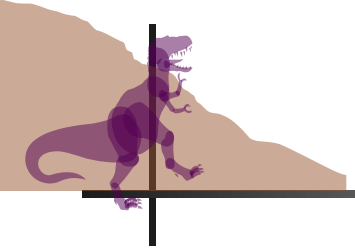


第五部分 存储管理

Part Five Storage Management

第11章 大容量存储器的结构
Mass-Storage Structure

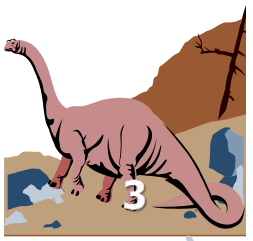
- 本章目标 CHAPTER OBJECTIVES
  - 描述不同二级存储设备的物理结构及设备结构对其使用的影响

    Describe the physical structures of various secondary storage devices and the effect of a device's structure on its uses
  - 说明大容量存储设备的工作特性

    Explain the performance characteristics of mass-storage devices

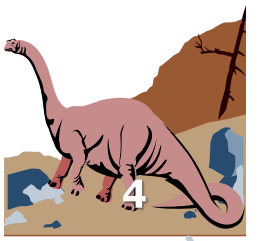
11.1 大容量存储器结构简介
Overview of Mass-Storage Structure

- 文件的存储设备主要有磁带、磁盘、光盘等。
- 存储设备的特性可以决定文件的存取方法。
- 下面介绍以磁带为代表的顺序存取设备和以磁盘为代表的直接存取设备。

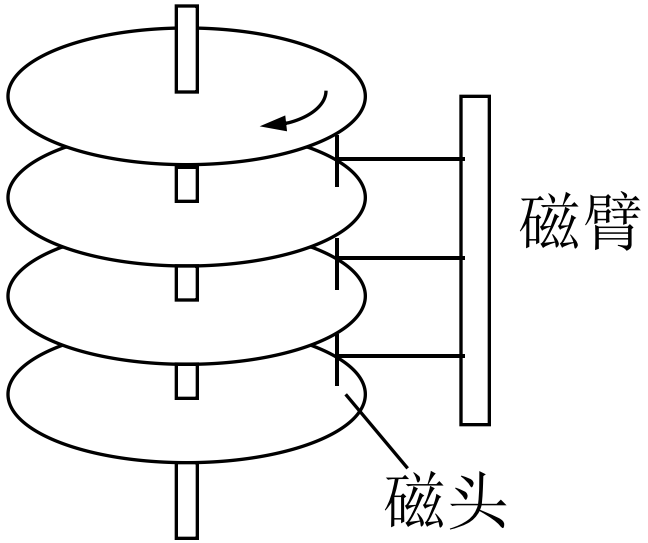


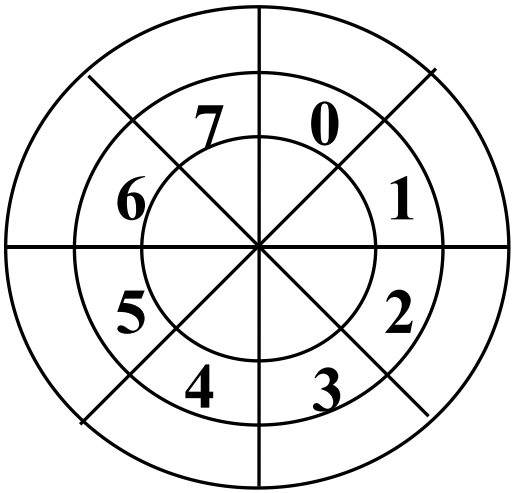
11.1.1 硬盘 Hard Disk

- 磁盘是典型的直接存取设备。
- 磁盘一般由若干磁盘片组成，可沿一个固定方向高速旋转。每个盘面对应一个磁头，磁臂可沿半径方向移动。
- 磁盘上的一系列同心圆称为磁道（track），磁道沿径向又分成大小相等的多个扇区（sector），与盘片中心有一定距离的所有磁道组成一个柱面（cylinder）。
- 磁盘上的每个物理块可用柱面号，磁头号和扇区号表示。



