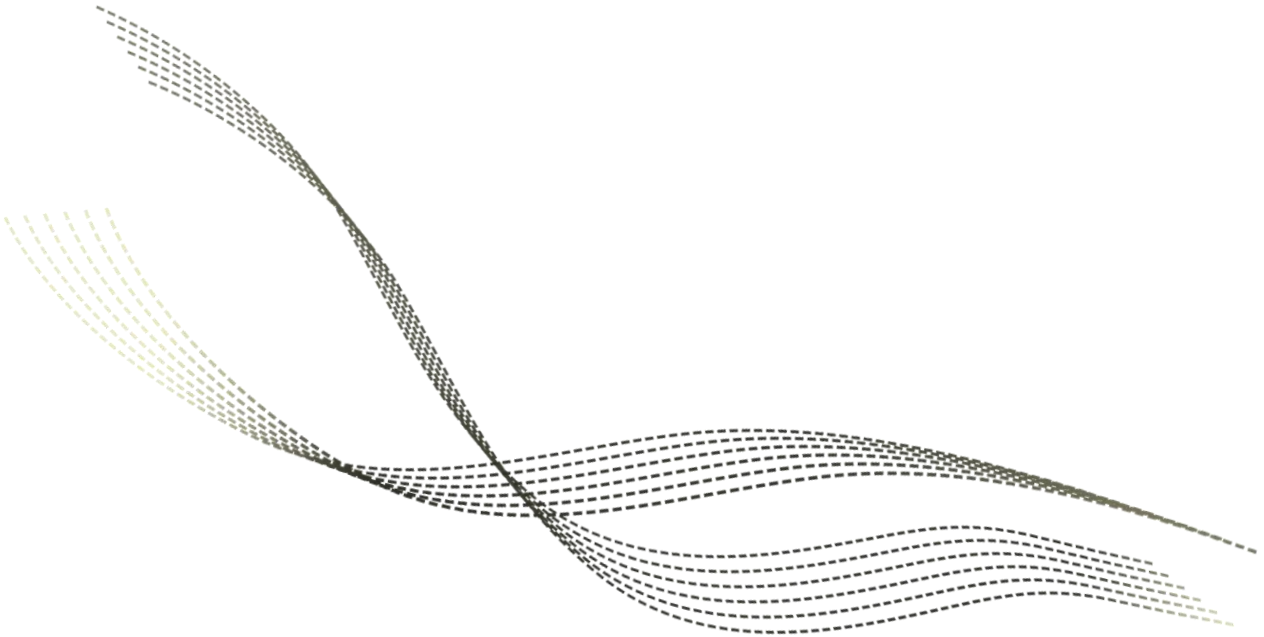
高级计算机体系结构

Advanced Computer Architecture



数据存储

沈明华

目录

CONTENTS

PART 01 磁盘与接口

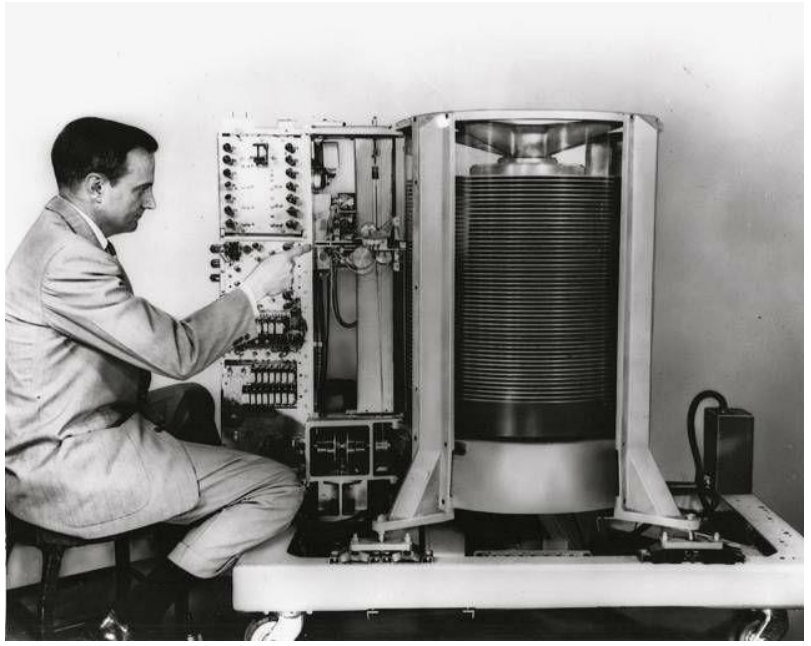
■ 数据存储的问题

- 计算速度增长与数据读取速度增长不匹配
  - 当前CPU处理速度
    - intel Core i5-14600KF@5.3Ghz，95GFLOPS
    - 支持最大带宽89.6GB/s
  - 同代内存数据读取速度
    - DDR5 5600Mhz，单通道，64bit
    - 5600 * 2 * 64 / 8 = 87.5GB/s
  - 与此同时，同代磁盘数据读取速度
    - PCIE4.0总线速度仅7GB/s
    - IO端性能增长每年不到10%

■ 磁盘

◘ IBM 305 RAMAC

- 世界上第一块商用磁盘(1956)
  - ➢ 包含50块24英寸的铝制盘片，每面包含100个磁道，每条磁道有1000个扇区，每个扇区仅能容纳7bit数据
  - ➢ 容量不到5MB，1200转/分钟

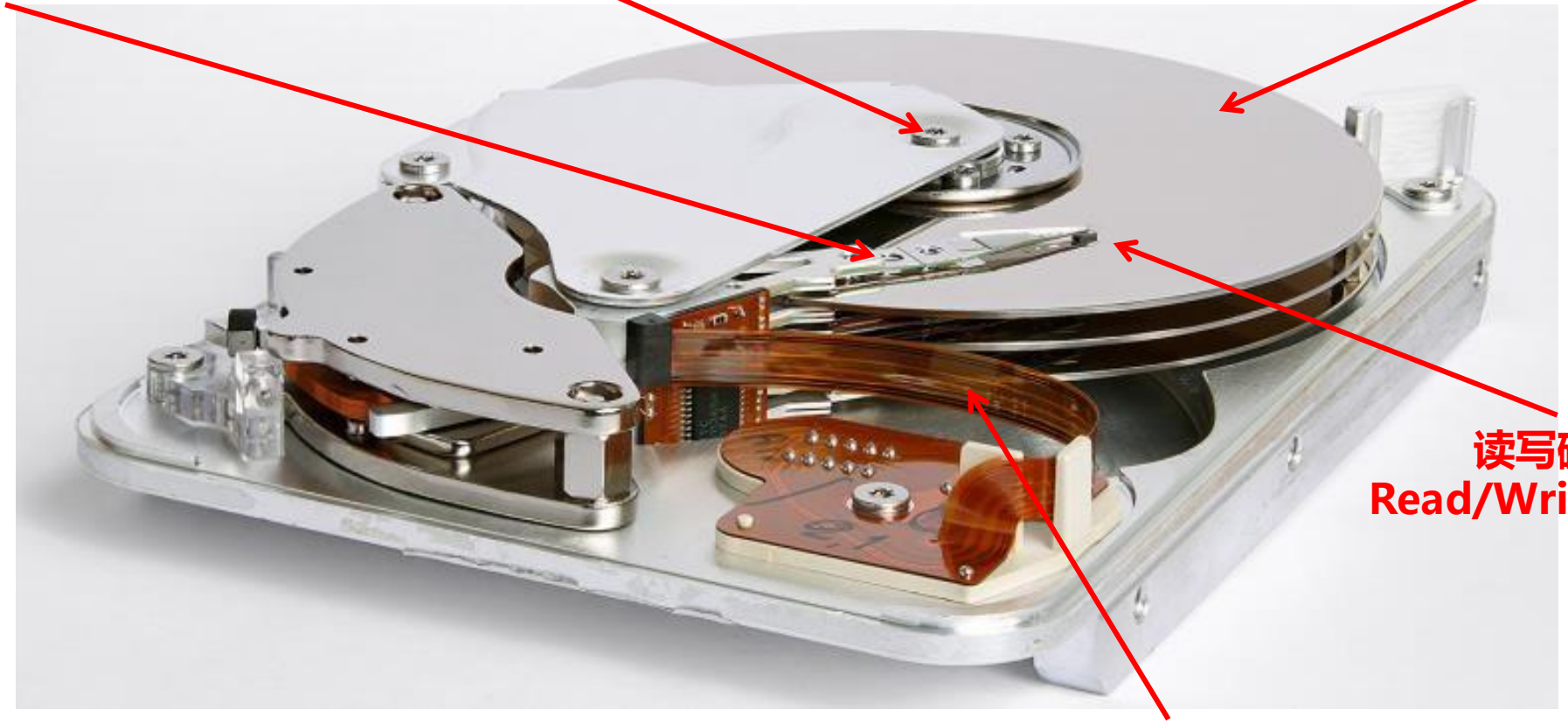


■ 磁盘

□ 磁盘发展历史

- 1956年：第一块商用磁盘(IBM)
- 1980年：第一块容量超过1GB磁盘(IBM)
- 1986年：SCSI接口标准诞生
- 1988年：第一块2.5英寸磁盘(PrairieTek)
- 2002年：突破137GB寻址上限
  - 早期磁盘接口寻址只有28bit，加上每个扇区512字节，寻址范围37bit，大约是137GB
- 2003年：SATA接口问世
- 2007年：第一块超过1TB容量磁盘(Hitachi)
- 2009年：第一块超过2TB容量磁盘(西数)

