存储器层次结构（2）

王晶

jwang@ruc.edu.cn，信息楼124

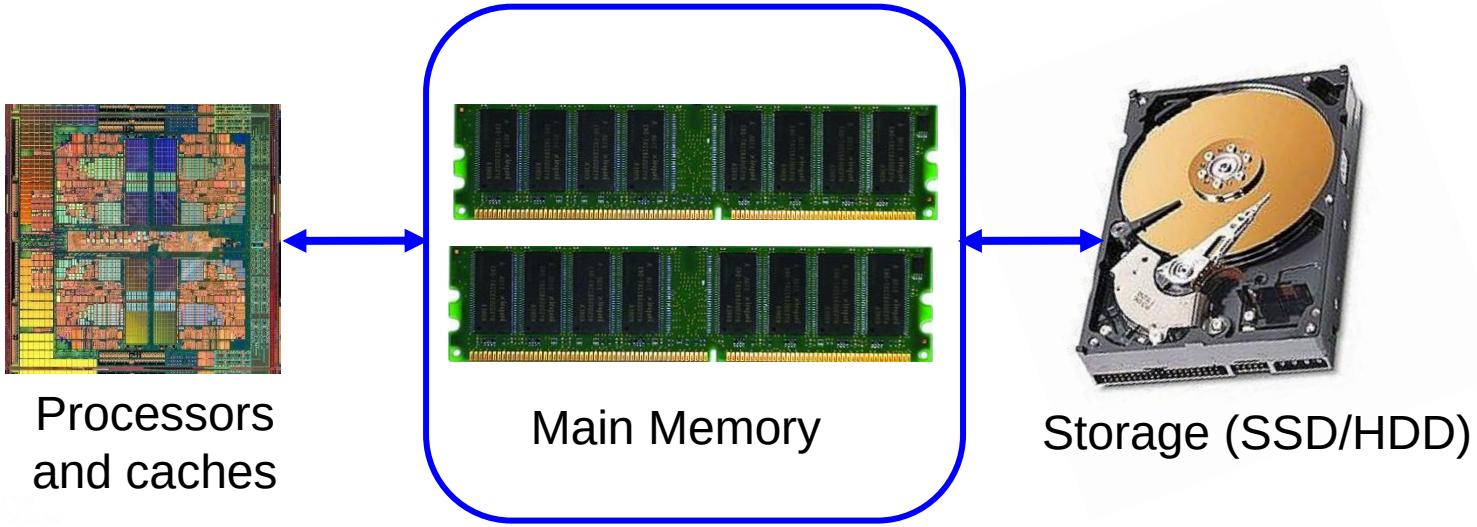
2024年12月

静态和动态存储器芯片特性

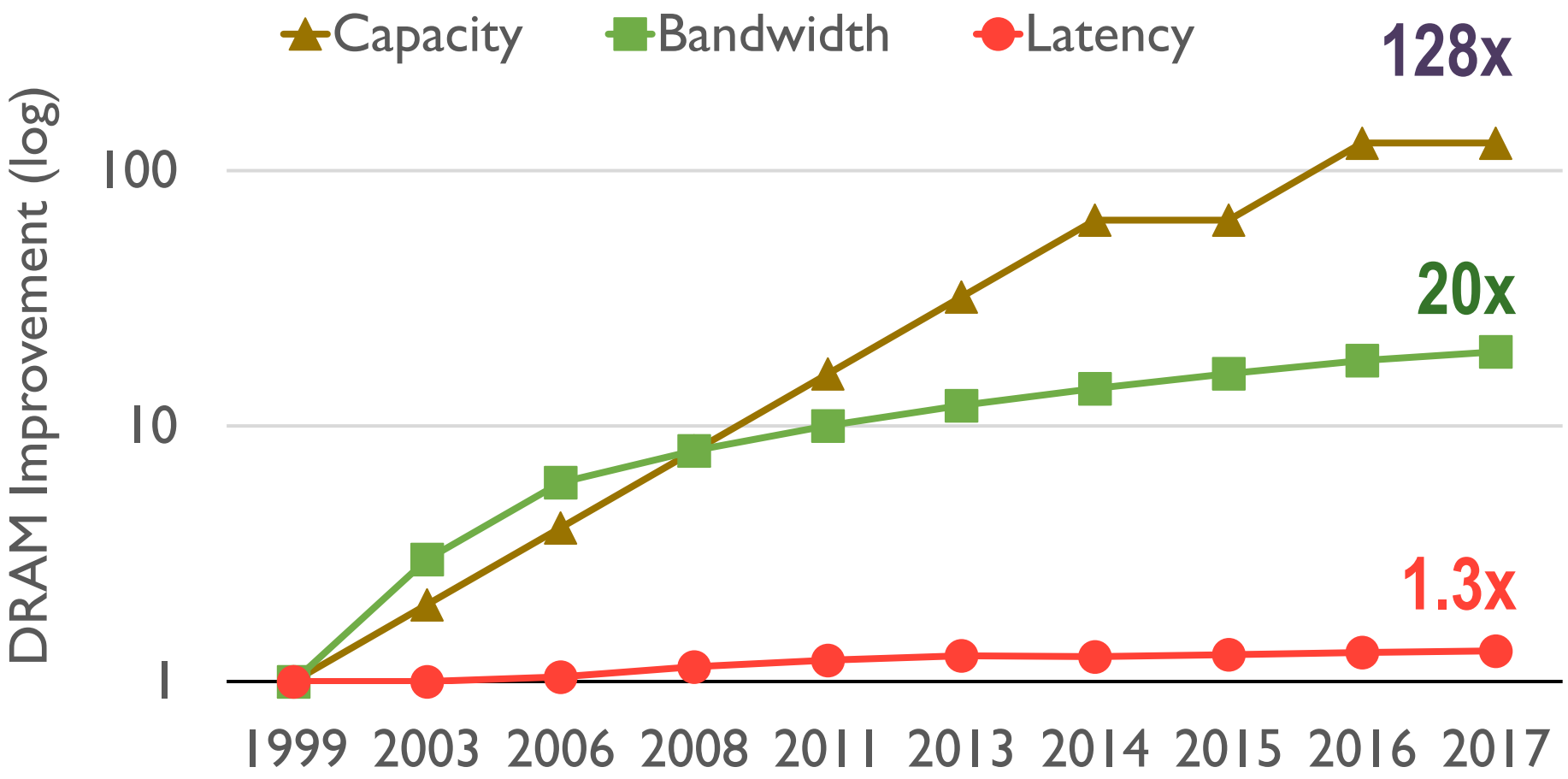
	SRAM	DRAM
存储信息	触发器	电容
破坏性读出	非	是
需要刷新	不要	需要
送行列地址	同时送	分两次送
运行速度	快	慢
集成度	低	高
发热量	大	小
存储成本	高	低

主存

- 主存是计算系统的关键组件：服务器平台、移动和嵌入式设备、桌面计算机以及传感器
- 主存储系统必须在尺寸、工艺、效率、成本和管理算法上可扩展，以维持性能增长和技术发展。
- 数据访问是一个主要的瓶颈。应用程序对数据的需求日益增长。
- 能耗是关键的限制因素。数据移动能耗占主导地位，尤其是对于芯片外到芯片内的数据移动。



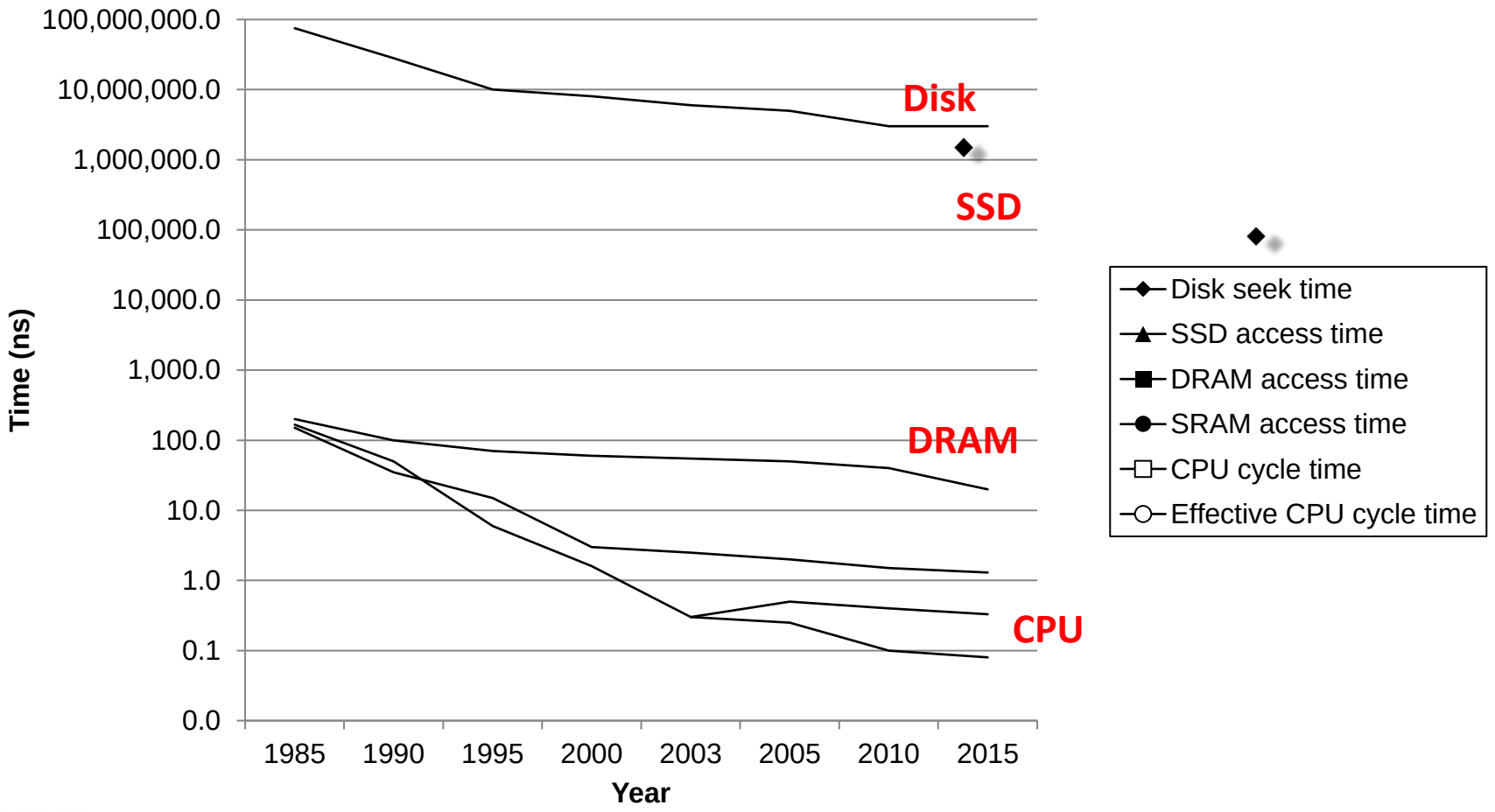
Capacity, Bandwidth & Latency



主存的延迟几乎没有变化

CPU和主存之间的鸿沟

The gap widens between DRAM, disk, and CPU speeds.



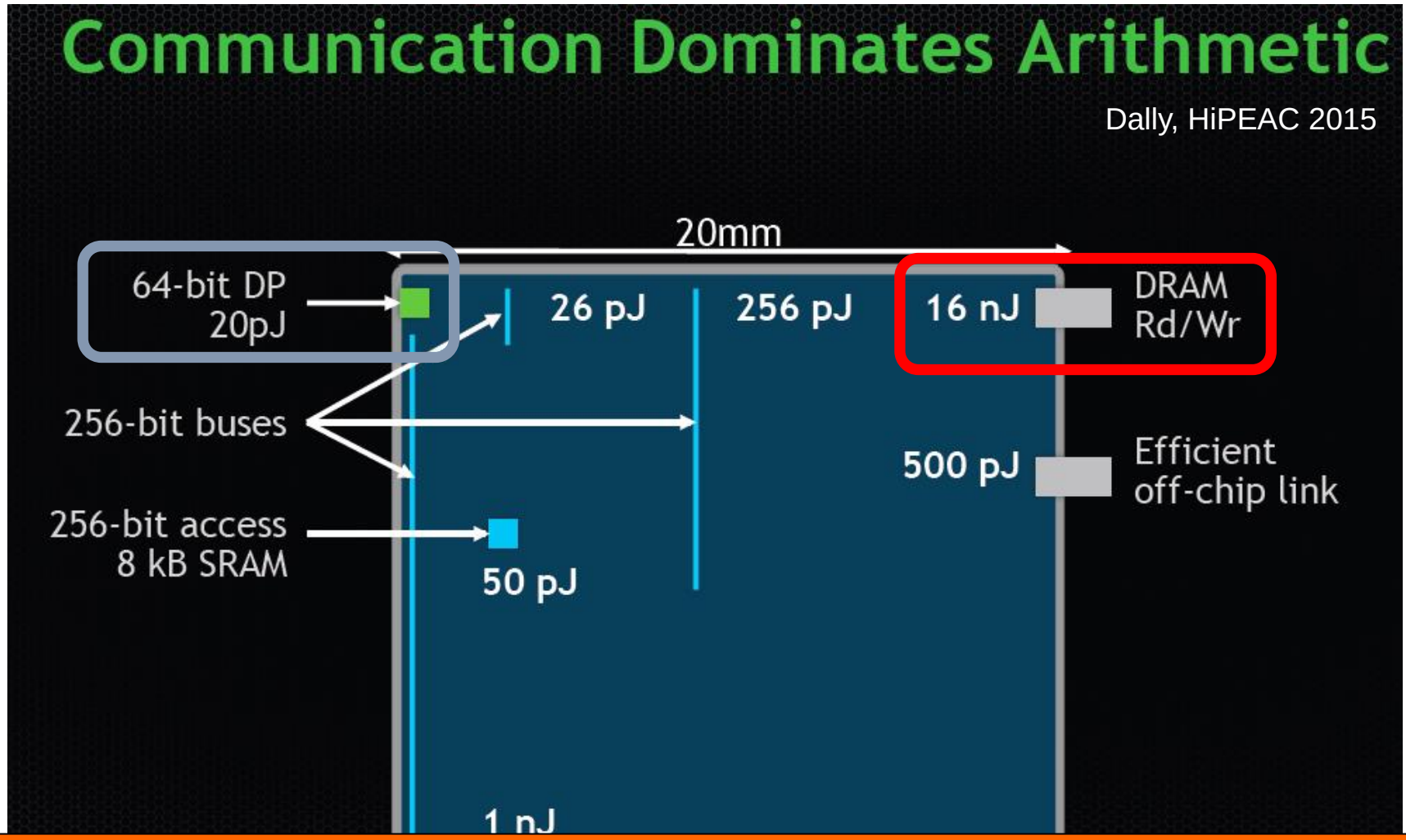
CPU Clock Rates

Inflection point in computer history
when designers hit the "Power Wall"

	1985	1990	1995	2003	2005	2010	2015	*2015:1985*
CPU	80286	80386	Pentium	P-4	Core 2	Core i7(n)	Core i7(h)	
Clock rate (MHz)	6	20	150	3,300	2,000	2,500	3,000	500
Cycle time (ns)	166	50	6	0.30	0.50	0.4	0.33	500
Cores	1	1	1	1	2	4	4	4
Effective cycle time (ns)	166	50	6	0.30	0.25	0.10	0.08	2,075