磁盘数据组织和格式示意图





磁道 第i扇区

…	间隙	标识字段	间隙	数据字段	间隙	…

逻辑扇区

- 将一个磁盘上的扇区从0柱面开始，按柱面顺序依次编号，所得到的编号称为逻辑扇区号。
- 逻辑扇区号=柱面号*每柱面扇区数+
  磁头号*每磁头扇区数+
  扇区号

其中，柱面号、磁头号、扇区号都从0开始

磁盘访问时间

- 磁盘访问时间由三部分组成：
  - 寻道时间（seek time）：指将磁头从当前位置移动到指定磁道所经历的时间。由启动磁臂时间和磁头移动多条磁道的时间构成。
  - 旋转延迟时间（rotational latency）：指扇区移动到磁头下面所经历的时间。平均旋转延迟时间是每转所需时间的一半。
  - 传输时间（tranfer time）：指从磁盘上读出数据或向磁盘写入数据所经历的时间。
- 由于这三部分操作均涉及机械运动，故磁盘块的访问时间约为0.01～0.1s之间，其中寻道时间所占的比例最大。

11.1.2 磁带 Magnetic tape

- 磁带是一种典型的顺序存取设备。由于磁带机的启动和停止要花费一定的时间，因此在磁带的相邻物理块之间设计有一段间隙将它们隔开，如下所示。

磁带

…	间隙	第i块	间隙	第i＋1块	间隙	…

磁带（续）

- 磁带的存取速度与信息密度（字符数/英寸）、磁带带速（英寸/秒）和块间间隙有关。
- 如果带速高、信息密度大且所需块间隙（磁头启动和停止时间）小，则磁带存取速度高。反之，若磁带带速低、信息密度小且所需块间隙（磁带启动和停止时间）大，则磁带存取速度低。

存储设备、存取方法和物理结构的关系1

- 文件的物理结构与文件存储器的特性和存取方法密切相关。
- 磁带是一种顺序存取设备，适合采用顺序结构存放文件，相应的存取方法通常是顺序存取法。若采用其他文件结构或采用直接存取方式都不太合适。

存储设备、存取方法和物理结构的关系2

- 磁盘属于直接存取存储设备，前述的几种物理结构都可以采用。存取方法也可以多种多样。
- 如果采用顺序存取法则前述的几种文件结构都可以采用。如果采用直接存取法，则索引文件效率最高，顺序文件效率居中，串联文件效率最低。

存储设备、存取方法和物理结构间的关系 3

存储设备	磁 盘			磁 带
物理结构	顺序结构	链接结构	索引结构	顺序结构
存取方法	顺序、直接	顺序	顺序、直接	顺序

11.2 磁盘调度
Disk Scheduling

- 磁盘是可以被多个进程共享的设备。当有多个进程都请求访问磁盘时，应采用一种适当的调度算法，以使各进程对磁盘的平均访问时间（主要是寻道时间）最短。
- 下面介绍几种磁盘调度算法。

11.2.1 先来先服务调度
FCFS Scheduling

- 先来先服务算法按进程请求访问磁盘的先后次序进行调度。
- 特点：简单合理，但未对寻道进行优化。

先来先服务调度例

下一磁道号	移动距离
55	45
58	3
39	19
18	21
90	72
160	70
150	10
38	112
184	146

从100号磁道开始，磁盘访问请求为：55、58、39、18、90、160、150、38、184

平均寻道长度为：55.3

最短寻道时间优先调度
SSTF Scheduling

- 最短寻道时间优先算法选择从当前磁头位置所需寻道时间最短的请求作为下一次服务的对象。

  Selects the request with the minimum seek time from the current head position.