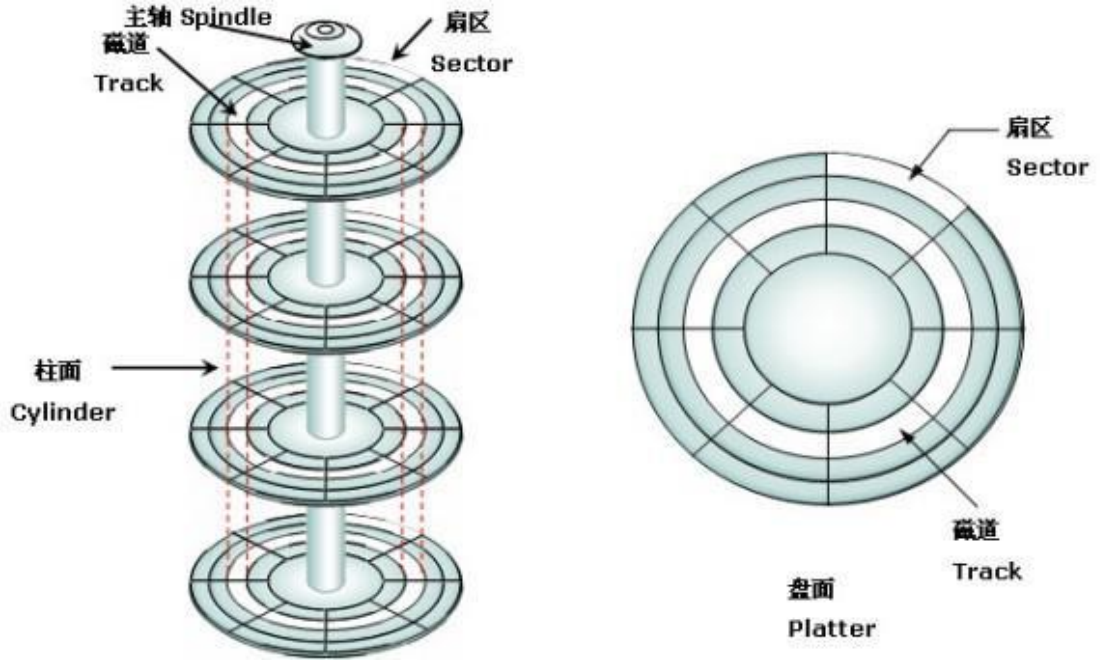
■ 磁盘组成部分


磁臂
Actuator Arm
电机
Motor
磁盘
Disk platter
读写磁头
Read/Write Head
柔性导线
Flex cable

■ 磁盘

◘ 磁盘寻址基本概念

- 盘面(platter)：非磁性材料盘片的表面
- 磁道(track)：盘面的一个环形切片
- 扇区(sector)：一条磁道的一部分
- 柱面(cylinder)：一组垂直方向上重叠的磁道



■ 磁盘

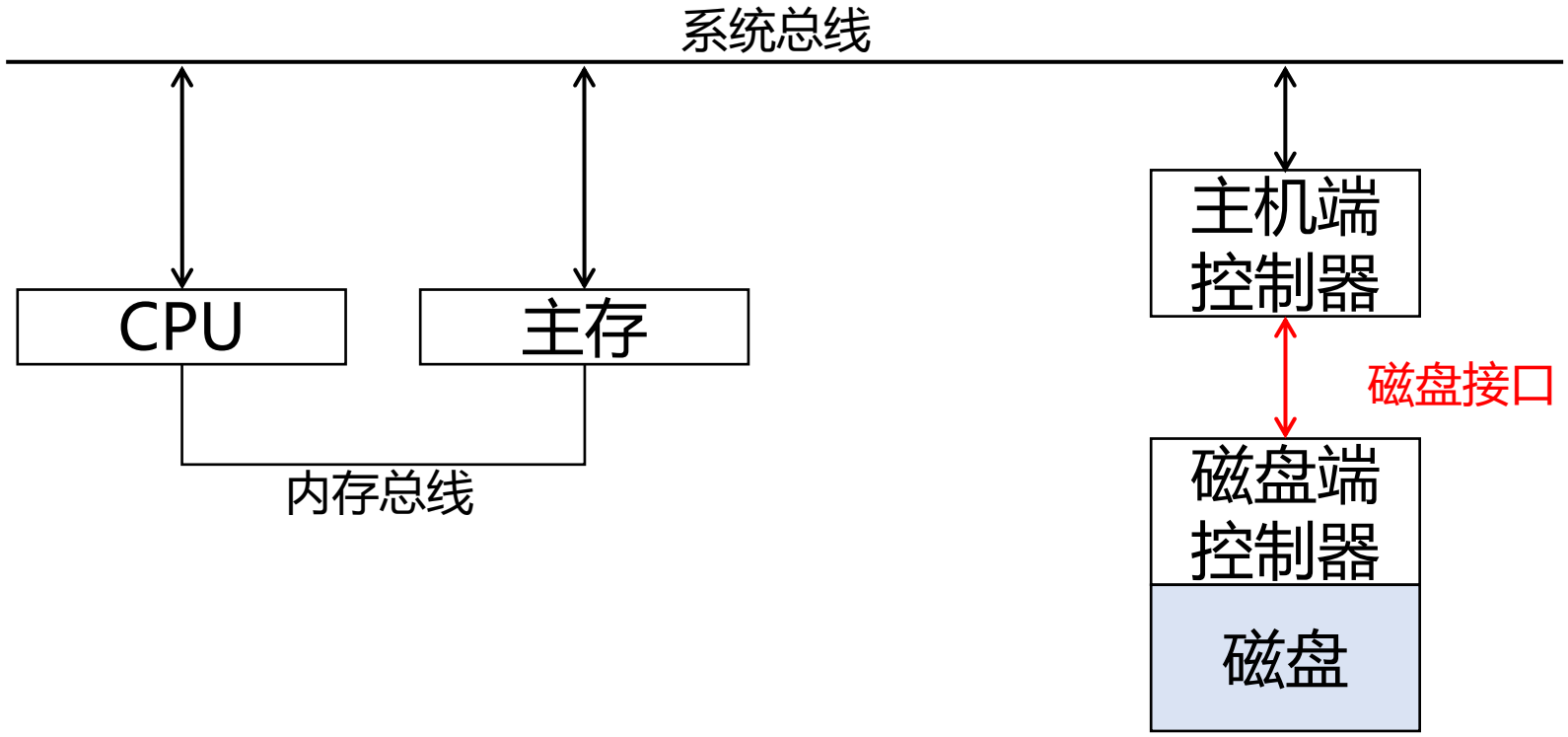
□ 访问延迟

- 可以分为三个部分
  - ➢ 寻道时间(seek time)：磁头定位到正确磁道用时
  - ➢ 旋转延迟(rotational Latency)：磁道上正确扇区到达磁头位置所需时间
  - ➢ 数据传输时间(data transfer time)：读取数据所需时间
  - ➢ 其他因素
    - 例如控制延迟、排队延迟等，和磁盘调度算法有关的延迟

■ 磁盘接口

□ 什么是磁盘接口

- 磁盘接口是连接主机与磁盘的“桥梁”
  - ➢ 传输IO请求的通道
  - ➢ 允许数据通过接口读取和写入磁盘



■ 磁盘接口

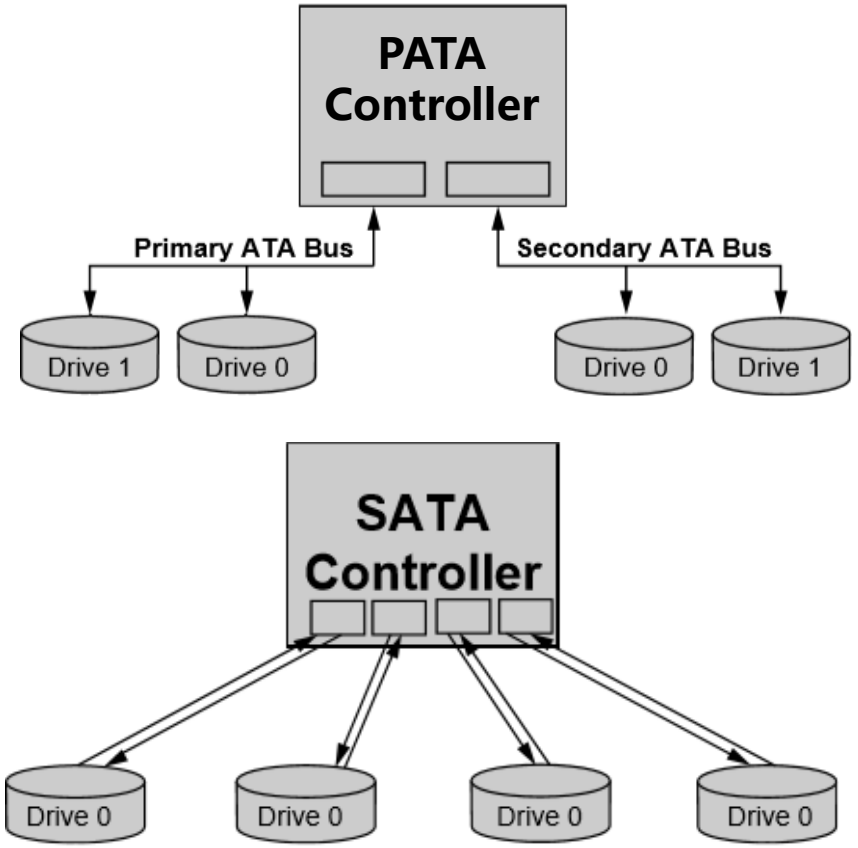
□ 理想磁盘接口应具备的特质

- 简单的控制协议
  - ➢ 更少的握手次数意味着更少的通信开销
- 高度的自动化
  - ➢ 意味着可以减少CPU的介入，降低计算开销
- 高数据传输率　　如果这一条不满足会发生什么？
  - ➢ 至少要比一般媒体所需的数据率高
- 重叠IO指令
  - ➢ 提高磁盘各个部件利用率，对于并发访问的磁盘十分重要
- 合适的IO指令调度策略
  - ➢ 合理安排IO执行顺序，有助于提高磁盘的吞吐量

■ 磁盘接口——ATA

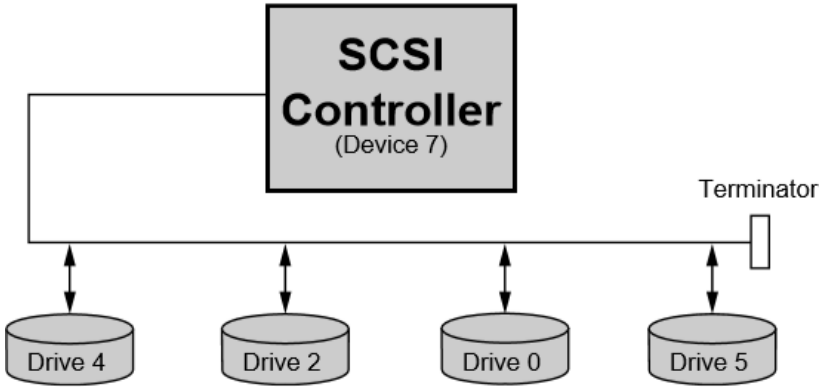
□ 高级技术附件(**ATA**, Advanced Technology Attachment)

- 分为并行高级技术附件(PATA)和串行高级技术附件(SATA)两种
- PATA
  - ➢ 并行接口，通常用于较早期的电脑
  - ➢ 连接硬盘主板的数据线缆不可超过18英寸(大约46厘米)
  - ➢ 速率在133MB/s
  - ➢ 支持多种不同数据传输格式
- SATA
  - ➢ 串行接口，主要应用于最新家用电脑
  - ➢ 线缆长度可超过1米，速率在600MB/s
  - ➢ 较好的后向兼容性



■ 磁盘接口——SCSI

- 小型计算机系统接口(**SCSI**, small computer system interface)
  - 相比ATA接口进一步改进
    - 例如同时支持串行/并行两种模式

