10 Mass Storage System

- Overview
  - Secondary Storage
    - Save data permanently.
      - 永久保存数据。
    - Slower than memory.
      - 比内存慢
    - Cheaper and greater than memory.
      - 比内存更便宜，更强大。
    - Types of Secondary Storages
      - Sequential access devices - store records sequentially, one afterthe other
        - 顺序存取设备-按顺序存储记录，一个接一个
        - Relatively permanent and holds large quantities of data
          - 相对永久且保存大量数据
        - Access time slow
          - 访问时间慢
      - Direct access devices - store data in discrete and separate location with a unique address.
        - 直接存取装置-用唯一的地址在离散和分离的位置存储数据。
        - Nonvolatile memory used like a hard drive
          - 像硬盘一样使用的非易失性存储器
        - Less capacity but much faster than HDDs
          - 容量较小，但比hdd快得多
  - Moving-head Disk Mechanism移动头磁盘机构
    -

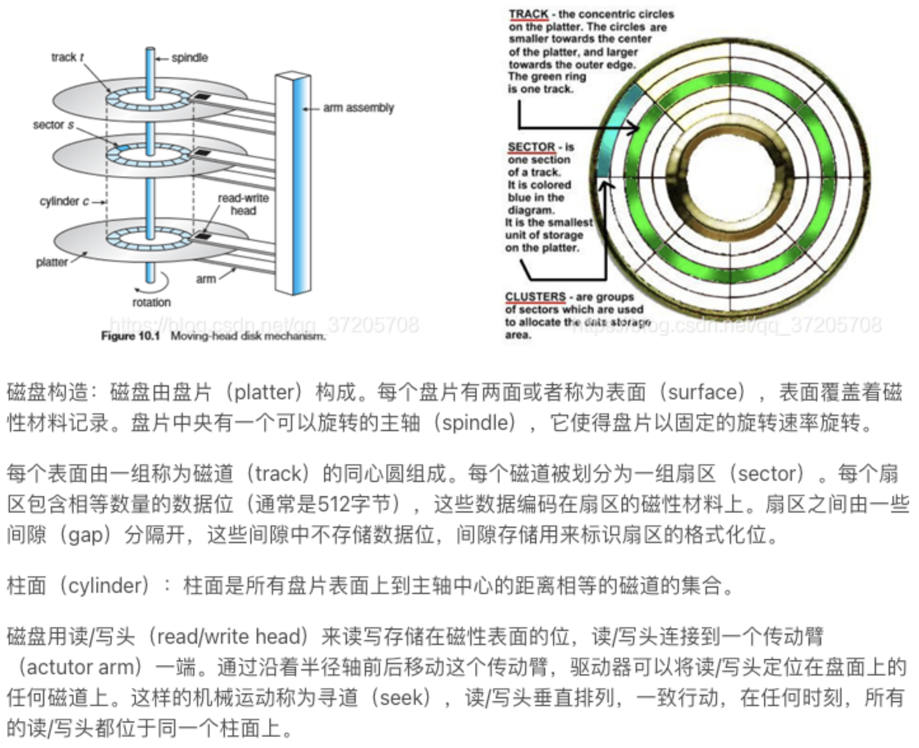
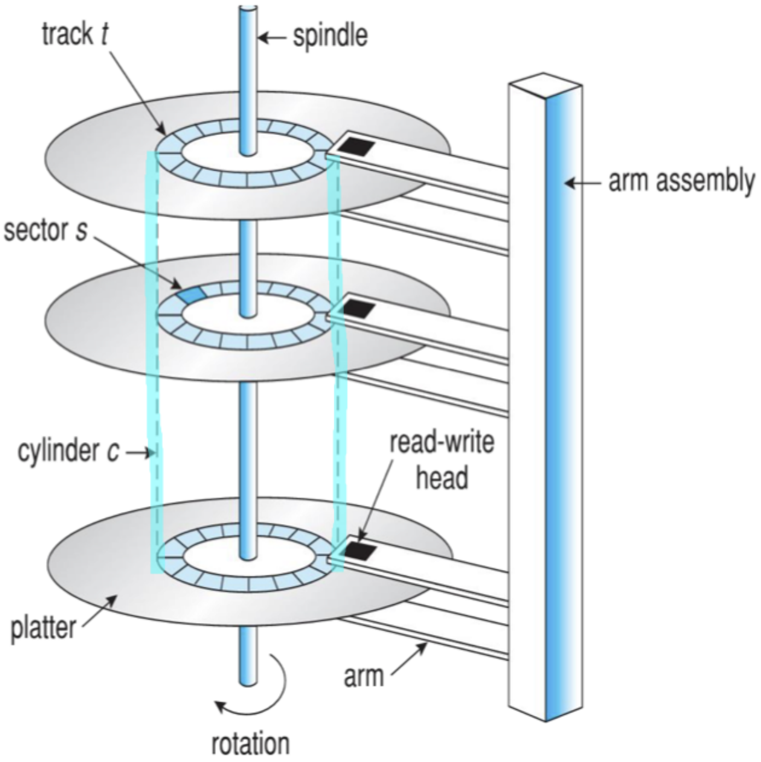

track t
spindle
arm assembly
sector s
cylinder c
read-write head
platter
arm
rotation
TRACK - the concentric circles on the platter. The circles are smaller towards the center of the platter, and larger towards the outer edge. The green ring is one track.
SECTOR - is one section of a track. It is colored blue in the diagram. It is the smallest unit of storage on the platter.
CLUSTERS - are groups of sectors which are used to allocate the data storage area.

Figure 10.1 Moving-head disk mechanism.