(n) Nehalem processor
(h) Haswell processor

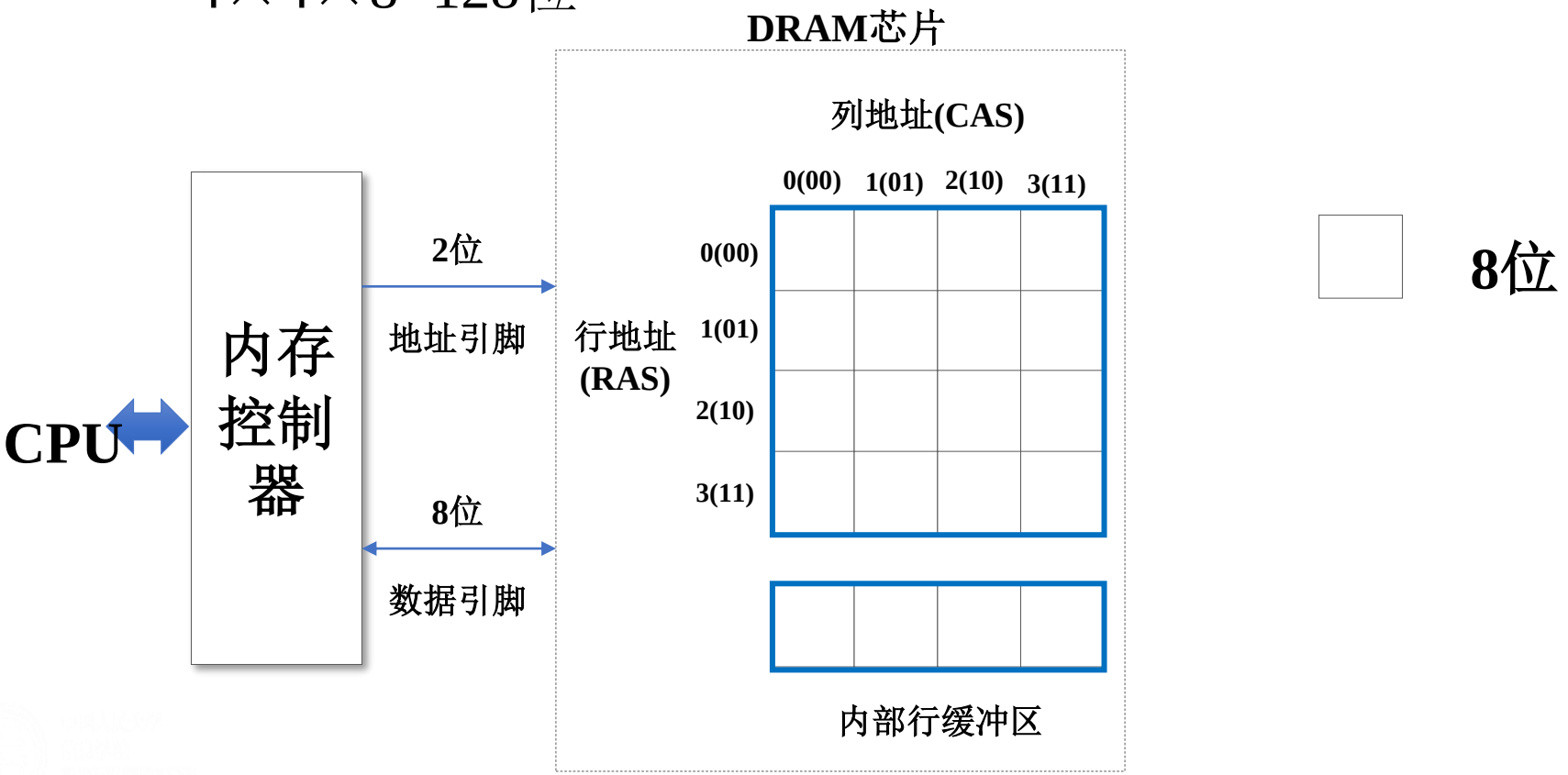
Data Movement vs. Computation Energy



主存的能耗大约是复杂加法的~1000X

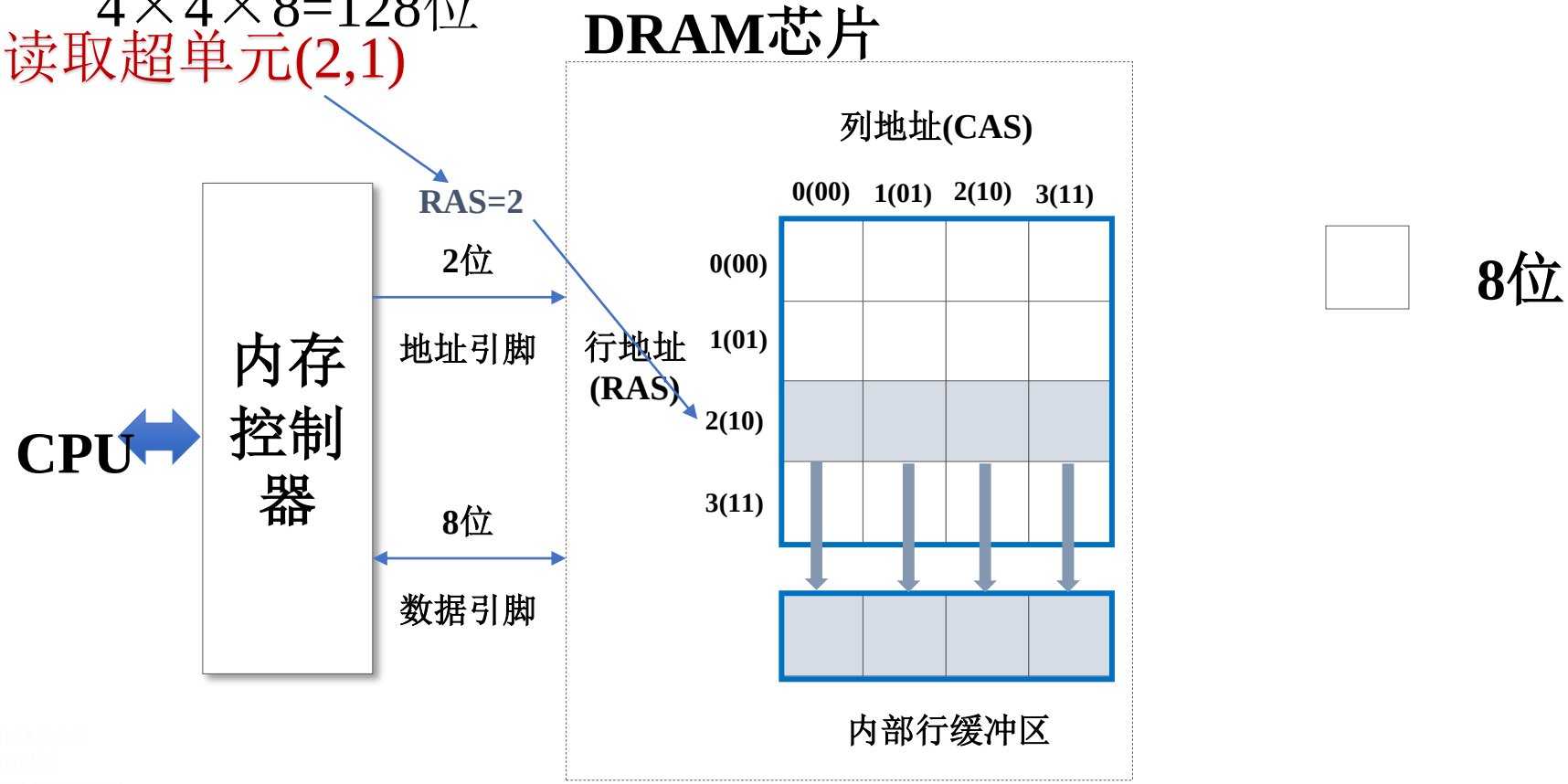
传统的DRAM

- 128位16×8的DRAM芯片视图
  - 芯片由*r*行×*c*列=*d*个超单元（suppercell）组成
  - 超单元由*w*位组成，芯片共有*d*×*w*位存储信息
  - 示例：4行、4列、8位存储单元总存储容量为：4×4×8=128位



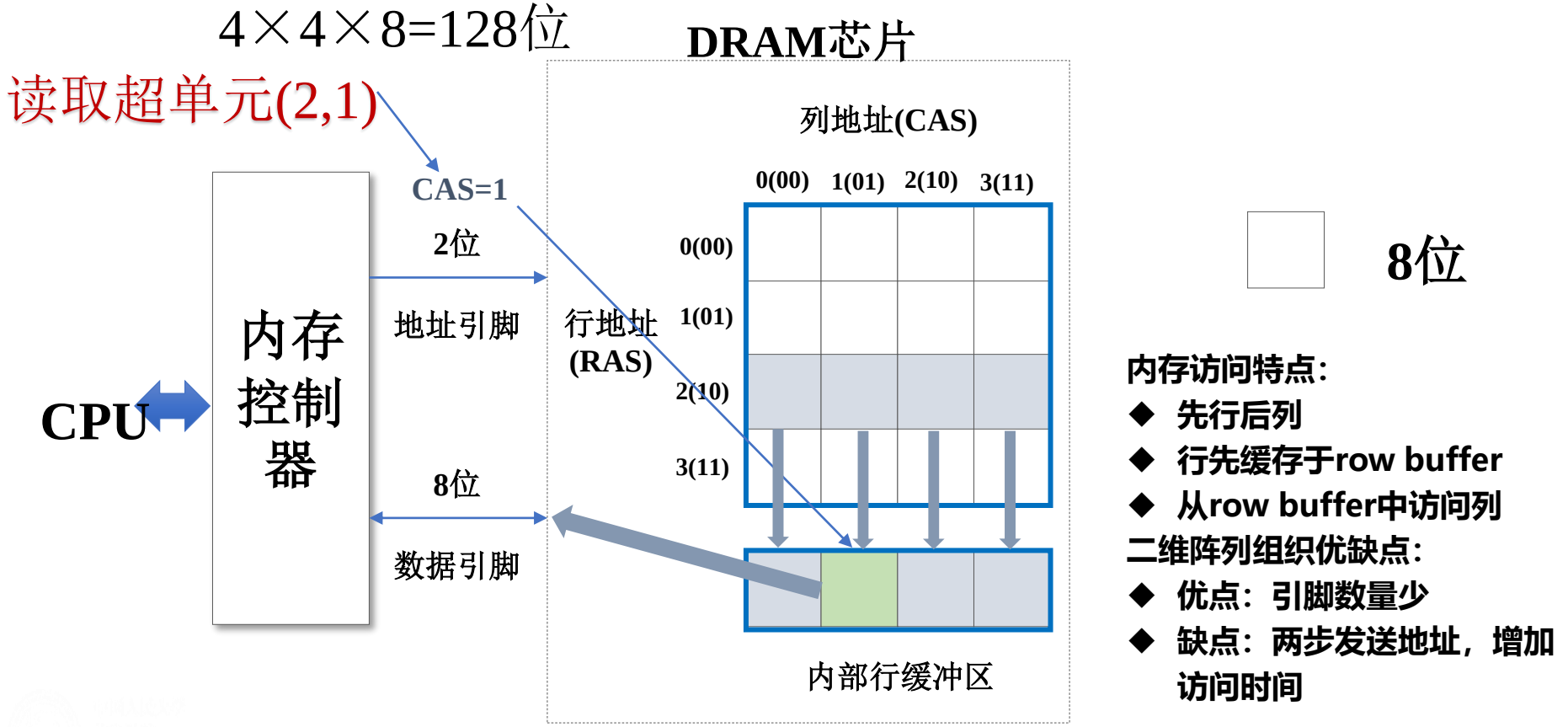
传统的DRAM

- 128位16×8的DRAM芯片视图
  - 芯片由*r*行×*c*列=*d*个超单元（suppercell）组成
  - 超单元由*w*位组成，芯片共有*d*×*w*位存储信息
  - 示例：4行、4列、8位存储单元总存储容量为：4×4×8=128位

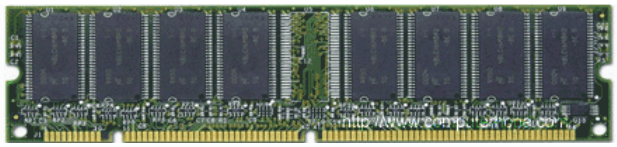


传统的DRAM

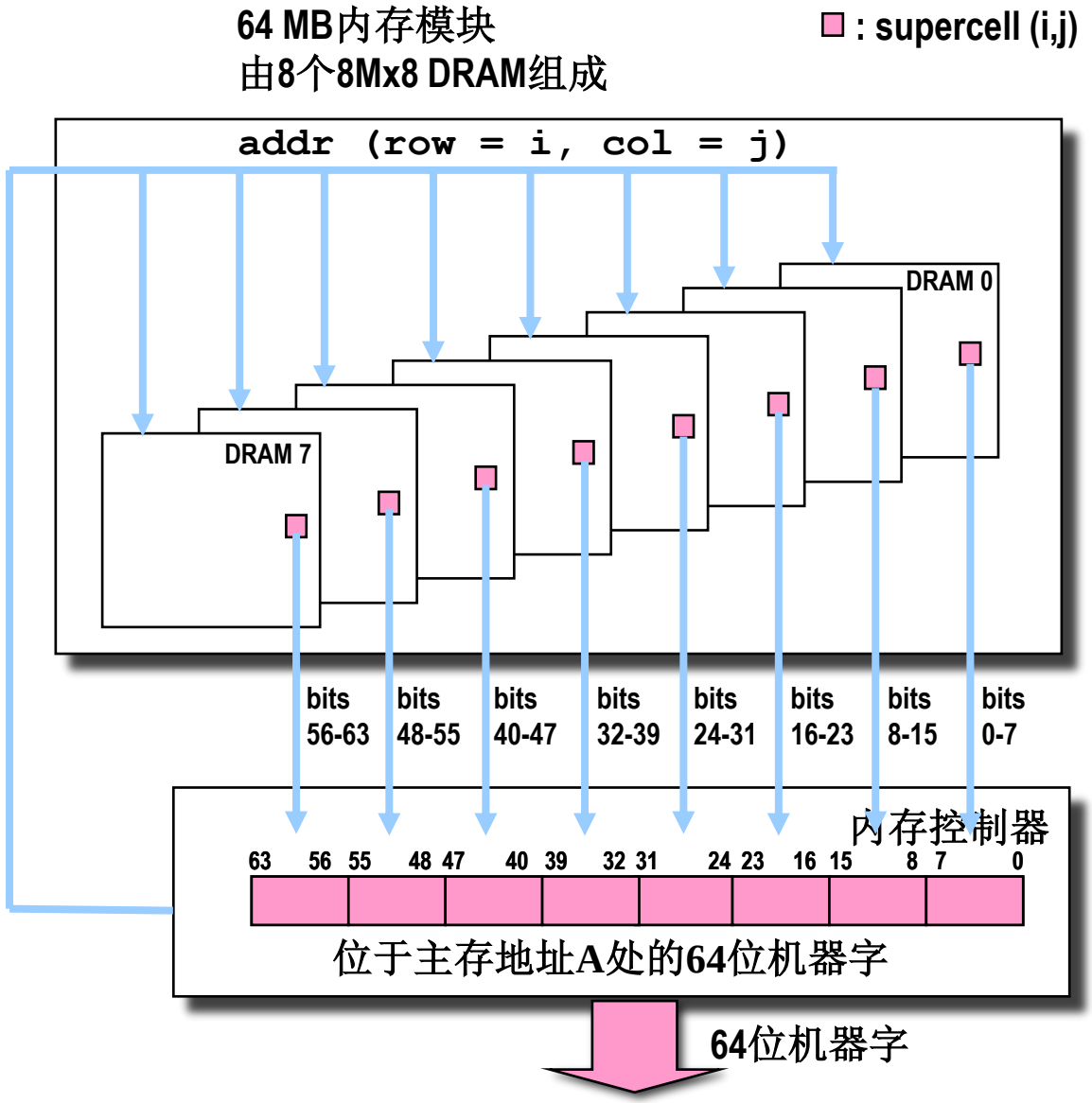
- 128位16×8的DRAM芯片视图
  - 芯片由*r*行×*c*列=*d*个超单元（suppercell）组成
  - 超单元由*w*位组成，芯片共有*d*×*w*位存储信息
  - 示例：4行、4列、8位存储单元总存储容量为：4×4×8=128位



内存模块



- DIMM：双列直插内存模块(Dual Inline Memory Module, DIMM)
- 以64位为块传送数据到内存控制器和从内存控制器传出数据
- 访问地址A处机器字：
  - 内存控制器将地址A转换为超单元地址(*i*,*j*)，发送到内存模块，
  - 内存模块将(*i*,*j*)广播到每个DRAM模块，每个DRAM模块输出(*i*,*j*)超单元内容
  - 模块中电路收集这些输出，合并成一个64位字，返回给内存控制器



填空题 3分

设置

1. 容量为1024B的存储空间，地址线需要 [填空1] 位

2. 某计算机字长16位，主存地址空间大小64KB，按字节编址，寻址范围 [填空2] 到 [填空3]

作答

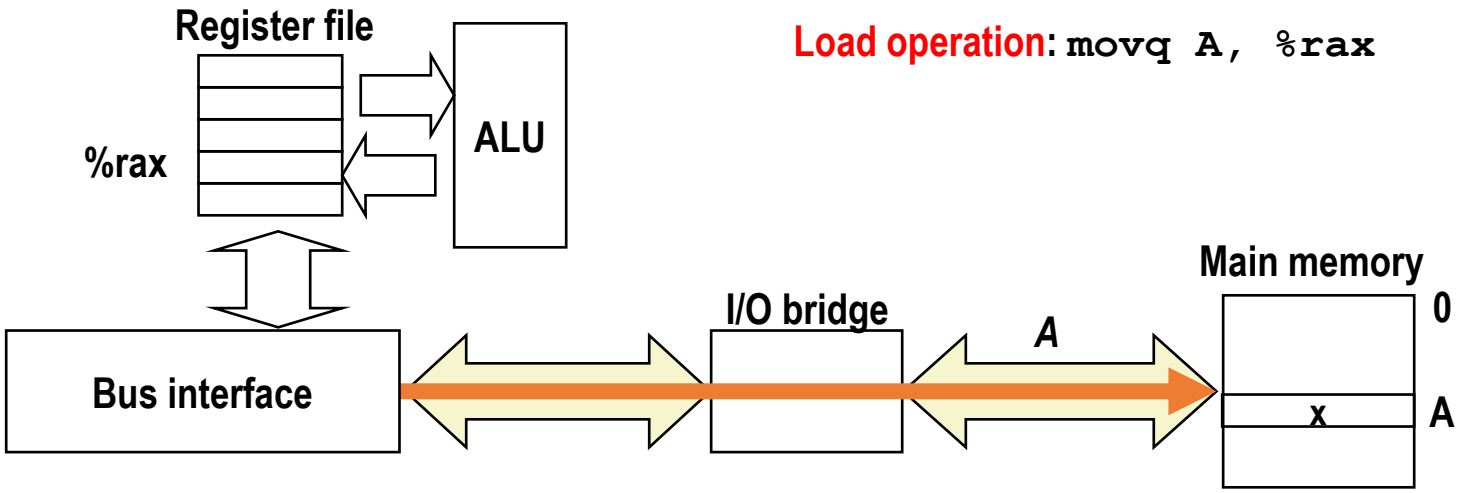
DRAM的地址

- r表示一个DRAM阵列中的行数，c表示列数，b_r表示行寻址需要的位数，b_c表示列寻址需要的位数。对于下面每个DRAM，确定2的幂数的阵列维数，使得$\max(b_r, b_c)$最小，$\max(b_r, b_c)$是对阵列的行寻址或列寻址所需的位数中较大的值
- 目标：使纵横比最小，使地址位数最小，数组越接近正方形，地址位数越少

组织	r	c	b_r	b_c	$Max(b_r,b_c)$
16×1	4	4	2	2	2
16×4	4	4	2	2	2
128×8	16	8	4	3	4
512×4	32	16	5	4	5
1024×4	32	32	5	5	5