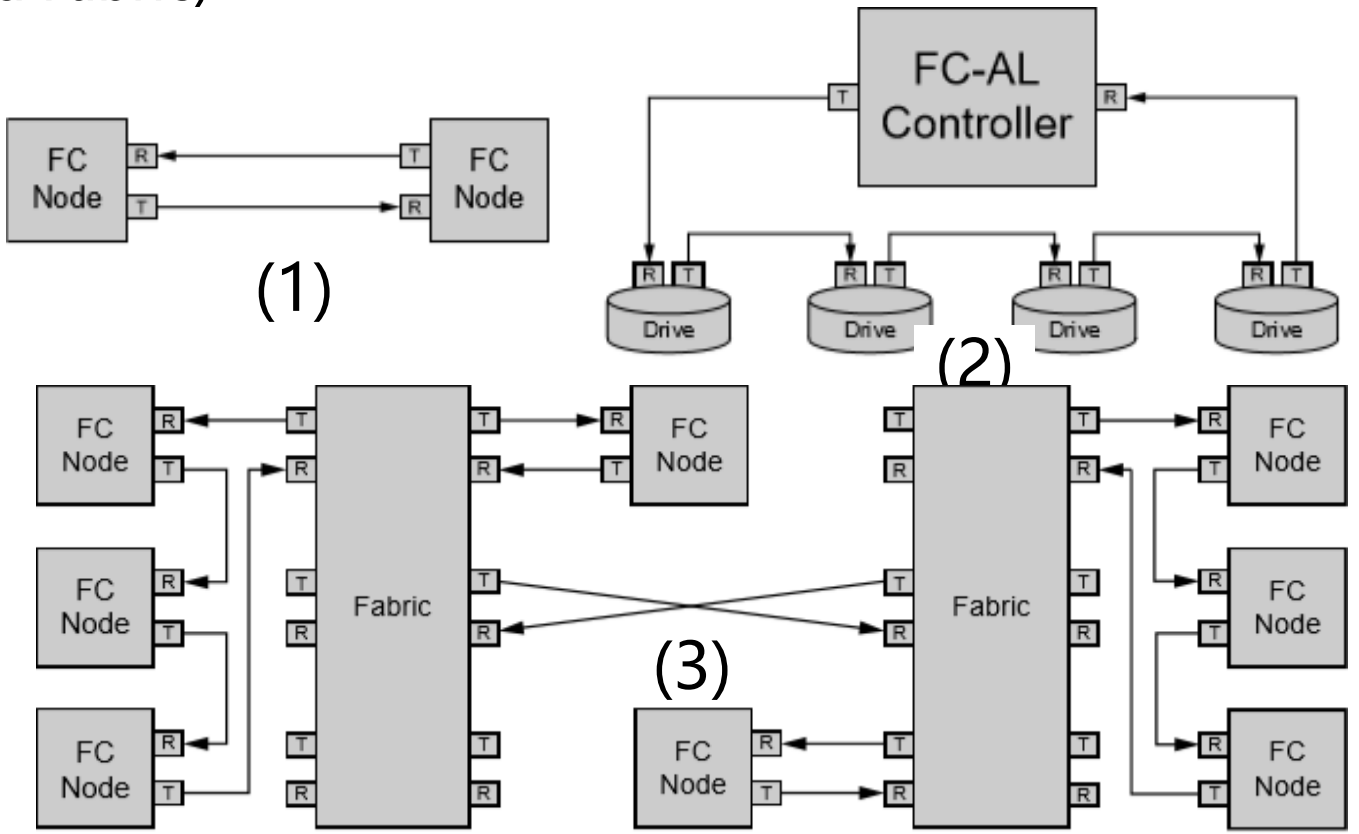


	SATA	Serial Attached SCSI (SAS)
优势	价格低、大容量、能耗更低	高速、高可靠、更长的线缆
应用	个人电脑	专业服务器系统

■ 磁盘接口——FC

□ 光纤通道(**FC**, fibre channel)

- 光纤通道是一种高速、功能丰富的串行接口
  - ➢ 有三种形态：1、点对点(Point to Point)；2、环路(Arbitrate Loop)；3、交换(Switched Fabric)



PART 02 磁盘阵列

■ 背景

◘ 问题

- 日益增长磁盘容量和速度需求与落后的访问速度和容量之间矛盾
  - 如720B全精度大模型动辄1TB数据存储需求，加上TB级数量的训练数据
  - 训练过程中需要高速访问数据
  - 当前市面上消费端单块磁盘容量最大20TB。而且提供的访问速度不足以支持LLM应用需求
- 磁盘技术提升边际效应明显，需要堆磁盘数量满足应用需要
  - 既要容量、又要速度、还要安全

■ 磁盘阵列

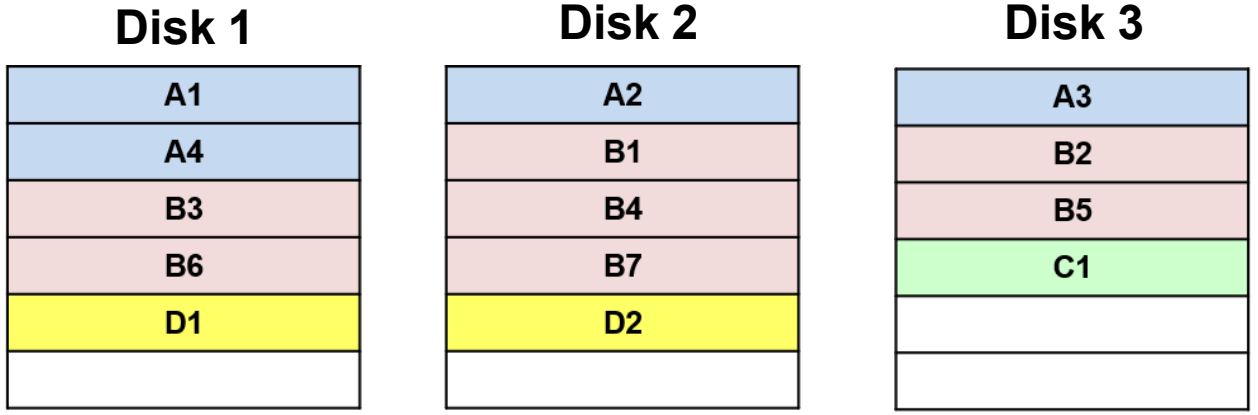
◘ 组合多块磁盘的策略

- JBOD(just bunch of driver)策略：逻辑上绑定一大串磁盘
  - 在操作系统层面把物理上的多块磁盘看作一个磁盘
  - 数据从第一块磁盘开始存放
  - **优势**：实现容易、成本低廉
  - **弊端**：数据安全性低，一块磁盘损坏可能导致所有数据无法读取
- 另一种策略是RAID策略
  - 依赖于数据条带化和数据镜像技术
  - **优势**：高访存性能、数据安全性高

■ 磁盘阵列

□ 数据条带化(data striping)

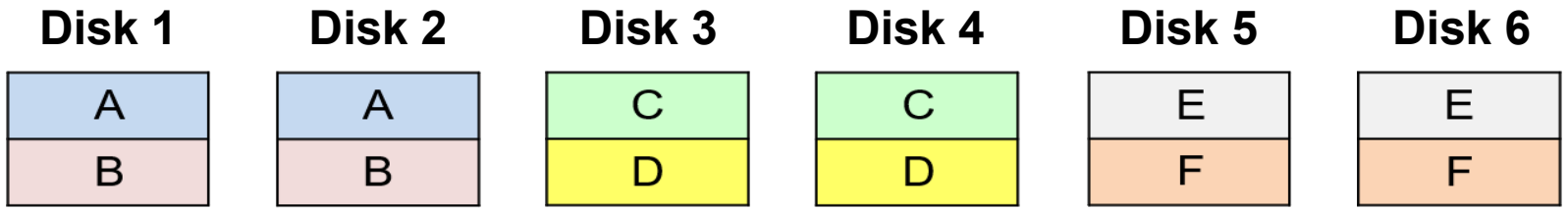
- 一种把连续数据划分为大小一致的数据块的技术
- 用三个参数控制：
  - 条带宽度(stripe factor/width)：指参与组建逻辑磁盘的磁盘数量
  - 条带单元(stripe unit)：固定大小的数据块
  - 条带大小(stripe size/depth)：数据块的大小
- 例如下图是
  - 条带宽度为3，4个用户数据的示例，每位用户文件大小不一



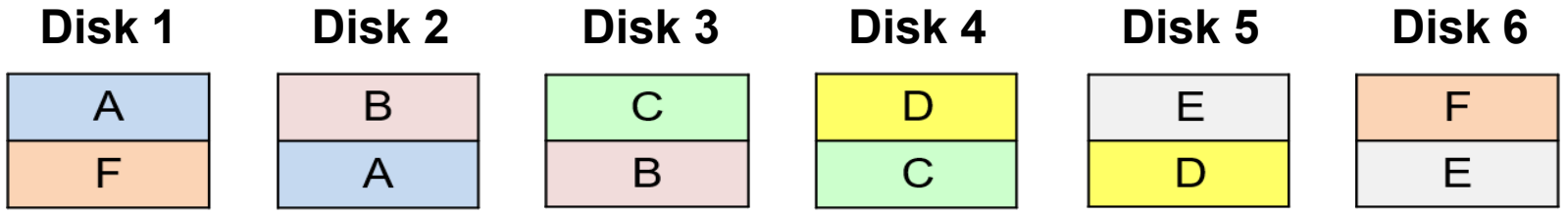
■ 磁盘阵列

□ 数据镜像(data mirror)

- 数据复制一份到另一个磁盘上，增加冗余提高数据安全性
- 普通数据镜像，6块磁盘时



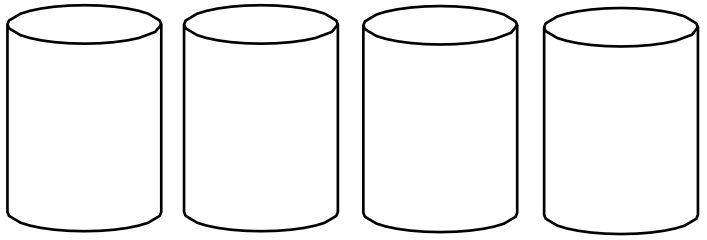
- 链式数据镜像，6块磁盘时
  - ➢ 对比普通镜像策略，链式镜像策略支持奇数块磁盘，因此使用更加灵活