> 磁盘构造：磁盘由盘片（platter）构成。每个盘片有两面或者称为表面（surface），表面覆盖着磁性材料记录。盘片中央有一个可以旋转的主轴（spindle），它使得盘片以固定的旋转速率旋转。
>
> 每个表面由一组称为磁道（track）的同心圆组成。每个磁道被划分为一组扇区（sector）。每个扇区包含相等数量的数据位（通常是512字节），这些数据编码在扇区的磁性材料上。扇区之间由一些间隙（gap）分隔开，这些间隙中不存储数据位，间隙存储用来标识扇区的格式化位。
>
> 柱面（cylinder）：柱面是所有盘片表面上到主轴中心的距离相等的磁道的集合。
>
> 磁盘用读/写头（read/write head）来读写存储在磁性表面的位，读/写头连接到一个传动臂（actutor arm）一端。通过沿着半径轴前后移动这个传动臂，驱动器可以将读/写头定位在盘面上的任何磁道上。这样的机械运动称为寻道（seek），读/写头垂直排列，一致行动，在任何时刻，所有的读/写头都位于同一个柱面上。

- Group of tracks is called a cylinder
  - 一组轨道称为圆柱体
- Aluminum platters with magnetic coating.
  - 带有磁性涂层的铝盘。
- A stack of 16 platters is about the maximum will find in modern drives.
  - 16个盘片的堆栈是现代驱动器中最大的。
- A track is logically divided into sectors.
  - 磁道在逻辑上被划分为扇区。
- The sectors are the smallest unit of data that a disk drive will transfer.
  - 扇区是磁盘驱动器将传输的最小数据单位。
- Disk address can be specified by the cylinder, head and sector numbers, or CHS addressing.
  - 磁盘地址可以通过柱体、磁头和扇区号或CHS寻址来指定。
- A disk with C cylinders, H heads, and S sectors per track has C x H x S sectors in all一个有C个圆柱、H个磁头和S个扇区的磁盘，每个磁道总共有C × H × S个扇区



- Disk speed
    - Transfer time = the time for data transfer / the time between the start of the transfer and the completion of the transfer.
        - 传输时间=数据传输的时间/传输开始到传输完成的时间。
        - For example, if it needs to read some data from the disk, then the transfer time is the time taken to transfer the data from the disk to the application.
            - 例如，如果它需要从磁盘读取一些数据，那么传输时间就是将数据从磁盘传输到应用程序所花费的时间。
    - Seek time = the time taken by the disk head to move from one cylinder to another / the time it will take to reach a track.
        - 寻道时间=磁头从一个圆柱体移动到另一个圆柱体所花费的时间/到达磁道所花费的时间。
    - Rotational latency = the time taken to rotate the platter and bring the required disk sector under the read-write head.
        - 旋转延迟=旋转盘片并将所需的磁盘扇区置于读写头下所花费的时间。
    - Positioning time / Random access time = seek time + rotational latency
        - 定位时间/随机访问时间=寻道时间+旋转延迟
    - Disk Access Time = the time to perform any operation on the disk.
        - 磁盘访问时间=对磁盘执行任何操作的时间。
        - seek time + rotational latency + transfer time.
            - 寻道时间+旋转延迟+传输时间。
    -

寻道时间：为了读取某个目标扇区的内容，传动臂首先将读/写头定位到包含目标扇区的磁道上。移动传动臂所需要的时间称为寻道时间。T_{seek}依赖于读/写头以前的位置和传动臂在盘面上移动的速度。现代驱动器中平均寻道时间$T_{avg\ seek}$是通过几千次对随机扇区的寻道求平均值来测量的，通常为3~9ms。

旋转时间：一旦读/写头定位到了期望的磁道，驱动器等待目标扇区的第一个位旋转到读/写头下。这个步骤的性能依赖于读/写头到达目标扇区时盘面的位置和磁盘的旋转速度。在最坏的情况下，读/写头刚刚错过了目标扇区，必须等待磁盘转一整圈。因此，最大旋转延迟（以秒为单位）是

$$T_{max\ rotation} = \frac{1}{RPM} \times \frac{60\,secs}{1\,min} \quad \text{（RPM:转每分钟）}$$

平均旋转时间就是max的一半

传送时间：当目标扇区的第一个位位于读/写头下时，驱动器就可以开始读或者写该扇区的内容了。一个扇区的传送时间依赖于旋转速度和每条磁道的扇区数目。因此，可以粗略地估计一个扇区以秒为单位的平均传送时间如下

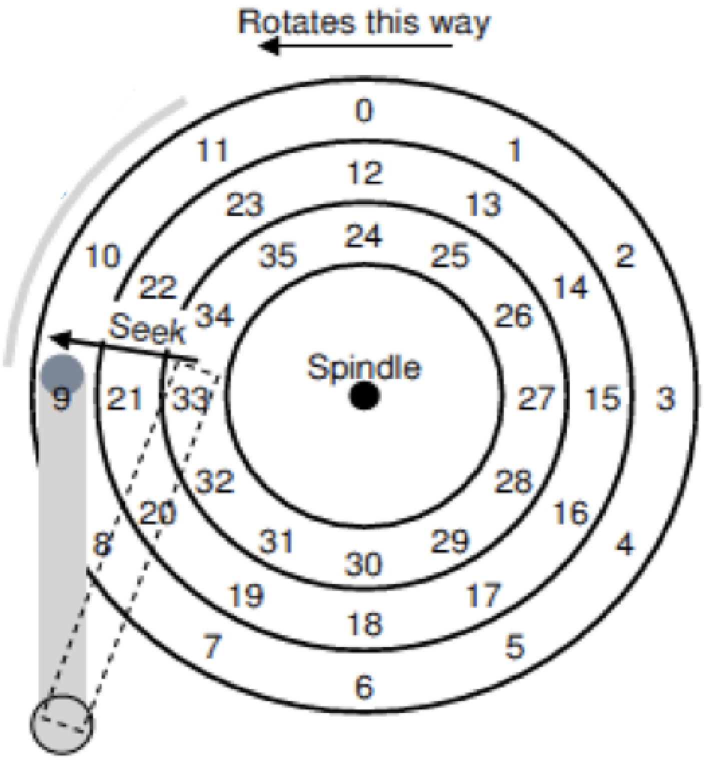
$$T_{avg\ transfer} = \frac{1}{RPM} \times \frac{60\,secs}{1\,min} \times \frac{1}{(\text{平均扇区数}/\text{磁道})},$$

