

# 计算机组成和体系结构

## 第十讲

四川大学网络空间安全学院

2020年4月27日

封面来自logicalincrements.com

# 版权声明

---

课件中所使用的图片、视频等资源版权归原作者所有。

课件原创内容采用 [创作共用署名—非商业使用—相同方式共享4.0国际版许可证\(Creative Commons BY-NC-SA 4.0 International License\)](#) 授权使用。

Copyright@四川大学网络空间安全学院计算机组成与体系结构课程组，2020



# 上期内容回顾

---

- 冯诺依曼模型的第三部分：输入输出系统
  - 子系统性能与计算机系统整体性能的关系：阿姆达尔定律
- I/O子系统组成和体系结构
  - I/O子系统示例
  - 如何控制I/O设备：五种控制方法、字符和块I/O、总线控制与时序图
- 数据传输模式：并行传输和串行传输

# 本期学习目标

---

- 持久存储介质
  - 硬盘、固态硬盘
  - 光盘
  - 磁带
- 独立磁盘冗余阵列
- 分布式存储简介

# 中英文缩写对照表

英文缩写	英文全称	中文全称
I/O	Input/Output	输入输出
DMA	Direct Memory Access	直接内存访问
RDMA	Remote Direct Memory Access	远程直接内存访问
HDD	Hard Disk Drive	硬盘
SSD	Solid State Drive	固态硬盘
RAID	Redundant Array of Independent Disks	独立磁盘冗余阵列
RS	Reed-Solomon	里德-所罗门编码



# I/O 3: 存储介质

# 持久存储介质

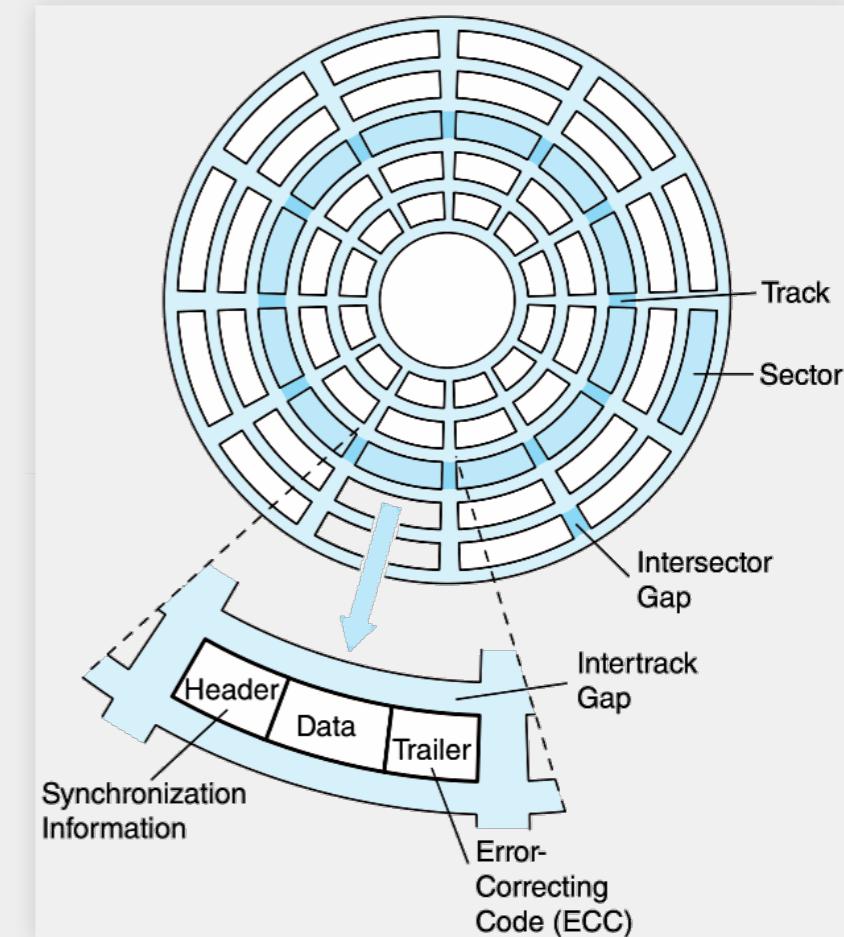
---

存储系统是计算机最常见的I/O设备，这一节中介绍的存储介质包含下面一些特征：

- 持久的数据(durable/persistent storage)保存能力：断电之后数据能否继续保存
- 性能指标包含了访问速度、有效保存时间、访问方式和单位成本等
- 针对不同的应用场景，根据特定的性能需求选择对应的存储介质

# 磁盘/机械硬盘

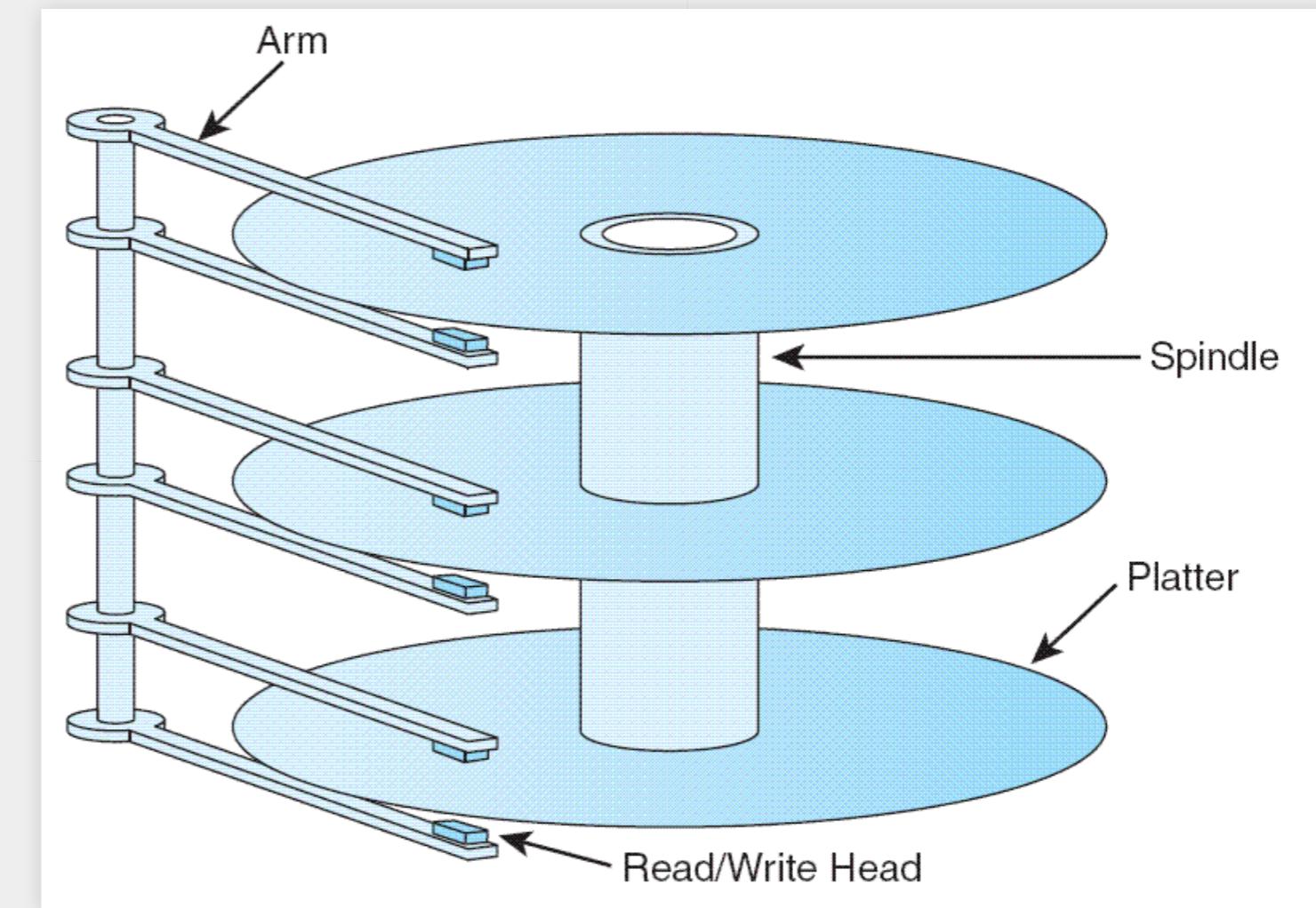
- 机械硬盘是一种**大容量、随机访问**的持久数据存储
- 数据采用磁性介质对应的磁极表示(北极表示0,南极表示1)
- 机械硬盘的逻辑视图如右图所示
  - 磁道(Track): 磁盘上的同心圆, 从外向里编号
  - 扇区(Sector): 一个磁道上分为多个扇形区域, 所有扇区的数据存储量相同, 且通常为512位
  - 每个扇区分为三段: 头部、数据和尾部