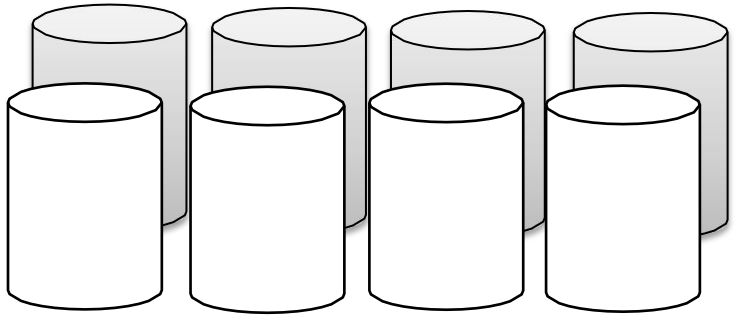
■ 磁盘阵列

- 廉价硬盘冗余阵列(**RAID**, redundant array of inexpensive drivers)
  - 为一组硬盘提供错误容忍度，保障数据安全
  - RAID有多种不同的策略
    - 其中RAID0和JBOD相似，没有数据冗余度。但因为条带化，把一个文件分散到多个磁盘上，并行读取提高读取速度
    - RAID1约等于两个RAID0互为镜像备份，提高数据安全性。但开销最大，假设4TB空间使用RAID1，实际可用仅有2TB

RAID0



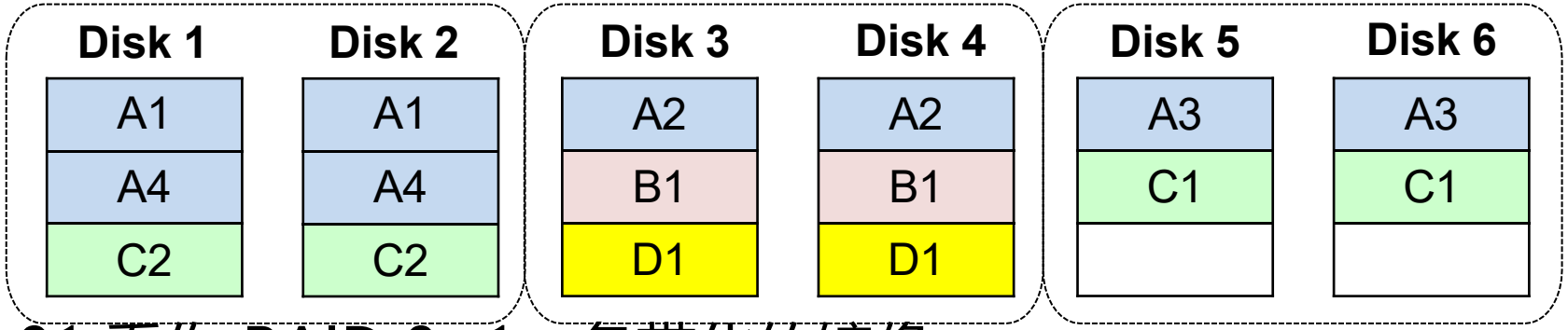
RAID1



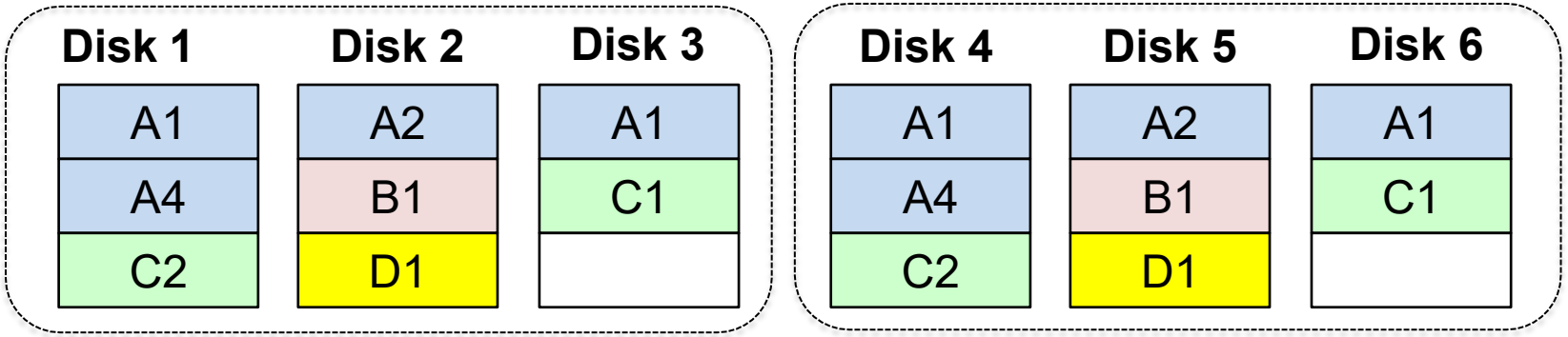
■ 磁盘阵列

◘ RAID 10 v.s. RAID 01

- RAID 10 看作 RAID 1+0，镜像的条带化



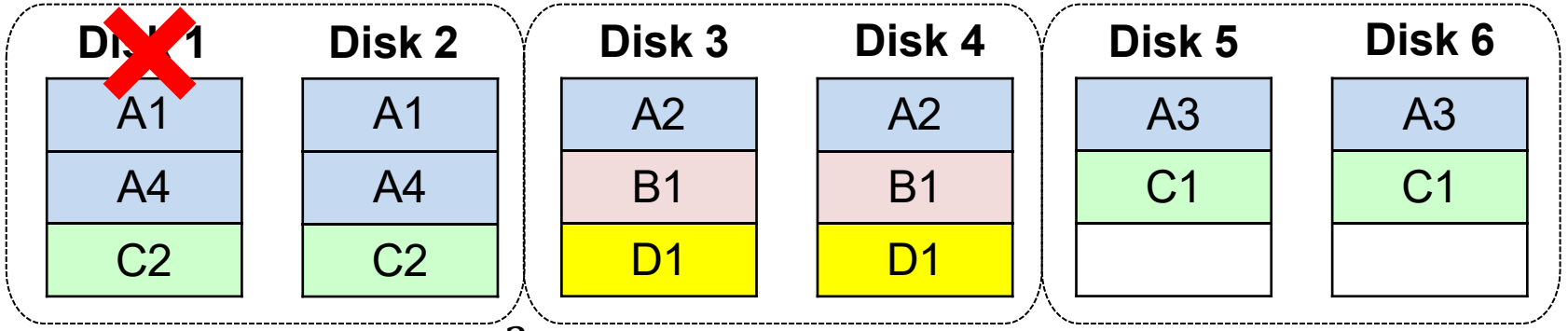
- RAID 01 看作 RAID 0+1，条带化的镜像



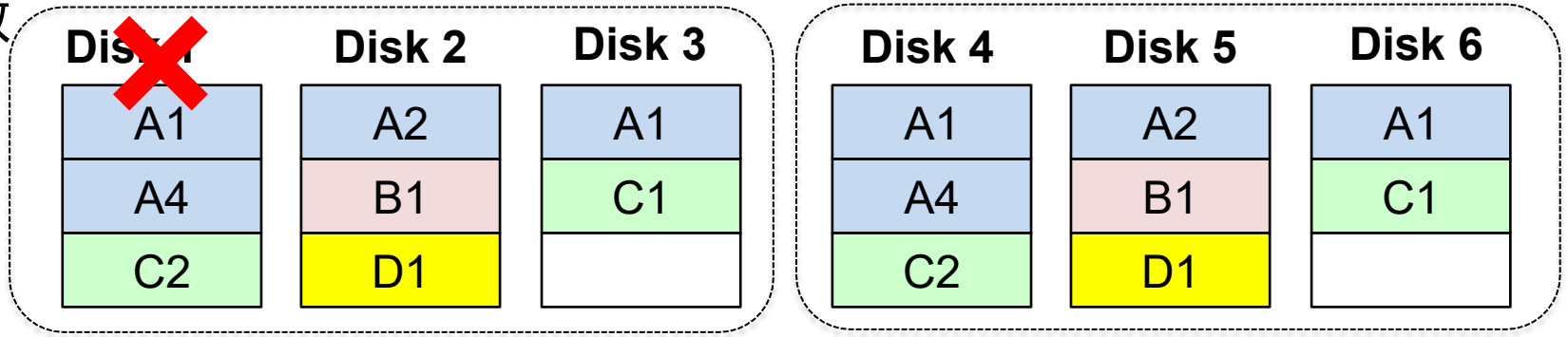
问题：如果每块磁盘损坏概率一致，是RAID01安全性高还是RAID10?

■ 磁盘阵列

- 答：RAID 10安全性更高。假设Disk 1已经损坏
  - RAID 10必须同一组的Disk 2也损坏，整个RAID才失效。失效概率是$\frac{1}{5}$



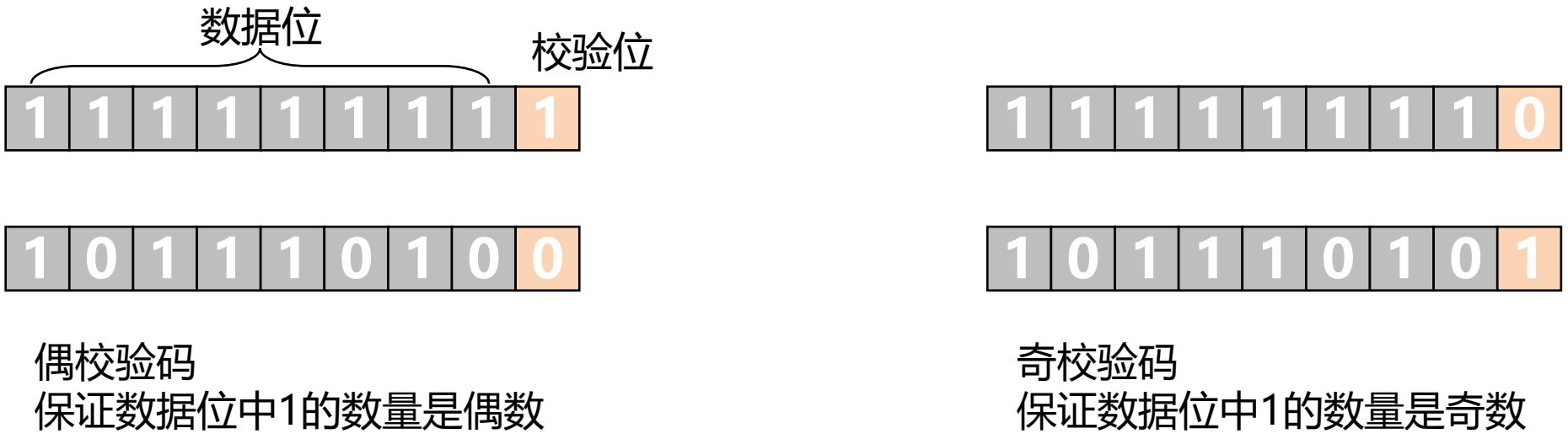
  - RAID 01的失效概率则是$\frac{3}{5}$
    - Disk 1失效后，Disk 2和Disk 3的数据不完整，失效。4~6再任意损坏一个，整个RAID失效



■ 磁盘阵列——校验码

◘ 错误容忍机制(fault tolerance)