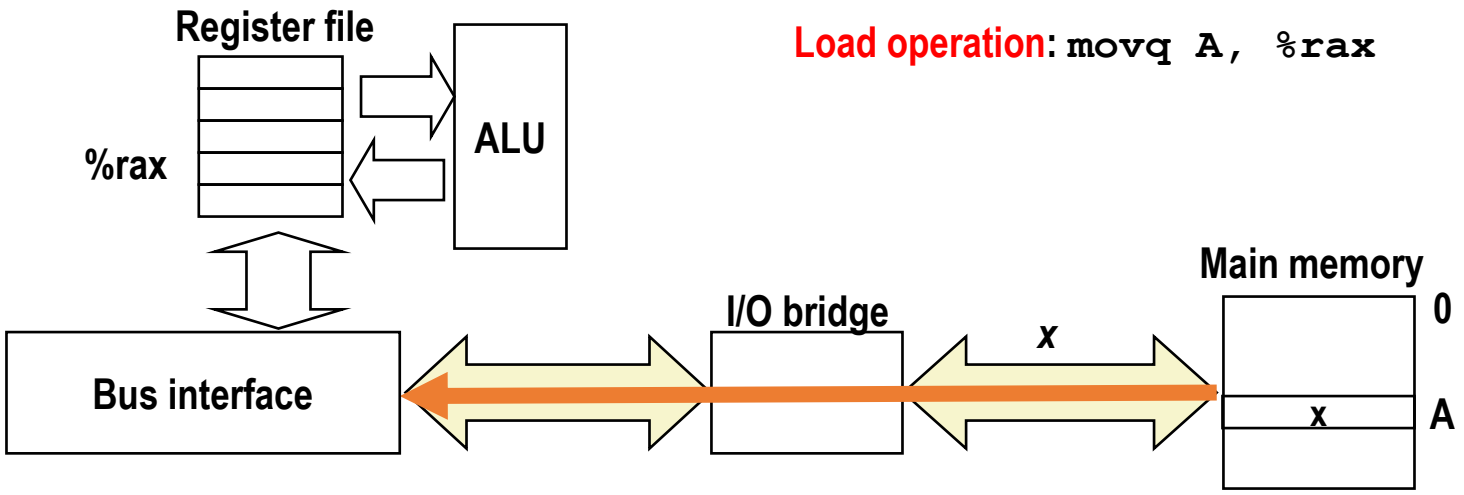
内存读操作

- Movq A, %rax
- 读事务的三个操作步骤：
  - CPU将地址A放到系统总线上



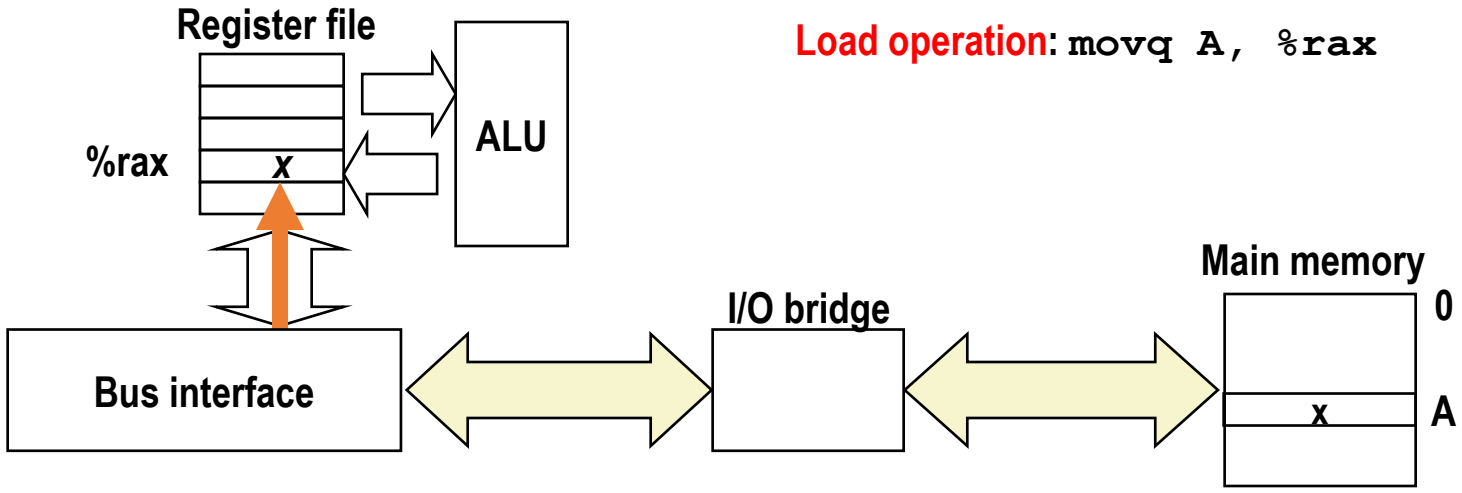
内存读操作

- Movq A, %rax
- 读事务的三个操作步骤：
  - CPU将地址A放到系统总线上
  - 主存从内存总线读地址A，从DRAM取出数据，并写到内存总线



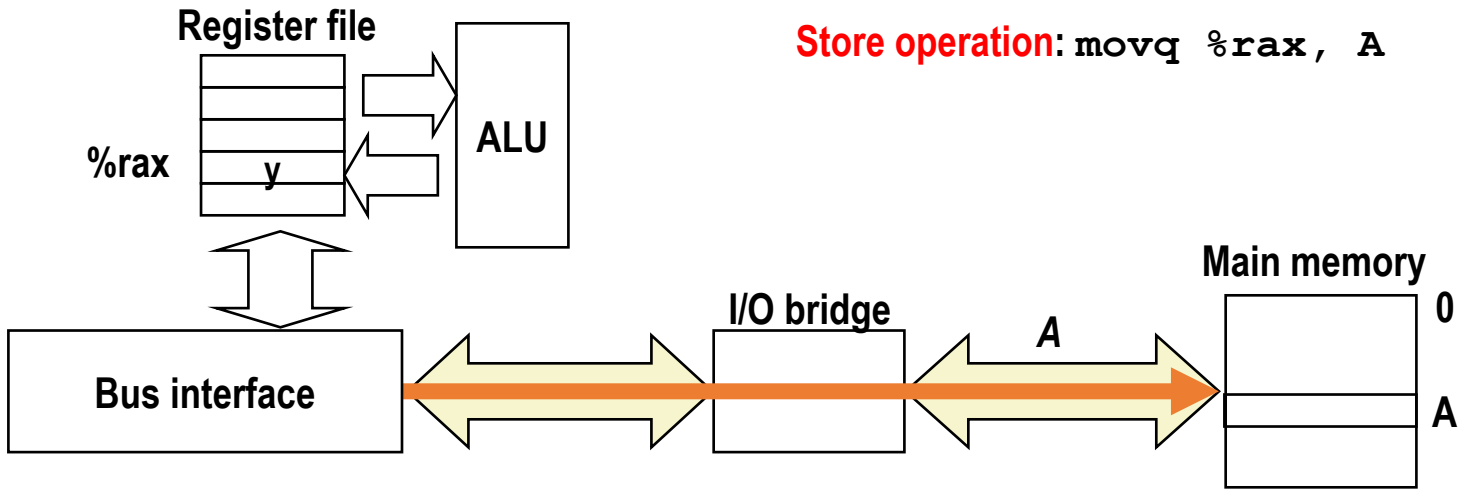
内存读操作

- Movq A, %rax
- 读事务的三个操作步骤：
  - CPU将地址A放到系统总线上
  - 主存从内存总线读地址，从DRAM取出数据，并写到内存总线
  - CPU从系统总线上读数据，并复制到寄存器%rax中



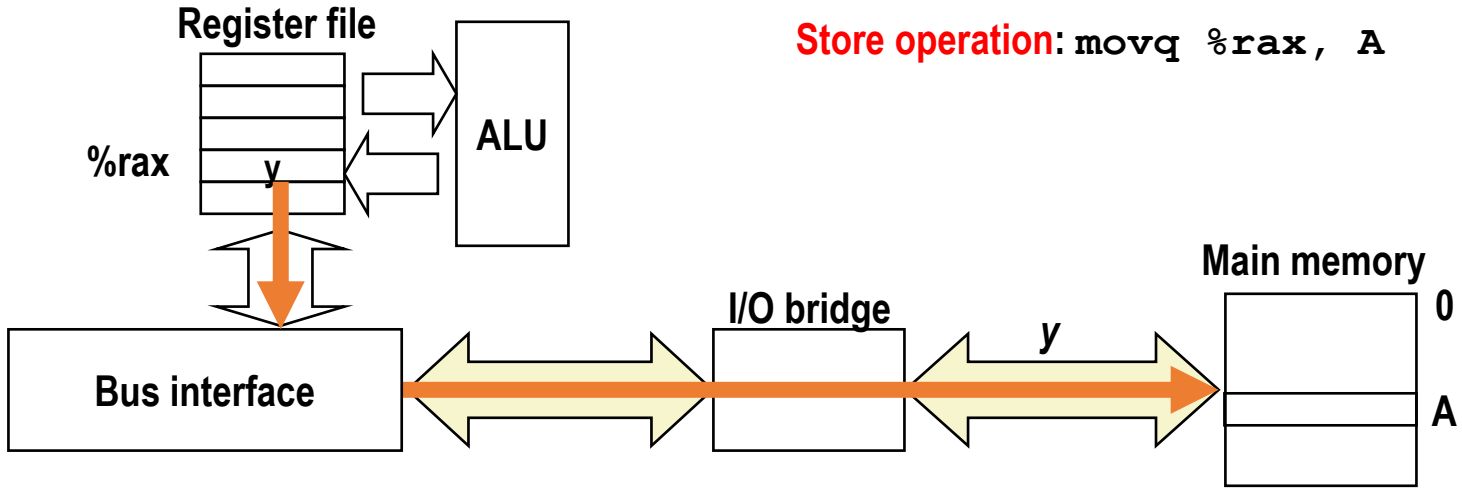
内存写操作

- Movq %rax, A
- 写事务的三个操作步骤：
  - CPU将地址A放到系统总线上



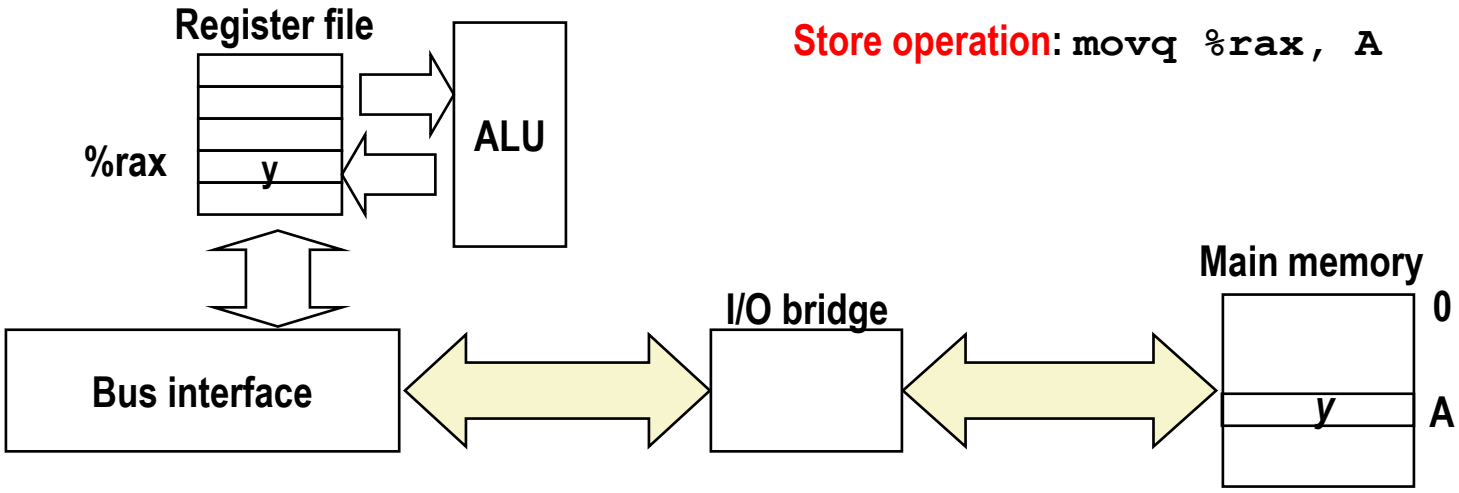
内存写操作

- Movq %rax, A
- 写事务的三个操作步骤：
  - CPU将地址A放到系统总线上
  - CPU将%rax中的数据复制到系统总线



内存写操作

- Movq %rax, A
- 写事务的三个操作步骤：
  - CPU将地址A放到系统总线上
  - CPU将%rax中的数据复制到系统总线
  - 主存从内存总线读出数据，并将这些位存储到DRAM中



更强的DRAMs

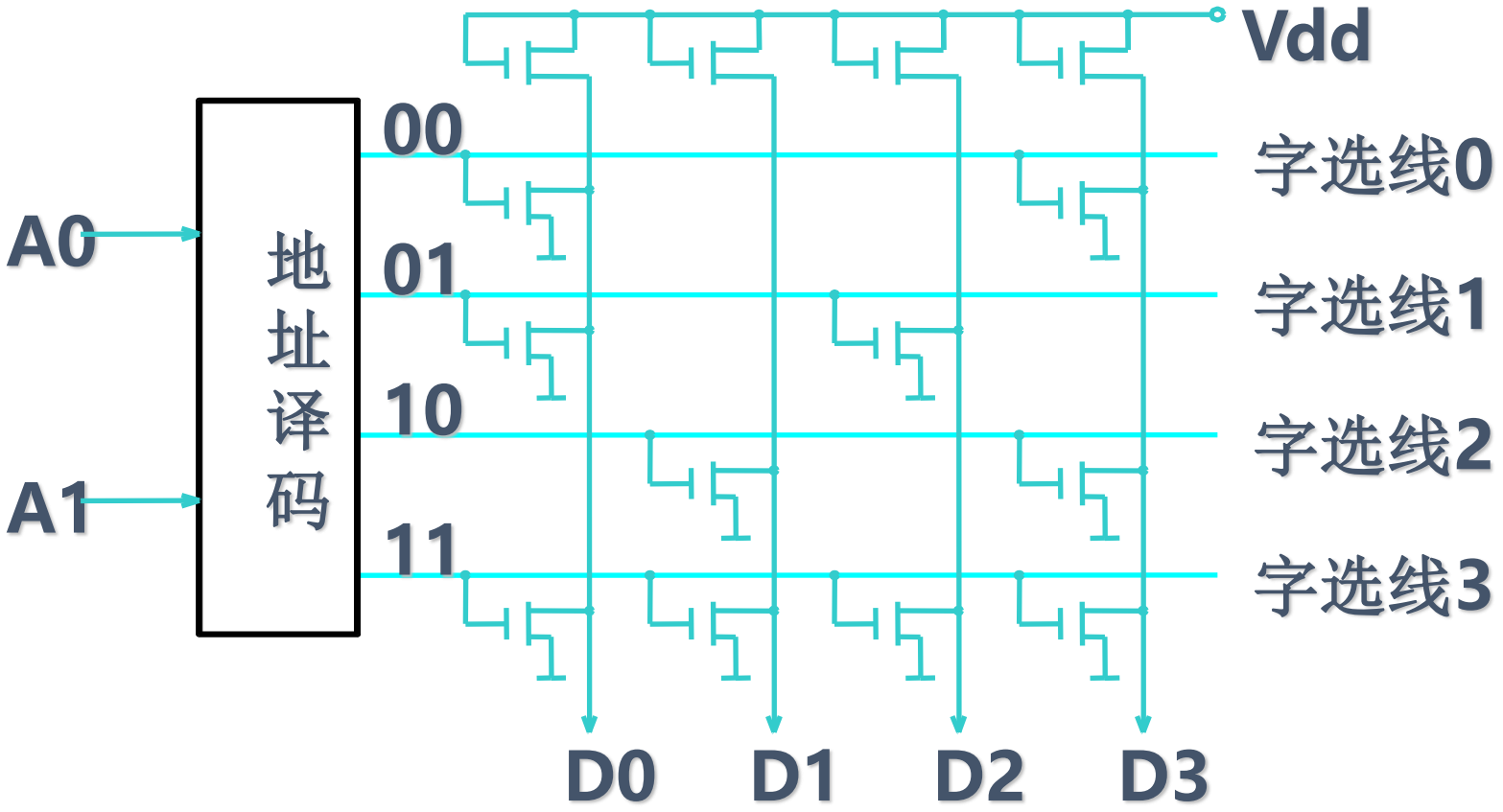
- 基本的DRAM单元自1966年发明以来就没有改变.
  - 1970年由英特尔商业化
- 具有更好接口逻辑和更快I/O的DRAM核心
  - 同步动态随机存取存储器(SDRAM)
    - 使用传统的时钟信号而不是异步控制
    - 允许重用行地址
  - 双倍数据速率同步动态随机存取存储器(DDR SDRAM)
    - 双倍边缘时钟每个周期每个引脚发送两个比特
    - 不同类型的区别在于小预取缓冲区的大小:
      - DDR (2 bits), DDR2 (4 bits), DDR3 (8 bits), DDR4 (16 bits)
    - 到2010年，成为大多数服务器和桌面系统的标配
    - 英特尔酷睿i7支持DDR3和DDR4 SDRAM

非易失存储

- 动态随机存取存储器（DRAM）和静态随机存取存储器（SRAM）是易失性存储器
  - 断电后会丢失信息
- 非易失性存储器即使断电也能保持数据
  - 只读存储器(ROM):在生产过程中编程
  - 可编程ROM (PROM):可以编程一次
  - 可擦除可编程ROM (EPROM):可以批量擦除（紫外线、X射线）
  - 电可擦除可编程ROM(EEPROM):具备电子擦除能力
  - 闪存: EEPROMs.具备部分（块级）擦除能力的EEPROM
    - 大约经过10万次擦除后会磨损
  - 3D XPoint（英特尔Optane）和新兴的非易失性存储器（NVMs）
- 非易失性存储器的用途
  - 存储在ROM中的固件程序(BIOS,磁盘控制器、网络卡、图形加速器、安全子系统)
  - 固态硬盘（取代U盘、智能手机、MP3播放器、平板电脑、笔记本电脑中的磁盘）
  - 磁盘缓存

掩模式只读存储器(ROM)

在ROM的制作阶段，通过“掩模”工序将信息写到芯片内的。　　容量为4×4位的电路情况。



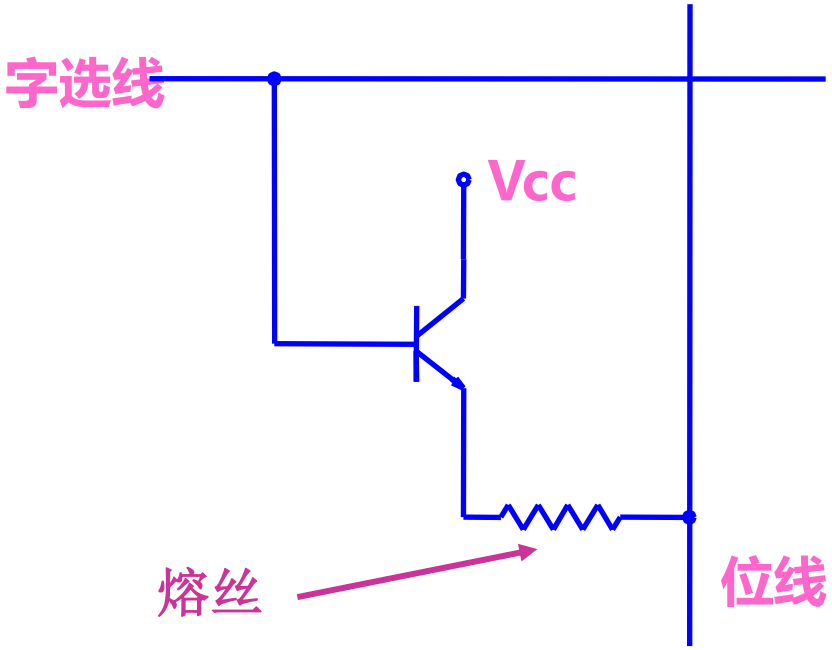
有管子的读出为“0”，无管子的读出为“1”

