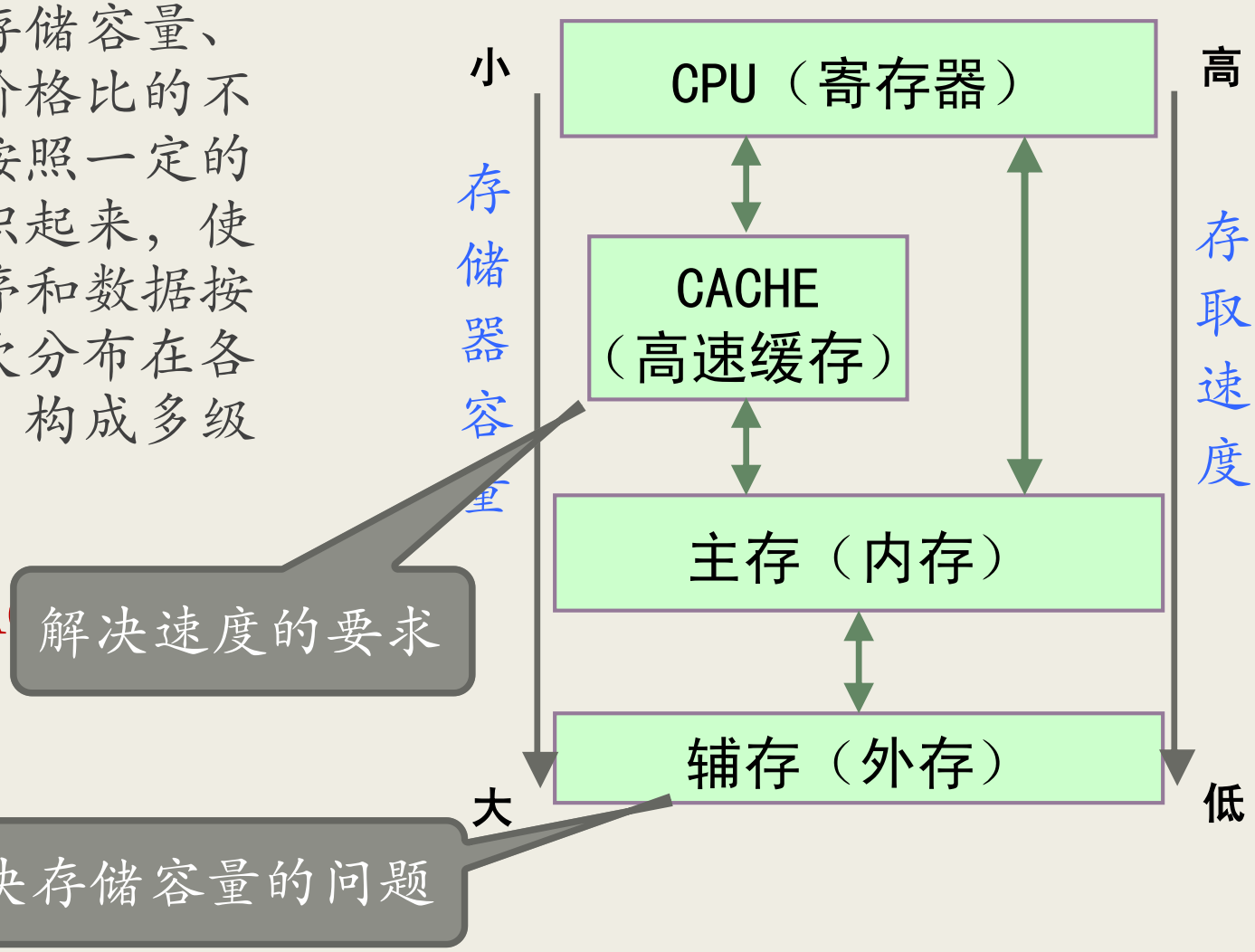
计算机系统中，根据各种存储器的存储容量、存取速度和价格比的不同，将它们按照一定的体系结构组织起来，使所存放的程序和数据按照一定的层次分布在各种存储器中，构成多级存储体系。

寄存器：

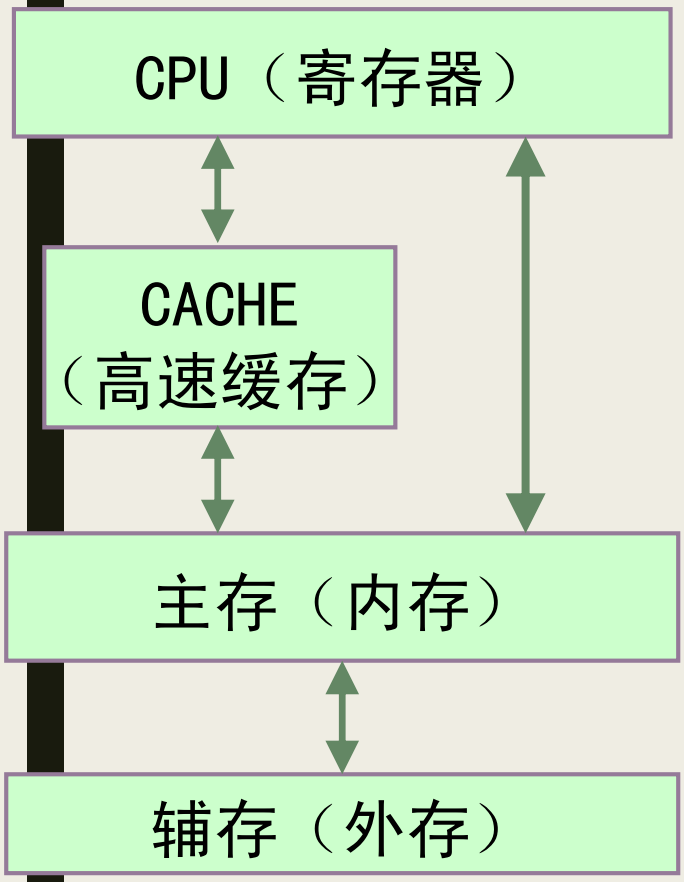
高速缓存(CA

主存：

辅存：

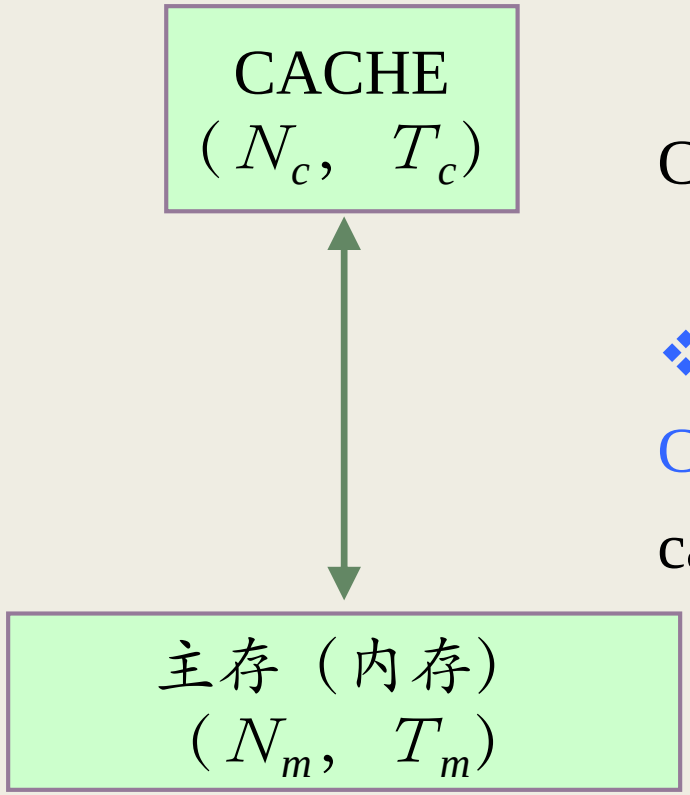


整体而言，存储系统主要有两个层次，即Cache-主存层次和主存-外存层次



- Cache-主存层次：主要解决CPU和主存之间的速度差异问题。在CPU和主存之间设置存取速度最快、容量小的高速缓冲存储器，能较好地解决存取速度问题，提高整机的运算速度。
- 主存-外存层次：解决的是存储器的大容量要求和低成本之间的矛盾。现代操作系统的形成和发展使得程序员摆脱了主存、辅存之间的地址人工定位，通过软件、硬件结合，把主存和辅存统一成了一个整体，程序员可以利用比主存实际容量大得多的逻辑地址编写程序。随着这种系统的发展和完善，逐渐形成了广泛使用的虚拟存储系统。

3. 多级存储体系的性能(考虑Cache-主存构成的两级存储)



cache命中率

$$H = N_c/(N_c + N_m)$$

CPU访问的平均时间

$$T_a = H \cdot T_c + (1 - H)T_m$$

❖ Cache和主存的存取周期直接影响CPU的平均访问时间

cache-主存系统的效率

$$e = \frac{T_c}{T_e} = \frac{1}{H + (1-H)\frac{T_m}{T_c}}$$

【课后习题18】某计算机系统的内存储器有Cache和主存构成，Cache的存取周期为45*ns*，主存的存取周期为200*ns*。已知在一段给定的时间内，CPU功访问内存4500次，其中340此访问主存。问：

①Cache的命中率是多少?