该磁道整个扇区全部传送完就是转一圈的时间，所以传送1个扇区所需的时间就估计为转1圈的时间除以扇区数。

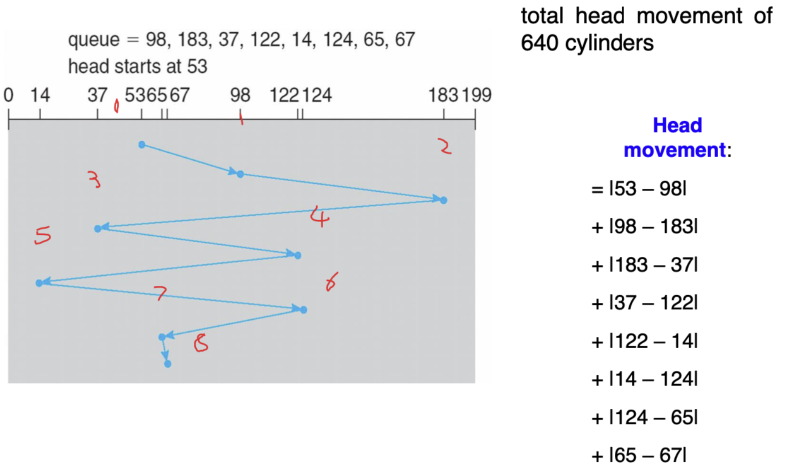
可以得出如下结论：

访问一个磁盘扇区中512字节的时间主要是寻道时间和旋转延迟。访问扇区的第一个字节用了很长时间，但是访问剩下的字节几乎不用时间。

- Disk Structure
  - Disk is addressed as a one-dimension array of logical sectors
    - 磁盘被寻址为逻辑扇区的一维阵列
    - Logical sector 0: the first sector of the first track of the first surface.
    - 逻辑扇区0:第一个表面的第一个磁道的第一个扇区。
  - Disk controller maps logical sector to physical sector identified by track #, surface # and sector #
    - 磁盘控制器将逻辑扇区映射到由磁道#、表面#和扇区#标识的物理扇区
  - 三轨+一个头(带Seek)



- Disk Attachment
  - Computer systems can access disk storage in two ways
    - 计算机系统可以通过两种方式访问磁盘存储
  - :➢ via I/O ports and this is common on small systems:
    - 通过I/O端口，这在小型系统上很常见:
    - ➢Host-attached storage. The most common interfaces areIntegrated Drive Electronics IDE, Advanced TechnologyAttachment ATA, USB each of which allow up to two drives per host controller.
      - ➢Host-attached存储。最常见的接口是集成驱动电子IDE，先进技术附件ATA, USB每个允许多达两个驱动器每个主机控制器。
  - ➢ via a remote host in a distributed file systems:
    - ◆通过远程主机在分布式文件系统中实现:
    - ➢Storage Area Network: fibre channels the most common interconnect, and InfiniBand (high speed connection)
      - (四)存储区域网络:光纤通道最常用的互联方式，InfiniBand(高速连接)
    - ➢ Network-Attached Storage: connection over TCP/IP,UDP/IP or host attached protocol like ISCSI)
      - (网络附加存储:通过TCP/IP,UDP/IP或主机附加协议(如ISCSI)连接)
- Disk Scheduling
  - Goal
    - minimize the positioning time
      - 最小化定位时间
    - scheduling is performed by both O.S. and disk itself

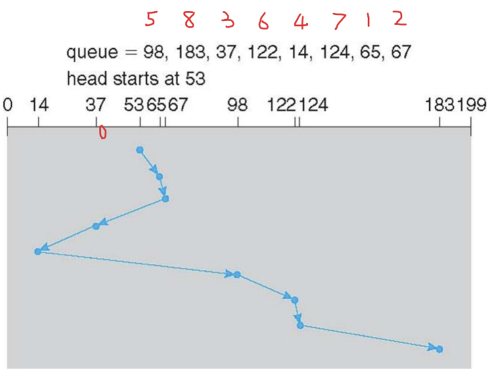
        - 调度由操作系统和磁盘本身执行
    - • O.S. can control: – Sequence of workload requests
        - •操作系统可以控制:-工作负载请求的顺序
    - • Disk knows: – Geometry, accurate positioning times (the time to move the hard disk arm to desired cylinder (seek time) and time fordesired sector to rotate under the disk head (rotational latency)).
        - •磁盘知道:-几何，精确的定位时间(将硬盘臂移动到所需圆柱体的时间(寻道时间)和所需扇区在磁盘磁头下旋转的时间(旋转延迟))。
- 前提
    - consider a disk queue with I/O requests on the following cylinders in their arriving order:
        - 考虑一个磁盘队列，其中I/O请求按到达顺序排列在以下柱面上:
    - 98, 183, 37, 122, 14, 124, 65, 67
    - The disk head is assumed to be at cylinder 53.The disk consists of total 200 cylinders (0-199) .
        - 磁盘磁头假定在53号气缸上。磁盘由总共200个气缸(0-199)组成。
    - to
        - Minimize arm movement
            - 尽量减少手臂活动
        - Minimize mean response time
            - 最小化平均响应时间
- FCFS
    - 

total head movement of 640 cylinders

Head movement:

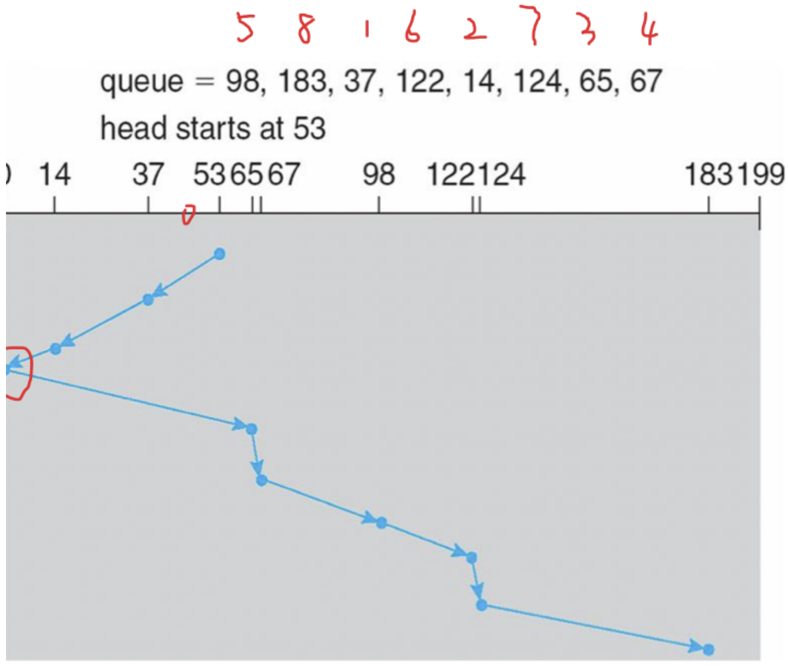
= |53 – 98|
+ |98 – 183|
+ |183 – 37|
+ |37 – 122|
+ |122 – 14|
+ |14 – 124|
+ |124 – 65|
+ |65 – 67|

- SSTF
    -

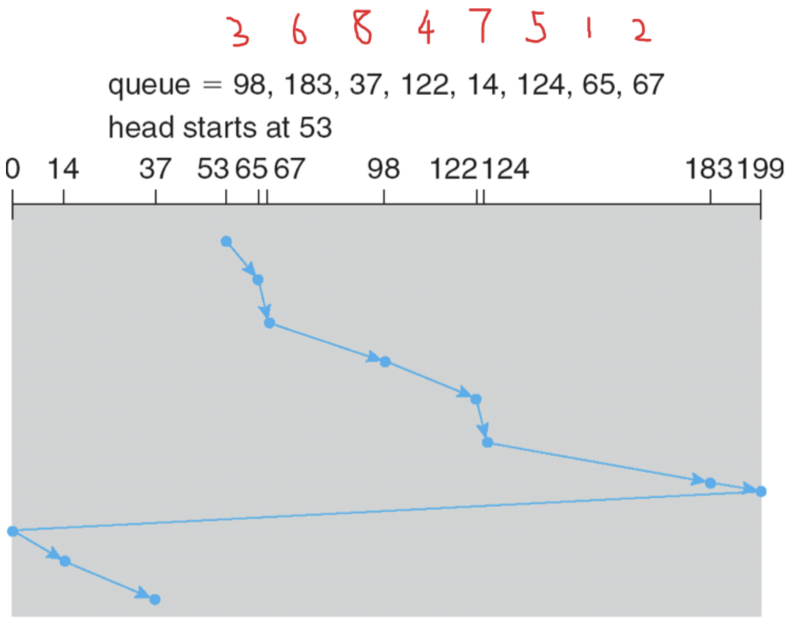


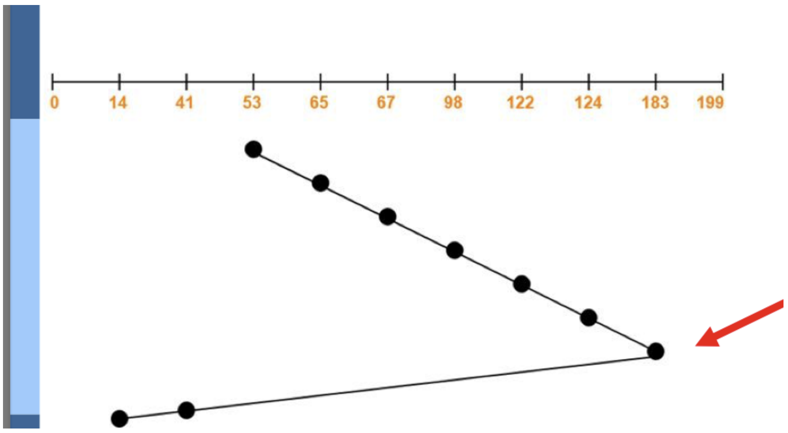
- selects the request with the minimum seek time (**the next shortest distance**) from the current head position.
- **total head movement of 236 cylinders**

- SCAN Elevator
  - The disk arm starts at one end of the disk, and moves toward the other end, servicing requests until it gets to the other end of the disk, where the head movement is reversed, and servicing continues.磁盘臂从磁盘的一端开始，向另一端移动，为请求提供服务，直到它到达磁盘的另一端，在那里磁头的运动反向，服务继续。
  - 该算法需要知道磁头的当前位置及移动方向
  - 各请求等待时间往往不平衡
  -