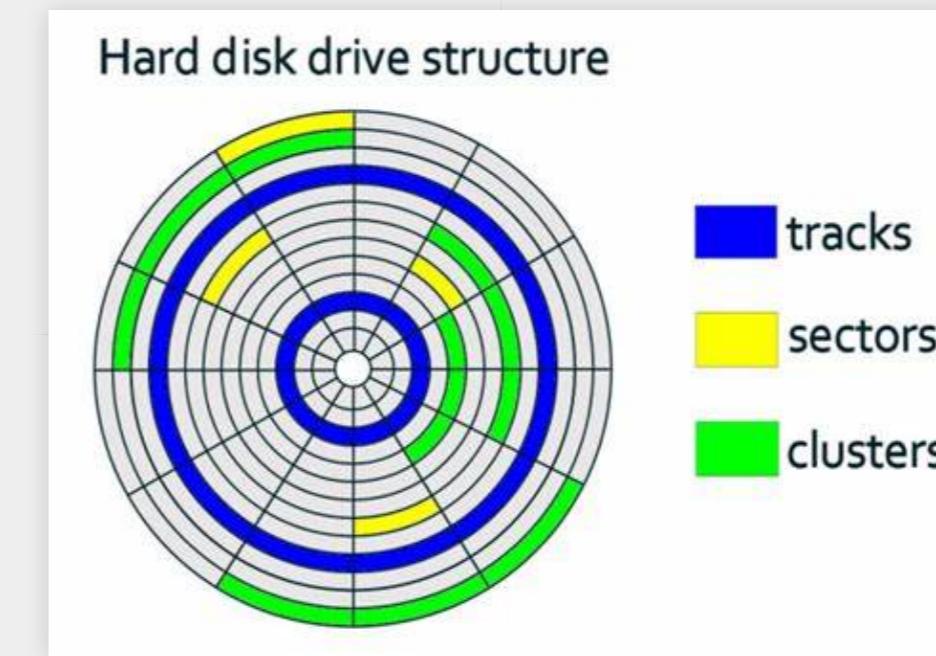
资料来源: [https://web.mit.edu/2.972/www/reports/floppy\\_drive\\_read\\_write/floppy\\_drive\\_read\\_write.html](https://web.mit.edu/2.972/www/reports/floppy_drive_read_write/floppy_drive_read_write.html)

# 机械硬盘的物理结构

- 一个机械硬盘通常包括多个磁盘盘片(Platter)和一个转轴(Spindle)
- 每个盘片都具有一个或两个读写磁头(Read/Write Heads)
- 读写磁头连接在驱动臂(Arm)上
- 所有片面上编号相同的磁道组成一个圆柱面(Cylinder)



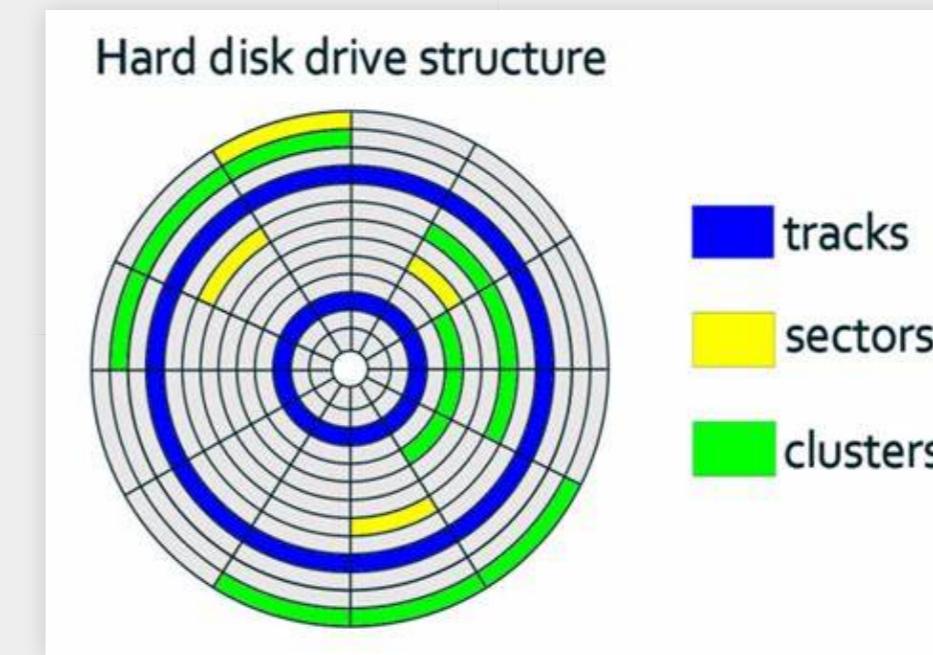
# 机械硬盘上的数据读写



图片来源：<https://www.slashcam.com/news/single/Why-should-a-maximum-of-80-of-hard-disk-space-be-14071.html>

# 机械硬盘上的数据读写

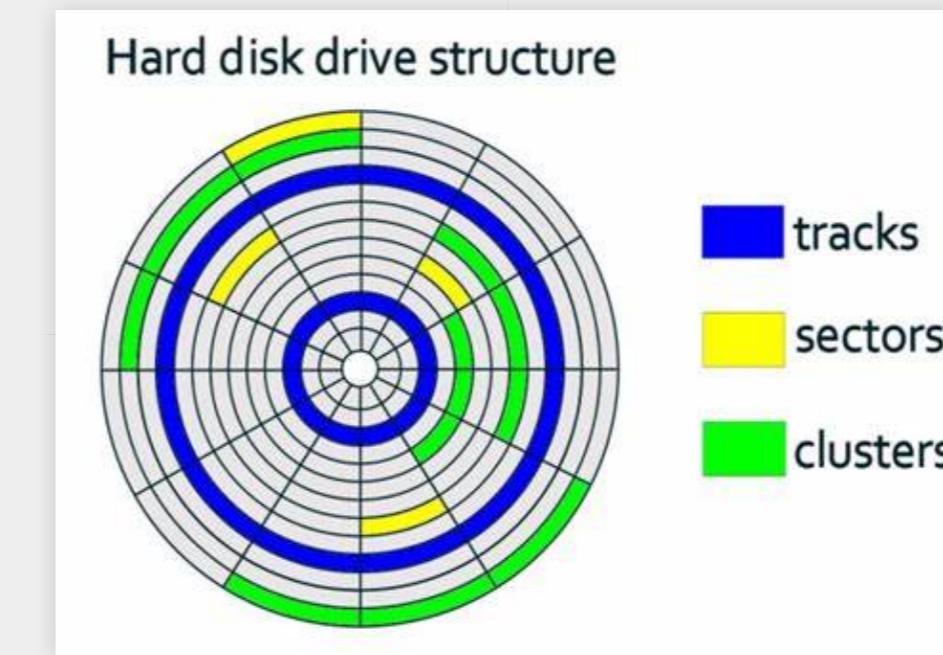
- 机械硬盘的数据读写是通过读写磁头感应、修改每个存储单元的磁极实现的



图片来源：<https://www.slashcam.com/news/single/Why-should-a-maximum-of-80-of-hard-disk-space-be-14071.html>

# 机械硬盘上的数据读写

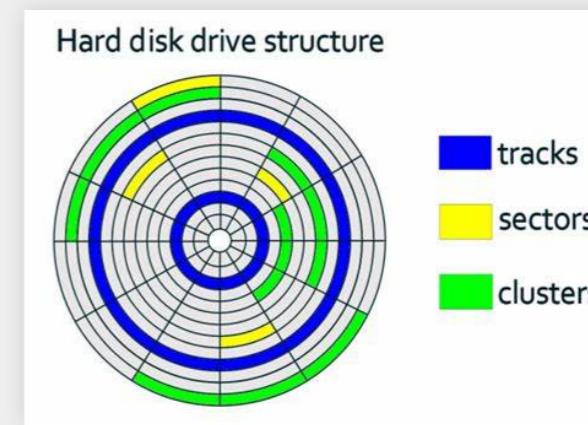
- 机械硬盘的数据读写是通过读写磁头感应、修改每个存储单元的磁极实现的
- 通常为了减少扇区寻址的开销和文件碎片，通常将多个相邻的扇区组合为一簇(cluster)



图片来源：<https://www.slashcam.com/news/single/Why-should-a-maximum-of-80-of-hard-disk-space-be-14071.html>