- 对存储介质的数据可靠性要求很高
  - ➢ 用数据备份的方法简单且十分有效，但非常昂贵
  - ➢ **纠错码**(error correcting coding)的方式相对有效，但成本低廉
- 例如采用奇偶校验码识别错误

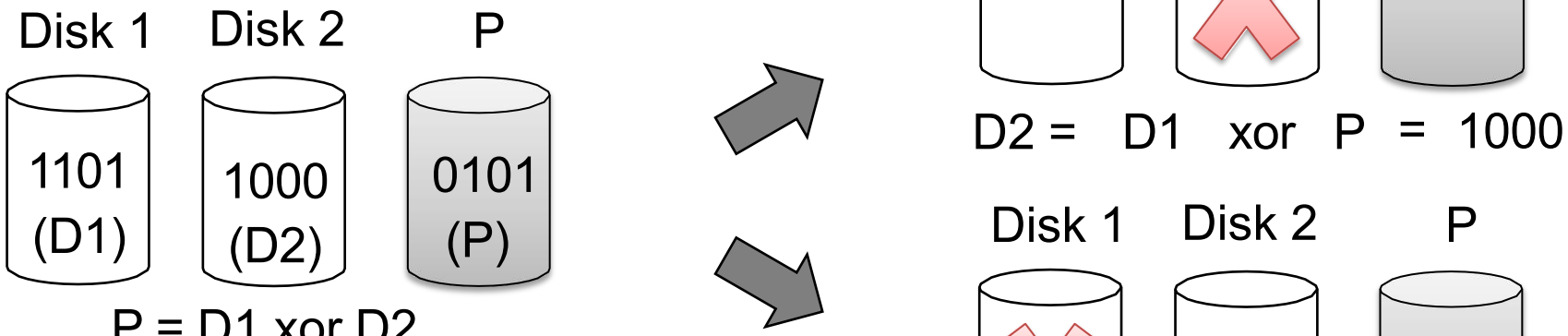


■ 磁盘阵列——校验码

□ 异或冗余校验码

- 利用异或计算的性质设计冗余备份
  - ➢ 仅需三块磁盘达成RAID 0的效果

XOR	0	1
0	0	1
1	1	0



■ 磁盘阵列

◘ RAID 2、3、4、5、6

- RAID 2：bit位级数据条带化，带汉明码纠错
  - 每个bit都分开存储
  - 在工业界已经绝迹，开销太大，复杂度很高
- RAID 3：Byte级数据条带化，带冗余校验位
  - 异或校验码
  - 在工业界极少使用，不具备纠错能力
- RAID 4：数据块级条带化，带冗余校验位
  - 和RAID 3相似，条带化从Byte级升到数据块级
  - 不常使用

■ 磁盘阵列

◘ RAID 2、3、4、5、6

- RAID 3和RAID 4的校验位单独存放于一块磁盘
  - 在写入数据时需要更新校验位，这块存放校验位的磁盘容易成为瓶颈
- RAID 5：类似RAID 4，校验位同数据一样条带化并分布到多个磁盘上
  - 允许阵列中一块磁盘出现错误
- RAID 6：在RAID 5上进一步改进，同时采用两种校验码
  - 允许阵列中两块磁盘出现错误
  - 在对数据安全和成本均要求较高的情况下采用
  - 复杂的控制逻辑提高相应硬件成本

■ 磁盘阵列

◘ RAID 2、3、4、5、6

- 实际上RAID 5损坏的事故最容易见到
  - RAID 5的开销最小，可以最大化存储空间
  - 用户过分相信RAID 5，没有留好数据备份
  - 组RAID 5的磁盘大概率是同时购入的，因此出现一块磁盘损坏时，其他磁盘的寿命也所剩无几
  - RAID 5校验恢复过程加重其余磁盘的负担，然后第二块磁盘也坏了，整个RAID失效

■ 磁盘阵列

- RAID不等于备份
  - RAID不能保证数据安全

PART 03 NAS与SAN

■ 背景

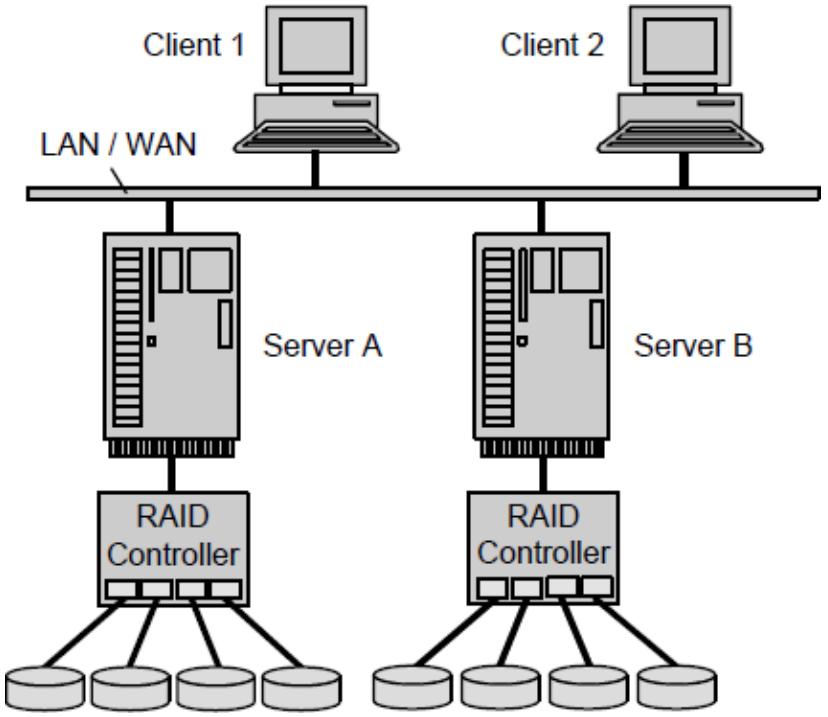
□ 一组磁盘仍然不能满足需求

- 一台服务器能承载的磁盘数量有限
  - ➢ 即一台服务器能提供的存储空间有上限，而且上限增长缓慢
  - ➢ 但应用的需求是无限的
- 这时需要把一系列服务器组成集群，然后共享各自的存储空间

■ DAS

□ 直接附加存储(**DAS**, direct access storage)

- 磁盘直接连接在服务器上
- 通过网络访问不同服务器上的数据



■ DAS

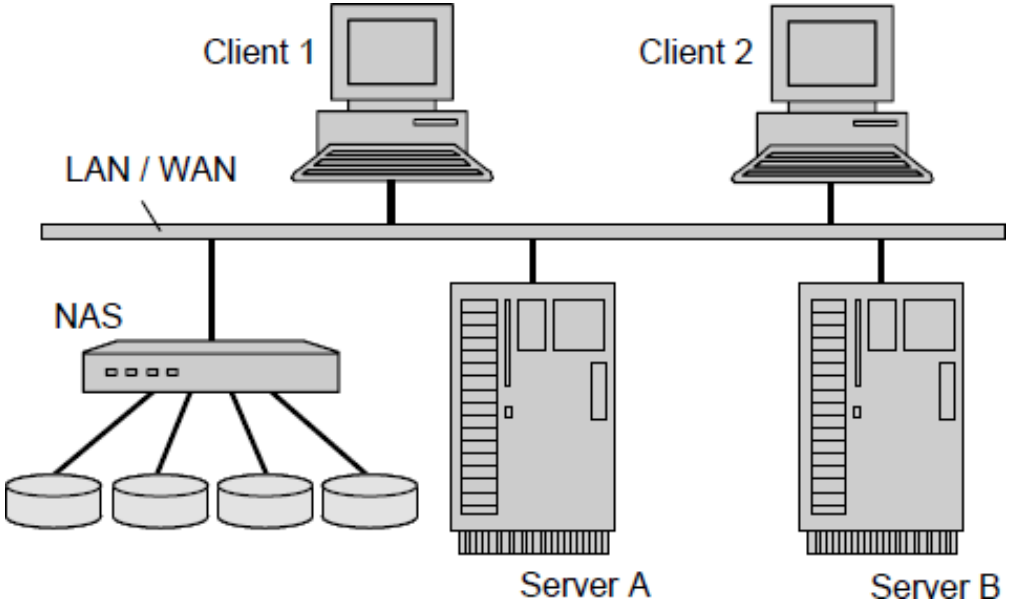
□ DAS

- 优势十分明显
  - ➢ 改造所需成本最低，不需要引入额外组件，且实现容易
- 弊端
  - ➢ 通过网络访问数据，读写性能一般，且延迟较高
  - ➢ 通过网络传输数据，会挤占其他正常的网络通信带宽
  - ➢ 如果服务器A关闭，那么无法访问服务器A下的数据

■ NAS

□ 网络连接存储(**NAS**, network attached storage)

- 解决DAS中服务器A关闭，无法访问服务器A下的数据问题
- 把计算和存储分离
  - 服务器只关心计算，不再负责数据的存储
  - NAS作为一种特殊的设备，包含最精简的操作系统、配置一般的处理器。只负责向网络上其他设备提供文件存储服务



■ NAS

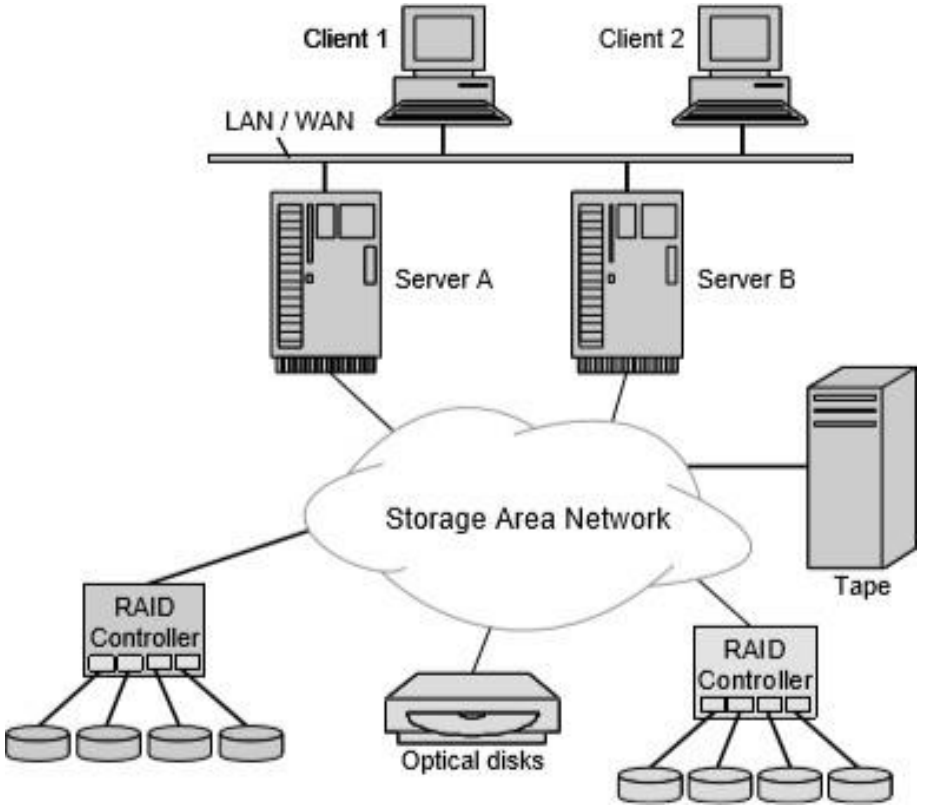
◘ NAS

- 优势
  - ➢ 十分经济实惠方式提供共享存储资源的服务
  - ➢ 容易配置，且NAS设备可以硬件支持RAID功能
  - ➢ 提高存储资源利用率，不再受限于服务器是否启动
- 弊端
  - ➢ 数据仍然通过网络传输，和正常网络通信抢占带宽的问题没解决
  - ➢ 延迟和读写性能受制于网络状况
  - ➢ 单台NAS能提供的容量有上限，多台NAS间协作困难