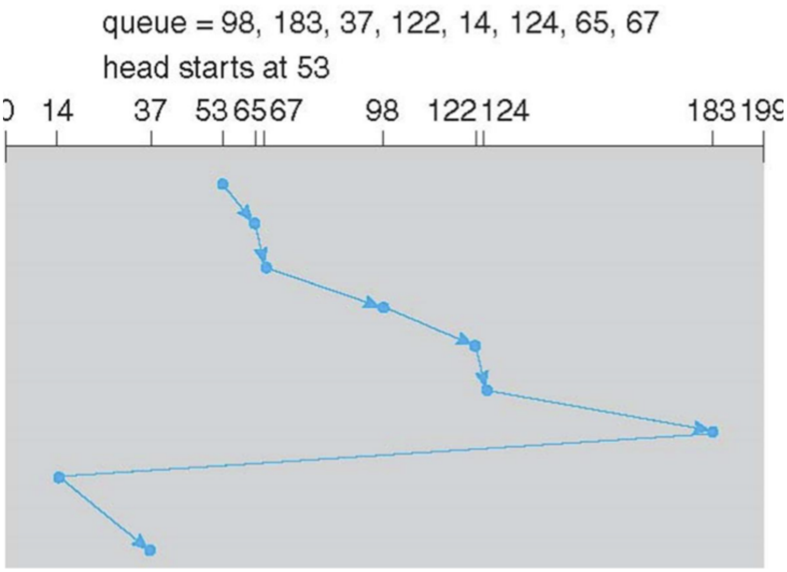
- Circular-SCAN / C-SCAN
  - The head moves from one end of the disk to the other, servicing requests as it goes.
    - 磁头从磁盘的一端移动到另一端，一边移动一边处理请求。
  - When it reaches the other end, however, it immediately returns to the beginning of the disk, without servicing any requests on the return trip.
    - 但是，当它到达另一端时，它立即返回到磁盘的起点，在返回过程中不处理任何请求。
    - 磁头通常会先移动到最大柱面号处，然后再向最小柱面号移动
  -



- LOOK
    - It is a variation of SCAN.
        - 它是SCAN的一种变体。
    - The disk head goes as far as the last request and reverses its direction.
        - 磁头移动到最后一个请求的位置并反转其方向。
        - 

- C-LOOK
    - a version of C-SCAN
        - C-SCAN的一个版本
    - ▪ Arm only goes as far as the last request in each direction, then reverses direction immediately, without first going all the way to the end of the disk
        - ▪Arm只执行每个方向上的最后一个请求，然后立即反转方向，而不会首先执行到磁盘的末端
    - ▪ the head moves till the last request instead of last cylinder
        - ▪头移动到最后一个请求，而不是最后一个气缸
    - C-LOOK that treats the request queue as circular.
        - 将请求队列视为循环的C-LOOK

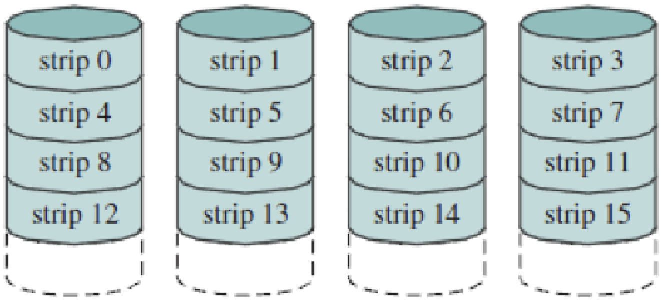
        - 与look区别,look是向反向走,而c-look是到0最近的点继续

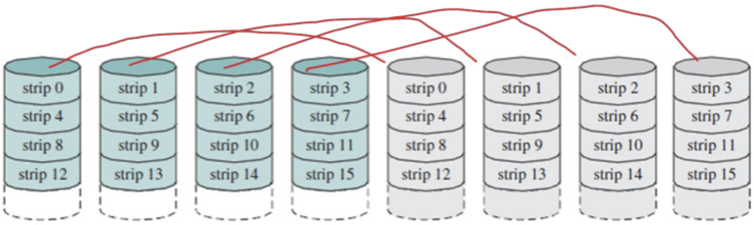


    - eve
        - FCFS works well with light loads; but as soon as the load grows, service time becomes unacceptably long.
        - • SSTF is quite popular and intuitively appealing. It works well with moderate loads but has the problem of localization under heavy loads.
        - • SCAN works well with light to moderate loads and eliminates the problem of indefinite postponement. SCAN is similar to SSTF in throughput and mean service times.
        - • C-SCAN works well with moderate to heavy loads and has a very small variance in service times.
        - FCFS在轻负荷下工作良好;但是，一旦负载增加，服务时间就会变得长得令人无法接受。
        - •SSTF很受欢迎，直观上很吸引人。它在中等负载下工作良好，但在重载下存在局部化问题。
        - •SCAN在轻到中等负载下工作良好，消除了无限期延迟的问题。SCAN在吞吐量和平均服务时间上与SSTF相似。
        - •C-SCAN在中等到重型负载下工作良好，服务时间变化很小。
- Disk Management
    - Low-level formatting, or physical formatting — create sectors on a blank platter
        - 低级格式化或物理格式化-在空白盘片上创建扇区,使用特殊结构填充
        - □Each sector can hold header information, plus data, plus error correction code (ECC)
        - □ Usually, 512 bytes of data but can be selectable
            - □每个扇区可以保存报头信息，加上数据，加上纠错码(ECC)
            - □通常为512字节，但也可以选择
            - 写入时,ecc根据数据区域所有的字节而计算新值更新