②CPU访问内存的平均时间是多少*ns*?

③Cache-主存系统的效率是多少?

④CPU访存的平均时间与哪些因素有关?

5.2 半导体静态存储器

内部存储器

为与CPU速度匹配，内存采用速度较快的半导体存储器。

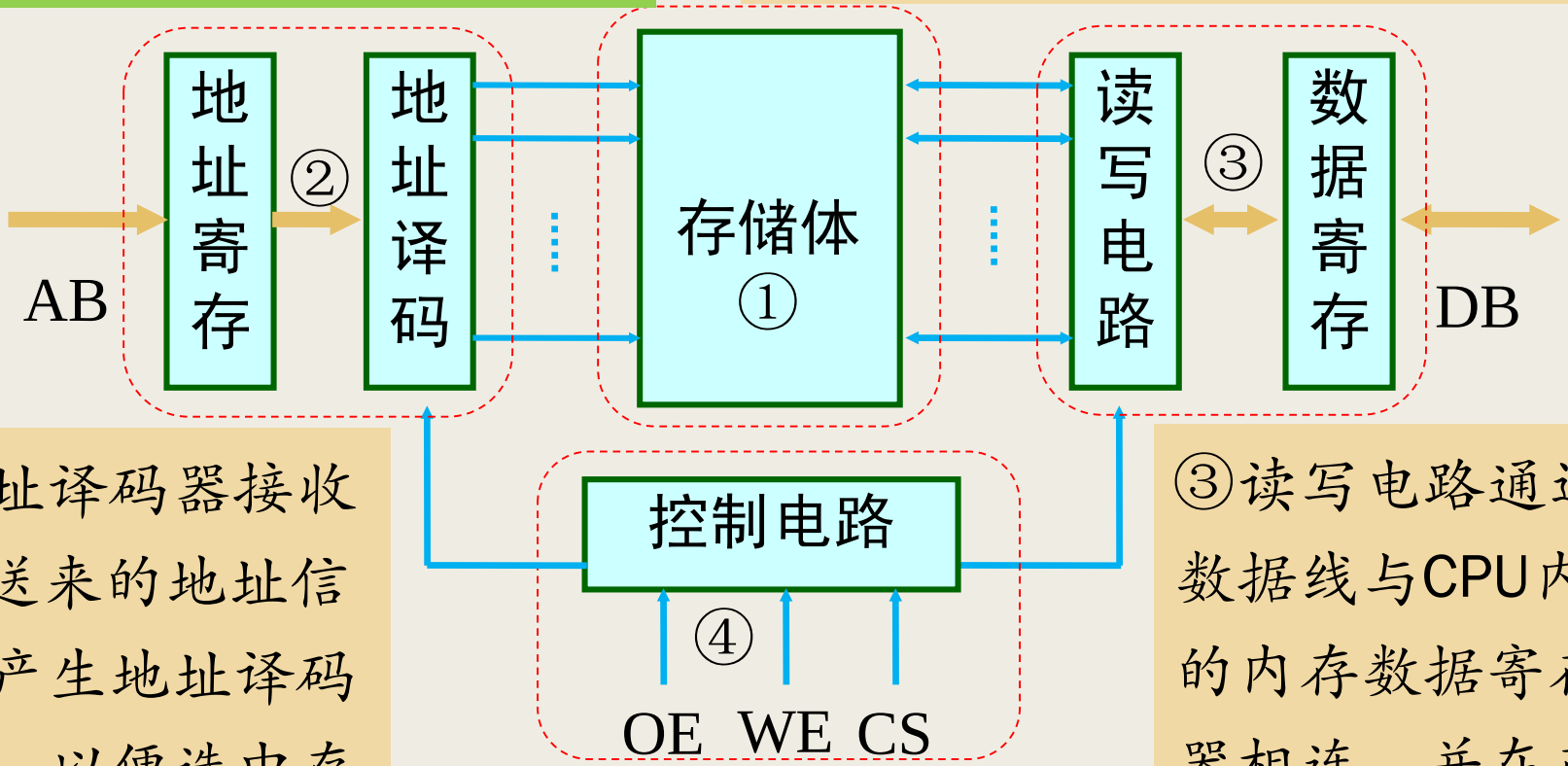
- 随机存取存储器(RAM)　易失性半导体存储器
  - 双极型
  - MOS型
    - 静态(SRAM)
    - 动态(DRAM)
- (根据数据的保存原理和读写过程的不同)
- 只读存储器(ROM)　非易失性半导体存储器
  - 掩膜ROM(MROM)
  - 可编程ROM(PROM)
  - 可擦除PROM(EPROM)
  - 电可擦除PROM(EEPROM)
  - 闪速存储器(Flash Memory)

SRAM与各种类型的ROM都属于半导体静态存储器

内存的基本组成

存储体、地址译码器、
数据缓冲器、控制逻辑电路

①存储体是能够存储二进制信息的基本存储单元的集合。按照一定的方式编址，配制成存储矩阵



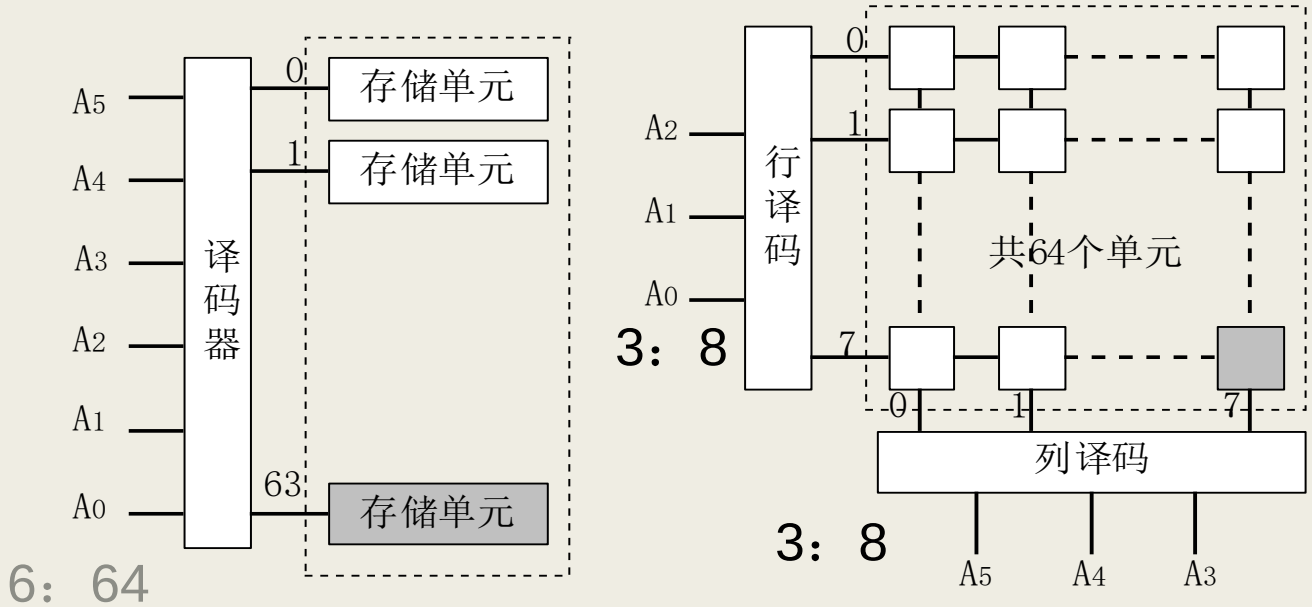
②地址译码器接收CPU送来的地址信号，产生地址译码信号，以便选中存储器矩阵中的某个存储单元。