可编程只读存储器(PROM)

可编程ROM (PROM):可以编程一次

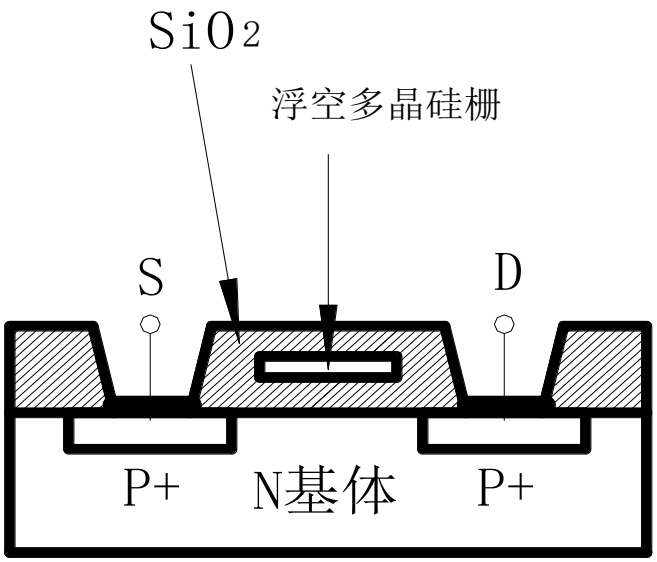
出厂时熔丝完好，表示存储元全为“1”。

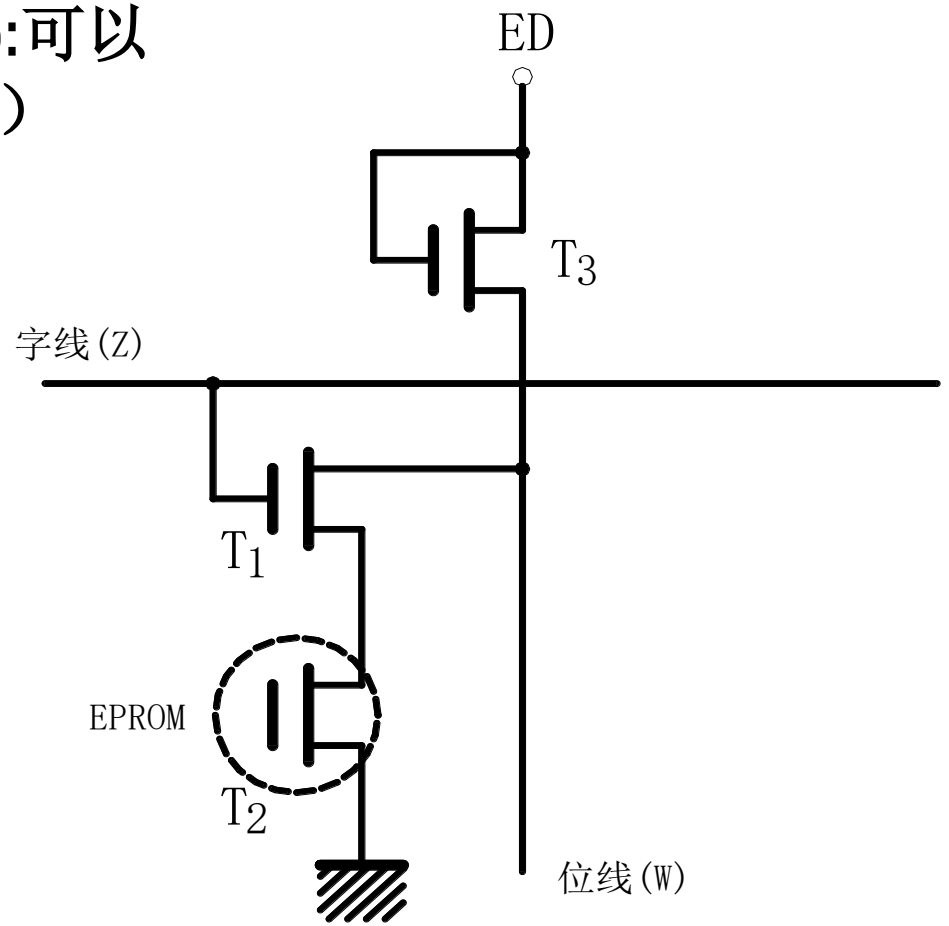
写“0”时，通大电流将熔丝烧断。

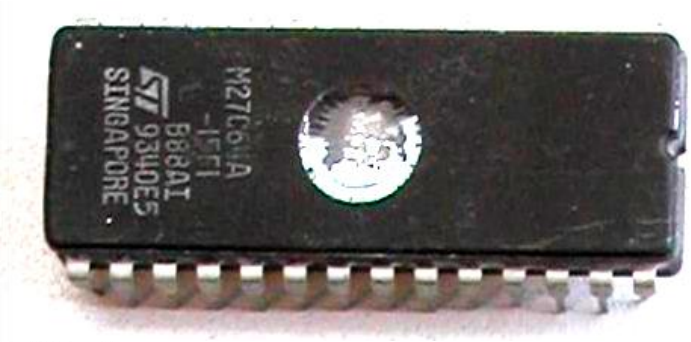


EPROM结构示意图

可擦除可编程ROM (EPROM):可以批量擦除（紫外线、X射线）

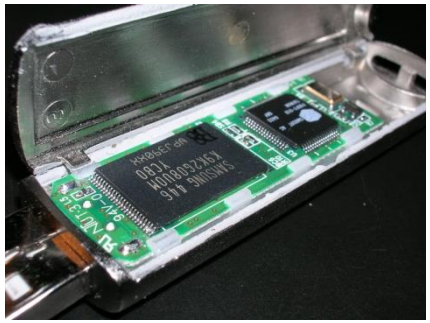






快擦写可编程只读存储器 FLASH

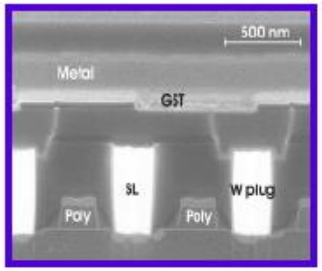
- Intel公司推出的一种半导体存储器
- 具有高密度、低成本、非易失性等特点的可读可写存储器。
- 可用电气方式快速擦写
- 既有DRAM的高集成度大容量的特点，又有可在线改写、断电信息不丢失的优点。
- 取代EPROM和EEPROM来保存主板等部件的BIOS，广泛应用于便携式计算机的PC卡存储器、USB电子盘、数码相机等。



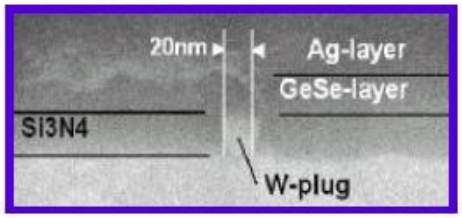
Emerging Nonvolatile Memories

FERAM

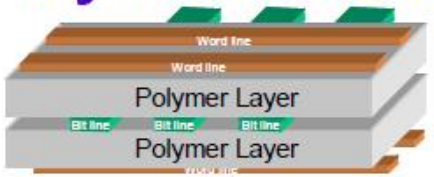
PCM



PMC RRAM

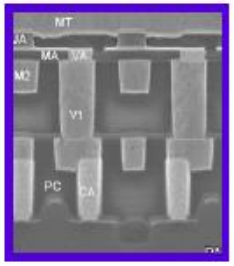


Polymer FeRAM

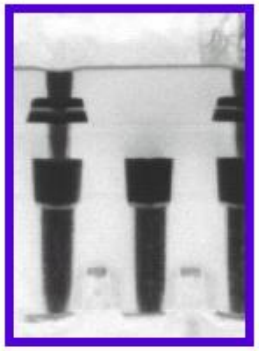


CNT

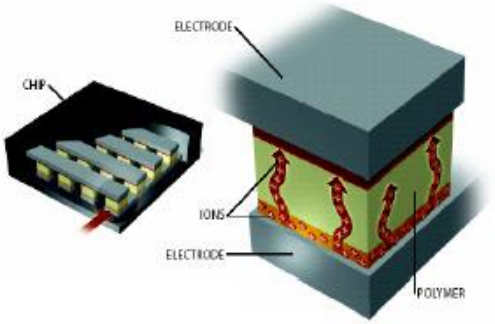
MRAM



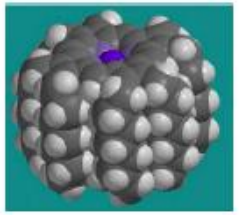
MOx-RRAM



Polymer RRAM



Molecular

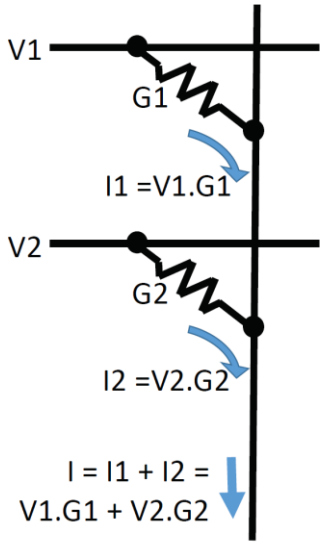


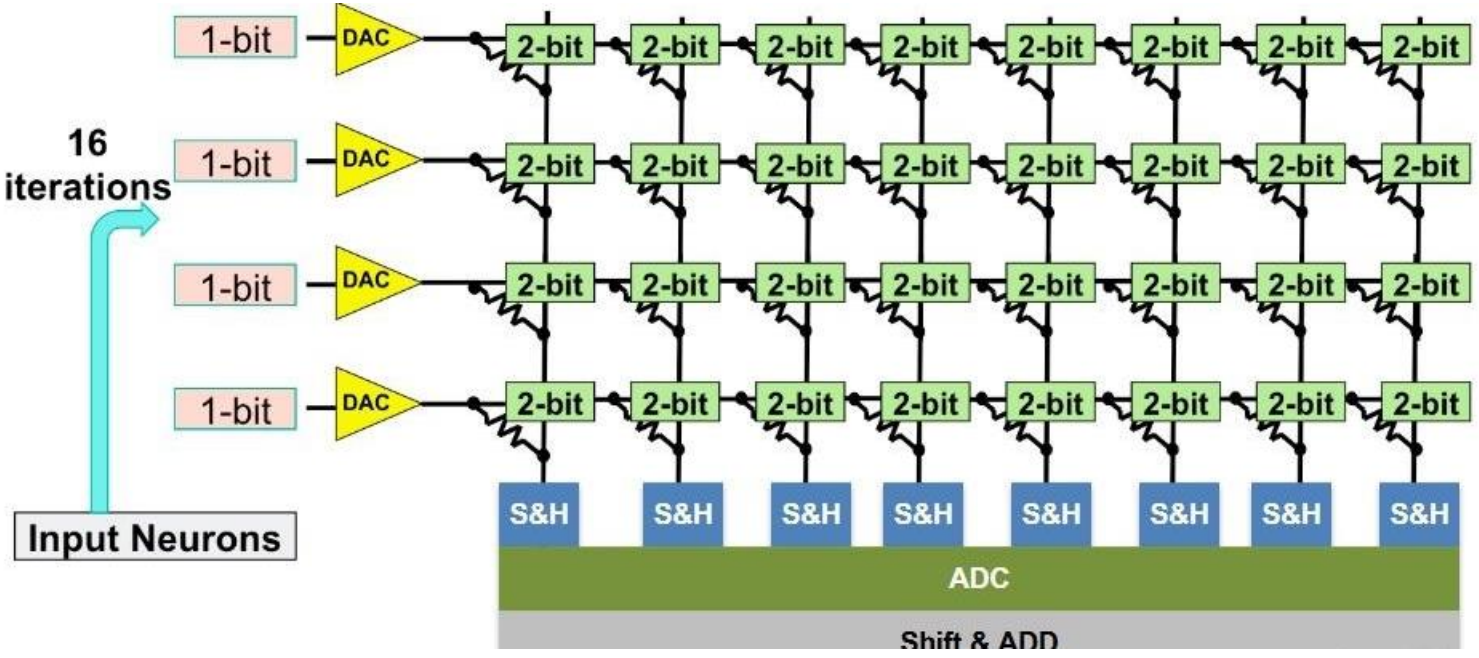
"Explosion" of "New" Memory Concepts

- New Storage Materials, New Storage Concepts
- Many Ideas, Varying Functionality/Cost, Many Still Unproven

Resistive Random Access Memory（RRAM）

- 多层网络中的大量存储和计算导致了大量的矩阵乘累加操作。
- ReRAM天然地适合进行乘累加操作。
- 因此，基于ReRAM的加速器被广泛用于DNN的部署。



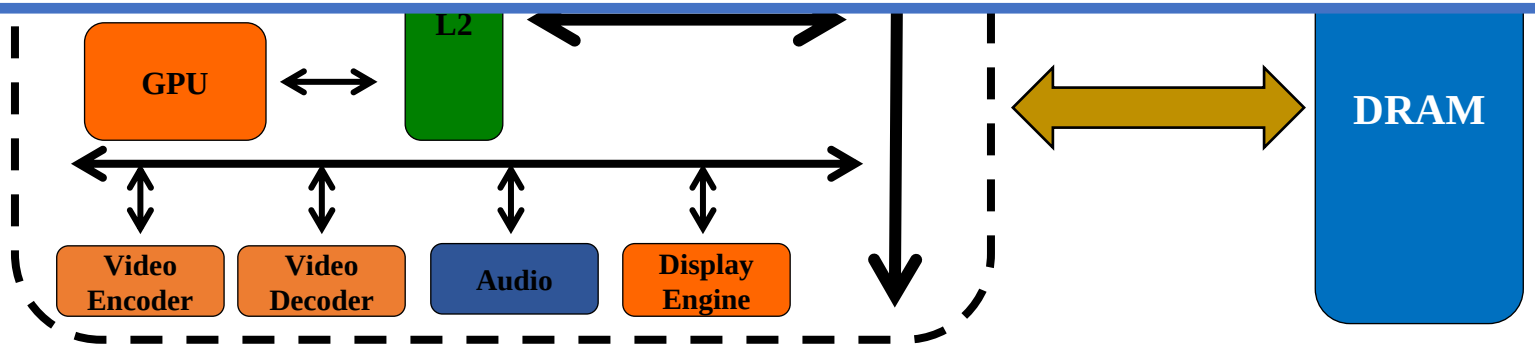


计算机系统中的数据搬运

- 数据移动是系统性能和能耗的主要瓶颈
  - 能耗是加法操作的**115x**
  - 执行浏览器应用时，占系统能耗的**41%**

计算机系统 应该 以数据为中心

存算一体 提出 算随数走

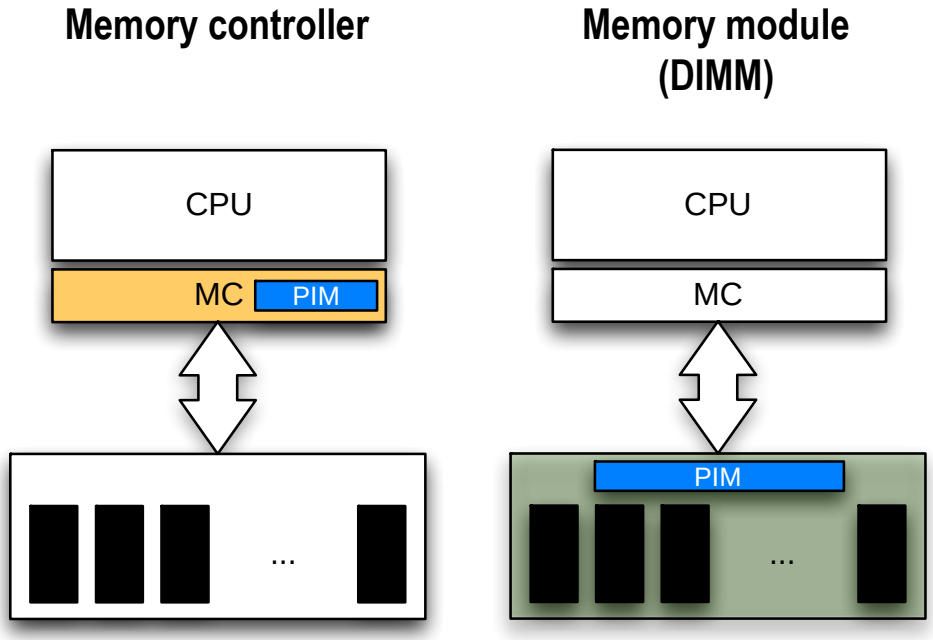


*Reducing data Movement Energy via Online Data Clustering and Encoding (MICRO'16)