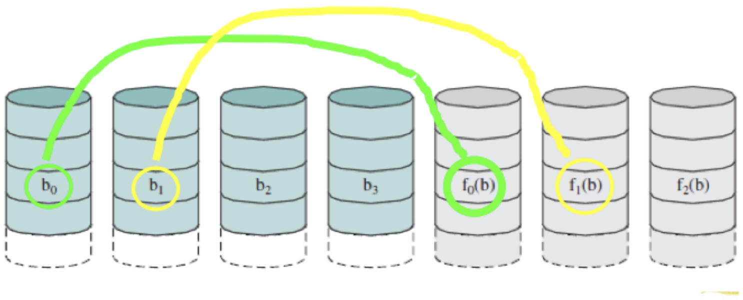
    - Partition organize disk in one or more groups of cylinders
        - 在可以使用磁盘存储文件之前,分区将磁盘组织成一组或多组圆柱体
    - Logical formatting write file system data structures
        - 逻辑格式化写入文件系统的数据结构
    - Boot block initializes system
        - 为了运行计算机,比如打开或者重启,需要初始程序
        - 引导块初始化系统
        - The bootstrap is stored in ROM (Read Only Memory)
        - Bootstrap loader program stored in boot blocks of boot partition
            - 引导存储在ROM(只读存储器)中。只读
            - 引导加载程序存储在引导分区的引导块中
- Swap-Space Management
    - Swap-space — Virtual memory uses disk space as an extension of main memory
        - 交换空间——虚拟内存使用磁盘空间作为主内存的扩展
    - Configure Swap-space
        - 配置交换空间
        - on a swap file in a file system
            - 在文件系统中的交换文件上,采用普通文件系统来创建命名分配
            - changing the size of a swap file is easier.
                - 更改交换文件的大小比较容易。会增加时间和额外的磁盘访问
        - on a separate swap partition
            - 在单独的交换分区上,不存放文件系统和目录结构,通过单独的交换空间存储器,优化速度
            - swap partition is faster but difficult to set it up (how much swap space your system requires?)
                - 交换分区更快，但设置起来比较困难(您的系统需要多少交换空间?),内部碎片增加
    - Solution: start with a swap file and create a swap partition when it knows what the system requires.
        - 解决方案:从交换文件开始，并在知道系统需要什么时创建交换分区。
        - Swap-space management
            - Kernel uses swap maps to track swap-space use
                - 交换空间管理
                - 内核使用交换映射来跟踪交换空间的使用情况
- RAID Structure
    - 概念

  - RAID: Redundant Arrays of Independent Disks.
    - RAID:独立磁盘冗余阵列。
  - RAID is a system of data storage that uses multiple hard disk drives to store data.
    - RAID是一种使用多个硬盘驱动器来存储数据的数据存储系统。
  - RAID is a set of physical drives viewed by the operating system as asingle logical drive.
    - RAID是一组被操作系统视为单个逻辑驱动器的物理驱动器。
  - There are several different storage methods, named levels.
    - 有几种不同的存储方法，称为级别。
  - RAID controller is used for controlling a RAID array. It may be hardware- or software-based.
    - RAID控制器用于控制RAID。它可能是基于硬件或软件的。
- RAID uses 3 main techniques:
  - ❑ Mirroring is copying data to more than one drive.
    - 镜像是将数据复制到多个驱动器上。
    - ➢ If one disk fails, the mirror image preserves the data from the failed disk.
      - 当其中一块硬盘故障时，镜像可保留故障硬盘上的数据。
  - ❑Striping breaks data into “chunks” that are written in succession to different disks.
    - 条带化将数据分割成“块”，连续地写入不同的磁盘。
    - ➢Striping provides high data-transfer rates, this improves performance because your computer can access data from more than one disk simultaneously.
      - 条带化提供高数据传输速率，这提高了性能，因为您的计算机可以同时从多个磁盘访问数据。
  - ❑Error correction redundant data is stored, allowing detection and possibly fixing of errors.
    - 存储冗余数据，允许发现并可能修复错误
- RAID Level 0
  - Level 0 does not provide redundancy.
    - 级别0不提供冗余。
  - ▪ Raid 0 treats multiple disks as a single partition
    - ▪Raid 0将多个磁盘视为单个分区
  - ▪ Files are Striped across disks, no redundant info.
    - ▪文件在磁盘上是条纹的，没有多余的信息。
  - ▪ High read throughput.
    - ▪高读吞吐量。
  - ▪ Any disk failure results in data loss.
    - ▪任何磁盘故障都会导致数据丢失

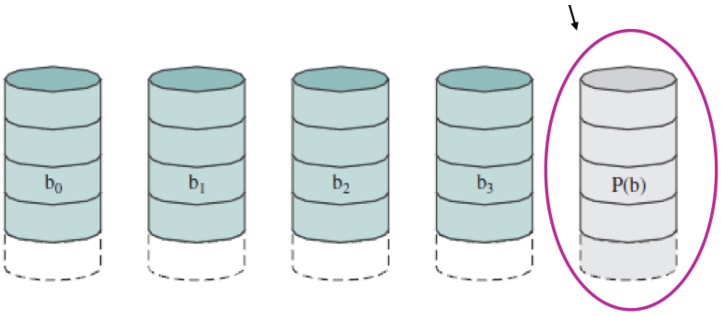
  - 

- RAID Level 1
  - disk mirroring
    - 磁盘镜像
  - - uses striping
    - - is called a mirrored configuration because it provides redundancy by having a duplicate set of all data in a mirror array of disks, which acts as a backup system in the event of hardware failure.
      - -被称为镜像配置，因为它通过在磁盘镜像阵列中拥有所有数据的副本集来提供冗余，在硬件故障时充当备份系
      - 