④控制电路：选中存储器芯片，执行读写操作

③读写电路通过数据线与CPU内的内存数据寄存器相连，并在存储体与MDR之间传递信息。

译码结构

地址译码器的功能是：根据输入的地址编码，选中芯片内某个特定的存储单元。芯片内的地址译码可采用：单译码结构(线性排列)和双译码结构 (矩阵形式排列)。



（a） 单译码结构　　（b） 双译码结构

图 8.4 存储器芯片内部地址结构

5.2.1 SRAM存储器

1、基本存储电路：通常由6个MOS管组成双稳态触发器来构成。

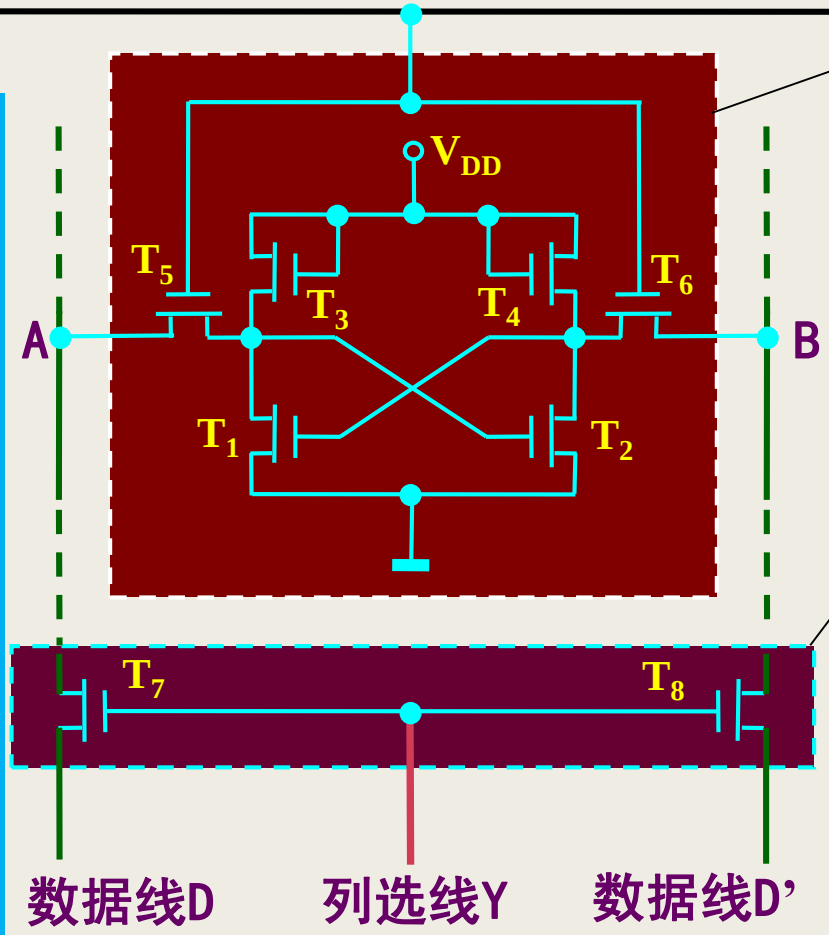
2、特点：

- 集成度高于双极型，但低于动态RAM
- 不需要刷新，可省去刷新电路
- 功耗比双极性的低，但比动态RAM高
- 易于用电池作为后备电源
- 存取速度较动态RAM快。

输入信息存储于T1、T2之栅极。当输入信号、地址选通信号消失后，T5~T8截止，能保持所存储信息，故不用刷新（即信息不用再生）。
D与D’对外只用一条输出端接到外部数据线上，这种存储电路读出是非破坏性的。

行选线X

当行选X=1（高电平），T5、T6导通，A，B与D，D’相连；当这个单元被选中时，相应的列选Y=1，T7、T8导通（它们为一列公用），D，D’输出。



6管基本存储单元

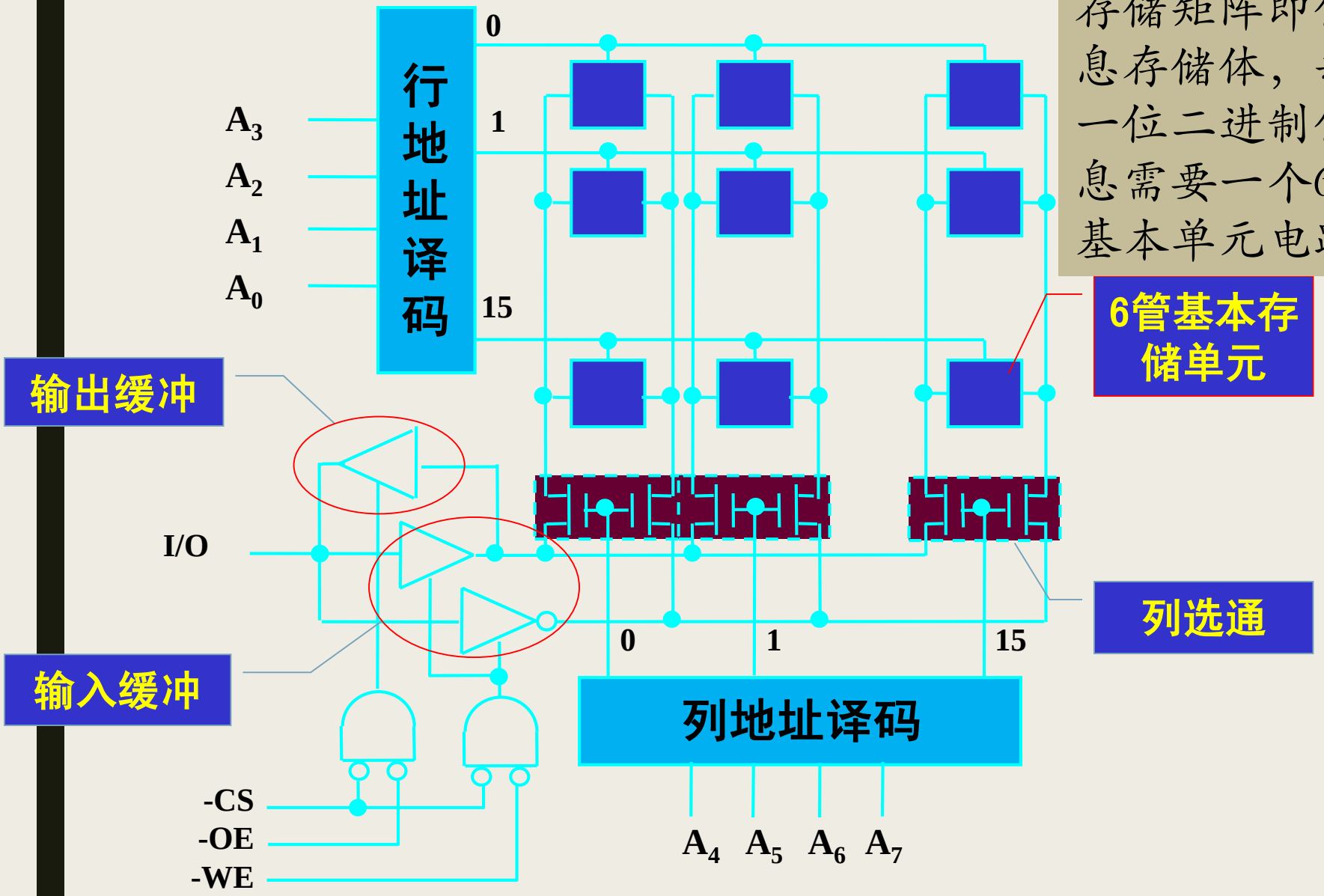
T_1，T_2—双稳态触发器，存储0与1;
T_3，T_4—负载管，为触发器补充电荷;
T_5，T_6—门控管，与数据线Di相连;

列选通

T_7，T_8—外部门控管，同一列的所有基本存储单元共有，通过列选线Y控制数据是否可以被送入或送出

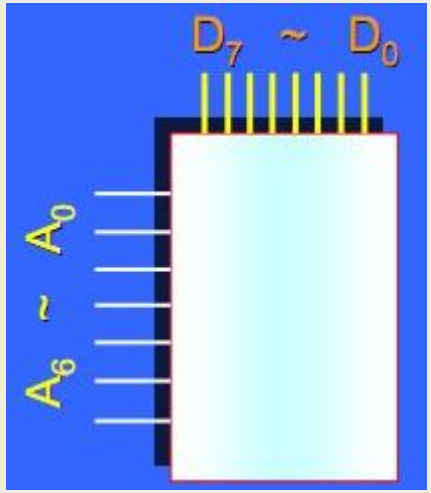
当写入时，写入信号自D（或D’）输入，此时，D=1，D’=0，T5、T6、T7、T8都导通（因为X=1，Y=1）
D→T7→T5→Q=1；D’→T8→T6→Q=0.