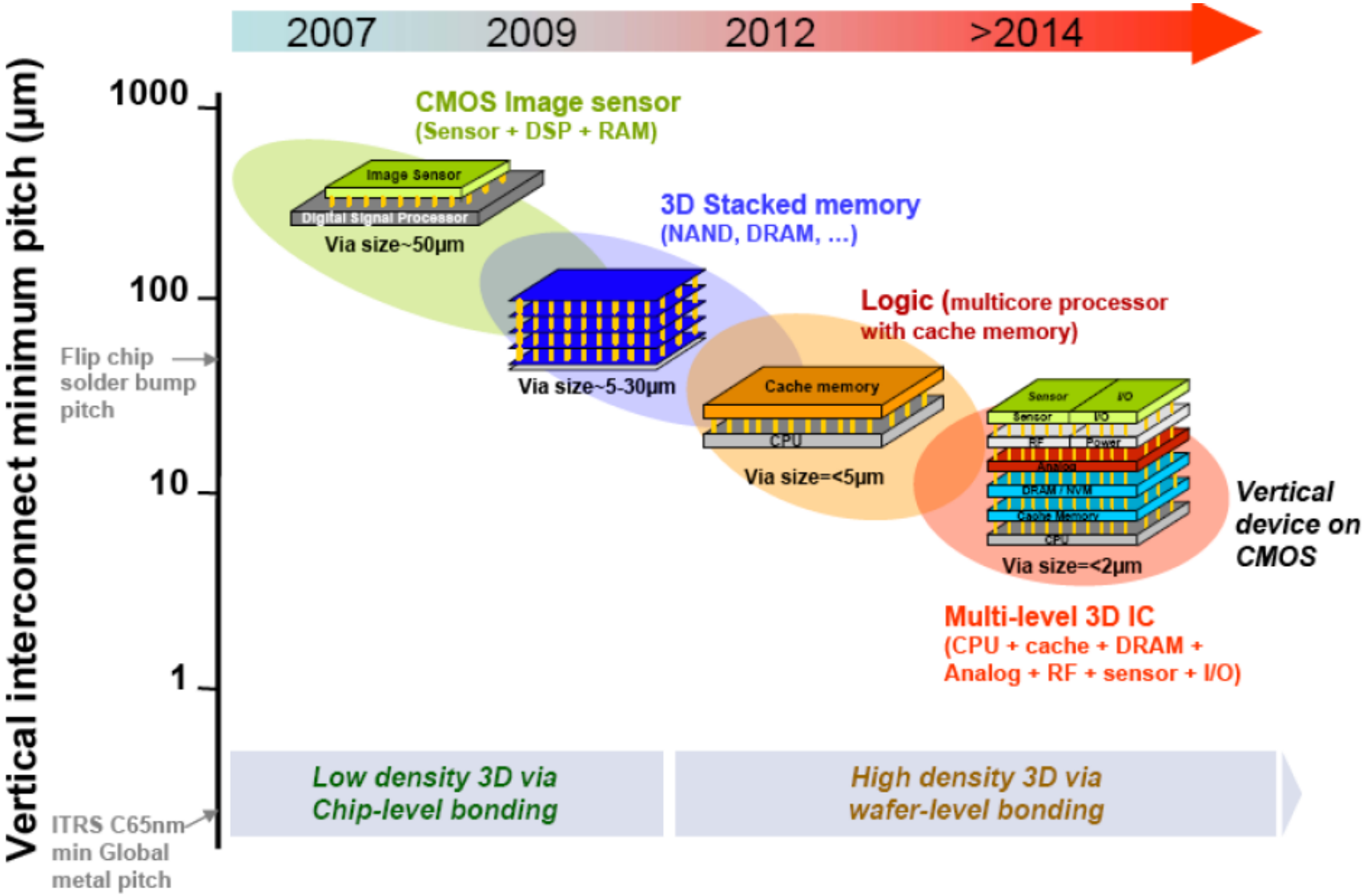
**Quantifying the energy cost of data movement for emerging smart phone workloads on mobile platforms (IISWC'14)

存算一体

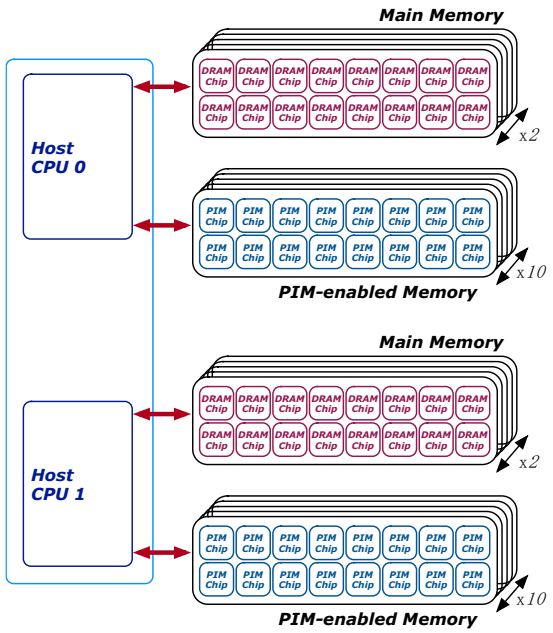
- 近数据处理或者存内计算
  - 数据在哪里，计算就在哪里



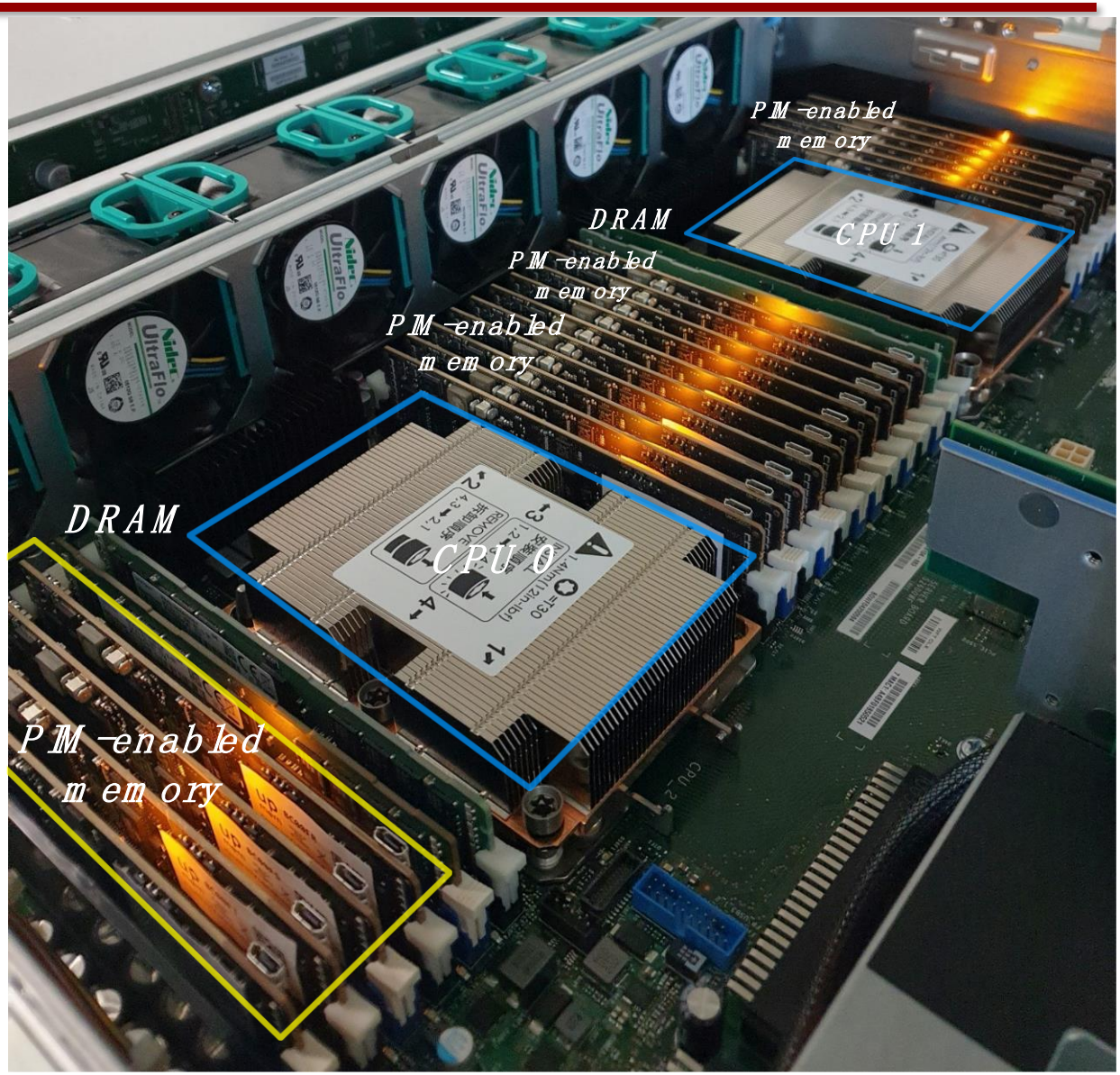
Application Trend of 3D



近存计算加速大模型推理



- 加速大模型推理
- 500M主频DPU
- 对标A6000



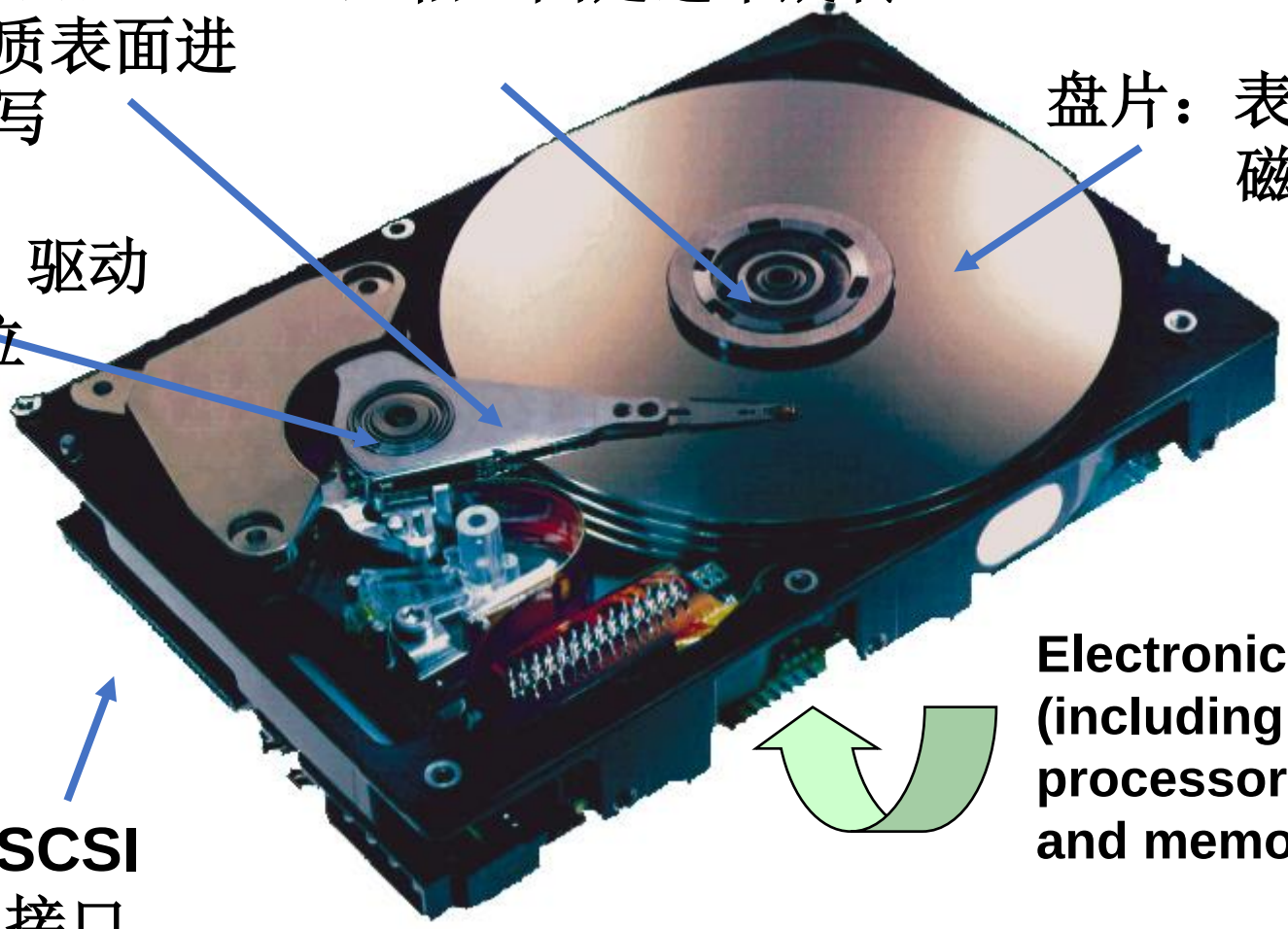
磁盘

磁头臂：操作硬盘磁头在介质表面进行数据读写

主轴：固定速率旋转

盘片：表面覆盖磁性材料

驱动电机：驱动磁盘臂定位

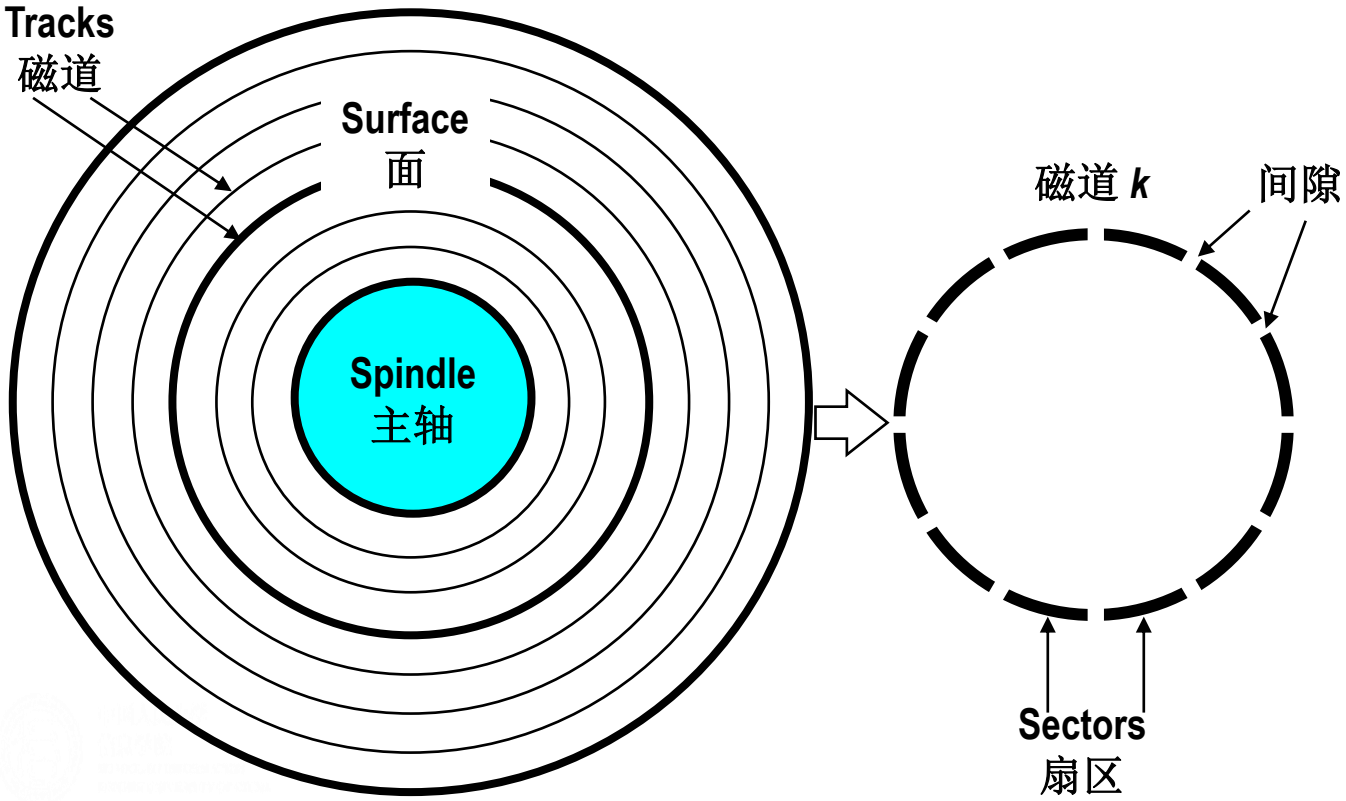


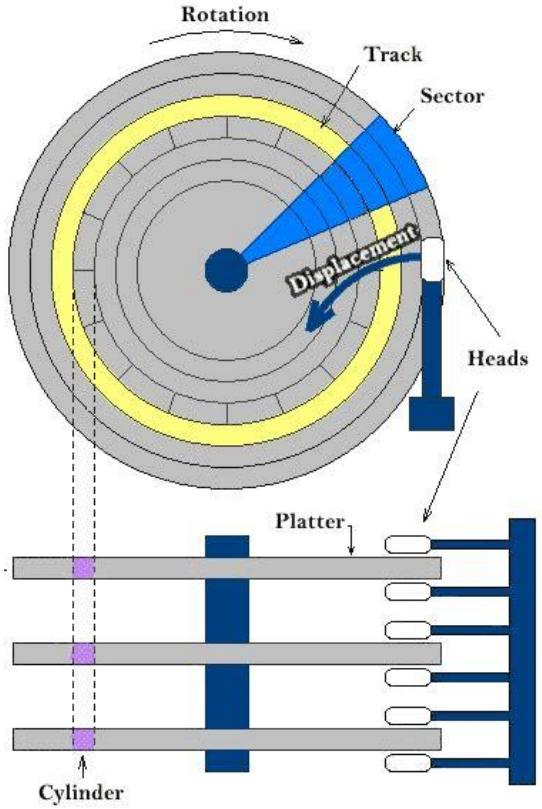
Electronics (including a processor and memory!)