- 特点：寻道性能比FCFS好，但不能保证平均寻道时间最短，还可能会使某些请求总也得不到服务。

最短寻道时间优先调度例

下一磁道号	移动距离
90	10
58	32
55	3
39	16
38	1
18	20
150	132
160	10
184	24

从100号磁道开始，磁盘访问请求为：55、58、39、18、90、160、150、38、184

平均寻道长度为：27.6

11.2.2 扫描调度
SCAN Scheduling（look）

- SSTF有可能引起某些请求的饥饿。

  SSTF scheduling may cause starvation of some requests.
- SCAN算法在磁头当前移动方向上选择与当前磁头所在磁道距离最近的请求作为下一次服务的对象。

扫描调度2
SCAN Scheduling（LOOK）

- 因这种算法中磁臂移动规律颇似大楼中电梯的运行，故又称为电梯调度算法。

  The SCAN algorithm is sometimes called the elevator algorithm，since the disk arm behaves just like an elevator in a building.
- 特点：具有较好的寻道性能，能避免进程饥饿，但不利于两端磁道的请求。

扫描算法例

下一磁道号	移动距离
150	50
160	10
184	24
90	94
58	32
55	3
39	16
38	1
18	20

从100号磁道开始，向磁道号增加方向移动。磁盘访问请求为：55、58、39、18、90、160、150、38、184

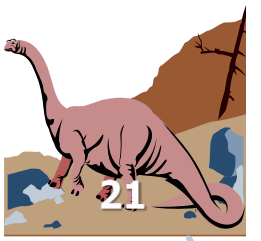
平均寻道长度为：27.8

11.2.3 循环扫描算法（CSCAN）
Circular SCAN Scheduling （C-LOOK）

- CSCAN算法是SCAN算法的变种，提供了一个更为均匀地等待时间。

  CSCAN scheduling is a variant of SCAN designed to provides a more uniform wait time than SCAN。
- 磁头从磁盘的一端向另一端移动，沿途响应请求。当它到了另一端，就立即回到磁盘的开始处，在返回的途中不响应任何请求。

  The head moves from one end of the disk to the other. servicing requests as it goes. When it reaches the other end, however, it immediately returns to the beginning of the disk, without servicing any requests on the return trip.



循环扫描算法2（CSCAN）
Circular SCAN Scheduling （C-LOOK）

- 特点：该算法消除了对两端磁道请求的不公平。

循环扫描算法例

下一磁道号	移动距离
150	50
160	10
184	24
18	166
38	20
39	1
55	16
58	3
90	32

从100号磁道开始，向磁道号增加方向移动。磁盘访问请求为：55、58、39、18、90、160、150、38、184

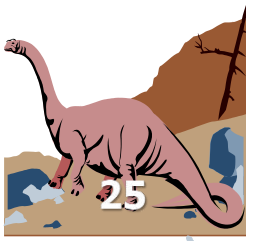
平均寻道长度为：35.8

N-Step-SCAN

- 若多个进程反复请求对某一磁道的访问，则磁臂可能停留在某处不动，这一现象称为磁臂粘着。
- N-Step-SCAN算法：将磁盘请求队列分成若干个长度为N的子队列，磁盘调度按FCFS算法依次处理这些子队列，而处理每个队列时按SCAN算法进行，一个队列处理完后，再处理其他队列。

FSCAN算法

- FSCAN算法是N-Step-SCAN算法的简化，它只将磁盘请求队列分成两个子队列。一个是当前所有请求磁盘I/O的进程形成的队列，由磁盘调度按SCAN算法进行处理，另一个队列则是在扫描期间新出现的磁盘请求。



11.2.4 磁盘调度算法的选择
Selecting a Disk-Scheduling Algorithm

- SSTF比较通用且很有吸引力。

  SSTF is common and has a natural appeal.
- SCAN和C-SCAN在重磁盘负载的系统中执行得较好。

  SCAN and C-SCAN perform better for systems that place a heavy load on the disk.
- 性能依赖于请求的数量和类型。

  Performance depends on the number and types of requests.