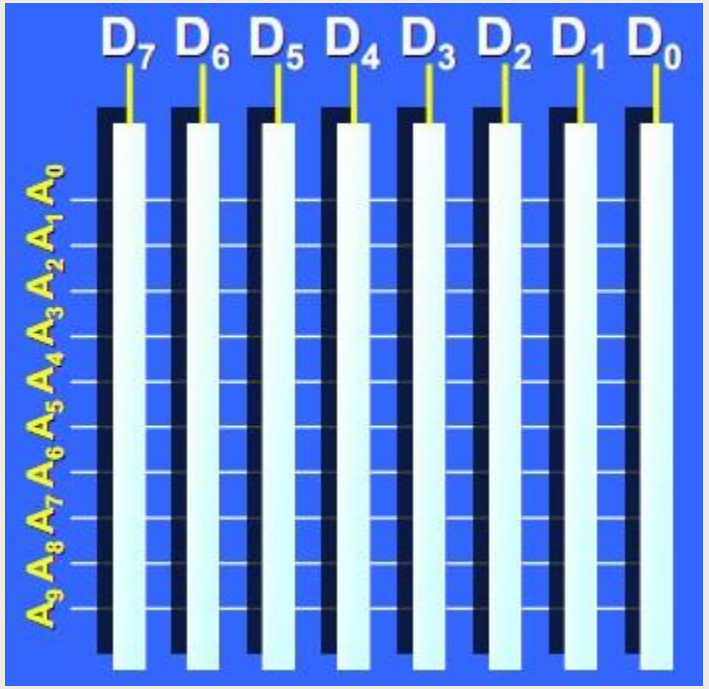
SRAM大多数都采用复合译码方式。一般把地址线分为行和列地址分别进行译码；



SRAM的外部特征

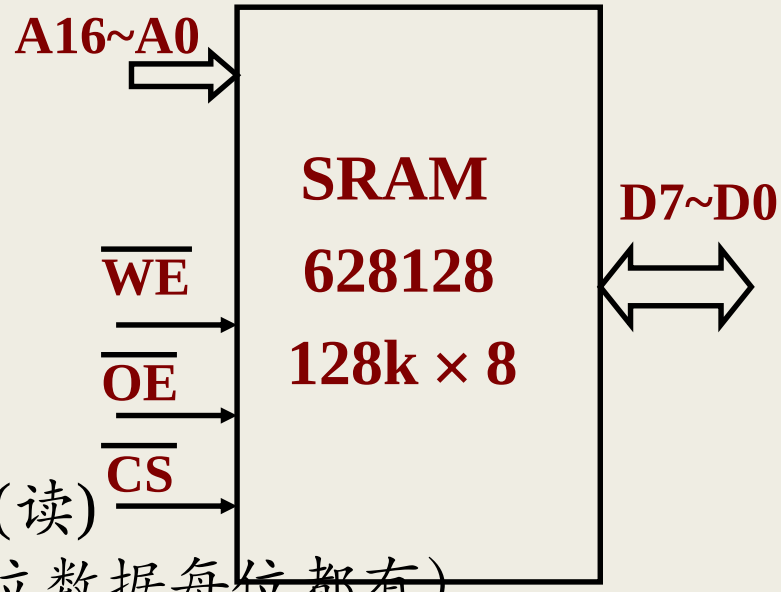
- 字结构：一个字节的8位制作在一块芯片上，选中芯片可一次性读/ 写8位信息，封装时引线较多。
- 位结构：1个芯片内的基本单元作不同字的同一位，8位由8块芯片组成。优点是芯片封装时引线少。





SRAM的引脚信号与读写操作

(芯片628128的引脚信号(128k×8))



A16~A0　地址线

D7~D0　双向数据线

CS　片选信号

WE　写允许信号

OE　输出允许信号(读)

这种芯片内部为字结构（即8位数据每位都有）

CPU总线与SRAM的连接方法：

①低位地址线、数据线、电源线直接相连

②高位地址线经译码后连接SRAM的片选信号CS（或CE）

③控制总线组合形成读/写控制信号WE或OE/WE

5.2.2 动态RAM存储器

特点:

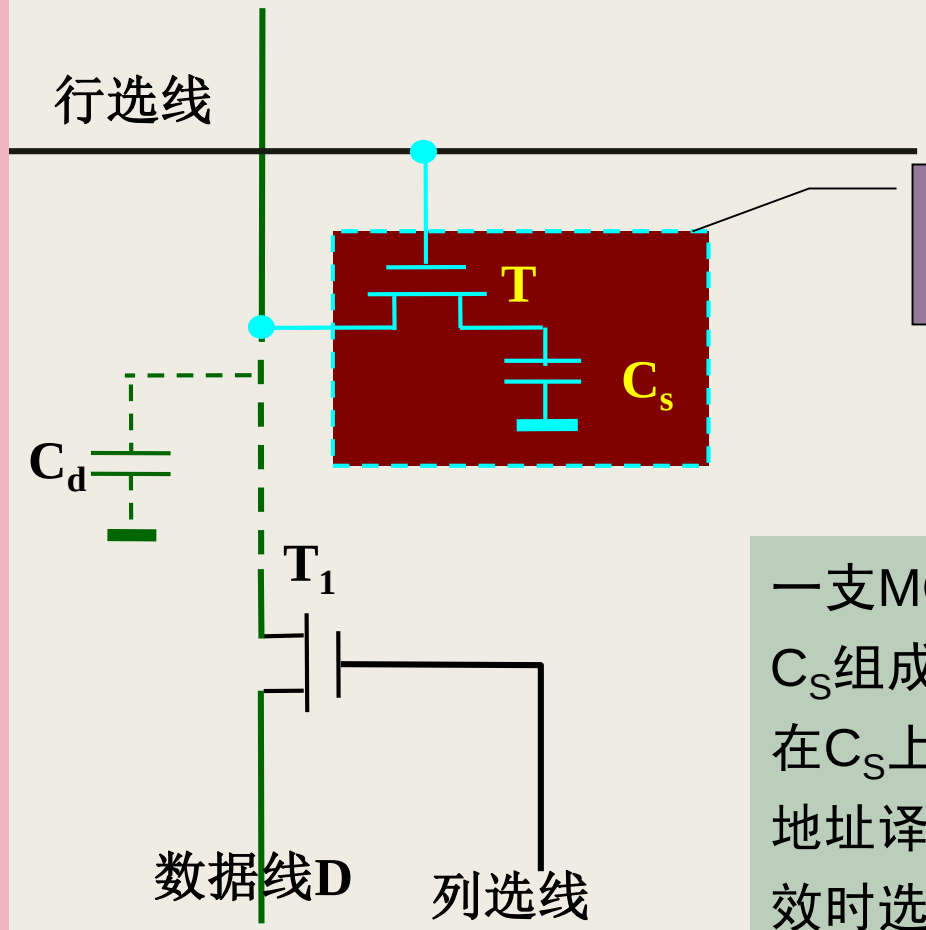
- 基本存储电路由一个晶体管及一个电容组
- 集成度高
- 比静态RAM的功耗更低
- 成本较低，另外耗电也少，
- 需要定时刷新，需要一个额外的刷新电路。

存储密度较高，适于在较大容量存储器的系统中作为随机存取存储器。

- DRAM 的基本存储单元是**单个场效应管**及其**极间电容**
  - 必须配备“**读出再生放大电路**”进行刷新
  - 每次同时对1行的存储单元进行刷新
  - 每个基本存储单元存储1位二进制数
  - 许多个基本存储单元形成行、列存储矩阵
- DRAM一般采用“位结构”存储体：
  - 每个存储单元存放 1 位
  - 需要 8 个存储芯片构成 1 个字节存储单元
  - 每个字节存储单元拥有 1 个唯一地址

单管动态RAM的基本存储电路

写入时，外部驱动数据线D，由D对C_S充/放电，改变所存储的信息；
读出时，C_S经D对数据线上的外部寄生电容C_d充/放电，改变外部C_d上的电压，读出所存储的信息。
因每次输出都会使Cs上原有的电荷泄放，存储的内容就会被破坏，所以读出是破坏性的。为此，每次读出后都需要进行再生（重新写入）以恢复Cs上的信息。