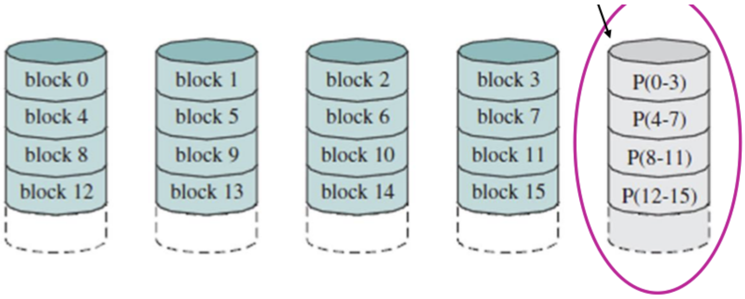
- RAID Level 2
  - memory-style error-correcting-code organization
    - 内存式纠错代码组织
  - - an error-correcting code is calculated across corresponding bits on each data disk, and the bits of the code are stored in the corresponding bit positions on multiple parity disks.
    - —在每个数据盘的对应位上计算一个纠错码，并将纠错码的位存储在多个校验盘的对应位上。
  - - uses very small strips (often the size of a word or a byte)
    - -使用非常小的条带(通常是一个字或一个字节的大小)



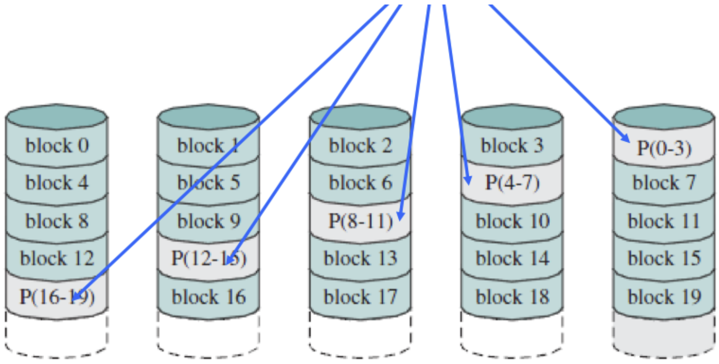
- RAID Level 3
  - bit-interleaved parity organization is a modification of Level 2and requires only a single redundant disk, no matter how large the disk array.
    - 位交错奇偶校验组织是对Level 2的修改，无论磁盘阵列有多大，都只需要一个冗余磁盘。
  - - single parity bit can be used for error correction / detection foreach strip, and it is stored in the dedicated parity disk.
    - —单个校验位可用于每个条带的纠错/检测，并存储在专用的校验盘中。
    - Suppose, strip X = {1010}, the parity bit is 0 as there are even number of 1s.
    - Suppose strip X = {1110}, the parity bit here is 1 as there are odd number of 1s
    - 假设条带X ={1010}，奇偶校验位为0，因为有偶数个1。
    - 假设条带X ={1110}，这里的奇偶校验位为1，因为有奇数个1
    -



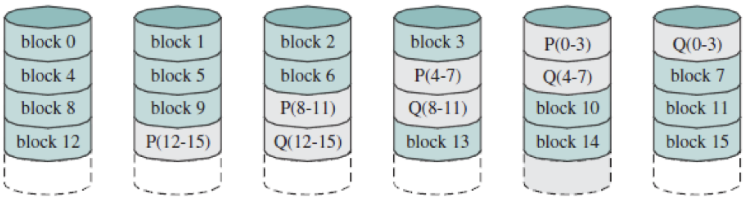
- RAID Level 4
  - -block-interleaved parity organization uses large-sized strips, and the data is striped as fixed-sized blocks;
    - -块交错奇偶校验组织采用大长度的条带，数据作为固定大小的块进行条带化;
  - -one block in size is 512 bytes by default but can be specified otherwise.
    - -一个块的大小默认为512字节，但可以另行指定。
  - -provides block-level striping (the same strip scheme found inLevels 0 and 1) and stores a parity block on a dedicated disk.
    - -提供块级条带化(与level0和level1中的条带方案相同)，并在专用磁盘上存储奇偶校验块。
    -



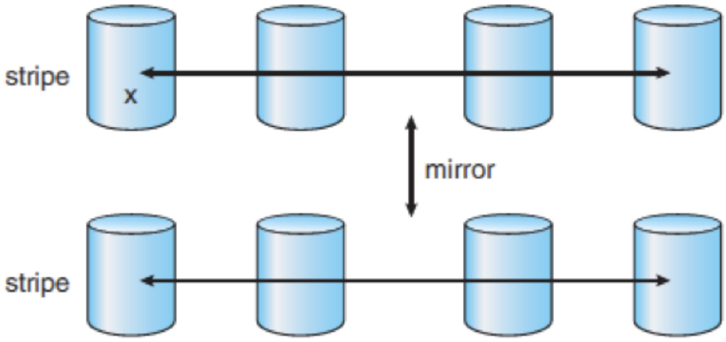
- RAID Level 5
  - block-interleaved distributed parity is a modification of Level4.
    - 块交错分布奇偶校验是Level4的修改。
  - the parity bits are not stored in a single disk
    - 奇偶校验位不存储在单个磁盘中
  - distributes the parity strips across the disks.
    - 在硬盘上分配校验条。
    -



- RAID Level 6
  - -P + Q redundancy scheme - independent data disks with double(dual) parity - extra degree of error detection and correction (parity check and Reed-Solomon codes).
    - - p + Q冗余方案-具有双(双)奇偶校验的独立数据磁盘-额外程度的错误检测和纠正(奇偶校验和里德-所罗门码)。
    - ▪ one calculation is the same as that used in Levels 4 and 5
      - ▪一次计算与第4级和第5级相同
    - ▪ the other is an independent data-check algorithm
      - ▪另一个是独立的数据检查算法
  - Both parities (P and Q) are distributed on separate disks across the array.T
    - 两个奇偶(P和Q)分布在整个阵列的单独磁盘上。T
  - he double parity allows for data restoration even if two disks fail.
    - 双奇偶校验允许数据恢复，即使两个磁盘故障。