磁盘调度算法的选择2
Selecting a Disk-Scheduling Algorithm

- 磁盘服务请求受到文件分配方式的影响。

  Requests for disk service can be influenced by the file-allocation method.
- 磁盘调度算法应该写成操作系统中的一个独立模块，在必要的时候允许用不同的算法来替换。

  The disk-scheduling algorithm should be written as a separate module of the operating system, allowing it to be replaced with a different algorithm if necessary.
- SSTF和LOOK都是缺省算法的合理选择。

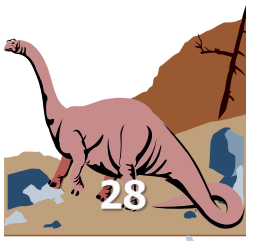
  Either SSTF or LOOK is a reasonable choice for the default algorithm.

11.3 磁盘结构 Disk Structure

- 磁盘设备编址为逻辑块的一维大数组。

  Disk drives are addressed as large 1-dimensional arrays of logical blocks.
- 一维逻辑块的数组按顺序映射到磁盘上的扇区。

  The 1-dimensional array of logical blocks is mapped into the sectors of the disk sequentially.



磁盘结构2 Disk Structure

- 0扇区是最外边柱面的第一个磁道的第一个扇区。

  Sector 0 is the first sector of the first track on the outermost cylinder.

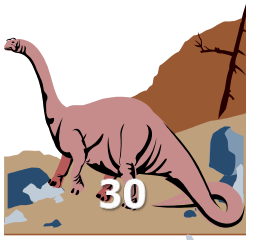
- 数据首先都映射到一个磁道，其余的数据映射到同一柱面的其他磁道，然后按照从外向里的顺序映射到其余的柱面。

  Mapping proceeds in order through that track, then the rest of the tracks in that cylinder, and then through the rest of the cylinders from outermost to innermost.

11.5 磁盘管理
Storage Device Management

- 这里讨论磁盘初始化，磁盘引导

  Here we discuss disk initialization , booting from disk



11.5.1 磁盘格式化 Disk Format

- 低级格式化，或物理格式化：把磁盘划分成扇区，以便磁盘控制器可以进行读写。

  Low-level formatting, or physical formatting: Dividing a disk into sectors that the disk controller can read and write.
- 每个扇区的数据结构通常由头、数据区域和尾部组成。

  Every sector consists of a header, a data area (usually 512 in size), and a trailer.
- 头部和尾部包含了一些磁盘控制器所使用的信息，如扇区号

  The header and trailer contain information used by disk controller, such as a sector number.

磁盘格式化2 Disk Format

- 为使用磁盘保存文件，操作系统还需要在磁盘上保存它自身的数据结构。这分为两步：

  To use a disk to hold files, the operating system still needs to record its own data structures on the disk.