单管基本存储单元

数据的读/写都有充、放电完成，集成度高，速度慢

一支MOS管T和电容C_S组成，信息存储在C_S上；
地址译码线（行）有效时选中该单元，使T管导通，C_S和数据线D 连通。

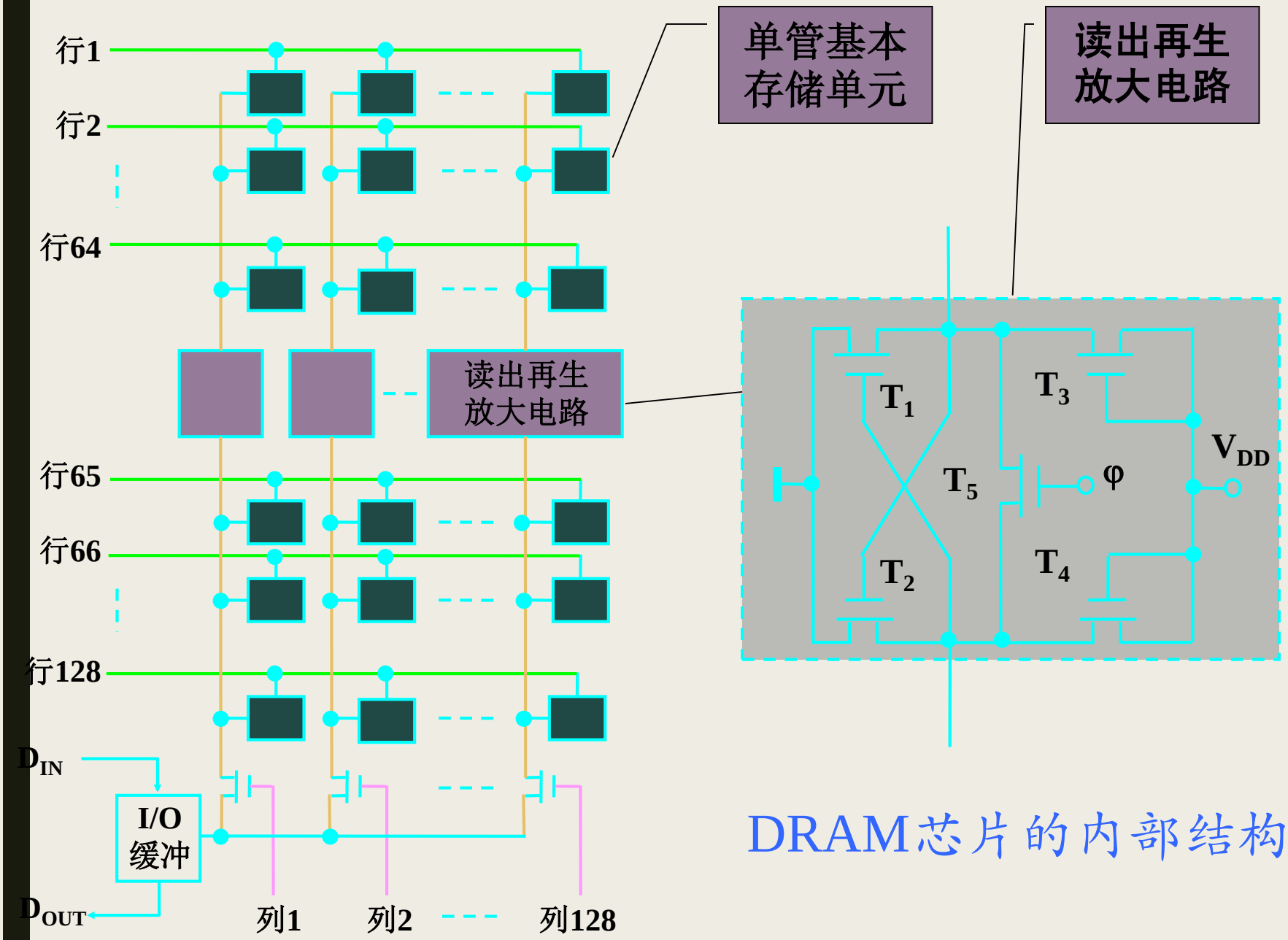
存储信息的原理：

①写操作：

行和列的选择信号为”1” →
基本存储单元被选中→
数据输入/输出线送来的信息通过刷新放大器和T管送到电容Cs
→数据写入存储单元；

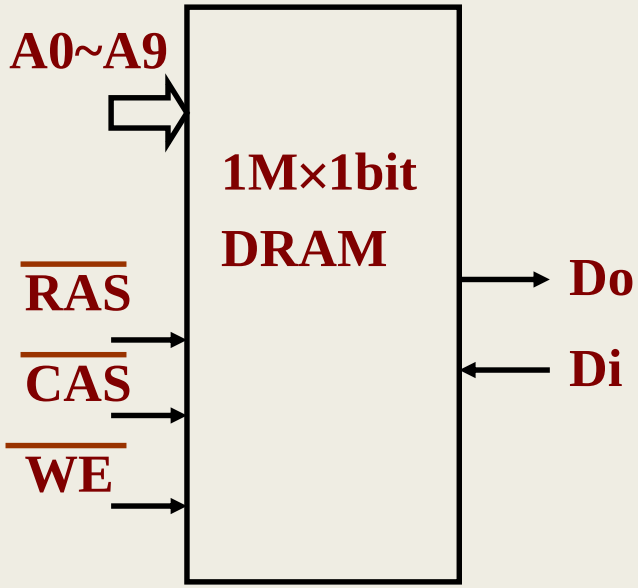
②读操作：

行地址译码使行选择信号为高电平→
行上管子T导通→
刷新放大器读取电容Cs上的电压值折合为”0”或”1”→
列地址译码使某列选通→行和列均选通的基本存储单元允许驱动→读出数据至Cd，即对Cd充电；



DRAM芯片的内部结构

5.3.2 DRAM的管脚信号与读/写操作



下图为1M×1bit的DRAM芯片：

WE：写允许信号

Di与Do为数据输入/输出信号

A0~A9：地址信号，∵ $1M=2^{20}$ ∴ 1Mb应有20位地址线，由于DRAM 的容量较大，又不希望有太多的引脚，所以大多数DRAM芯片都采用分时复用方式传输地址，将地址分为行地址和列地址两部分分时在地址线上传送。对本芯片用A0~A9先传送低10位地址，再传送高10位地址A10~A19。

读：存储地址需要分两批传送:

- 开始传送行地址，行地址选通信号-RAS有效，-RAS相当于片选信号
- 随后传送列地址，列地址选通信号-CAS有效
- 读写信号-WE读有效1
- 数据从D_{OUT}引脚输出

写：存储地址需要分两批传送:

- 开始传送行地址，行地址选通信号-RAS有效
- 随后传送列地址，列地址选通信号-CAS有效
- 读写信号-WE写有效0
- 数据从D_{IN}引脚进入存储单元

RAS: (Row Address Strobe)行地址选通信号，有效时在地址线上传送的是行地址(低10位)，用其后沿将低10位地址锁存到内部行地址锁存器。

CAS: (Column Address Strobe)列地址选通信号，有效时在地址线上传送的是列地址(高10位)，用其后沿将高10位地址锁存到内部列地址锁存器。

∴ DRAM芯片不需要片选CS。

❖ DRAM的刷新

由于基本单元电路简单，使DRAM的集成度(集成基本存储单元数)很高，但DRAM的附属电路较复杂。(需读出放大器，整形，刷新等电路)

为什么DRAM要不断地刷新?

由于DRAM是靠电容Cs存储信息的，Cs有电荷时为逻辑“1”，没有电荷时为逻辑“0”。

但由于任何电容都存在漏电，因此当电容Cs存有电荷时，过一段时间由于电容的放电会导致电荷流失，信息也会丢失。

解决的办法是刷新，即每隔一定时间(大约1~4ms)就要刷新一次，使原来处于逻辑“1”的电容的电荷又得到补充，而原来处于电平“0”的电容仍保持”0”。

注意：

①两次刷新的时间间隔与温度有关。

②DRAM的刷新是一行一行进行的，每刷新一行的时间称为刷新周期。

刷新策略有集中、分散、异步刷新方式三种。

刷新模式有-RAS有效，自动刷新两种。

1、DRAM的刷新策略

DRAM芯片有片内刷新，片外刷新。

①集中刷新

将整个刷新周期分为两部分,前一部分可进行读、写或维持(不读不写)，后一部分不进行读写操作而集中对DRAM刷新操作。这种方式控制简单。但在刷新过程中不允许读写，存在死时间。

②分散刷新(隐式刷新)

在每个读写或维持周期之后插入刷新操作，刷新存储矩阵的一行所有单元。这样把一个存储系统的周期分为两部分，读写、维持时间和刷新时间。优点是控制简单，不存在死时间；缺点是刷新时间占整个读写系统时间的一半，故只用于低速系统。