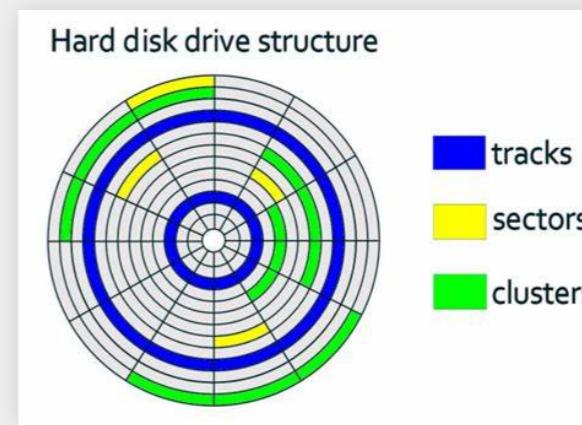
# 机械硬盘的访问速度

---



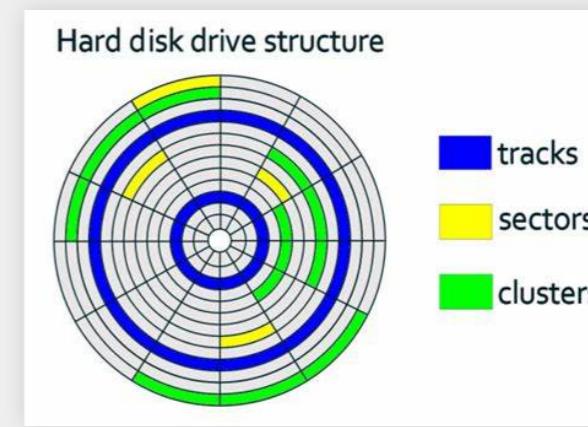
# 机械硬盘的访问速度

- 通过控制驱动臂的角度，将磁头移动到对应的磁道上，这个过程称为**寻道**



# 机械硬盘的访问速度

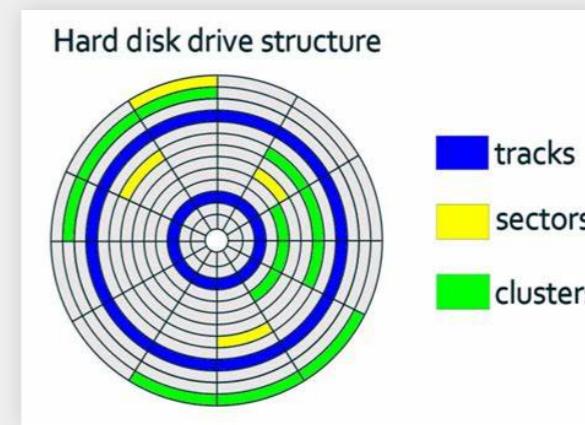
- 通过控制驱动臂的角度，将磁头移动到对应的磁道上，这个过程称为**寻道**
- 通过控制转轴，将数据对应的扇区旋转到磁头所在位置，这个过程称为**旋转**



# 机械硬盘的访问速度

- 通过控制驱动臂的角度，将磁头移动到对应的磁道上，这个过程称为**寻道**
- 通过控制转轴，将数据对应的扇区旋转到磁头所在位置，这个过程称为**旋转**
- 机械硬盘的平均旋转延迟(Average Latency): 扇区旋转到磁头所在位置需要的平均时间

$$\frac{\frac{60s}{r} \times \frac{1000ms}{s}}{2}, \text{其中 } r \text{ 代表硬盘转速}$$

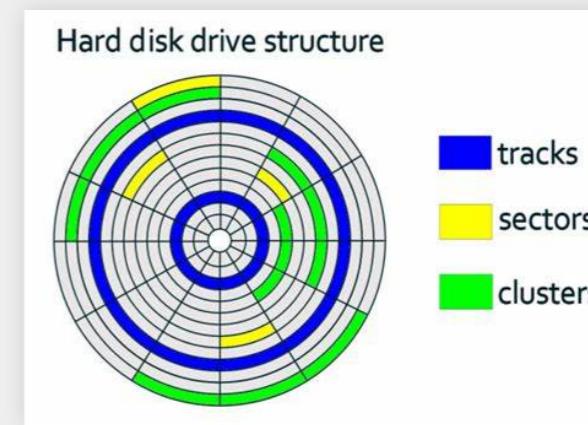


# 机械硬盘的访问速度

- 通过控制驱动臂的角度，将磁头移动到对应的磁道上，这个过程称为**寻道**
- 通过控制转轴，将数据对应的扇区旋转到磁头所在位置，这个过程称为**旋转**
- 机械硬盘的平均旋转延迟(Average Latency): 扇区旋转到磁头所在位置需要的平均时间

$$\frac{\frac{60s}{r} \times \frac{1000ms}{s}}{2}, \text{其中 } r \text{ 代表硬盘转速}$$

- 机械硬盘的访问时间 访问时间 = 寻道时间 + 旋转延迟

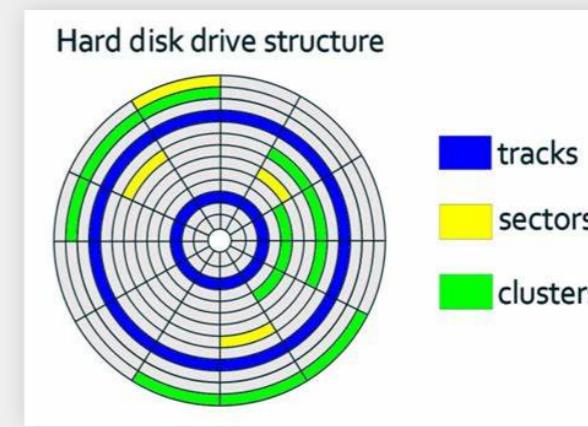


# 机械硬盘的访问速度

- 通过控制驱动臂的角度，将磁头移动到对应的磁道上，这个过程称为**寻道**
- 通过控制转轴，将数据对应的扇区旋转到磁头所在位置，这个过程称为**旋转**
- 机械硬盘的平均旋转延迟(Average Latency): 扇区旋转到磁头所在位置需要的平均时间

$$\frac{\frac{60s}{r} \times \frac{1000ms}{s}}{2}, \text{其中 } r \text{ 代表硬盘转速}$$

- 机械硬盘的访问时间 访问时间 = 寻道时间 + 旋转延迟
- 机械硬盘的传输时间  
传输时间 = 访问时间 + 数据传输时间



# 机械硬盘的缺点和固态硬盘

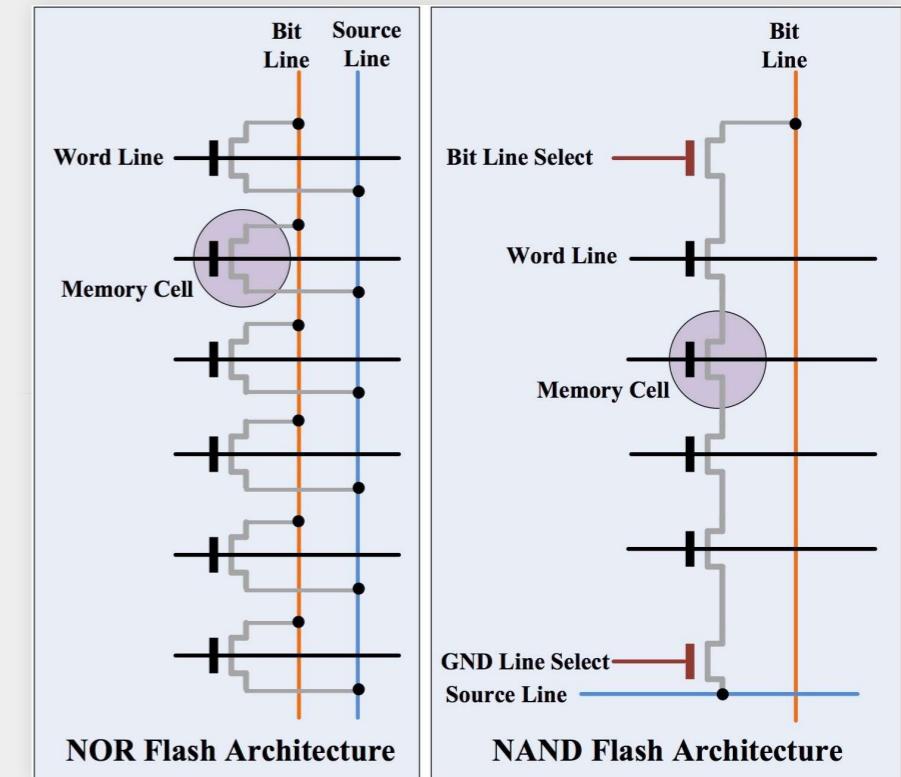
---

- 机械硬盘提供了大容量、可持久的、较为便宜的随机存储
- 但是存在**访问速度相对较慢，抗干扰能力较弱, 机械元件磨损**等问题
- 近年来，随着内存单元价格降低，采用大容量闪存(flash memory)的固态硬盘成为了机械硬盘的替代品

# 固态硬盘

- 采用SATA接口的固态硬盘可以替换现有机械硬盘
- 还有一些固态硬盘采用非易失性内存接口(Non-Volatile Memory Express, NVMe)
- 内部实现采用的是闪存，通常是NAND闪存(存储较大、读写较快)

Feature	NOR Flash		NAND Flash	
	General	S70GL02GT	General	S34ML04G2
Capacity	8MB – 256MB	256MB	256MB – 2GB	256MB
Cost per bit	Higher	$6.57 \times 10^{-9}$ USD/bit for 1ku	Lower	$2.533 \times 10^{-9}$ USD/bit for 1ku
Random Read speed	Faster	120ns	Slower	30μS
Write speed	Slower		Faster	
Erase speed	Slower	520ms	Faster	3.5ms
Power on current	Higher	160mA (max)	Lower	50mA (max)
Standby current	Lower	200μA (max)	Higher	1mA (max)
Bit-flipping	Less common		More common	
Bad blocks while shipping	0%		Up to 2%	
Bad block development	Less frequent		More frequent	
Bad block handling	Not mandatory		Mandatory	
Data Retention	Very high	20 years for 1K program-erase cycles	Lower	10 years (typ)
Program-erase cycles	Lower	100,000	Higher	100,000
Preferred Application	Code storage & execution		Data storage	