  - 把磁盘划分成分区

    Partition the disk

  - 逻辑格式化或“创建文件系统”。

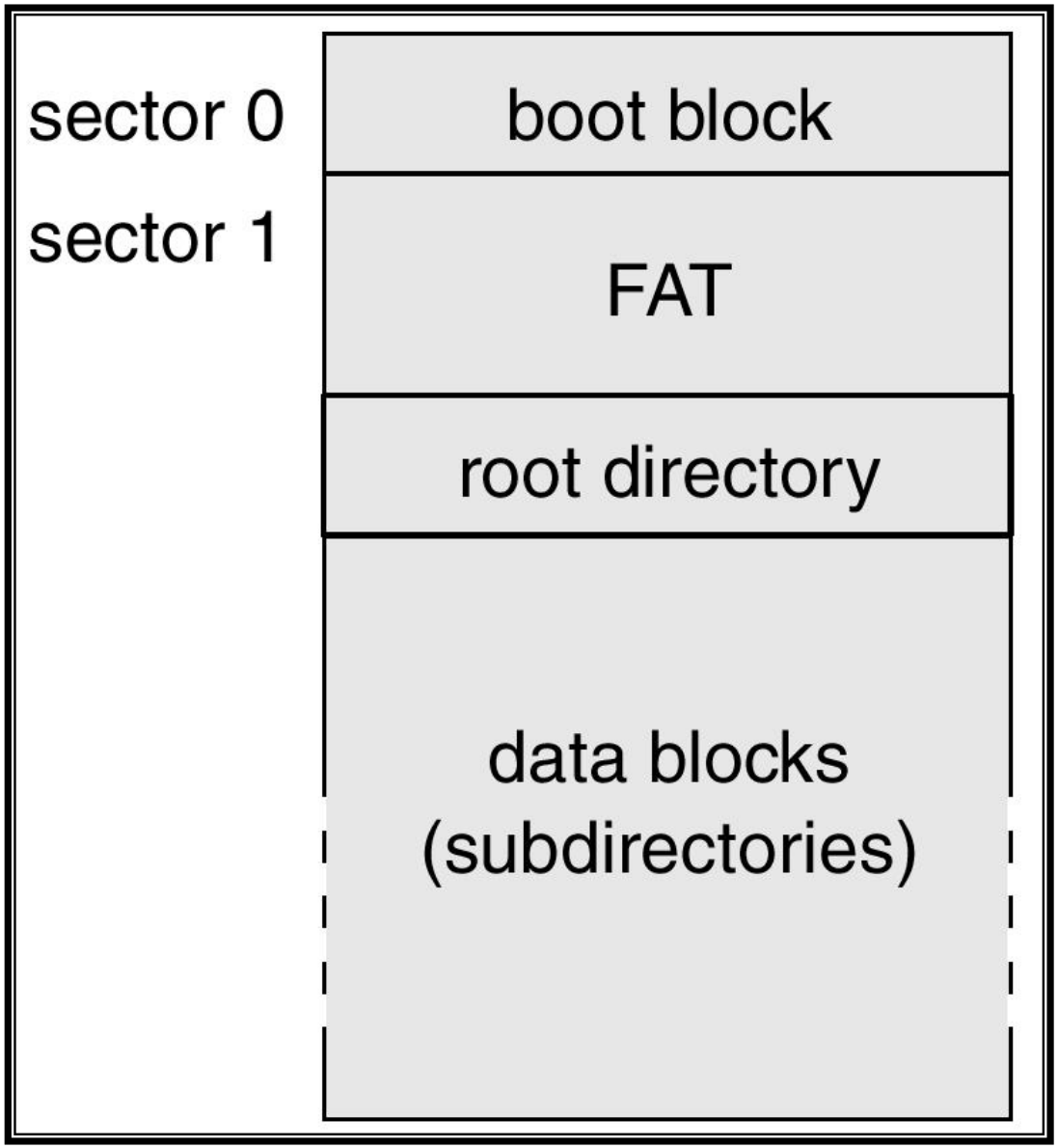
    Logical formatting or “making a file system”

- 也称高级格式化.

11.5.2 引导块 Boot Block

- 引导块位于磁盘的固定位置，如引导分区的第一扇区
- 引导过程 Booting process
  - **CPU自检 CPU self-testing**
  - **运行ROM中的自举程序 Run the bootstrap at the ROM (BIOS for PC)**
  - **从引导分区装入第一块 Load the first block from the bootable partition**
- 绝大多数系统只在启动ROM中保留一个很小的自举装入程序，其作用是进一步从磁盘上调入更为完整的自举程序。它能从磁盘上装入整个操作系统。

Windows 2000的磁盘引导
Booting from disk in Windows 2000



11.5.3 坏块 Bad Block

- 磁盘上的一个或多个扇区可能坏掉。
  - 对于简单磁盘， Format等程序可以标记坏扇区以通知分配程序不使用。
  - 更为复杂的磁盘，对坏块的处理更为智能化。如采用扇区备用或转寄方案，即低级格式化时留一些块作为备用，发现坏块时用备用块逻辑替代坏块；

11.6 交换空间管理
Swap-Space Management

- 交换空间：虚拟内存使用磁盘空间作为主存的扩展

  Swap-space ： Virtual memory uses disk space as an extension of main memory.