SCSI 接口

磁盘结构

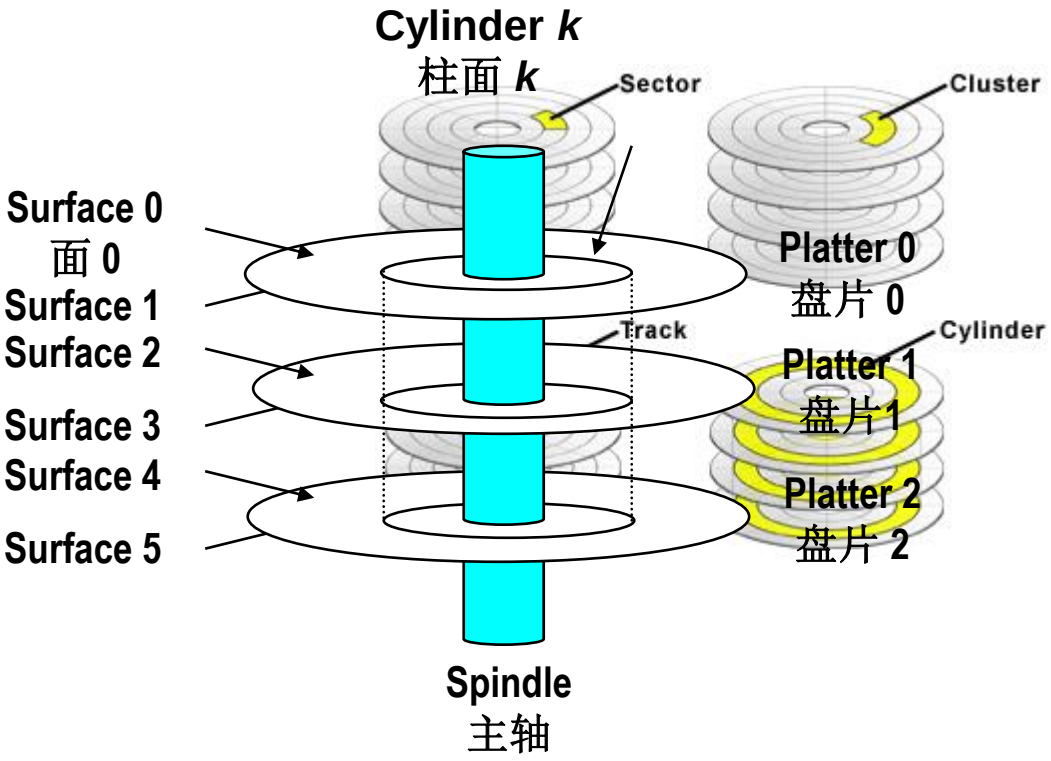
- 磁盘由一组盘片（platers）构成，每个盘片有两个存储面（surface）
- 每个盘面由一系列同心圆磁道（track）组成
- 每个磁道由一系列扇区（sectors）组成，扇区之间由间隙（gaps）分隔





Disk Geometry (Muliple-Platter View)

- 柱面(cylinder)：所有盘片表面上到主轴中心的距离相等的磁道的集合
  - 示例：柱面*k*是3个盘片、6个面上6个磁道*k*的集合



Computing Disk Capacity

$$\text{磁盘容量} = \frac{\text{字节数}}{\text{扇区}} \times \frac{\text{平均扇区数}}{\text{磁道}} \times \frac{\text{磁道数}}{\text{表面}} \times \frac{\text{表面数}}{\text{盘片}} \times \frac{\text{盘片数}}{\text{磁盘}}$$

- **示例：**
  - 512 bytes/sector
  - 300 sectors/track (on average)
  - 20,000 tracks/surface
  - 2 surfaces/platter
  - 5 platters/disk

$$\text{磁盘容量} = \frac{512\text{字节}}{\text{扇区}} \times \frac{300\text{扇区}}{\text{磁道}} \times \frac{20000\text{磁道}}{\text{表面}} \times \frac{2\text{表面}}{\text{盘片}} \times \frac{5\text{盘片}}{\text{磁盘}}$$

= 30,720,000,000

= 30.72 GB

练习题 计算磁盘容量

- 磁盘有4个盘片，2000个柱面，每条磁道平均3000个扇区，每扇区512字节

$$\text{磁盘容量} = \frac{512\text{字节}}{\text{扇区}} \times \frac{3000\text{扇区}}{\text{磁道}} \times \frac{2000\text{磁道}}{\text{表面}} \times \frac{2\text{表面}}{\text{盘片}} \times \frac{4\text{盘片}}{\text{磁盘}}$$

$$= 24{,}000{,}000{,}000$$

$$= 24\ \text{GB}$$

- 磁盘有2个盘片，10000个柱面，每条磁道平均400个扇区，每扇区512字节

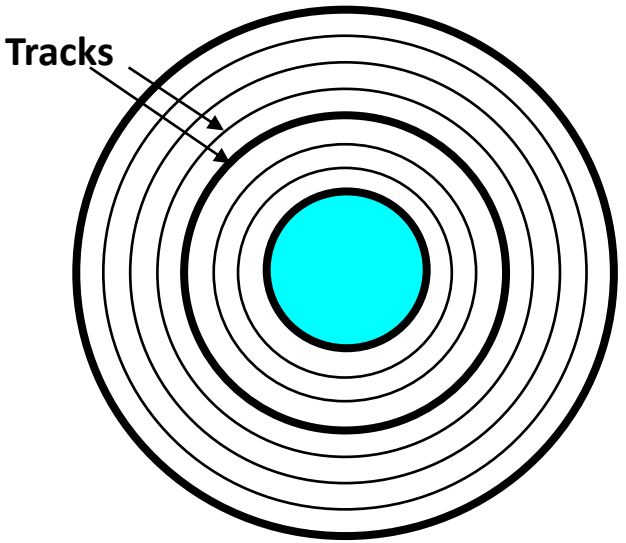
$$\text{磁盘容量} = \frac{512\text{字节}}{\text{扇区}} \times \frac{400\text{扇区}}{\text{磁道}} \times \frac{10000\text{磁道}}{\text{表面}} \times \frac{2\text{表面}}{\text{盘片}} \times \frac{2\text{盘片}}{\text{磁盘}}$$

$$= 8{,}192{,}000{,}000$$

$$= 8.192\ \text{GB}$$

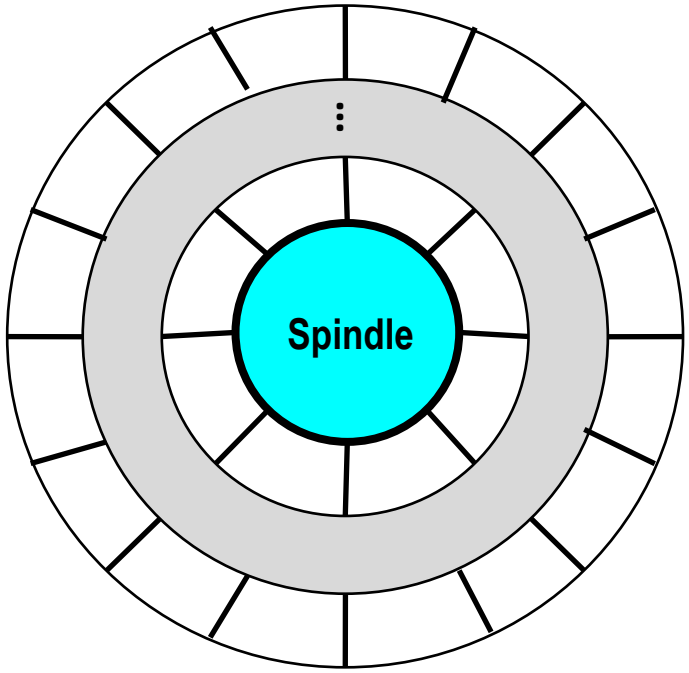
磁盘容量

- Capacity(容量）：能够存储的最多比特数.
  - 制造商通常用gigabytes (GB)作为单位
    1 GB = 10^9 Bytes.
- 决定容量的因素:
  - 记录密度(Recording density (位/英寸))： 磁道一英寸的段中可以放入的位数
  - 磁道密度(Track density (道/英寸))： 从盘片中心出发半径上一英寸的段内可以有的有磁道数
  - 面密度(Areal density (位/平方英寸))：
    记录密度与磁道密度的乘积



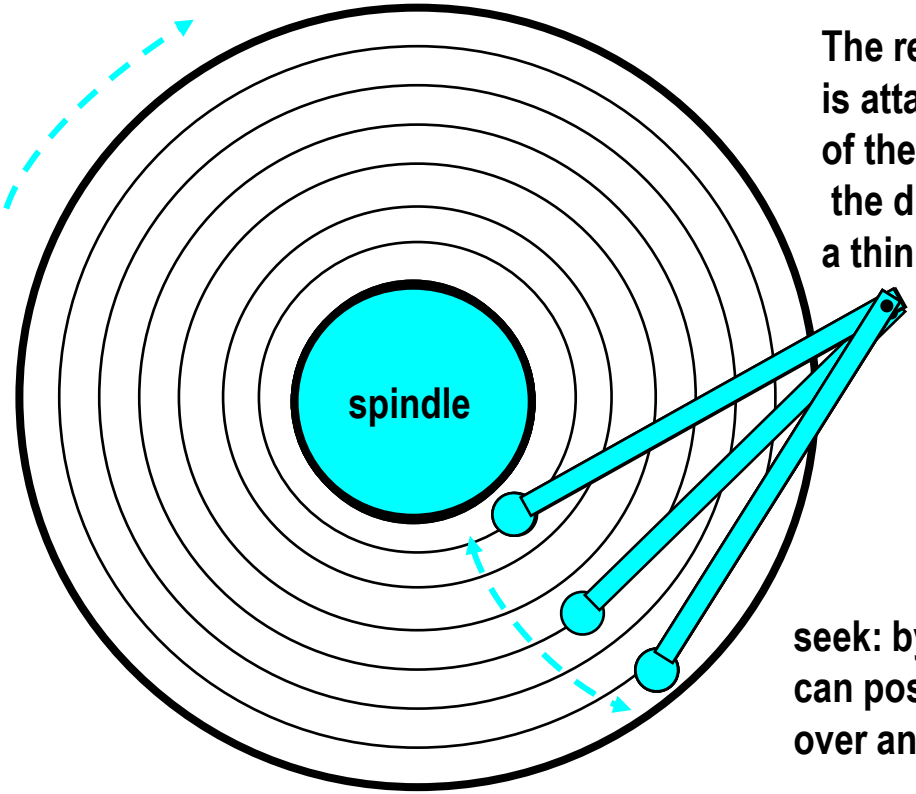
记录区Recording zones

- 初期：每个磁道分为数目相同的扇区，扇区数量由最靠内的磁道能记录的扇区数决定
  - 缺点：越靠外，扇区间隔越大
- 当前：柱面集合分割为不相交的子集合—记录区(recording zone)
- 每个记录区包含一组连续的柱面
- 一个区中每个柱面中的每条磁道都有相同数量的扇区
- 扇区数量由该区中最里面的磁道所能包含的扇区数量决定



Disk Operation (Single-Platter View)

The disk surface spins at a fixed rotational rate

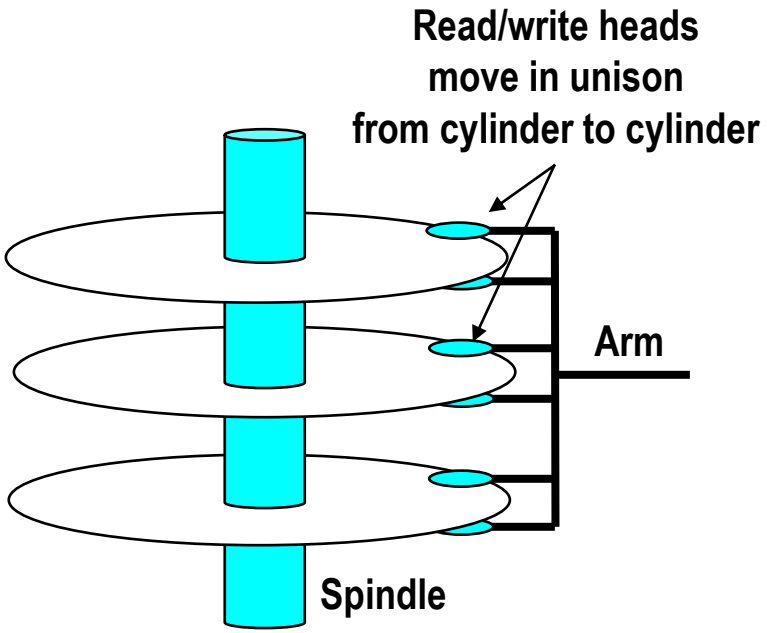


The read/write *head* is attached to the end of the *arm* and flies over the disk surface on a thin cushion of air.

seek: by moving radially, the arm can position the read/write head over any track.

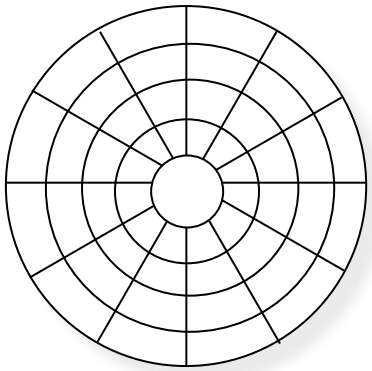
寻道：通过传动臂将磁头定位在指定磁道上

Disk Operation (Multi-Platter View)



读/写：读写头感知(读该位)或修改(写该位)磁道上的位值

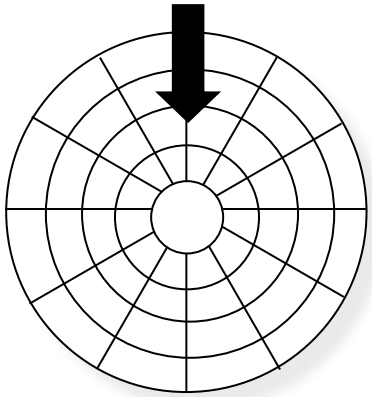
Disk Structure - top view of single platter