■ SAN

□ 区域存储网络(**SAN**, storage area network)

- 解决NAS存在的弊端
  - 具有较好可扩展性
  - 不同存储设备间可以协作，对外提供统一接口
- 为存储区域提供单独的网络带宽资源
  - 通常是配备额外的交换机、光纤模块等硬件



■ SAN

◘ SAN的优势

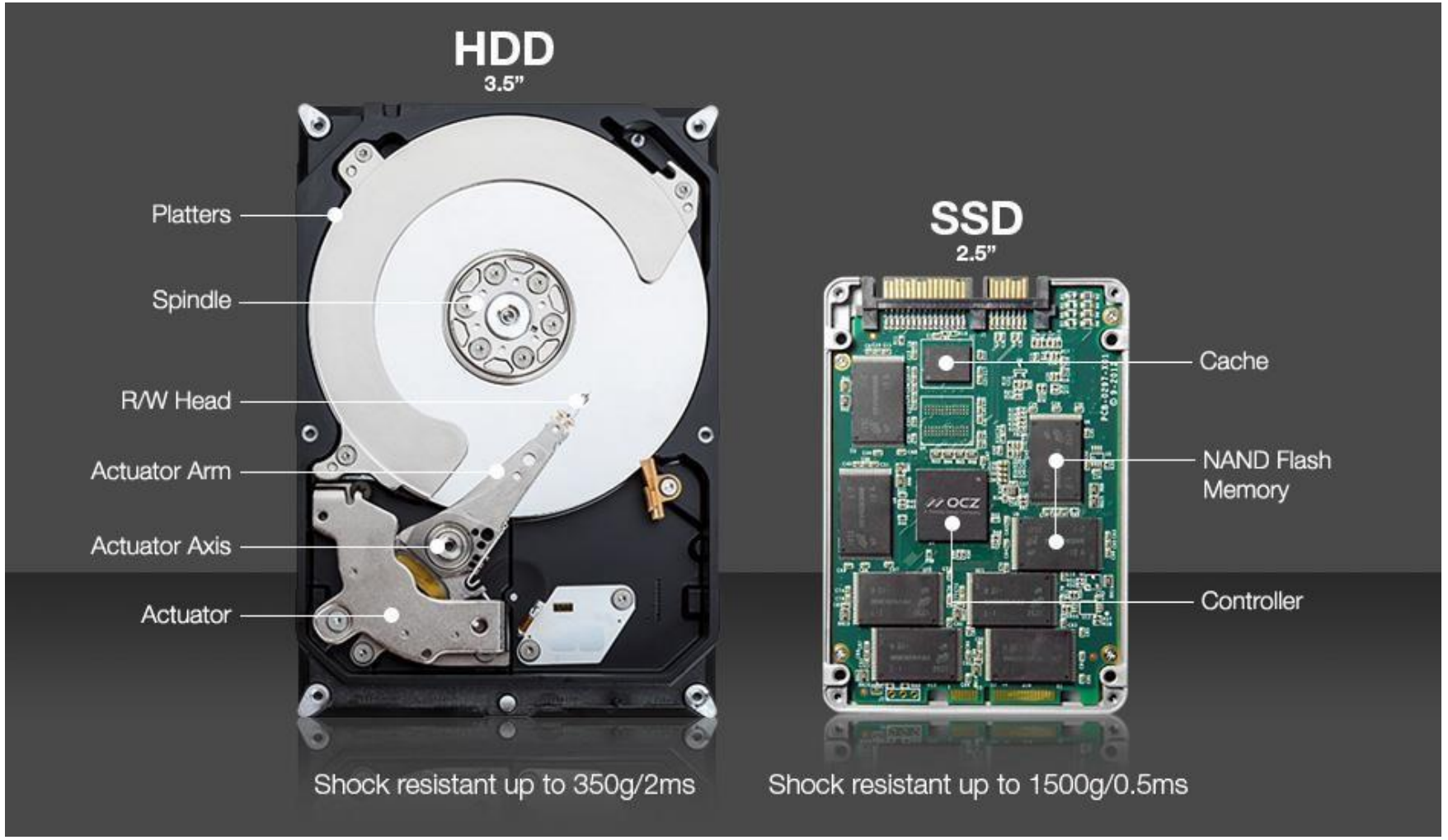
- 不占用正常网络访问带宽
- 支持异构存储设备
- 适合中心化管理
- 极高的硬件利用率和读写性能

	SAN	NAS
互联结构	光纤模块	以太网
数据访问方式	数据块	文件
价格	极高	便宜适中
性能	极高	一般
适用范围	数据中心、超算中心	家庭、个人

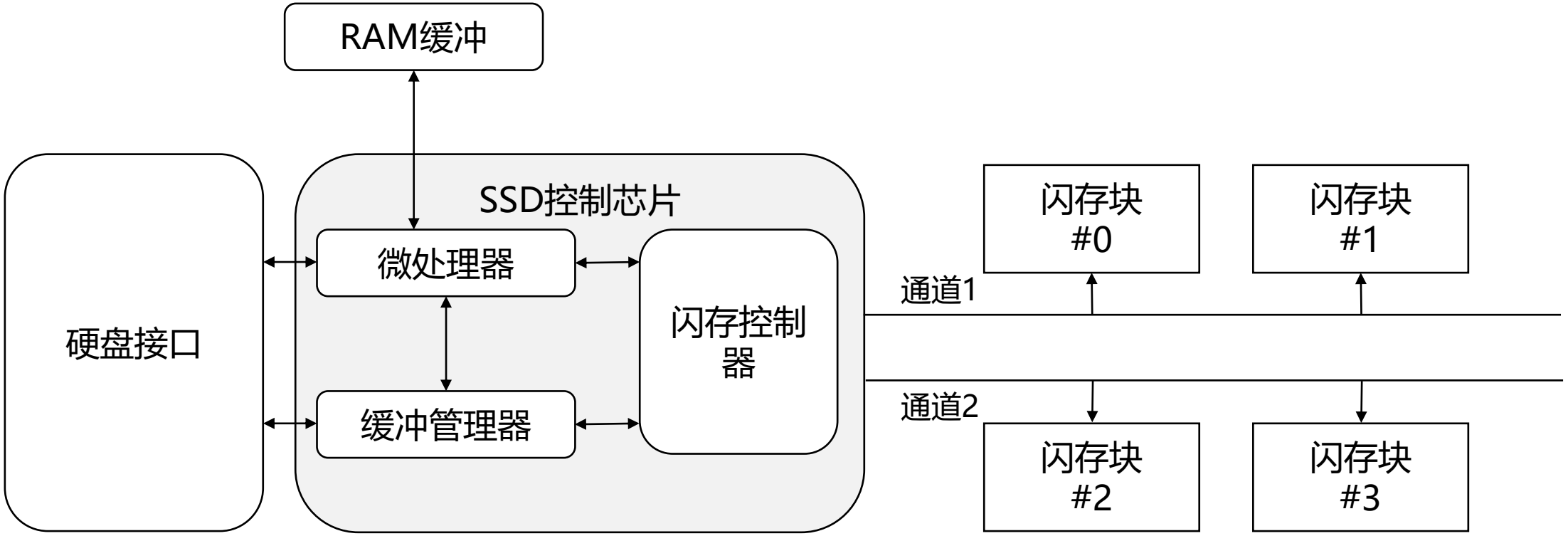
PART 04 闪存

■ 固态硬盘

□ 固态硬盘(SSD, solid-state drive)



■ 固态硬盘结构


RAM缓冲
SSD控制芯片
微处理器
缓冲管理器
闪存控制器
硬盘接口
通道1
通道2
闪存块 #0
闪存块 #1
闪存块 #2
闪存块 #3

■ 固态硬盘

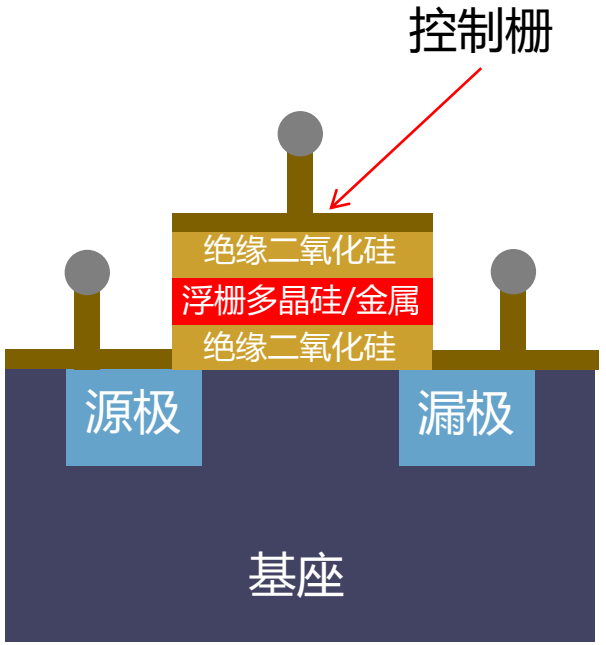
◘ 所见即所得

- 相比于包的严严实实的机械硬盘
- NVME接口的固态硬盘十分容易见到对应的元件
- 看上去很复杂，其实就两个关键原件
  - ➢ 用于存数据的闪存颗粒
  - ➢ 用于管理闪存的控制器
  - ➢ 高端的固态硬盘还配有独立DRAM缓冲

■ 固态硬盘原理

◘ 浮栅晶体管(floating gate transistor)