资料来源：<https://www.embedded.com/flash-101-nand-flash-vs-nor-flash/>

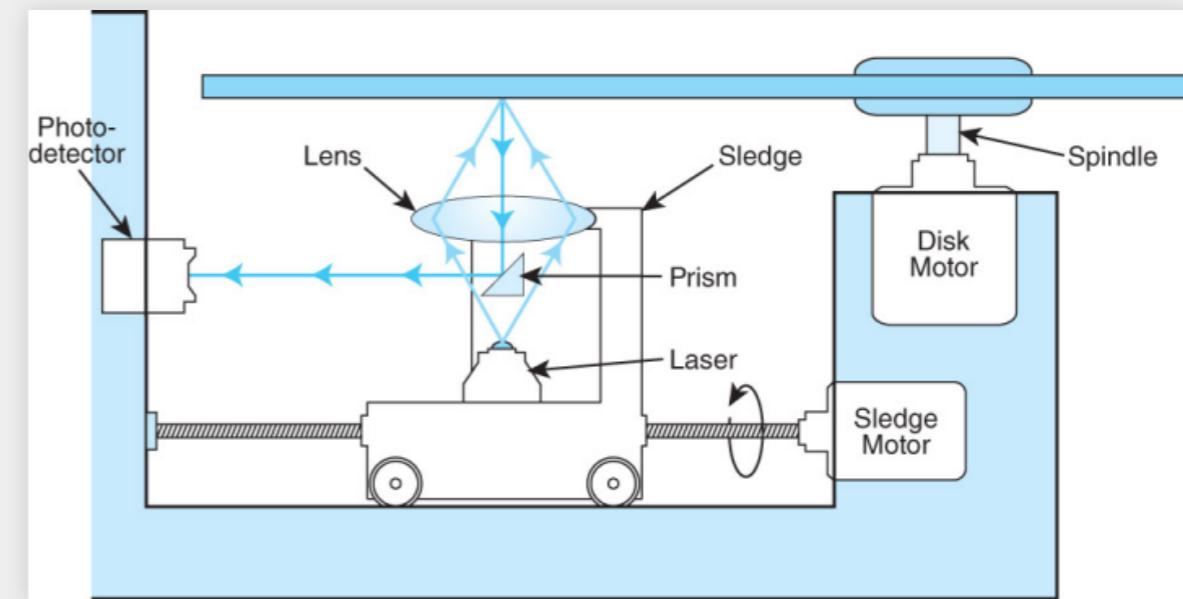
# 光盘

---

- 光盘(Optical Disk)可以提供非常便宜的大容量持久存储
- 光盘的有效存储时间可以达到100年，其它存储介质通常只能保证10年左右的有效时间
- 常见的光盘包含下面几类：CD-ROM、CD-RW、DVD和WORM(Write-Once Read-Many)

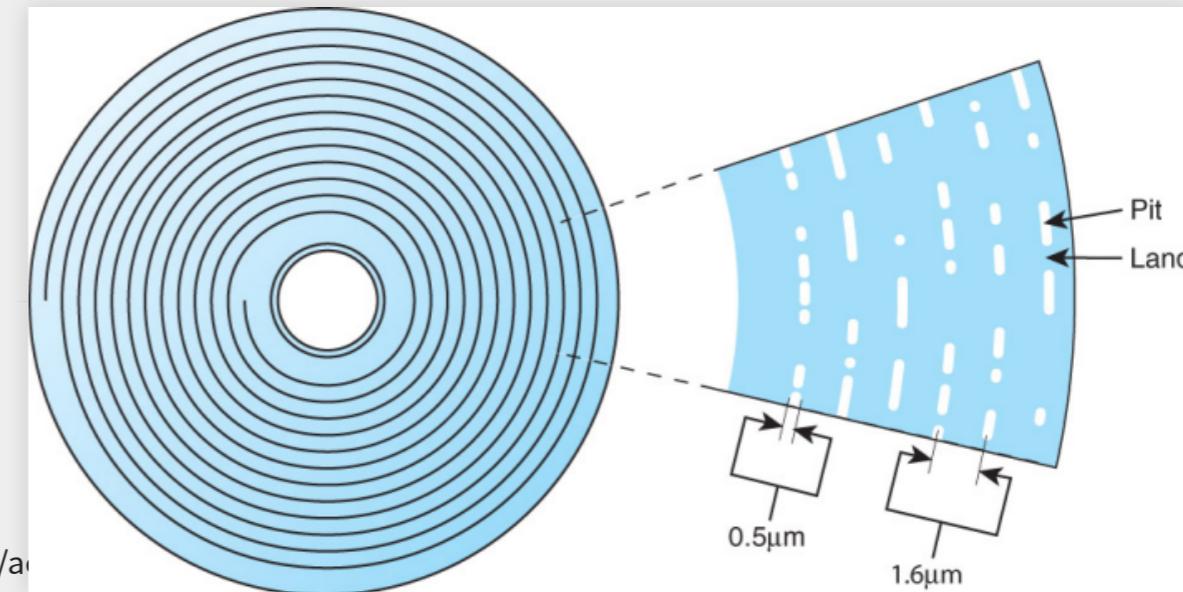
# 光盘的基本结构

- 光盘包含一个材料层和一个反射层
- 材料层从**中心到边缘**形成一条螺旋形的轨道
- 轨道上凹陷的区域称为凹坑(pit)，凹坑之间的区域称为平台(land)
- 凹陷的区域会因为光的干涉亮度变暗，平台会因为光的干涉亮度加强



图片来源：<http://www.ibtimes.co.in/karnataka-family-tries-sacrifice-1-year-old-girl-find-hidden-treasure-742759>

图片来源：<https://tx-land.com/land-for-sale-central-texas-owner-financing-texas-veteran-land-all-affordable-acreage/a>



# 光盘中的数据表示

---

- 光盘并不是用凹坑和平台直接表示0/1
- 采用NRZI(Non-return to Zero Inverted)编码
  - 在一个周期内，如果从凹坑变到平台，解释为1
  - 在一个周期内，如果从平台变到凹坑，解释为1
  - 在一个周期内，如果全是凹坑或者全是平台，解释为0

资料来源：<https://www.explainthatstuff.com/cdplayers.html#binary>

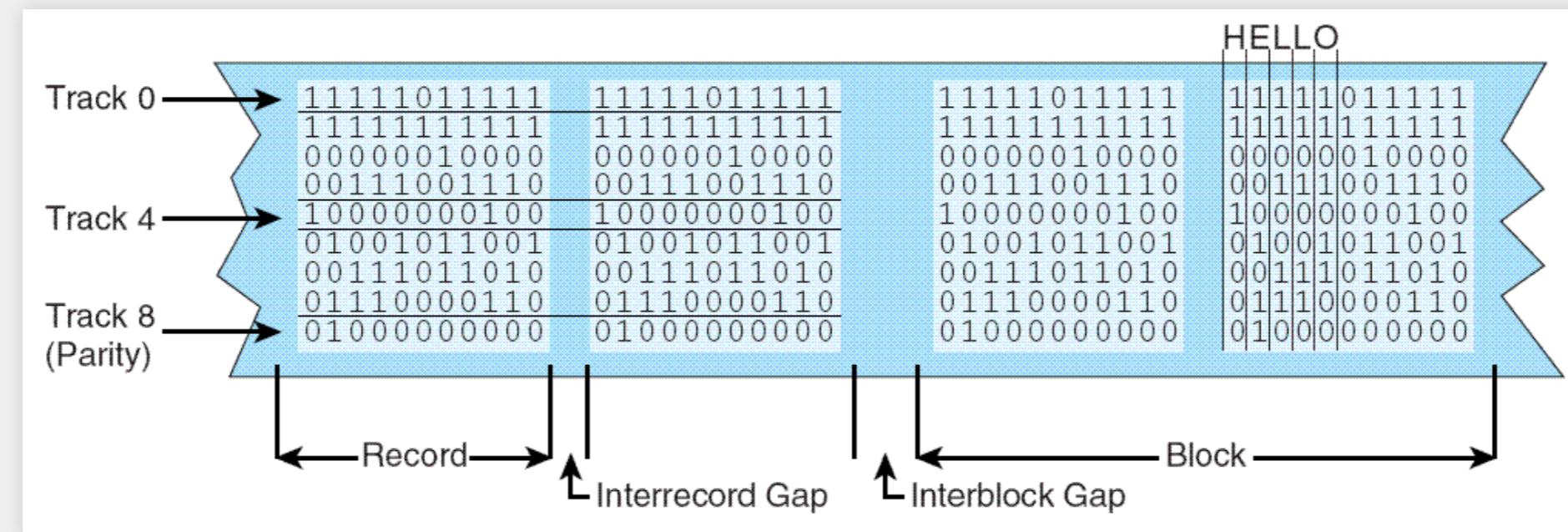
# CD、DVD和蓝光光盘

---

- 光学存储器的原理类似，但是采用的光波波长不同
- CD：750nm，DVD：650nm，蓝光：405nm
- 凹陷的宽度和长度不同、轨道间隔不同，最终导致同样大小的光盘上存储的容量不同
  - CD轨道间距 $1.6\mu m$ ，轨道长度约8km，容量650MB
  - DVD轨道间距 $0.74\mu m$ ，轨道长度约11.8km，容量单层单面4.78GB，双层双面17GB
  - 蓝光光盘单层单面容量25GB