- 交换空间的使用 Swap-space use
  - 保存整个进程映像，

    Hold an entire process image
  - 存储换出内存的页

    store pages that have been pushed out of main memory

交换空间管理2
Swap-Space Management

- 交换空间的位置
  - 交换空间创建在普通文件系统上。通常是文件系统内的一个简单大文件。这种方式实现简单但效率较低。
  - 交换空间创建在独立的磁盘分区上（如Unix/Linux）。
  - 有些OS较为灵活，可以由系统管理员来选择使用以上哪种方式。

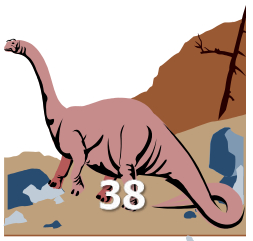
练习

- 11.13 Suppose that a disk drive has 5,000 cylinders, numbered 0 to 4,999. The drive is currently serving a request at cylinder 2,150, and the previous request was at cylinder 1,805. The queue of pending requests, in FIFO order, is:

  2,069; 1,212; 2,296; 2,800; 544; 1,618; 356; 1,523; 4,965; 3,681

  Starting from the current head position, what is the total distance (in cylinders) that the disk arm moves to satisfy all the pending requests for each of the following disk-scheduling algorithms?

- a. FCFS
- b. SCAN
- c. C-SCAN



选择题

- 共享设备磁盘的物理地址为（柱面号，磁头号，扇区号），磁头从当前位置移动到需访问柱面所用的时间称为 ① ，磁头从访问的柱面移动到指定扇区所用时间称为 ② 。

  A. 寻道时间　　B. 传输时间

  C. 旋转等待时间　　D. 周转时间

- 若进程P1访问199号柱面，磁头是从0号柱面移到199柱面的，且在访问期间依次出现了P2申请读299号柱面，P3申请写209号柱面，P4申请读199号柱面，访问完199号柱面以后，如果采用：先来先服务算法，将依次访问 ① ；最短寻道时间优先算法，将依次访问 ② ；扫描算法，将依次访问 ③ 。

  A. 299，199，209　　B. 299，209，199

  C. 199，209，299　　D. 209，199，299

选择题2

- 存放在磁盘上的文件 _____ 。

  A. 只能随机访问　　B. 只能顺序访问

  C. 既可随机访问，又可顺序访问

  D. 不能随机访问

- 用磁带作文件存储介质时，文件只能组织成 _____ 。

  A. 目录文件　　B. 链接文件

  C. 索引文件　　D. 顺序文件

填空题

- 活动头磁盘的访问时间包括 ①、② 和 ③。
- _____ 算法选择与当前磁头所在磁道距离最近的请求作为下一次服务的对象。

考研题1

- 假设磁头当前位于第105道，正在向磁道序号增加的方向移动。现有一个磁道访问序列请求为35、45、12、68、110、180、170、195，采用SCAN算法得到的磁道访问序列为（）。09