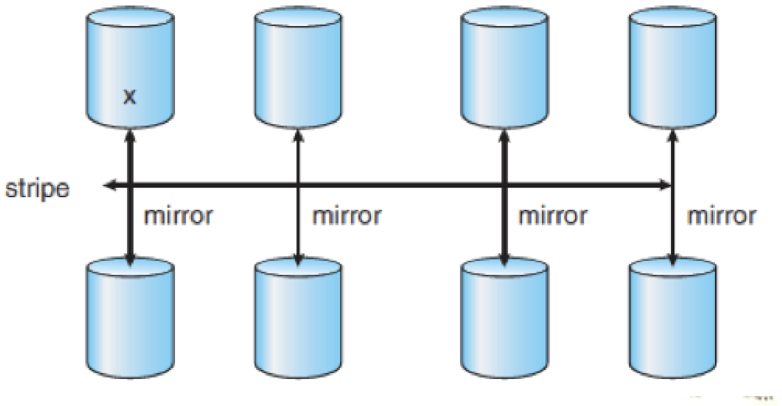


  - RAID level 6 may suffer in performance due to two parity disks.Better features→RAID levels combined.RAID级别6存在两个校验盘，可能导致性能下降。更好的功能→RAID级别组合。
- Better features→RAID levels combined.
  - RAID level 0 + 1
    - “Mirror of stripes” – is a combination of the striping of RAID 0 (a set of disks are striped) and mirroring of RAID 1 (the stripe is mirrored to another).
    - - a set of n disks are striped, and then the stripe is mirrored on n redundant disks.
    - “条带镜像”——是RAID 0的条带化(一组磁盘是条带的)和RAID 1的镜像(条带镜像到另一个磁盘)的组合。
    - —将n个磁盘进行分条，然后将分条镜像到n个冗余磁盘上。

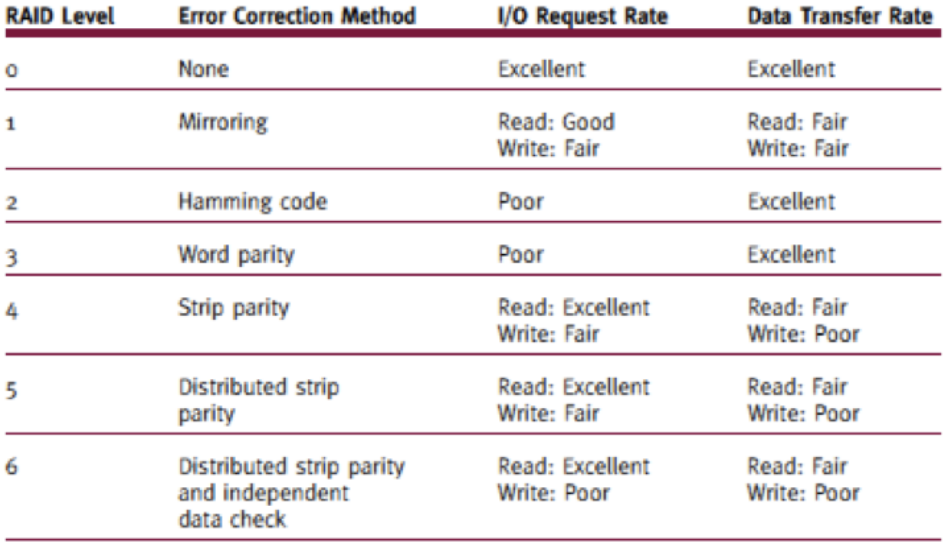


  - RAID level 1 + 0
    - “Stripe of mirrors” - combines the mirroring of RAID 1 (disks are mirrored for redundancy) with the striping of RAID 0 (stripes across disks for higher performance).
      - “分条镜像”——结合了RAID 1的镜像(磁盘被镜像以实现冗余)和RAID 0的分条(磁盘上的分条以获得更高的性能)。
    - The advantage in 1 + 0 is that in case of failure of a single disk, the mirror copy of the whole disk is available
      - 1 + 0的优点是，在单个磁盘出现故障的情况下，可以使用整个磁盘的镜像副本
    - is ideal for highly utilized database servers or any server that's performing many write operations.
      - 非常适合高利用率的数据库服务器或执行许多写操作的任何服务器。

- - gives the best performance, but it is also costly (requires twice as many disks as other RAID levels).
    - -提供最佳性能，但成本也很高(需要的磁盘数量是其他RAID级别的两倍)。



- RAID Conclusions
  - RAID is secure because mirroring duplicates all your data.
    - RAID是安全的，因为镜像复制了您的所有数据。
  - • RAID is fast because the data is striped across multiple disks; chunks of data can be read and written to different disks simultaneously.
    - •RAID是快速的，因为数据是跨多个磁盘的条纹;数据块可以同时被读写到不同的磁盘上。
  - • RAID is not a backup. A backup is a copy of data, which is stored somewhere else and is detached from the original data both in space and time.
    - •RAID不是备份。备份是数据的副本，它存储在其他地方，在空间和时间上都与原始数据分离。



RAID Level	Error Correction Method	I/O Request Rate	Data Transfer Rate
0	None	Excellent	Excellent
1	Mirroring	Read: Good Write: Fair	Read: Fair Write: Fair
2	Hamming code	Poor	Excellent
3	Word parity	Poor	Excellent
4	Strip parity	Read: Excellent Write: Fair	Read: Fair Write: Poor
5	Distributed strip parity	Read: Excellent Write: Fair	Read: Fair Write: Poor
6	Distributed strip parity and independent data check	Read: Excellent Write: Poor	Read: Fair Write: Poor