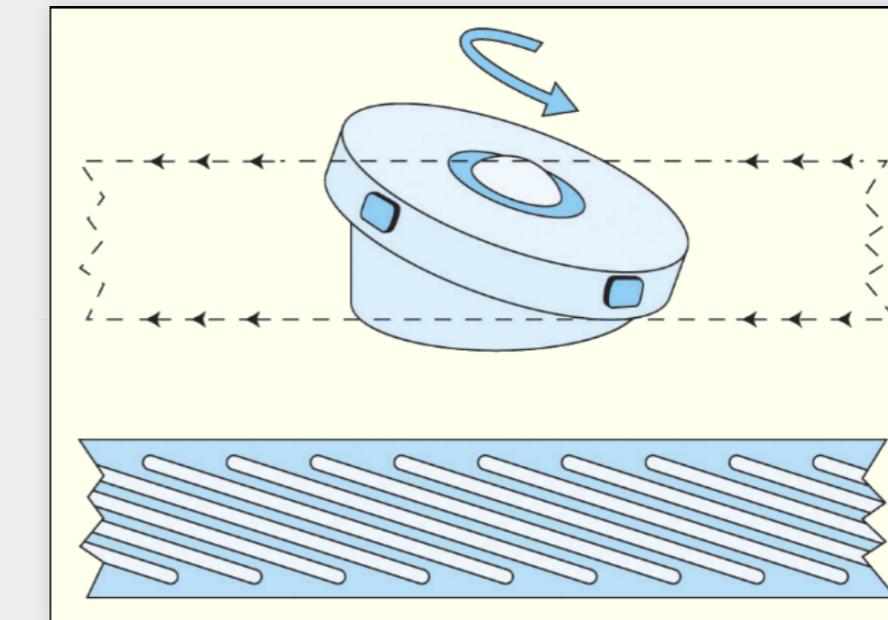
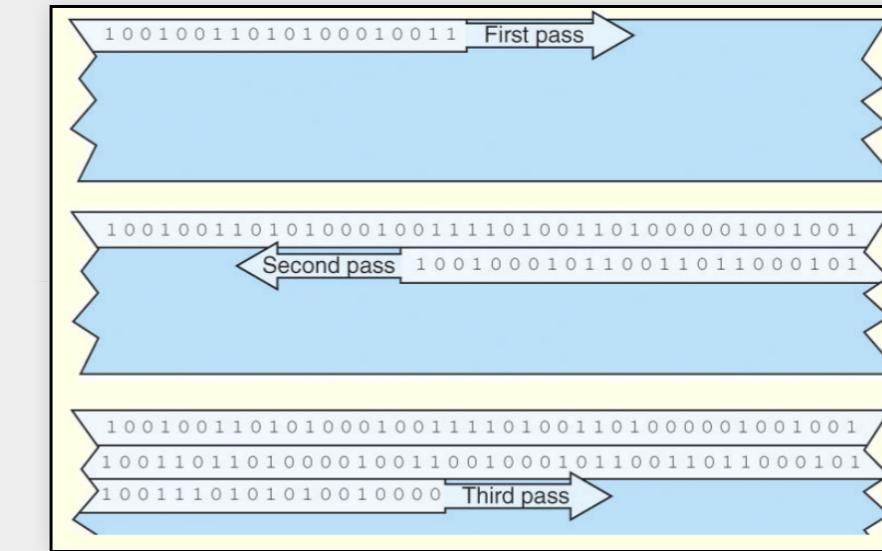
# 磁带

- 磁带提供了大容量、可持久的、较为便宜的顺序存储
- 最早的磁带容量为11MB
  - 磁带按字节写入
  - 每个字节对应9个磁道(track): 8位数据和1位校验位



# 新的磁带记录方法

- 蛇形记录(Serpentine)
- 螺旋扫描(Helical Scan)



# 独立磁盘冗余阵列

- 磁盘存储介质的数据可靠性受到环境的影响
- 发生数据错误的代价很大
- 两种解决思路
  - 大型、昂贵的单磁盘
  - 用多个便宜的普通磁盘组成可靠的存储系统

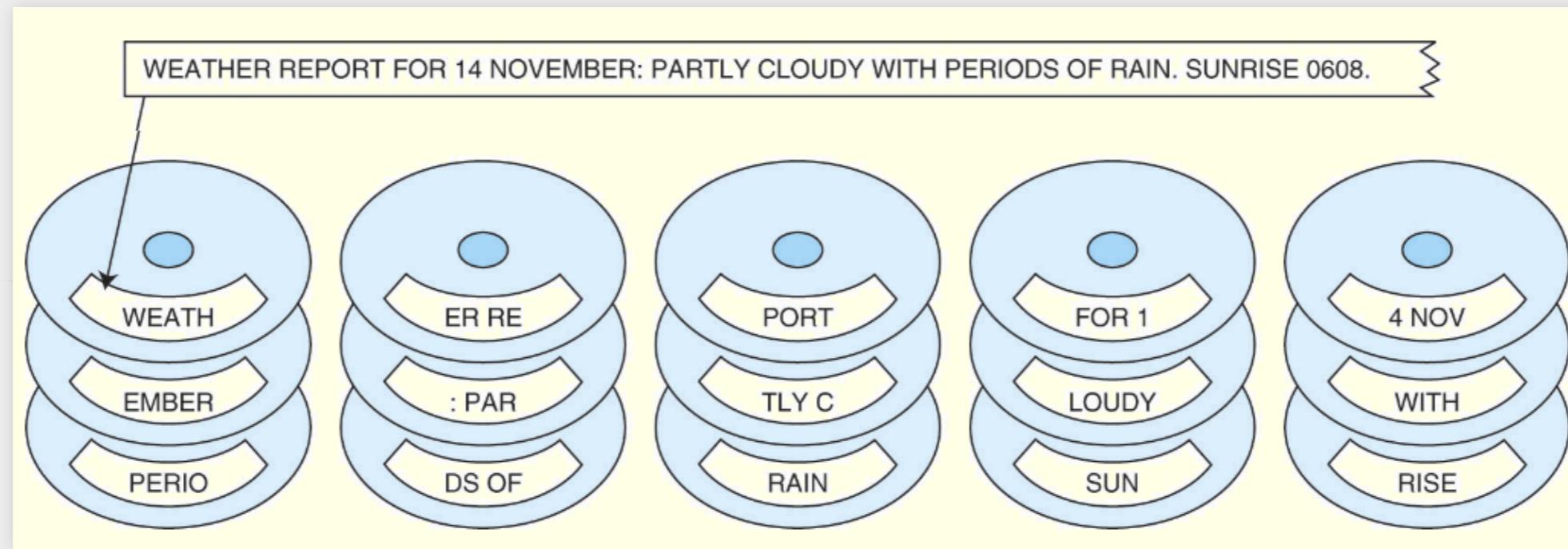
## A Case for Redundant Arrays of Inexpensive Disks (RAID)

*David A. Patterson, Garth Gibson, and Randy H. Katz*

Computer Science Division  
Department of Electrical Engineering and Computer Sciences  
571 Evans Hall  
University of California  
Berkeley, CA 94720  
(pattrsn@ginger berkeley.edu)

# RAID-0

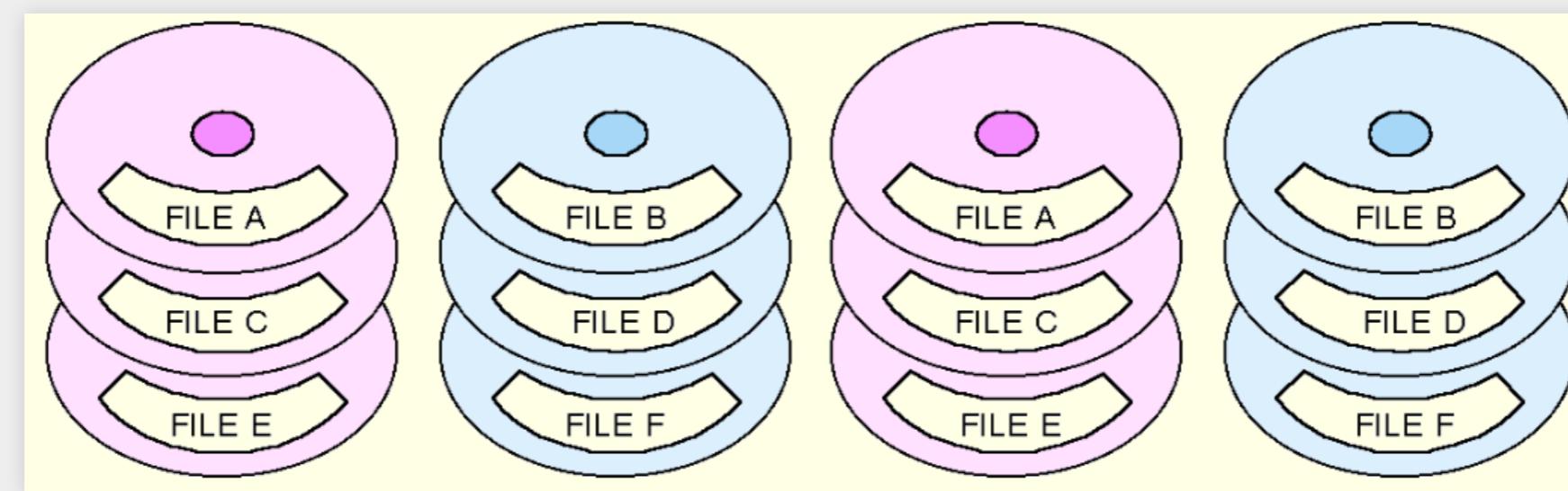
- RAID-0也称为**磁盘跨区**(Disk Spanning)
- RAID-0**没有冗余备份，不提高可靠性，但是可以提高性能**
  - 类比内存的低位交叉：并行读写