③异步刷新

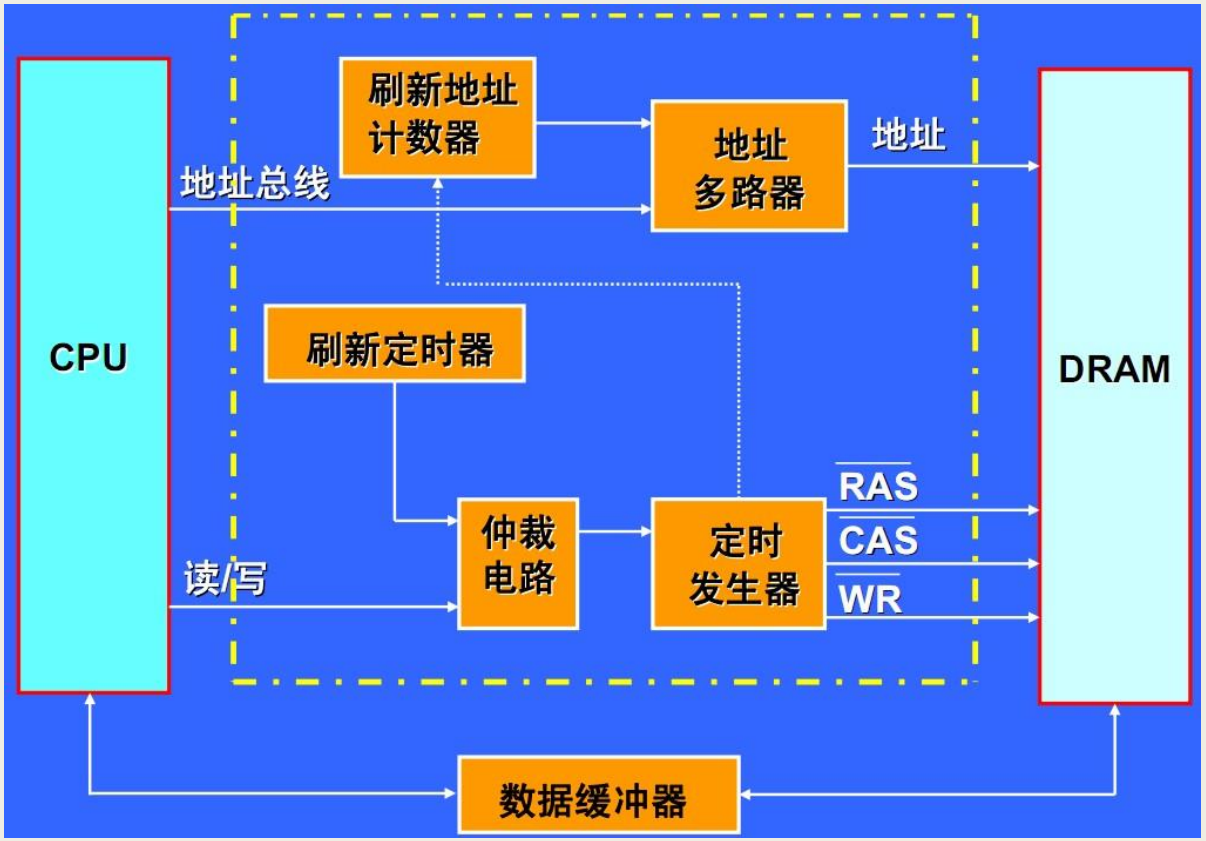
利用CPU不访问存储器的时间进行刷新操作。若按照预定的时间间隔应该刷新时，CPU正在访问存储器，刷新周期可以向后稍微延迟一段时间，只要保证在刷新周期内所有的行都能得到刷新即可。

这种方式优点是：对CPU访存的效率和速度影响小，又不存在死时间；缺点是：控制电路较复杂。

总之，可以在DMA控制器的控制下进行分散或异步刷新，也可在中断服务程序中进行集中或分散刷新。用DMA方式刷新比中断方式效率高。

5.3.4 DRAM控制器

CPU和DRAM之间的接口电路，把CPU的信号转换成适合DRAM芯片的信号，解决DRAM芯片地址两次打入和刷新控制等问题。



5.3.5 PC机的DRAM存储器

- PC机采用各种类型的DRAM作为可读写主存。
- 通过提高存储器芯片的密度，可以扩充存储器的容量。PC机的DRAM容量从早期的十几千字提高到目前的数百兆字节
- 存储器存取速度的提高，除了采用更高速度的器件外，主要是改进存储器的组织结构和访问方式

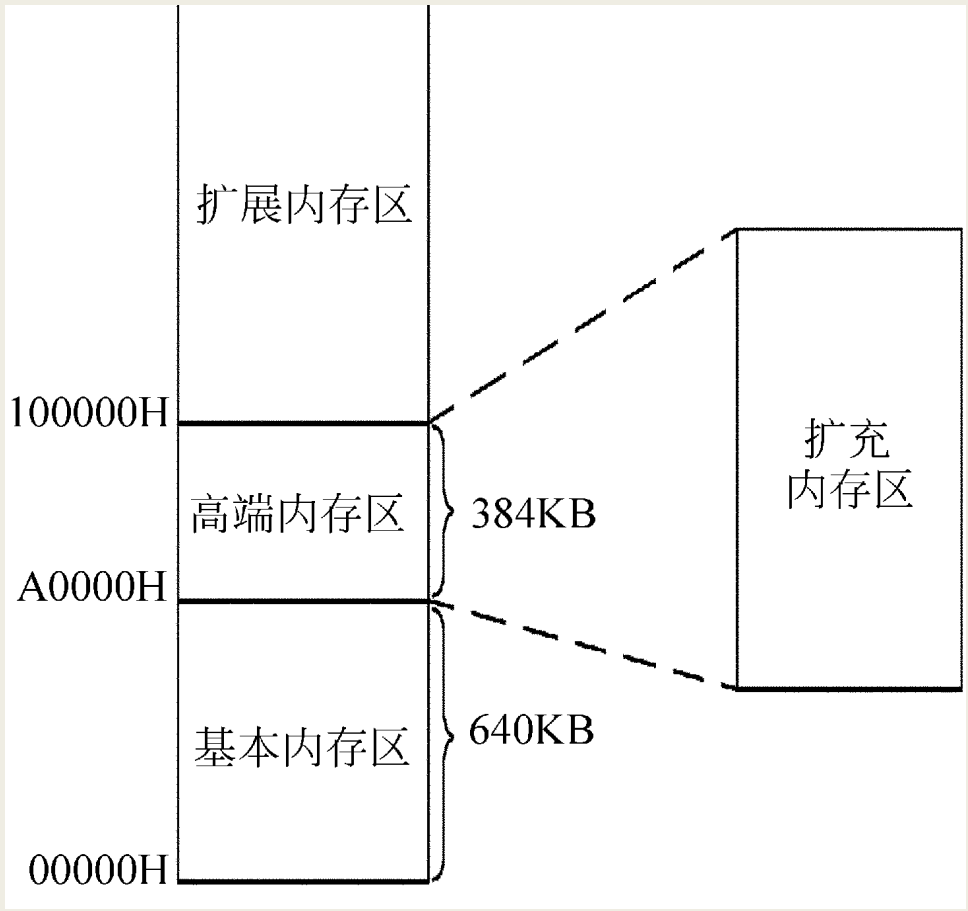
1、PC机随机存储器的演变
2、DRAM内存条的接口特性

- 为了便于存储器的扩充升级，一般将多片DRAM芯片塑封在一个长条形小电路插件板上，以DRAM存储条形式来构成具有32位或64位数据总线宽度的内存，电路板可以插入到主机板上的标准存储器插槽中——内存条。
- 目前大多使用的是168线双边接触内存模块（DIMM）插槽，分为缓冲型和非缓冲型。