Surface organized into tracks

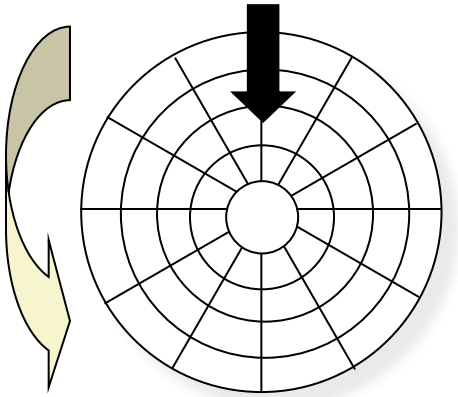
Tracks divided into sectors

Disk Access



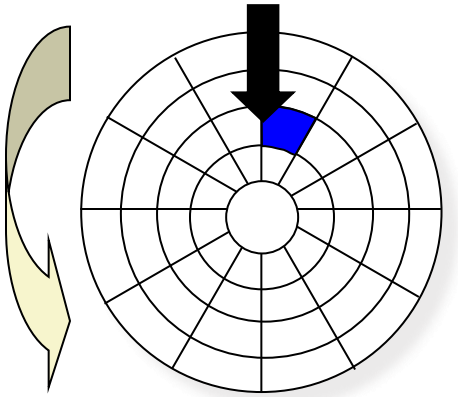
Head in position above a track

Disk Access



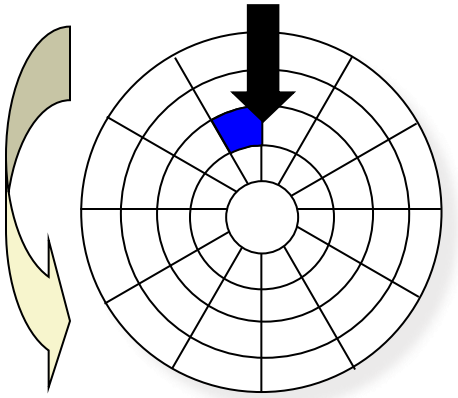
Rotation is counter-clockwise

Disk Access – Read



About to read blue sector

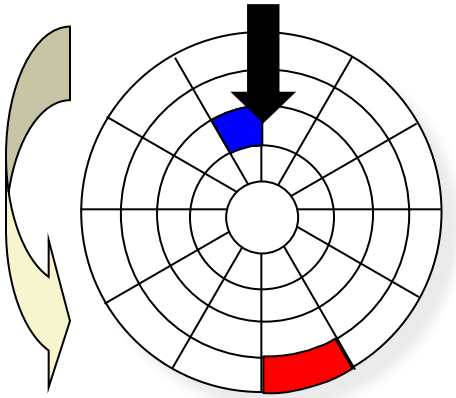
Disk Access – Read



After BLUE read

After reading blue sector

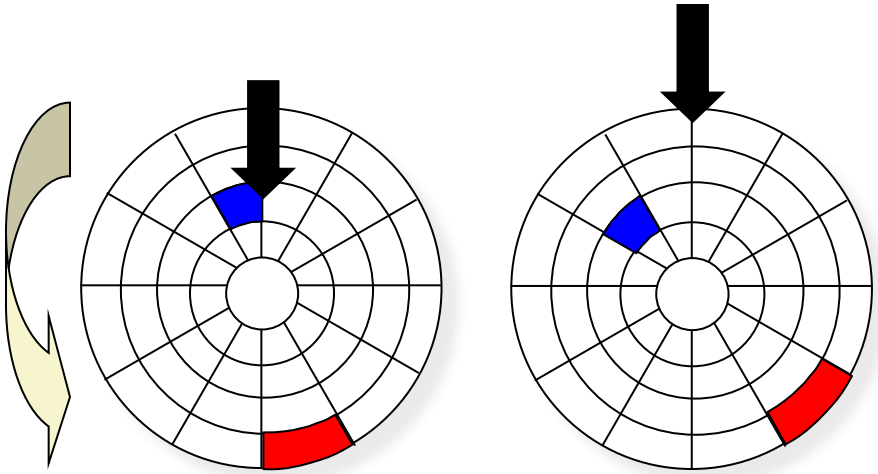
Disk Access – Read



After BLUE read

Red request scheduled next

Disk Access – Seek

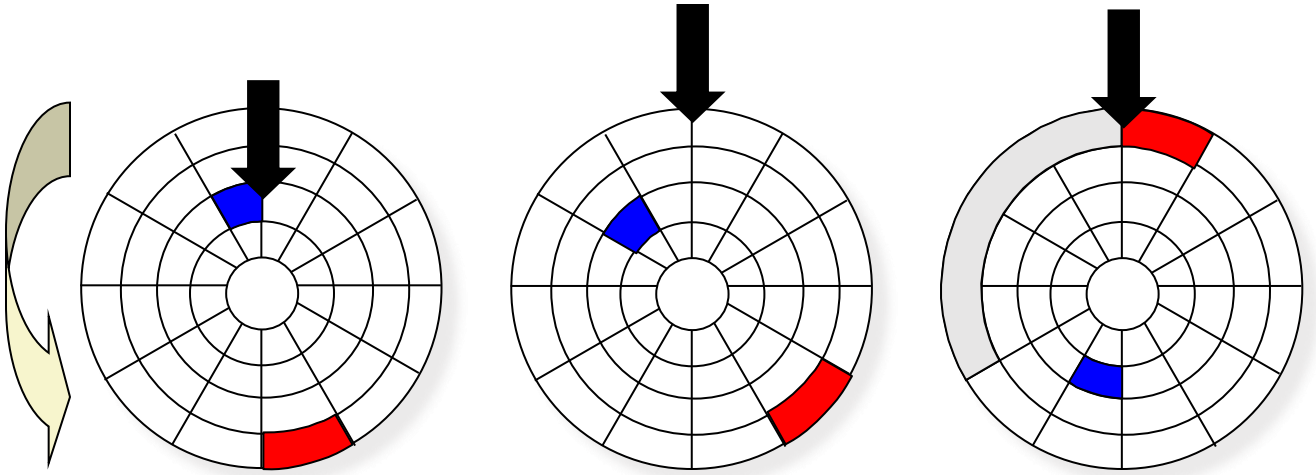


After BLUE read

Seek for RED

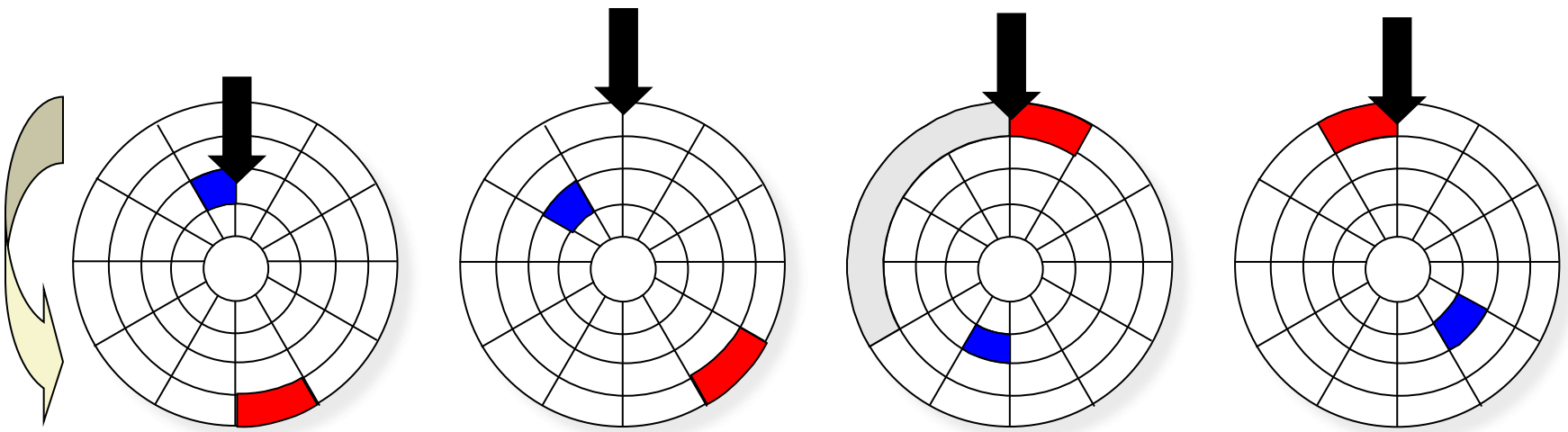
Seek to red's track

Disk Access – Rotational Latency



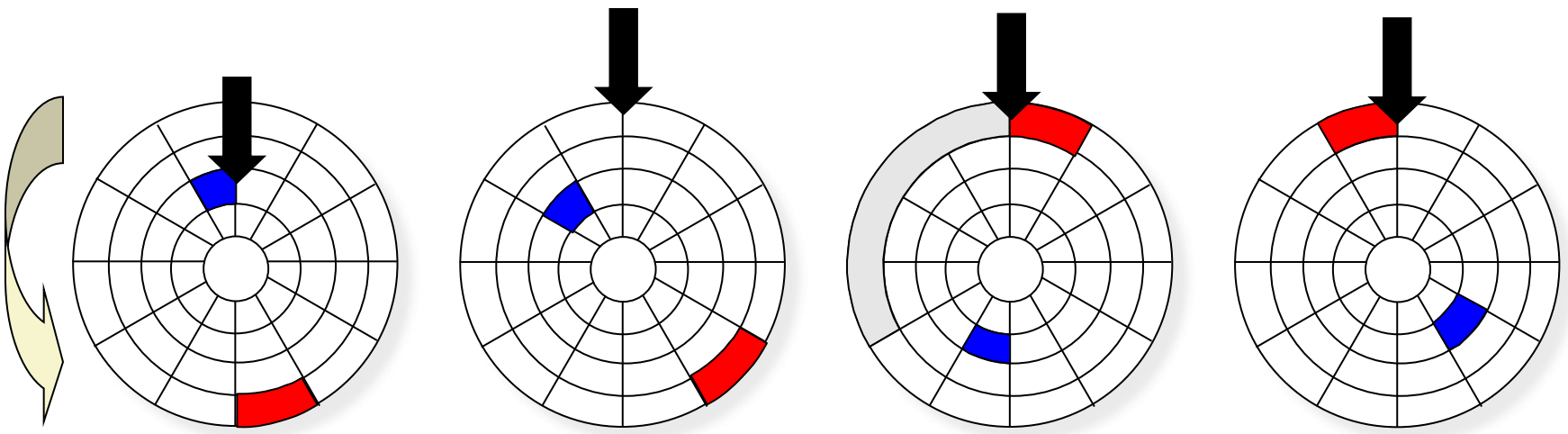
Wait for red sector to rotate around

Disk Access – Read



Complete read of red

Disk Access – Service Time Components



扇区访问时间

- 扇区访问时间由三部分组成：寻道时间、旋转时间和传送时间
  - $T_{access} = T_{avg\ seek} + T_{avg\ rotation} + T_{avg\ transfer}$
- 寻道时间$T_{avg\ seek}$：
  - 移动传动臂将读/写头定位到包含目标扇区磁道上所需的时间
  - 平均寻道时间为3~9ms
- 旋转时间：
  - 驱动器等待目标扇区的第一位旋转到读/写头下的时间
  - 最大旋转延迟为等待一整圈：$T_{MaxRotation} = \frac{1}{RPM} \times \frac{60s}{1min}$
  - 平均旋转延迟为最大旋转延迟一半：$T_{AvgRotation} = \frac{1}{2} \times \frac{1}{RPM} \times \frac{60s}{1min}$
  - 典型磁盘转速为7200RPM或15K RPM
- 传送时间：
  - 当目标扇区的第一个位位于读/写头下时，驱动器读或写该扇区的时间
  - 一个扇区以秒为单位的平均传送时间为：
  - $T_{AvgTransfer} = \frac{1}{RPM} \times \frac{60s}{1min} \times \frac{1}{\text{平均扇区数/磁道}}$

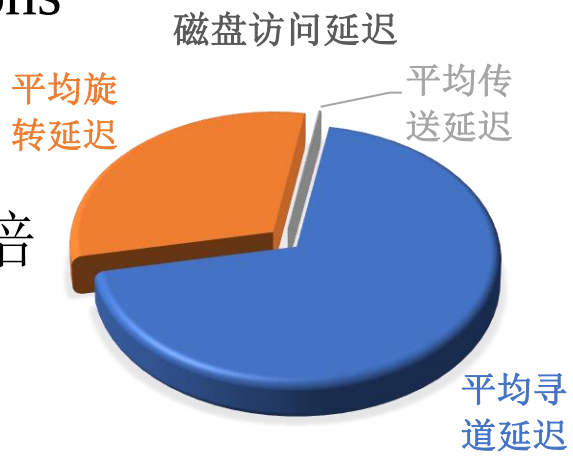
time for one rotation (in minutes)　　fraction of a rotation to be read

练习题

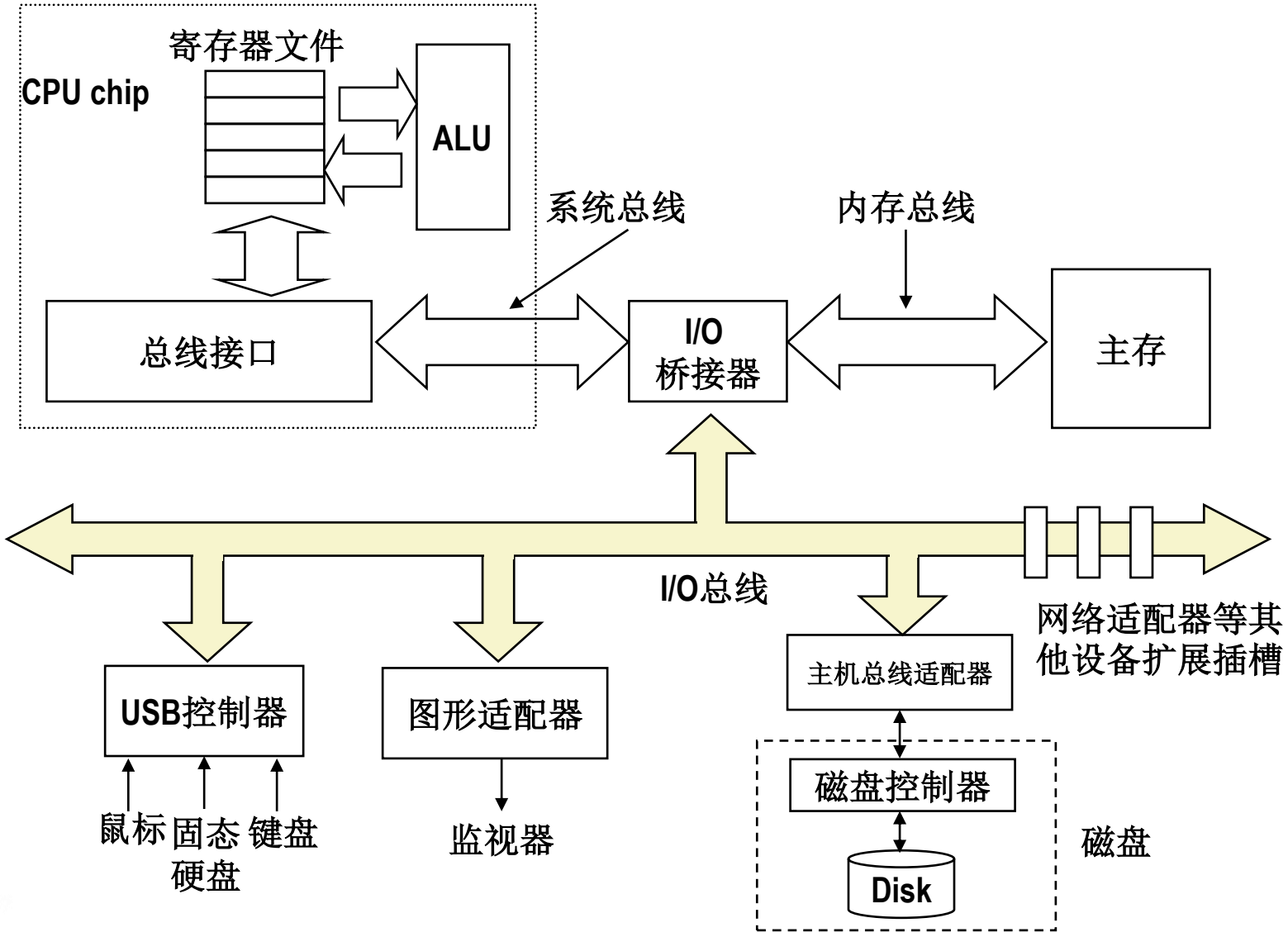
- 估算磁盘扇区访问时间(ms)
  - 旋转速率：15000RPM
  - 平均寻道时间：8ms
  - 每条磁道的平均扇区数：500
- 访问时间估算：
  - 平均旋转延迟：$T_{AvgRotation} = \frac{1}{2} \times \frac{1}{15000} \times \frac{60s}{1min} \times \frac{1000\text{ms}}{s} \approx 2ms$
  - 平均传送时间：$T_{AvgTransfer} = \frac{1}{15000} \times \frac{1}{500/1} \times \frac{60s}{1min} \times \frac{1000\text{ms}}{s} \approx 0.008ms$
  - 总访问延迟：$T_{Access} = T_{AvgSeek} + T_{AvgRotation} + T_{AvgTransfer} = 8ms + 2ms + 0.008ms = 10.008ms$

扇区访问时间示例

- 访问延迟
  - 磁盘大约10ms的访问延迟，主要是寻道延迟和旋转延迟
  - 访问扇区第一个位用时很长，访问其他字节时间极少
  - SRAM中一个cache line(64字节)访问时间为4ns
  - DRAM访问时间为60ns；
- 访问512字节大小
  - SRAM读512个字节大小的块时间为256ns
  - DRAM为4000ns
  - 磁盘访问时间为10ms
  - 是SRAM的40000倍，是DRAM的2500倍

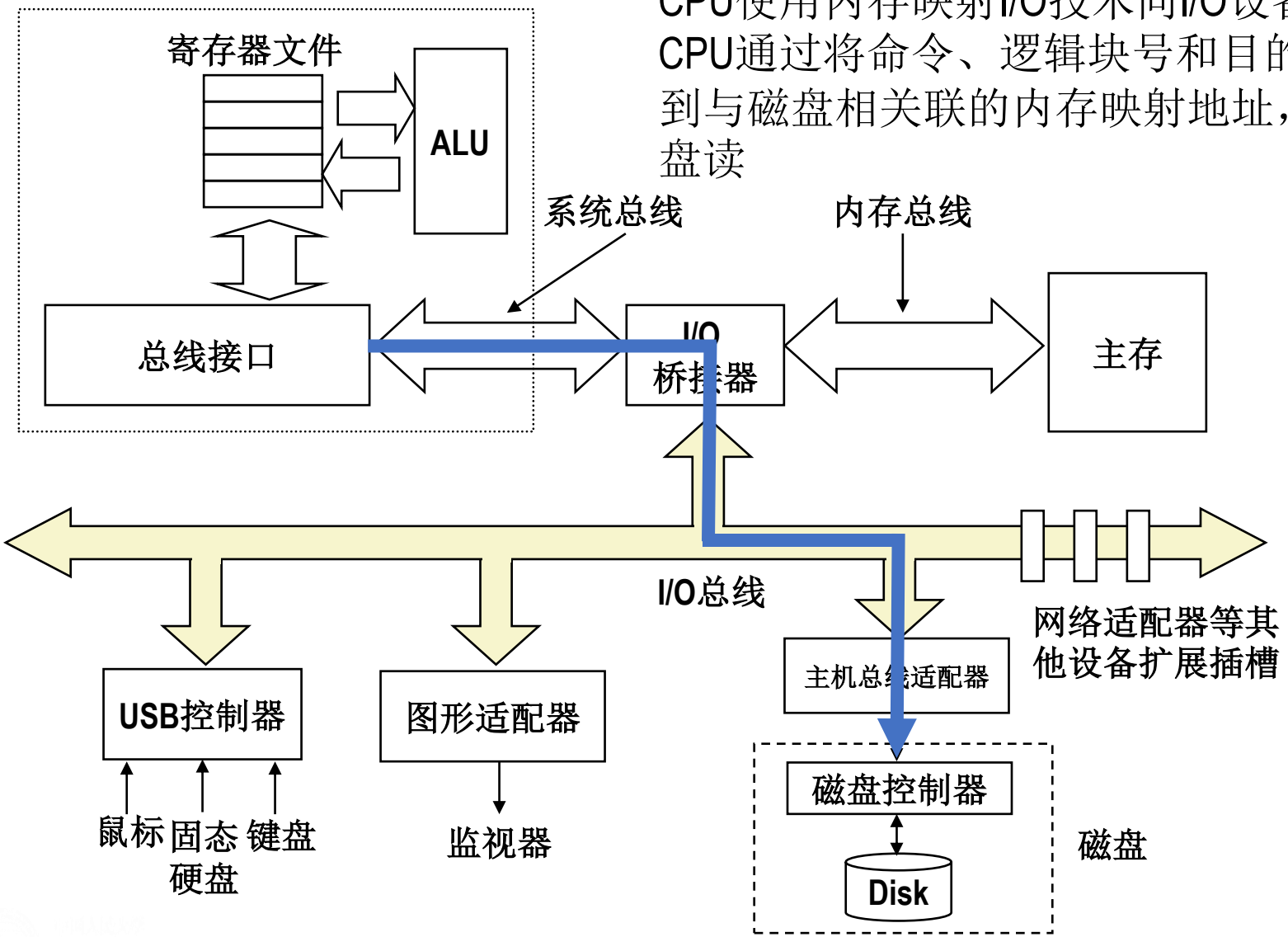


I/O 总线


寄存器文件
CPU chip
ALU
系统总线
内存总线
总线接口
I/O
桥接器
主存
I/O总线
网络适配器等其
他设备扩展插槽
USB控制器
图形适配器
主机总线适配器
磁盘控制器
Disk
磁盘
鼠标
固态
硬盘
键盘
监视器

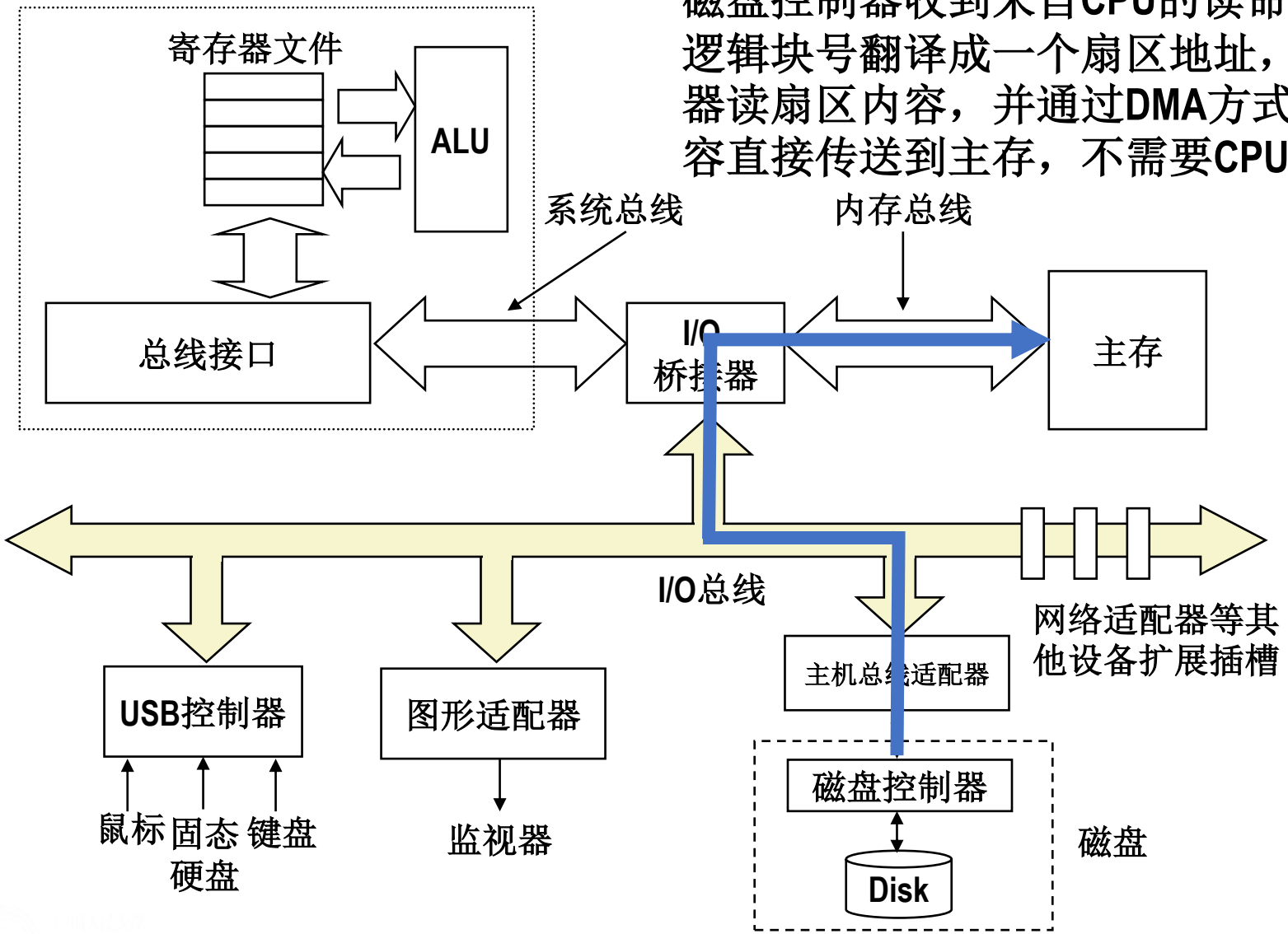
访问磁盘——从磁盘读数据(1)