# RAID-1

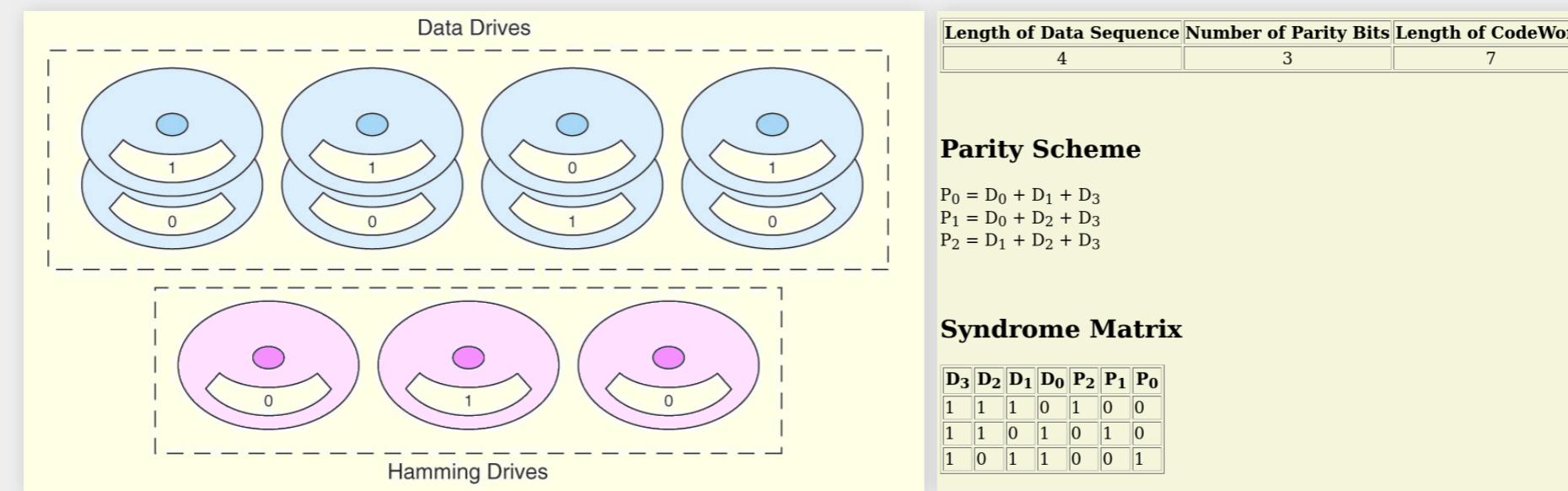
---

- RAID-1也称为**磁盘镜像**(Disk Mirroring)
- 每份数据保存在两个独立磁盘上
- RAID-1需要了100%的冗余存储



# RAID-2

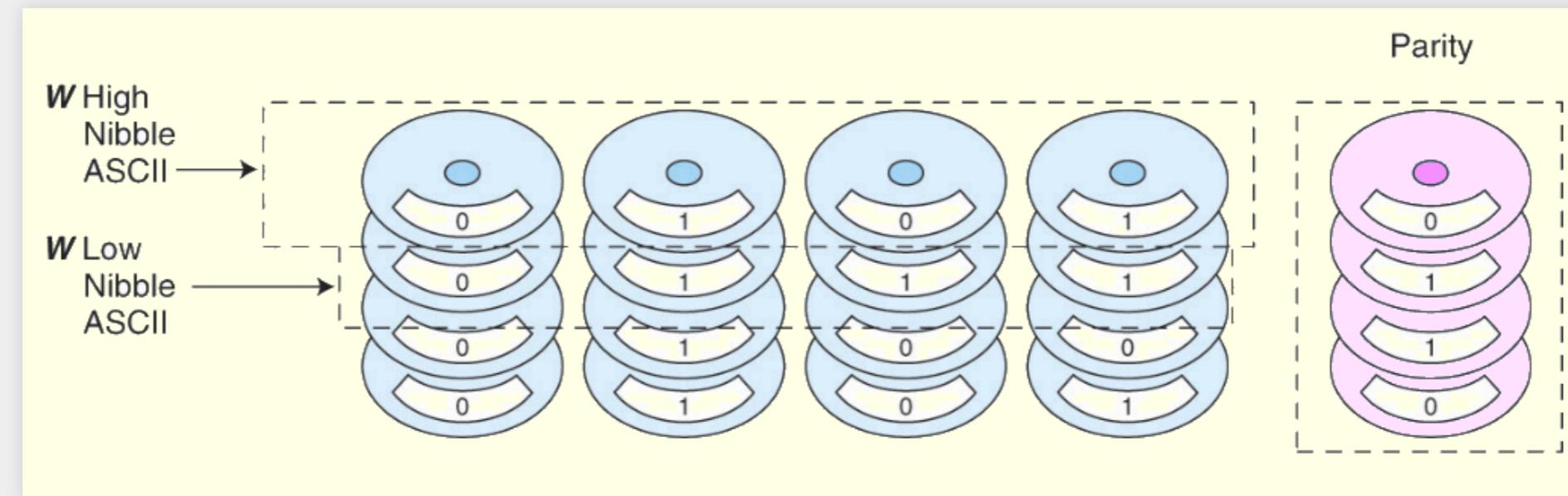
- RAID-2是一个理论上的RAID组成方式
- 将硬盘分为两个集合：数据驱动器和校验驱动器
- 每个数据驱动器写入1位，校验驱动器采用汉明码校验
  - 下面的例子中采用了(7, 4)汉明码



图片来源:<http://www.ecs.umass.edu/ece/koren/FaultTolerantSystems/simulator/Hamming/HammingCodes.html>

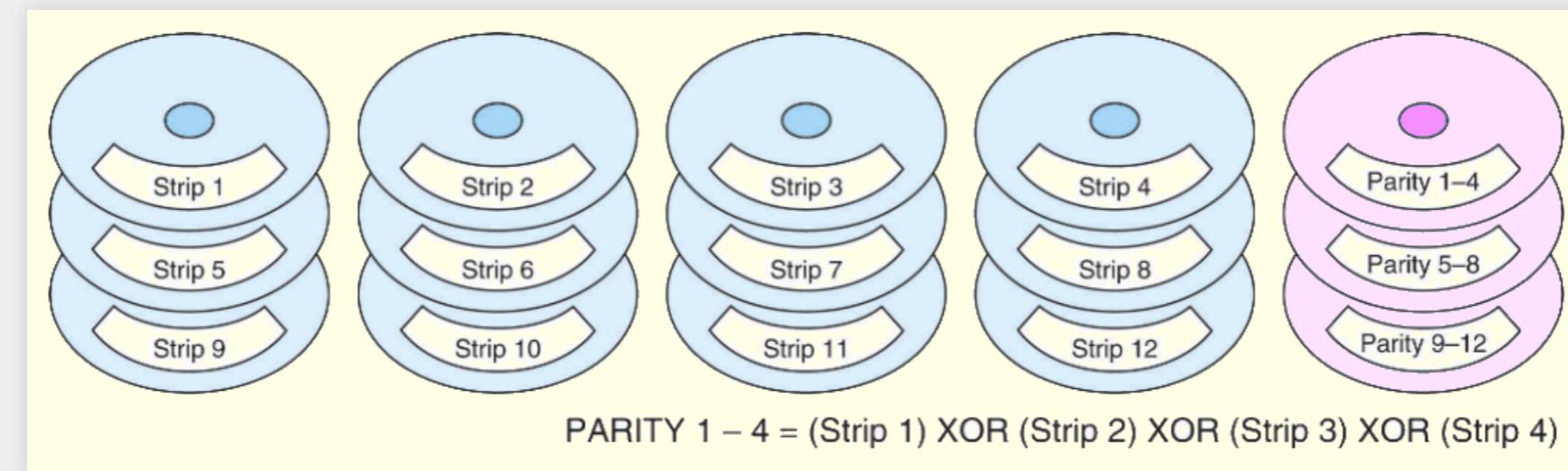
# RAID-3

- RAID-3也分为数据驱动器和校验驱动器
- 每个数据驱动器写入1位，只需要一个校验驱动器做奇偶校验(parity check)
- $\text{parity} = b_0 \text{ XOR } b_1 \text{ XOR } b_2 \text{ XOR } b_3$
- 替换掉数据驱动器后可以恢复数据：替换0号数据驱动器，则 $b_0 = b_1 \text{ XOR } b_2 \text{ XOR } b_3 \text{ XOR } \text{parity}$