  A、110、170、180、195、68、45、35、12

  B、110、68、45、35、12、170、180、195

  C、110、170、180、195、12、35、45、68

  D、12、35、45、68、110、170、180、195

- 下列选项中，不能改善磁盘I/O性能的是（）12

  A. 重排I/O请求次序

  B. 在一个磁盘上设置多个分区

  C. 预读和滞后写

  D. 优化文件物理块的分布

考研题2

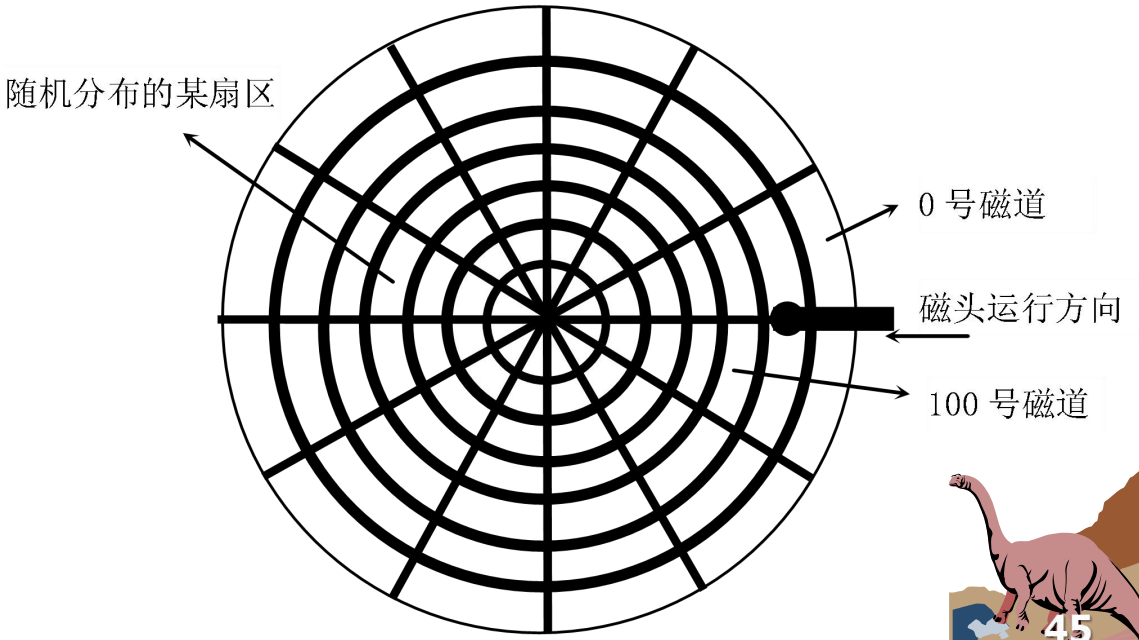
- 假设计算机系统采用CSCAN（循环扫描）磁盘调度策略，使用2KB的内存空间记录16384个磁盘块的空闲状态。
- （1）请说明在上述条件下如何进行磁盘块空闲状态管理。
- （2）设某单面磁盘旋转速度为每分钟6000转，每个磁道有100个扇区，相邻磁道间的平均移动时间为1ms。若在某时刻，磁头位于100号磁道处，并沿着磁道号增大的方向移动（如图所示），磁道号请求队列为50，90，30，120，对请求队列中的每个磁道需读取1个随机分布的扇区，则读完这个扇区点共需要多少时间？要求给出计算过程。

考研题2-2

- （3）如果将磁盘替换为随机访问的Flash半导体存储器（如U盘、SSD等），是否有比CSCAN更高效的磁盘调度策略？若有，给出磁盘调度策略的名称并说明理由；若无，说明理由。

考研题2-3

- （1）使用位示图法表示磁盘的空闲状态（1分），每一位表示一个磁道块是否为空闲，共需要16384/32=512个字=512×4个字节=2KB，正好可放在系统提供的内存中（1分）。



考研题2-4

- （2）采用CSCAN调度算法，访问磁道的顺序为120、30、50、90，则移动磁道长度为20+90+20+40=170，总的移动磁道时间为170×1ms=170ms（1分）。
- 每分钟6000转，则平均旋转廷迟为60/(6000×2)=5ms，总的旋转廷迟时间=5ms×4=20ms（1分）。
- 每分钟6000转，则读取一个磁道上一个扇区的平均读取时间为10ms/100=0.1ms,总的读取扇区的时间=0.1ms×4=0.4ms。
- 读取上述磁道上所有扇区所花的总时间=170ms+20ms+0.4ms=190.4ms（1分）。

考研题2-5

- （3）采用FCFS（先来先服务）调度策略更高效（1分）。因为Flash半导体存储器的物理结构不需要考虑寻道时间和旋转廷迟，可直接按I/O请求的先后顺序服务（1分）。
- 提示：Flash存储器（闪存）属可改写ROM，是一种长寿命的非易失性（在断电情况下仍能保持所存储的数据信息）的存储器，数据删除不是以单个的字节为单位而是以固定的区块为单位，区块大小一般为256KB到20MB。