CPU使用内存映射I/O技术向I/O设备发射命令。CPU通过将命令、逻辑块号和目的内存地址写到与磁盘相关联的内存映射地址，发起一个磁盘读



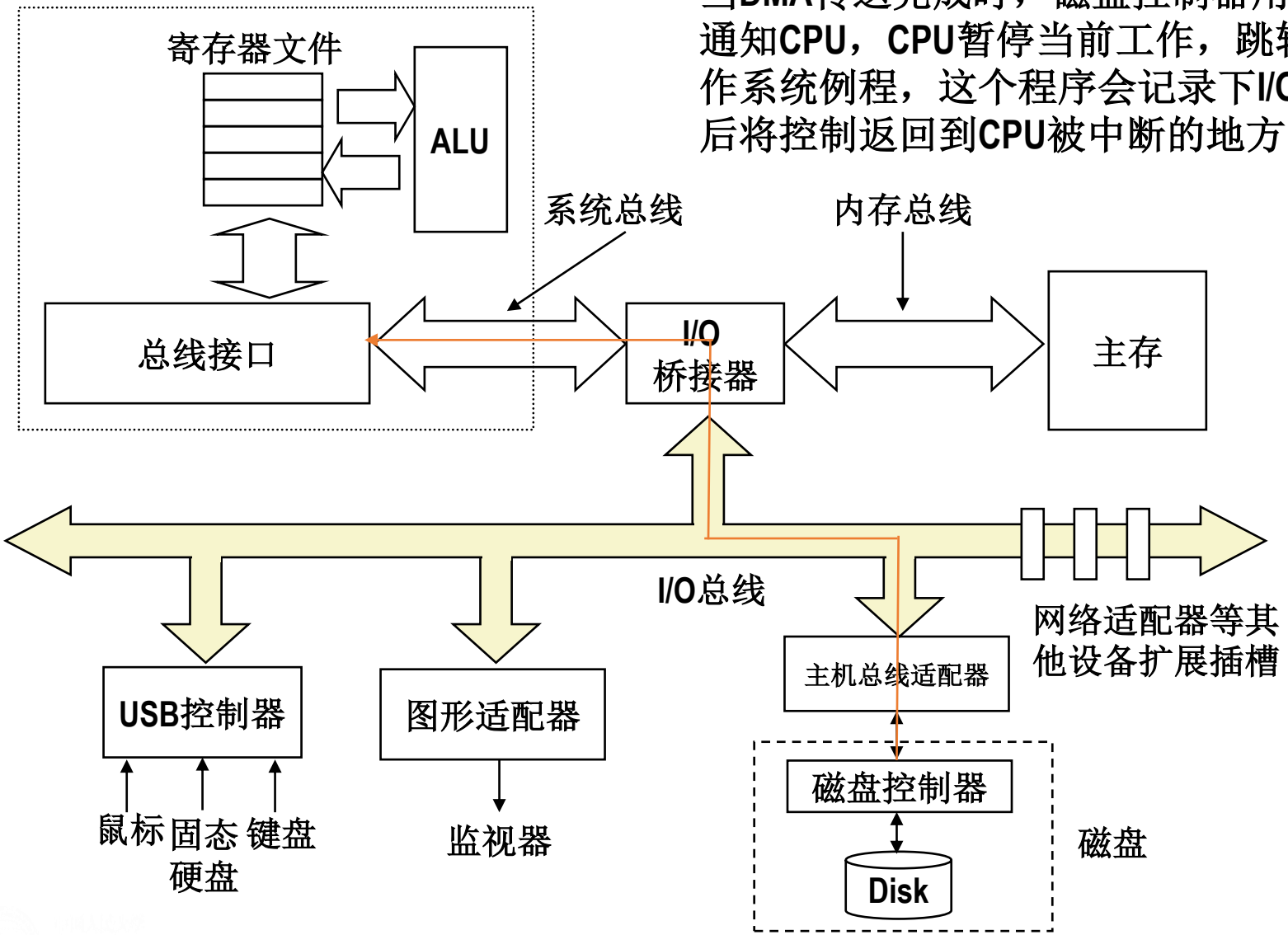
访问磁盘——从磁盘读数据(2)

磁盘控制器收到来自CPU的读命令后，将逻辑块号翻译成一个扇区地址，磁盘控制器读扇区内容，并通过DMA方式将扇区内容直接传送到主存，不需要CPU的干涉

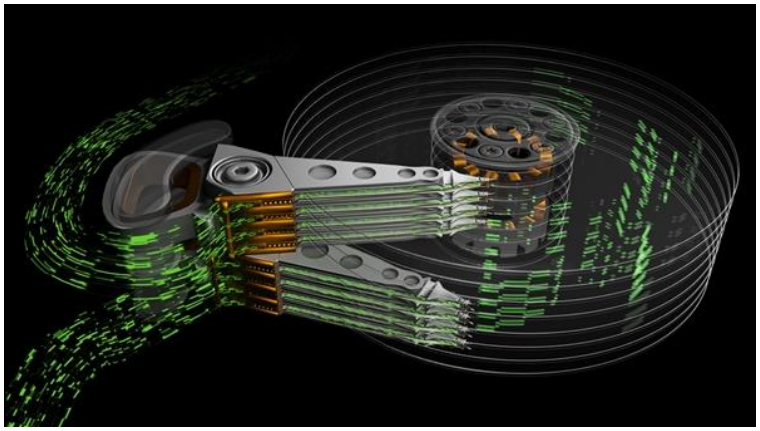


访问磁盘——从磁盘读数据(3)

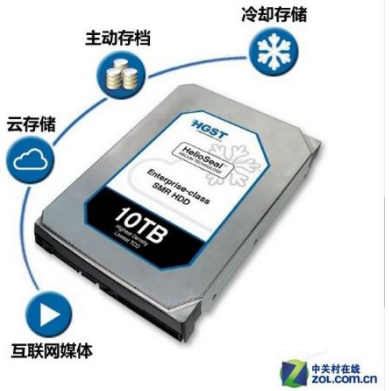
当DMA传送完成时，磁盘控制器用中断的方式通知CPU，CPU暂停当前工作，跳转到一个操作系统例程，这个程序会记录下I/O已完成，然后将控制返回到CPU被中断的地方



硬盘新技术：速度&容量

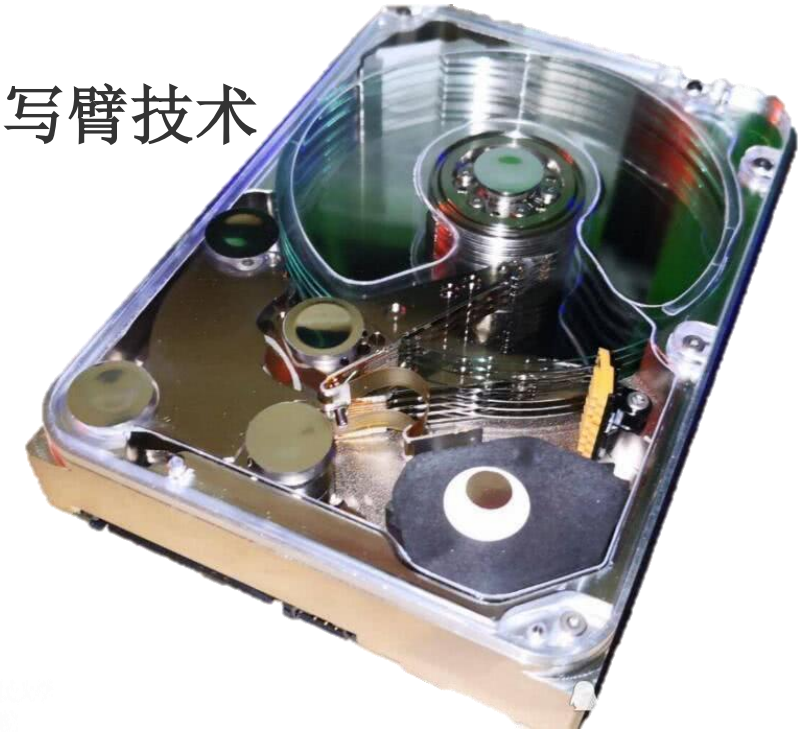


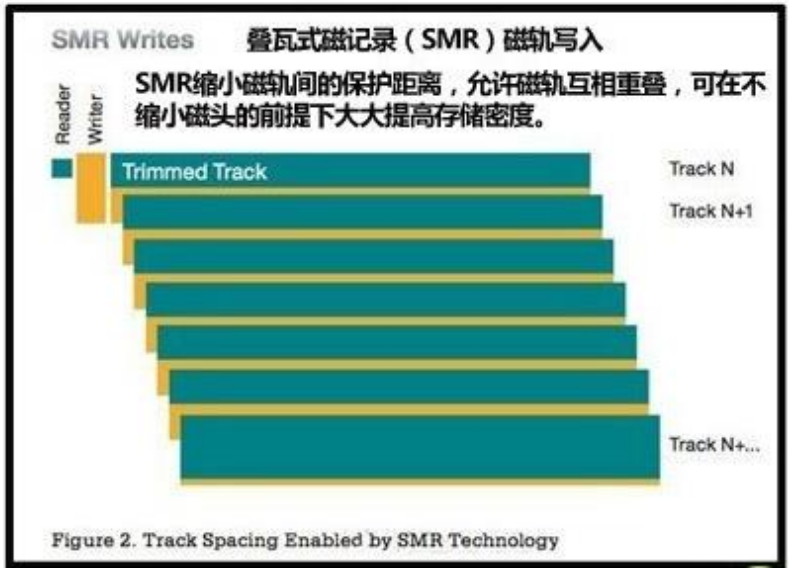
叠瓦式磁记录（SMR）技术



叠瓦式与传统磁轨写入对比

多读写臂技术



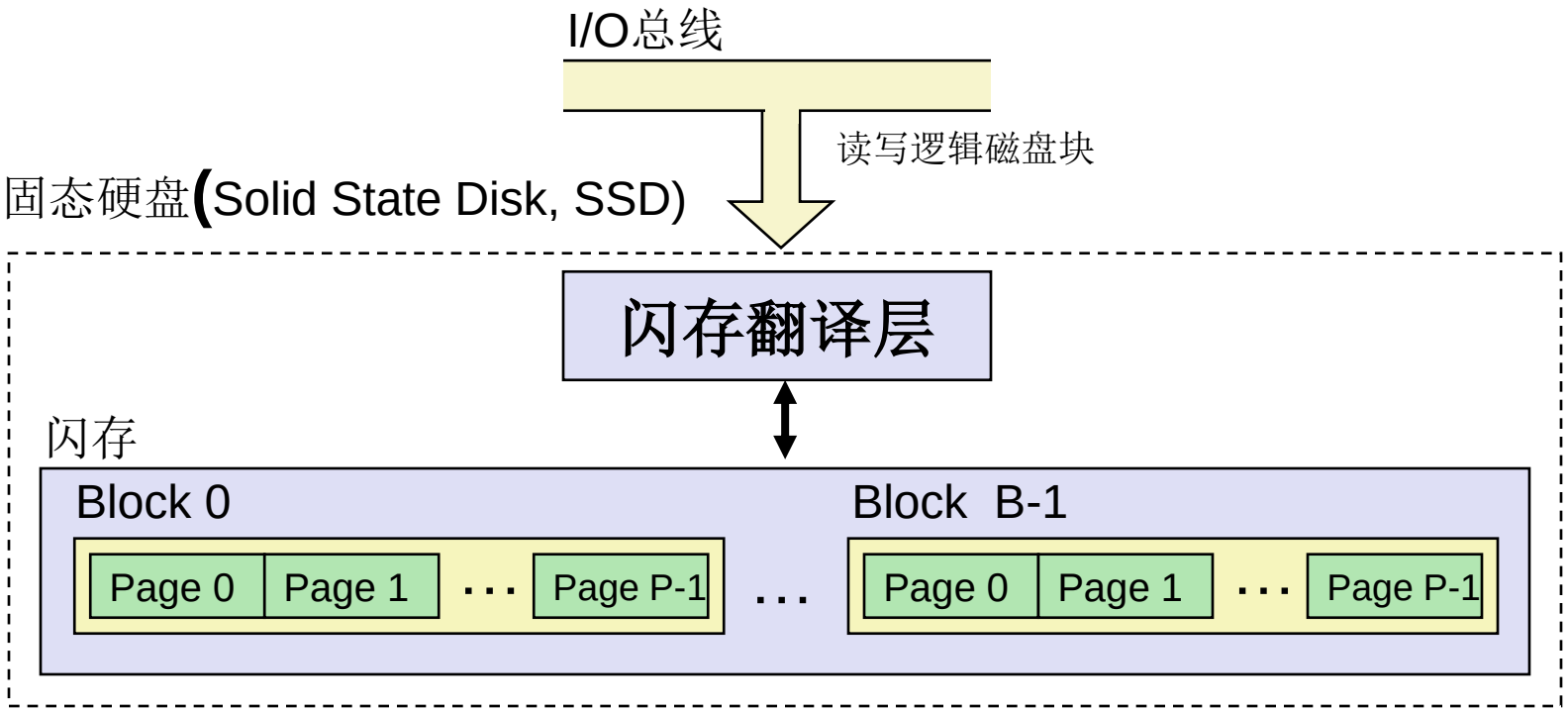


固态硬盘

Intel Optane系列SSD

固态硬盘是一种基于闪存的存储技术，SSD封装为M.2、PCIe、SATA、USB等接口插到I/O总线上

固态硬盘逻辑结构



- 页大小：512KB~4KB, 块: 32~128页
- 数据以页为单位读写
- 页所属的块被整个擦除后才能写该页
- 块磨损：经过大约100000次重复写后，块磨损坏，不可用

固态硬盘的特性

- 基准测试程序Samsung 940 EVO Plus

  https://ssd.userbenchmark.com/SpeedTest/711305/Samsung-SSD-970-EVO-Plus-250GB

顺序读的吞吐量	2,126 MB/s	顺序写吞吐量	1,880 MB/s
随机读的吞吐量	140 MB/s	随机写吞吐量	59 MB/s

  顺序访问比随机访问快
  - 存储层次结构总的通用规律
- 随机写入速度稍慢
  - 擦除块需要很长时间（~1ms）
  - 修改块页面需要将所有其他页面复制到新块
  - Flash转换层允许在进行块写入之前累积一系列小写入

固态硬盘（SSD）与旋转磁盘的对比

- 优点
  - 没有机械移动→ 更快，更节能，更耐用
- 缺点
  - 会磨损
    - 通过闪存转换层中的“磨损均衡逻辑”来减轻
    - 例如，三星940 EVO Plus保证在它们磨损之前可以进行600次写入/字节的写入
    - 控制器迁移数据以最小化磨损水平
  - 在2019年，每字节的成本大约是原来的4倍
    - 相对成本将持续下降
- 应用
  - MP3播放器、智能手机、笔记本电脑
  - 在台式机和服务器中越来越普遍

练习题 估算SSD寿命

- 假定SSD能够经得起128PB的写，根据工作负载估算SSD寿命
  - A. 顺序写最糟情况：以470MB/s速度持续写SSD
  - B. 随机写最糟情况：以303MB/s速度持续写SSD
  - C. 平均情况：以20GB/天速度写SSD
- 两种情况的估算寿命：
  - A. 顺序写最糟情况寿命：$\frac{128\times10^9}{470\times60\times60\times24\times365}\approx 8.64$年
  - B. 随机写最糟情况寿命：$\frac{128\times10^9}{303\times60\times60\times24\times365}\approx 13.4$年
  - C. 平均情况寿命：$\frac{128\times10^9}{20000\times365}\approx 17534.25$年
    - SSD的实际使用寿命通常都是大大超过官方标称值的

练习题6.6 磁盘价格估算

- 基于2005-2015年数据，估算哪一年可以以$500的价格买到1PB的磁盘？
  - 1PB磁盘$500对应的单位存储价格为0.0005$/GB
  - 2005至2015年间价格变化相对稳定，作为基准参考价格
  - 取2010、2015两阶段平均价格递减倍数13.33
  - 估算后几个5年间磁盘单位容量价格，其中2025年约为0.00017 $/GB，即在2025年之前可能达到$500的价格买1PB磁盘的需求

							价格	容量(GB)	$/GB	
							500	1000000	0.0005	
Year	1985	1990	1995	2000	2005	2010	2015	2020	2025	2030
$/GB	100000	8000	300	10	5	0.3	0.03	0.00225	0.00017	0.00001
递减倍数		12.50	26.67	30.00	2.00	16.67	10.00			
						平均倍数	13.33			

总结

- CPU、内存和大容量存储之间的速度差距继续扩大。
- 编写良好的程序具有一种称为局部性的特性。
- 基于缓存的内存层次结构通过利用局部性来缩小差距。