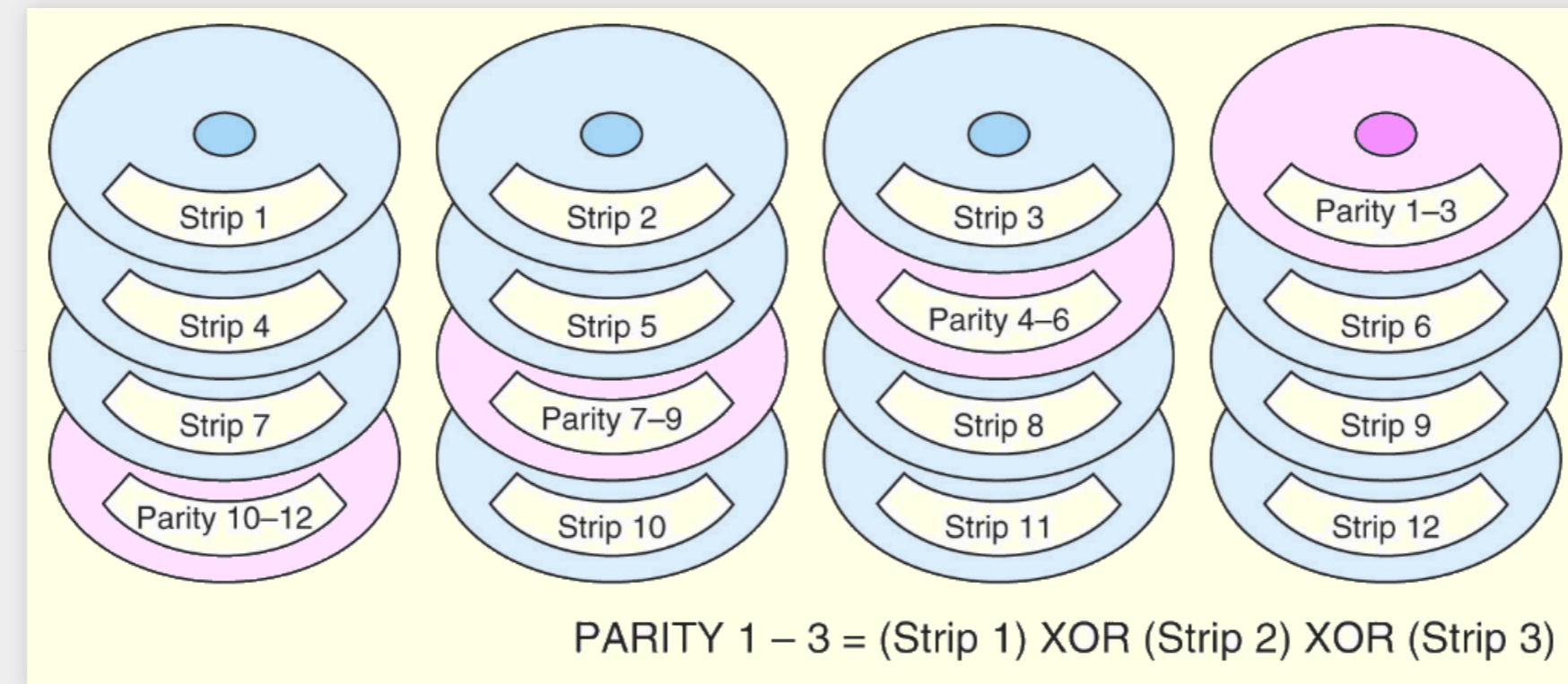
# RAID-4

- RAID-4也是一个理论上的RAID组合方式
- 对数据条带做奇偶校验
- 并行写入会导致奇偶校验驱动器的竞争，从而造成性能瓶颈



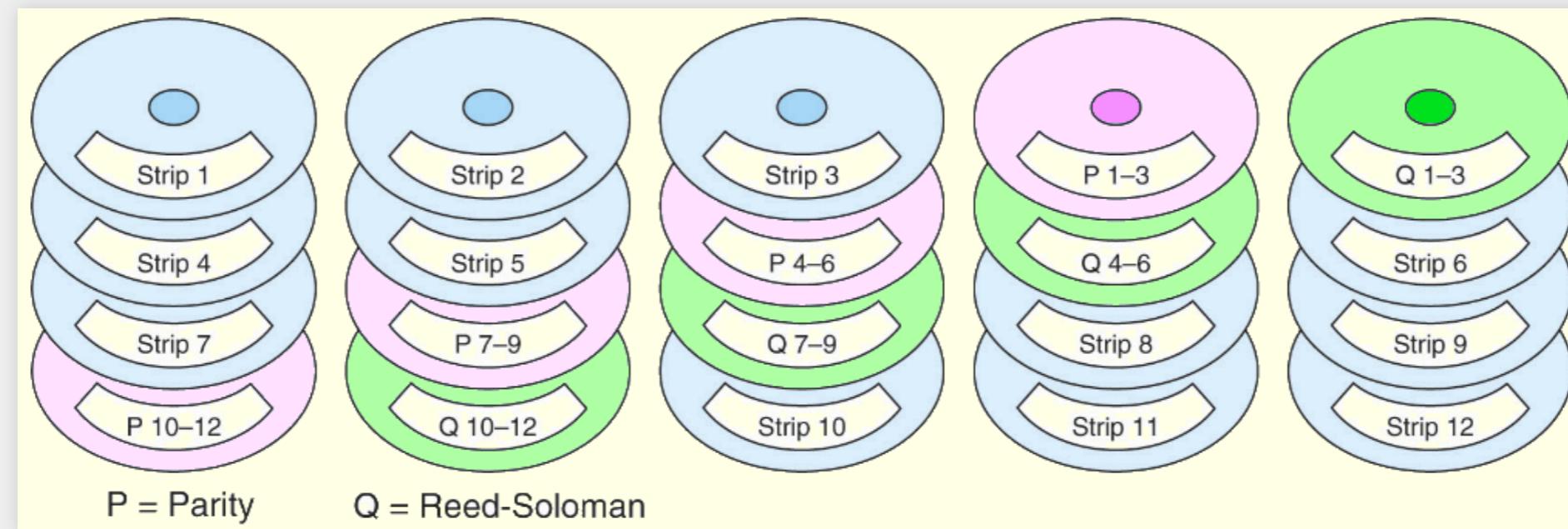
# RAID-5

- RAID-5的思路是避免奇偶校验驱动器的竞争：不要将奇偶校验放在同一个驱动器上
- RAID-5是应用最广泛的RAID系统



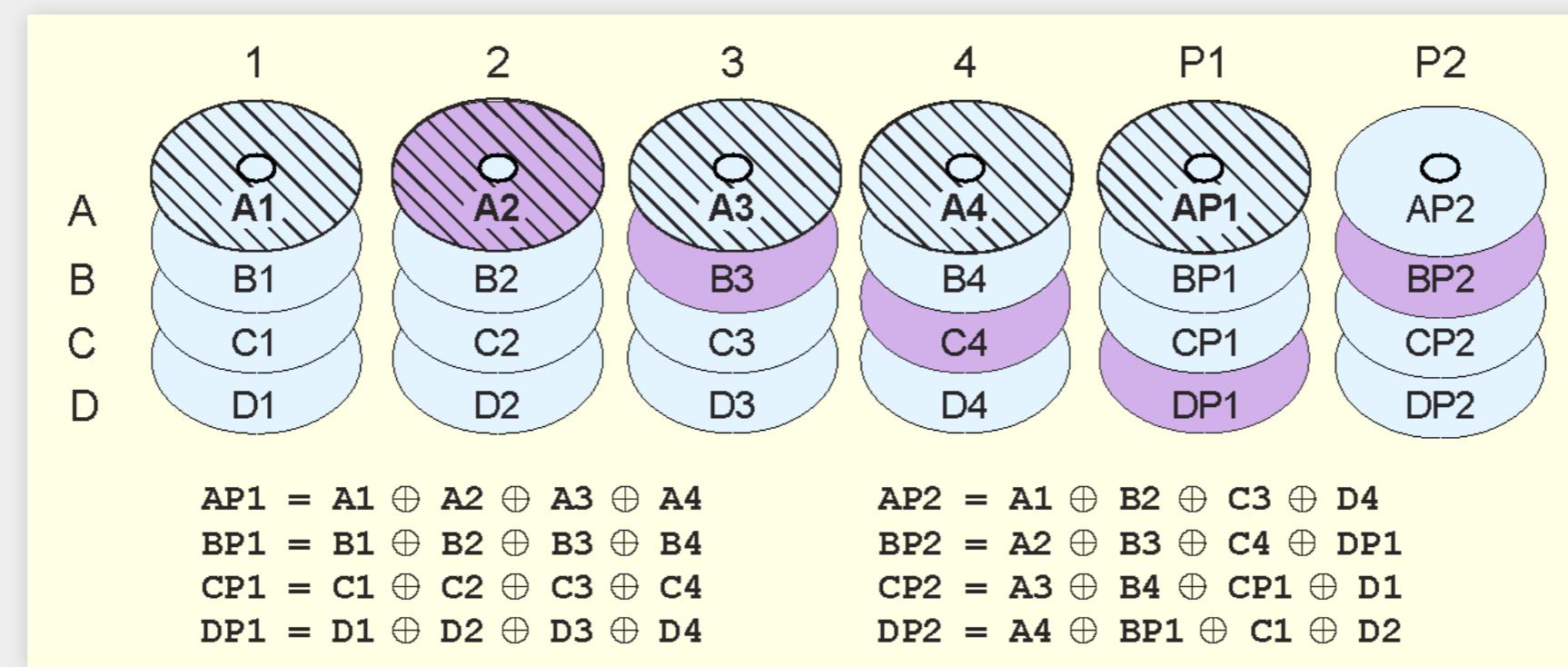
# RAID-6

- 上述的RAID系统只能处理最多一个磁盘出错的情况
- 实际中可能存在多个硬盘失效的情况：统一批次采购的设备、运行环境影响造成硬盘错误等
- RAID-6**支持最多2个硬盘错误**：采用两个校验码P和Q
  - P是奇偶校验码，Q是里德-所罗门校验码