3、内存的组织

①内存的分区结构

都是由DRAM组成的，用分区方式进行层次化组织；有利于软件的开发和系统的维护



基本内存区：主要供DOS操作系统使用

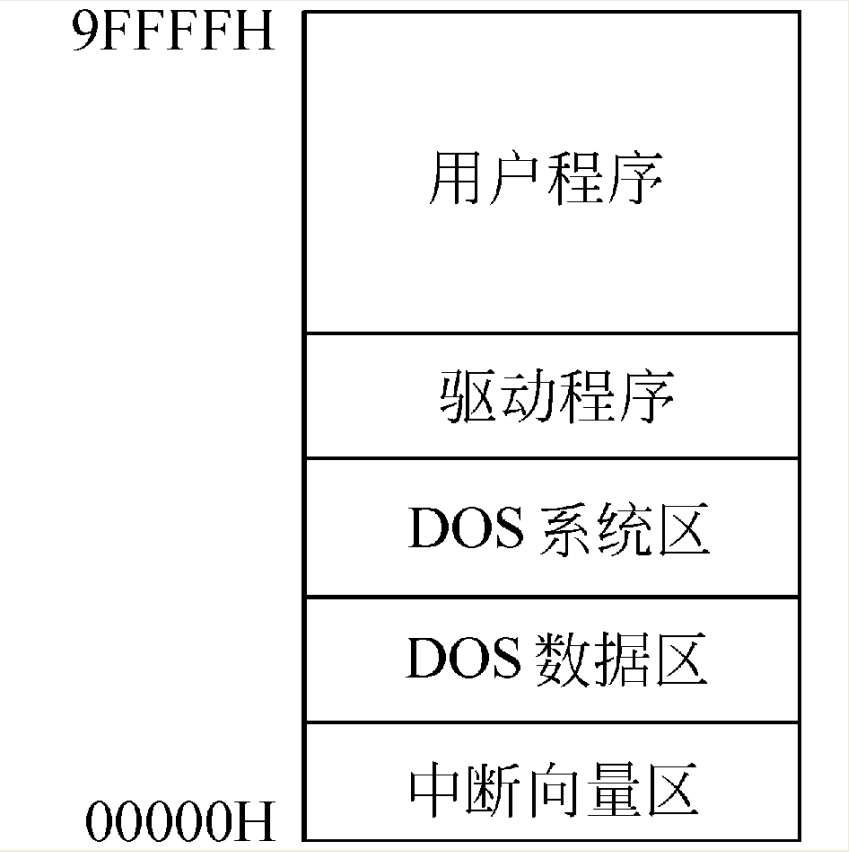
高端内存区：留给系统ROM和外部设备的适配卡缓冲区使用；

扩充内存区：通过在总线槽上插内存扩充卡来扩大内存空间；实际上是CPU直接寻址范围以外的物理存储器。

扩展内存区：32位微机系统中才有的内存区。是指1MB以上，但不是通过内存扩充卡映射来获得的内存空间；

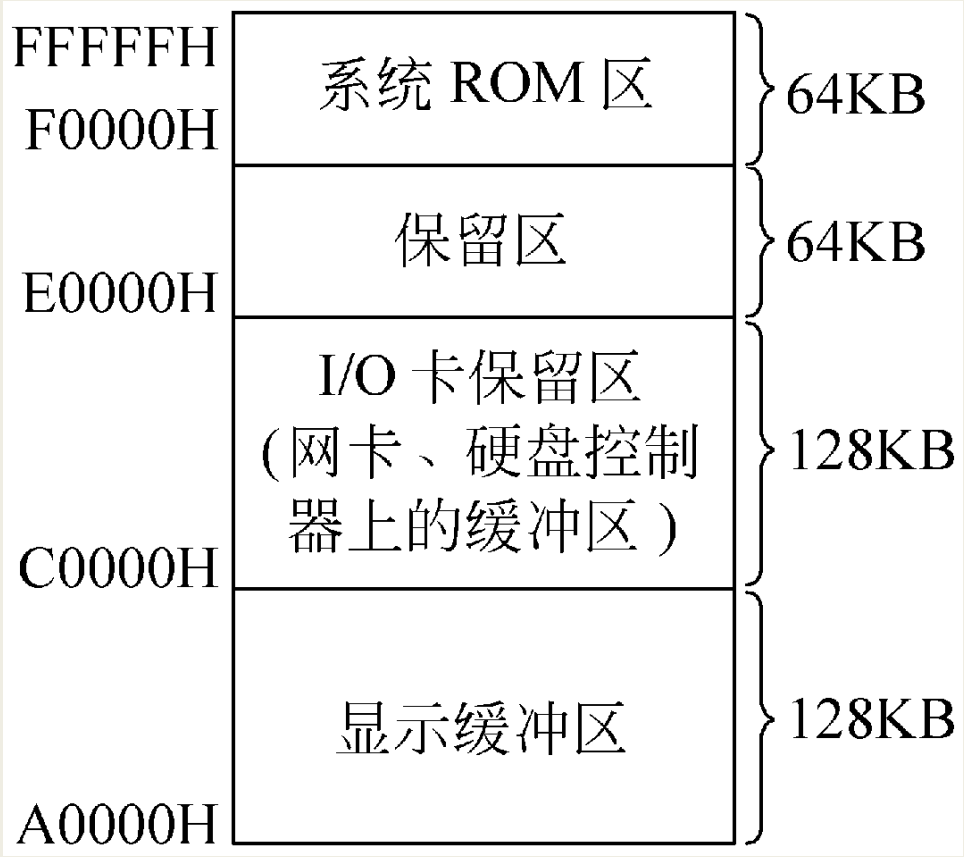
基本内存区的组织

- 用户不能动
- 640KB
- 主要供DOS操作系统使用，容纳了DOS操作系统，DOS运行需要的系统数据、驱动程序以及各种操作系统都要用到的中断向量表

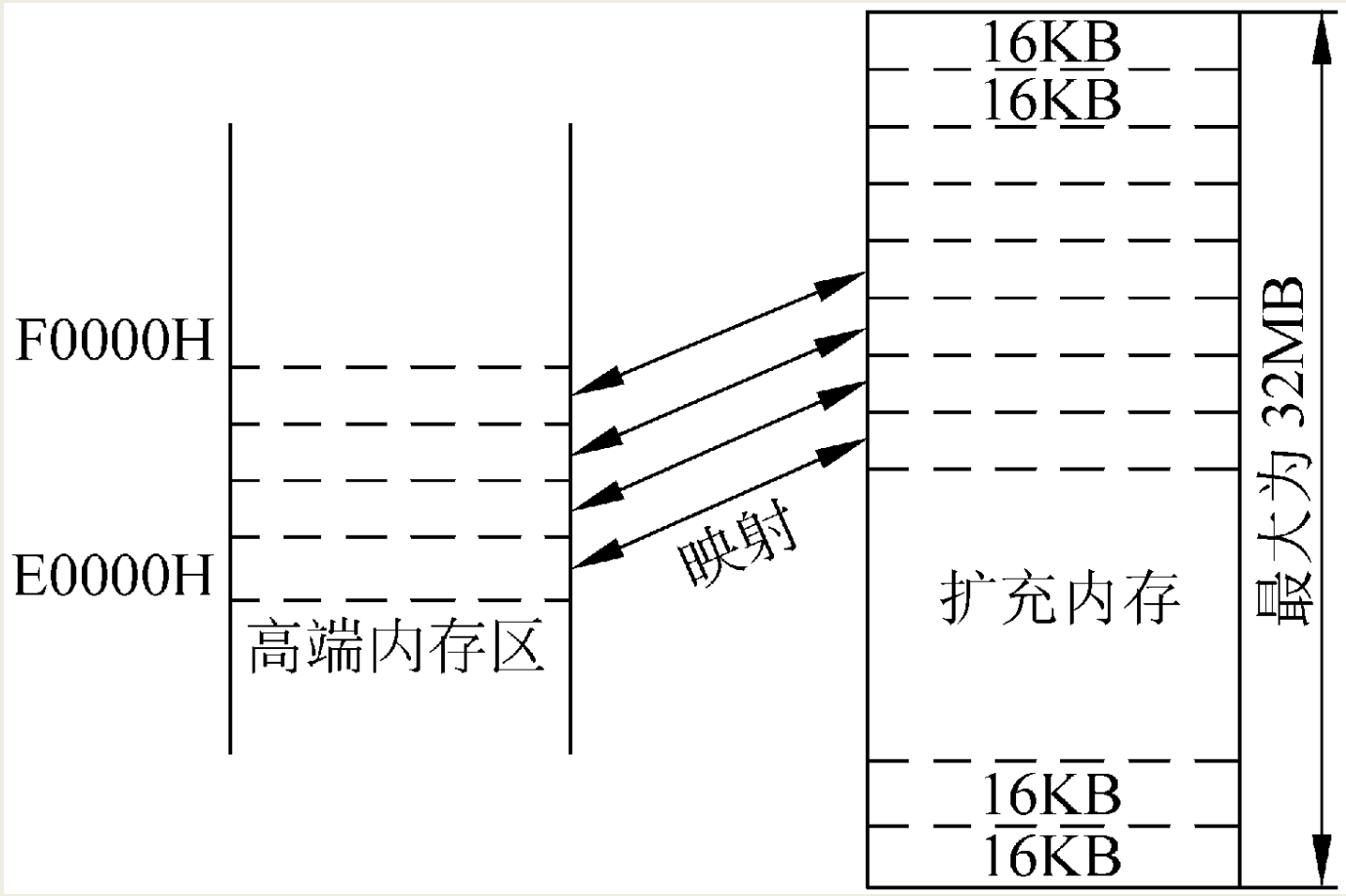


高端内存区的组织

- 384KB
- 留给系统ROM和外部设备的适配卡缓冲区使用

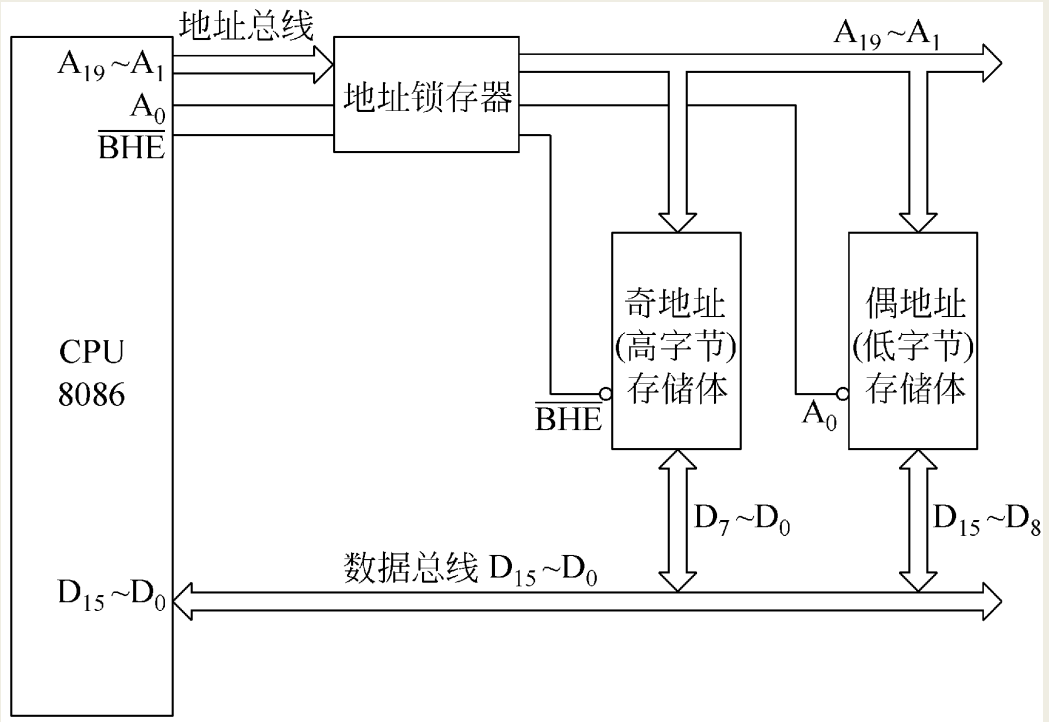


用高端内存区64KB映射扩充内存的1个页组



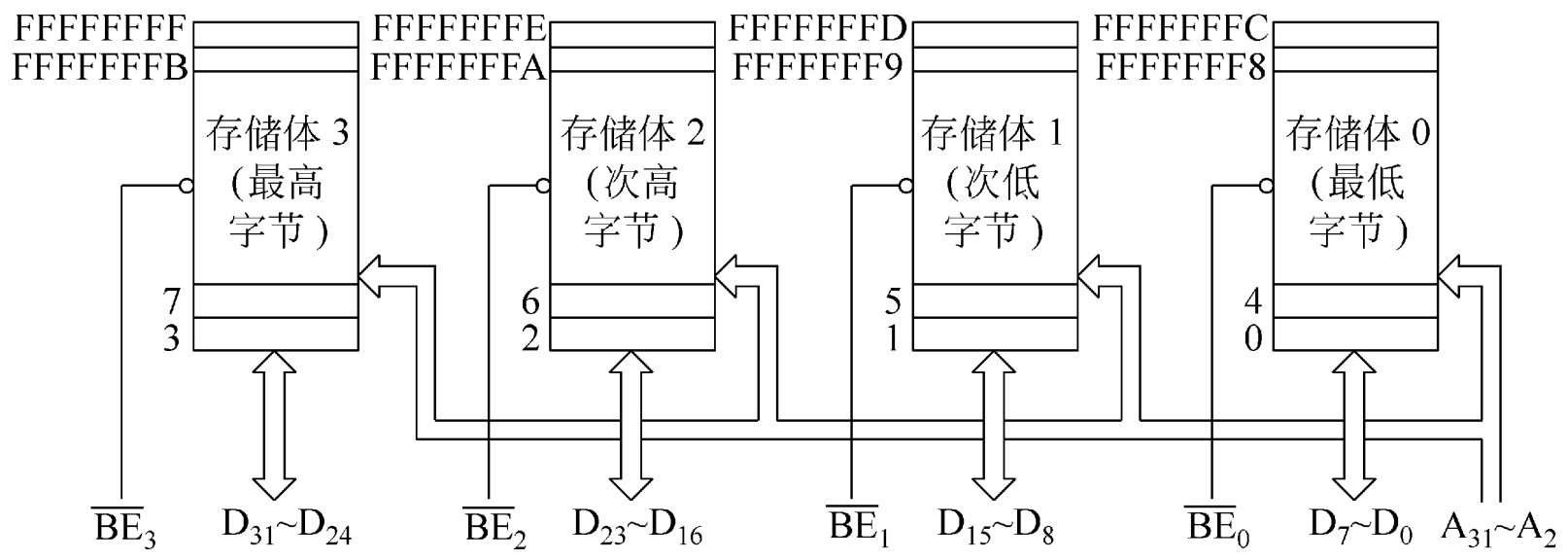
②16位微机系统的内存组织

- 8086用20位地址总线寻址1MB存储空间，(00000H~FFFFFH)
- 由两个存储体组成：均为512KB，由信号A_0和BHE作为存储体选择信号：
  - 奇地址存储体
  - 偶地址存储体
- 16位CPU对存储器的访问分为：
  - 按字节访问
  - 按字访问



③32位微机系统的内存组织

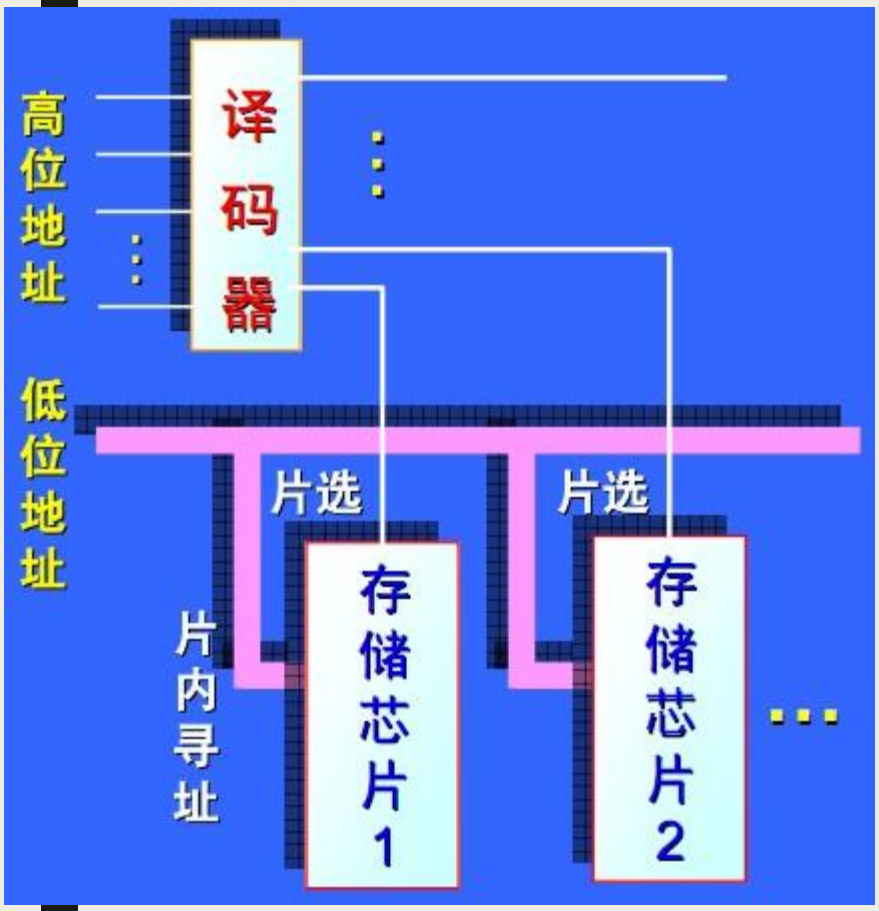