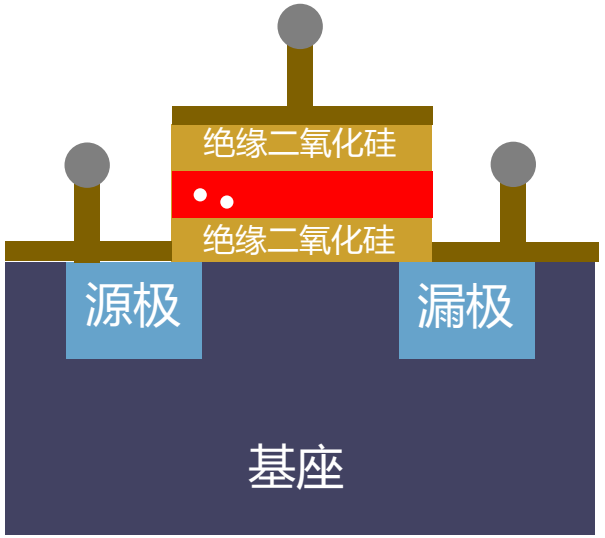
- 浮栅名字来源于其形状
  - ➢ 最顶上的铜触点称为控制栅，存储电荷的浮栅被绝缘体包裹。电子一旦进入浮栅，在没有外接高压电情况下无法逃逸，因此可以保存数据长达十年之久
- 是闪存的最基本组成元件



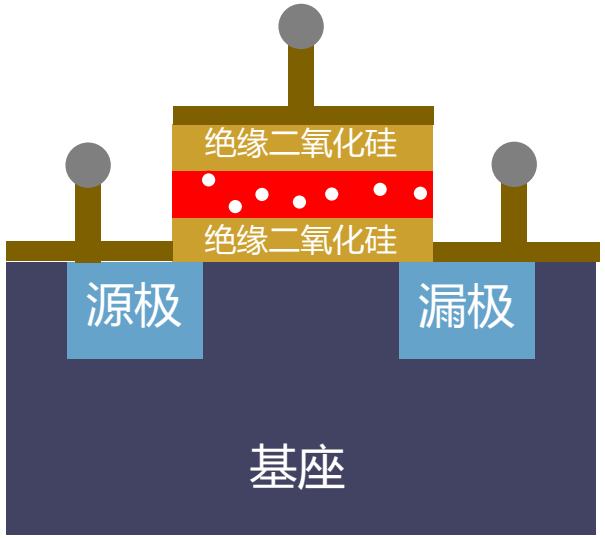
■ 固态硬盘原理

◘ 浮栅晶体管

- 能保存电子数量，就可以通过控制进入浮栅的电子数量，来表示0或1
  - ➢ 精确读取电子数量非常困难，但好在我们只需要知道电子数量是否超过一定阈值
  - ➢ 电子数较少时代表存储1，电子数较多时代表存储0



1

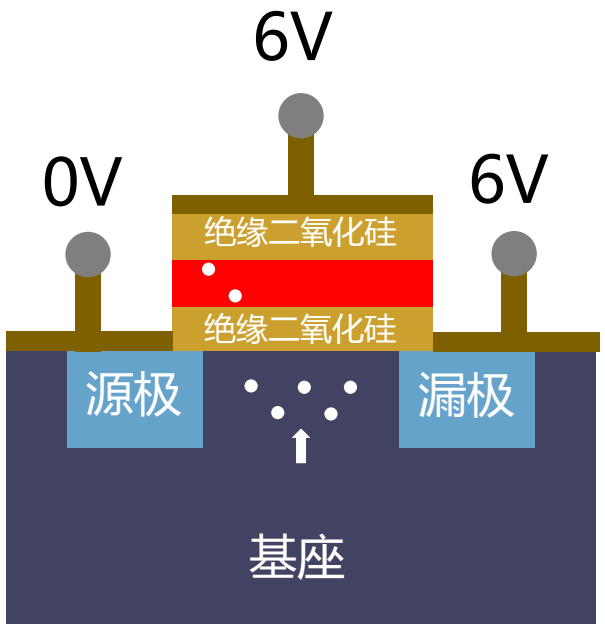


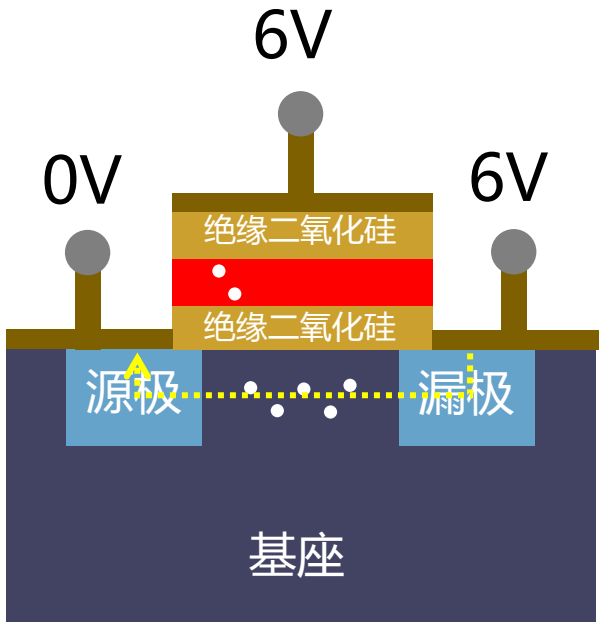
0

■ 固态硬盘原理

◘ 浮栅晶体管——读取

- 读取时给控制栅和漏极加一个高电压
  - ➢ 基座的电子受到控制栅吸引聚集在基座表面
  - ➢ 当控制栅的电压达到一定阈值，基座表面积累足够的自由电子。源极和漏极会导通，这个导通电压称为阈值电压。此时读取到1