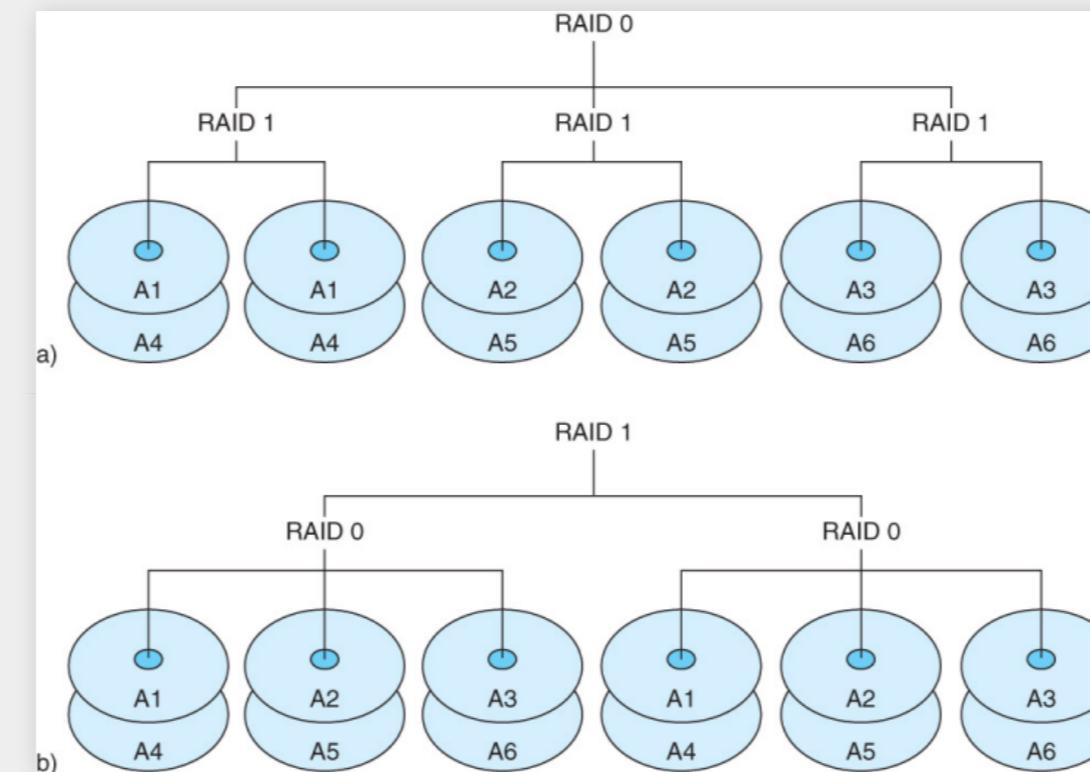
# RAID-DP

- RAID-DP可以恢复最多两个硬盘错误
- RAID-DP采用两个奇偶校验码
  - CP1是数据驱动器的水平奇偶校验码
  - CP2是数据驱动器和CP1的对角奇偶校验码



# 混合RAID系统

- 采用多级RAID，每一级RAID采用不同的RAID类型
- 命名法则：从普通硬盘开始，按照RAID组合方式命名
  - 先组成RAID-1,然后组成RAID-0,命名为RAID-10
  - 先组成RAID-0,然后组成RAID-1,命名为RAID-01



# 分布式存储

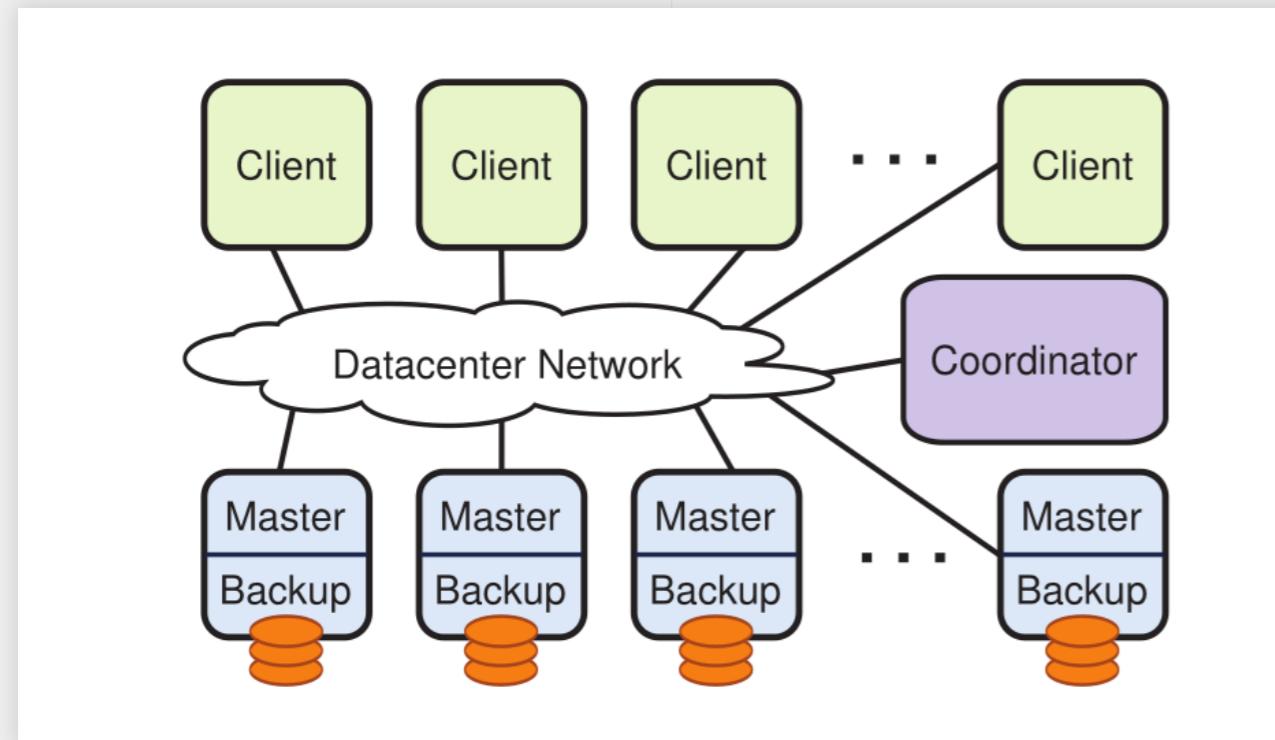
---

- RAID只能保证一台计算机上面的数据可靠性
- 现代计算模式导致计算机的集中度更高（例如数据中心），可能出现整机、机柜甚至整个房间出现故障的情况
- 分布式高可靠存储系统采用了类似RAID的方法，但是采用多台计算机上的硬盘进行数据备份和容灾

# 分布式存储系统中的容灾

采用数据备份：GFS、HDFS、RAMCloud等

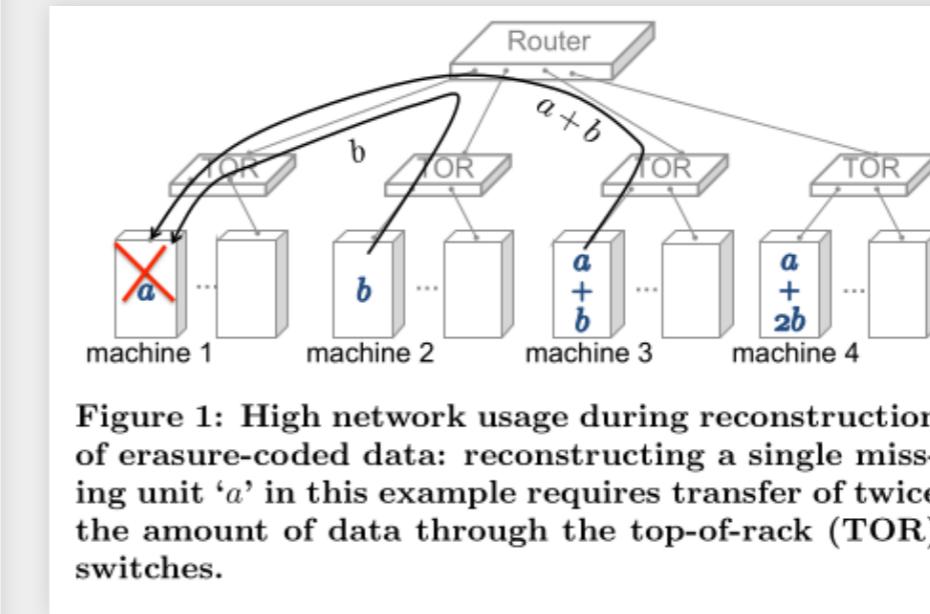
- 占用存储较多
- 恢复数据时网络传输量少



RAMCloud

采用Erasure Code（例如里德-所罗门编码，RAID-6中使用的编码）

- 占用存储较少
- 恢复数据时网络传输量大



# 第十讲结束

# 本期内容总结

---

- 持久存储介质
  - 硬盘、固态硬盘
  - 光盘
  - 磁带
- 独立磁盘冗余阵列
- 分布式存储简介



# Q & A