- 用32位地址总线寻址4GB的物理地址空间，(0~FFFFFFFFH)
- 分为4个存储体，每个均为1GB，字节允许信号BE_3~BE_0作为存储体选择信号分别连接一个存储体



4个存储体均与32位数据总线相连，也均与地址线A31~A2相连；字节允许信号BE3~BE0作为存储体选择信号分别连接一个存储体，当某个字节允许信号为有效电平时，便选中对应的存储体

5.4 存储器扩展与寻址

5.4.1 存储器的地址选择



容量、速度、功能、字长的要求:

存储器由多个芯片构成，CPU进行读/写操作时，首先应选中特定的芯片，称为片选，然后从该芯片中选择所要访问的存储单元。

存储器的寻址必须有两个部分:

低位地址线连到所有存储器芯片，实现片内寻址;

将高位地址线通过译码器或线性组合后输出作为芯片的片选信号，实现片间寻址。

5.4.2 存储空间的扩展

位扩展、字扩展、字位扩展三种方法。

1、位扩展

是指芯片的字满足要求，而位数不够，需要对存储单元的位数进行扩展。

【例】 采用64K×1的DRAM存储器芯片，扩展成64K×8的RAM存储器。

解：

【例】若用Intel 2114(1K×4)的存储芯片组成1K×8位的存储器，需要多少片芯片？并画出连接图。

2、字扩展

扩容时芯片的位数已符合要求，而字数不够，需要利用多个芯片扩充容量，也就是扩充了存储器地址范围。只是增加地址范围。

“字”的含义就是内存中存放的一个数据，可以是8/16/32位宽度。

【例】 用16K×8的芯片构成64K×8存储器。

解：

【例】若用Intel 6264(8K×8)存储芯片构成一个16K×8位的存储器系统，需要多少片芯片？并画出连接图。

3、字位扩展

存储芯片的容量和位数都需要进行扩展。

【例】 用1K×4的SRAM芯片2114构成4K×8的存储器。