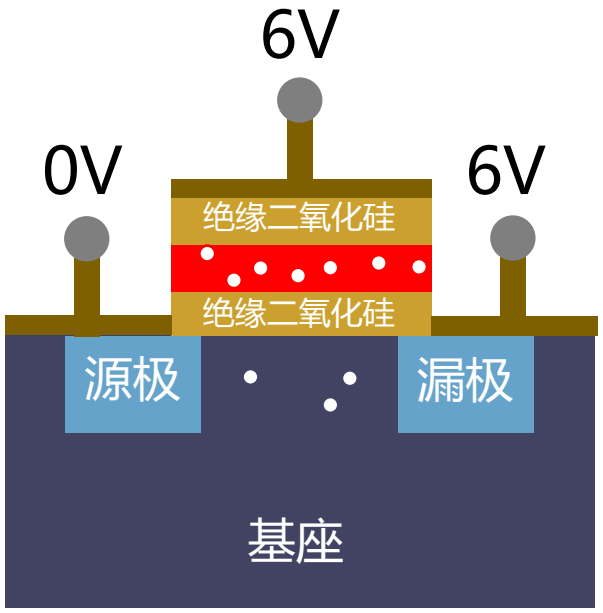


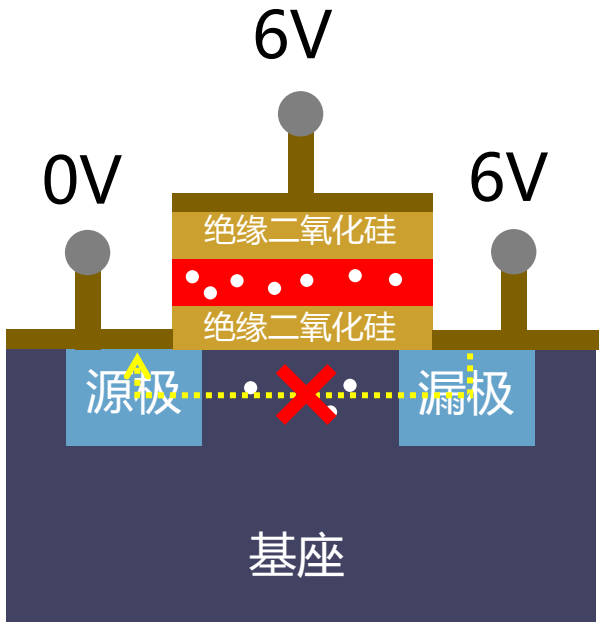
■ 固态硬盘原理

□ 浮栅晶体管——读取

– 反过来，如果浮栅晶体管内部有较多自由电子，会排斥靠近基座表面的电子

➢ 积累的自由电子不够多，无法导通源极和漏极。读到0

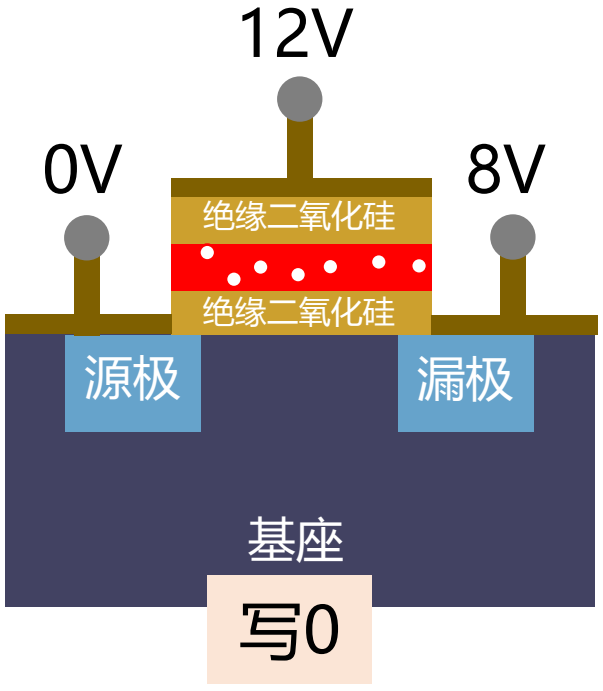


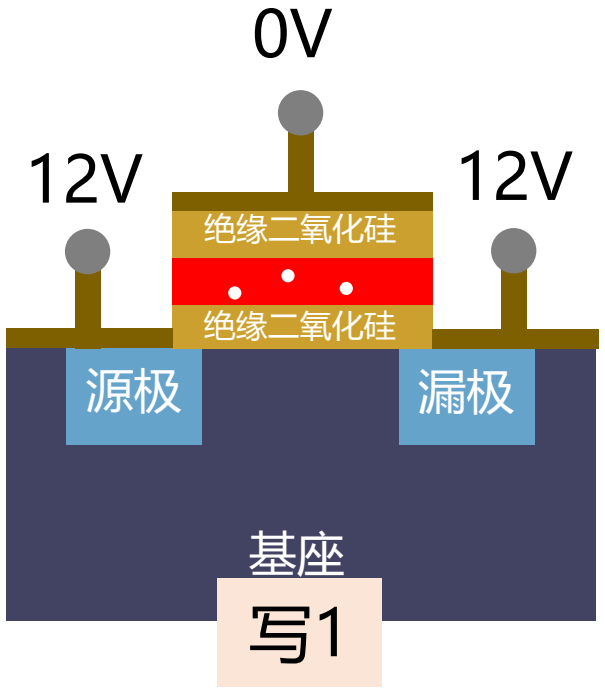


■ 固态硬盘原理

□ 浮栅晶体管——写入

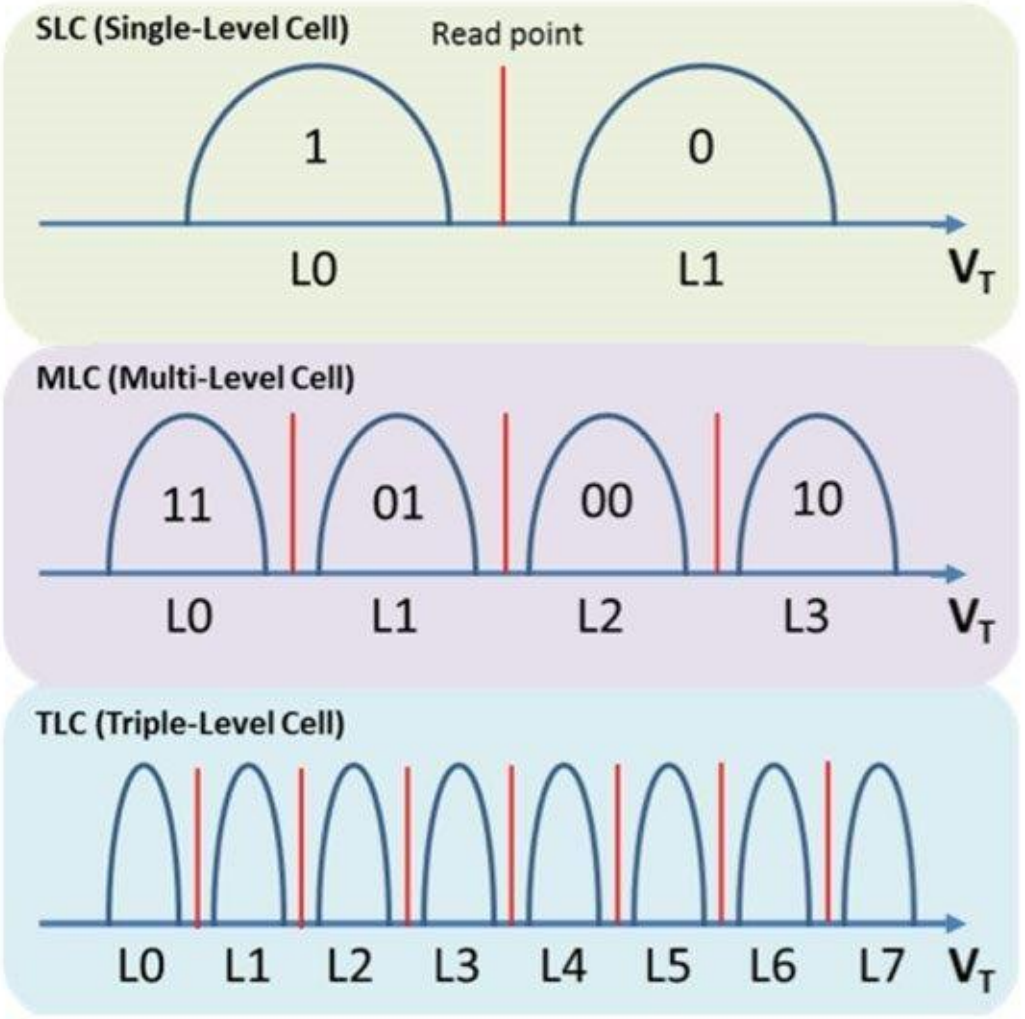
- 写0等于向浮栅存入电子，给控制栅加一个极高电压
  - ➢ 电子因为量子隧穿效应穿过绝缘的二氧化硅层
- 电子穿过绝缘二氧化硅层有损耗
  - ➢ 多次写入后浮栅会和基座接触，失去存放电子的能力。因此固态硬盘的闪存有写入寿命





■ 固态硬盘原理

- SLC(single-level cell)与MLC(multi-level cell)
  - 通过检测电子数量是否超过一个阈值，来读取晶体管存放的数据
    - SLC存储1bit数据，只存储0/1两种状态
      - 缺点是对于晶体管利用率不高，无法做到较大容量
      - 但读取速度非常快，量一下电压即可
    - MLC存储2bit数据，存储00/01/10/11
    - TLC(triple-level cell)存3bit数据
      - 提高晶体管利用率，能做到较大容量
      - 但读取相对缓慢



■ 固态硬盘原理

◘ SLC与MLC

- 目前市面上主流是TLC与QLC
  - QLC一个晶体管能存放4bit数据
  - 技术上PLC(存放5bit数据)也已经出现，但未进入市场
- 多bit数据如何存入一个晶体管
  - 通过不同电压区分不同数据
  - 例如SLC中1V阈值电压表示数据0，0V表示1
  - 例如MLC则是1V阈值电压表示00，0.75V表示10，0.5V表示01，0V表示11

■ 固态硬盘原理

◘ SLC→QLC，容量变大，读写速度下降

- 容量增加可以理解，一个晶体管能表示的bit数增多
- 读写下降是为什么，先说读
  - 假设一个MLC的晶体管，存放10数据，阈值电压是0.75V
  - 0.75V是阈值电压，1V也能导通这个晶体管
  - 如果在控制栅加1V电压，发现晶体管导通，不能说明当前晶体管存储00，有可能是10
  - 因此需要在控制栅依次加电压，记录首次导通的电压
  - 对于MLC需要扫描4次，TLC则是8次

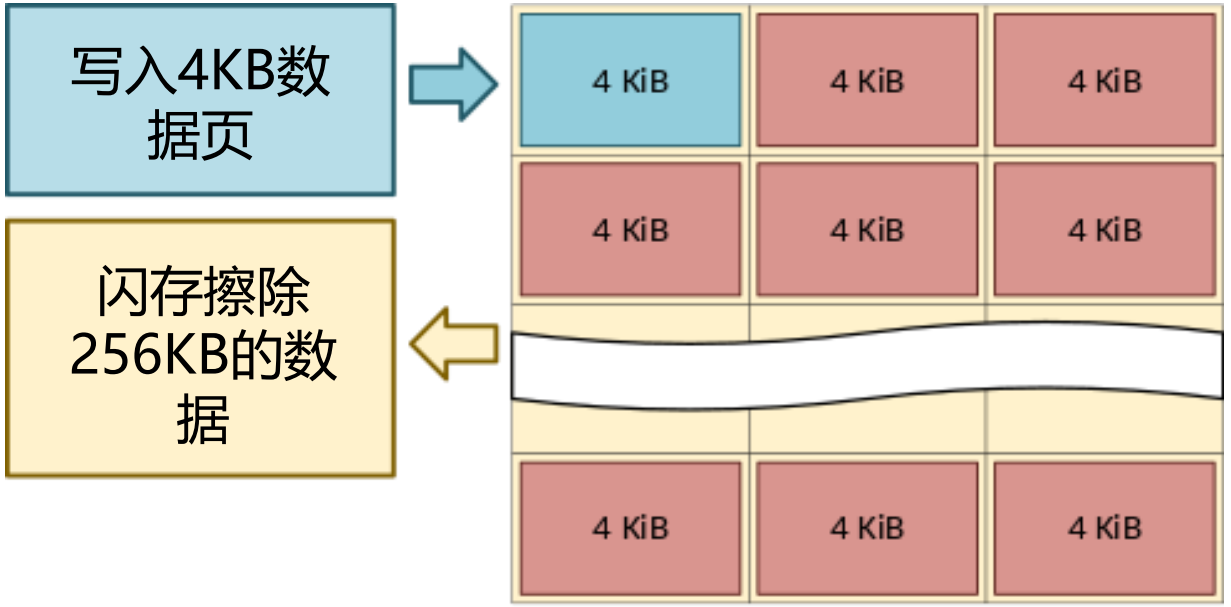
■ 固态硬盘原理

- SLC→QLC，容量变大，读写速度下降
  - 至于写
    - 不同阈值电压的本质是浮栅晶体管中存储的电子数量在一定范围内
    - 假定1V电压对应10000个电子
    - 对于TLC，只需要控制电子数量误差在1000以内，就不会出现错误
    - 对于QLC，则需要精准到500
  - 更高的精准度要求，增加写入控制复杂度
    - 为确保写入的正确，还需要回头验证写入结果

■ 固态硬盘原理

□ 写入放大效应(write amplification)

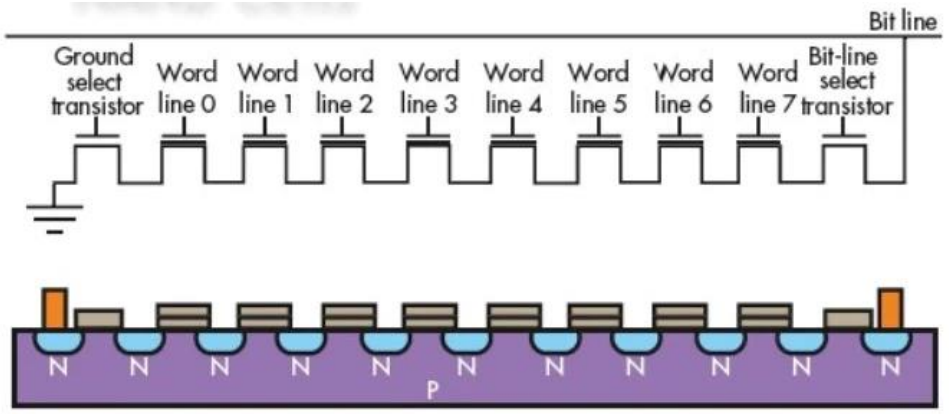
- 定义：实际向固态硬盘写入的数据数量远大于需求的现象
  - ➢ 例如只修改4KB的数据，但闪存实际修改了256KB的数据块
- 影响：增加对闪存的写入，降低寿命



■ 固态硬盘原理

□ 写入放大效应

- 原因是人们为了进一步提高浮栅晶体管的密度
- 不断把晶体管做得越来越小，寸土寸金的内部空间不允许为每个单独的晶体管设置单独的地址线
  - ➢ 需要多晶体管共享一个地址线，不能精确控制写入的晶体管



■ 小结

◘ 即便是TLC

- 在访问延迟上也远低于机械硬盘
  - 没有寻道时间，没有旋转延迟。受物理限制，机械硬盘这两部分延迟不可能无限降低
- 在读写速度上超过机械硬盘
  - 7000MB/s (致钛 Ti600)
  - 机械硬盘读取普遍在300MB/s
  - 尤其是大量小文件读写，优势更明显
- 比机械硬盘更加安全稳定
  - 机械结构少，抗摔防振

■ 小结

◘ SSD在现代存储系统

- 混合：磁存储介质+SSD缓冲
  - ➢ 更快启动，更少加载时间
  - ➢ 更大容量，价格便宜
- 纯闪存
  - ➢ 追求极致性能
  - ➢ 无机械结构更加安全
  - ➢ 需要平衡价格和容量、性能

感谢！