

Mass Storage — Chapter 12

2022年12月

薛哲

School of Computer Science (National Pilot Software Engineering School)



北京邮电大学

Performance Indicators

Reliability Speed Capacity Cost Access Mode

Storage

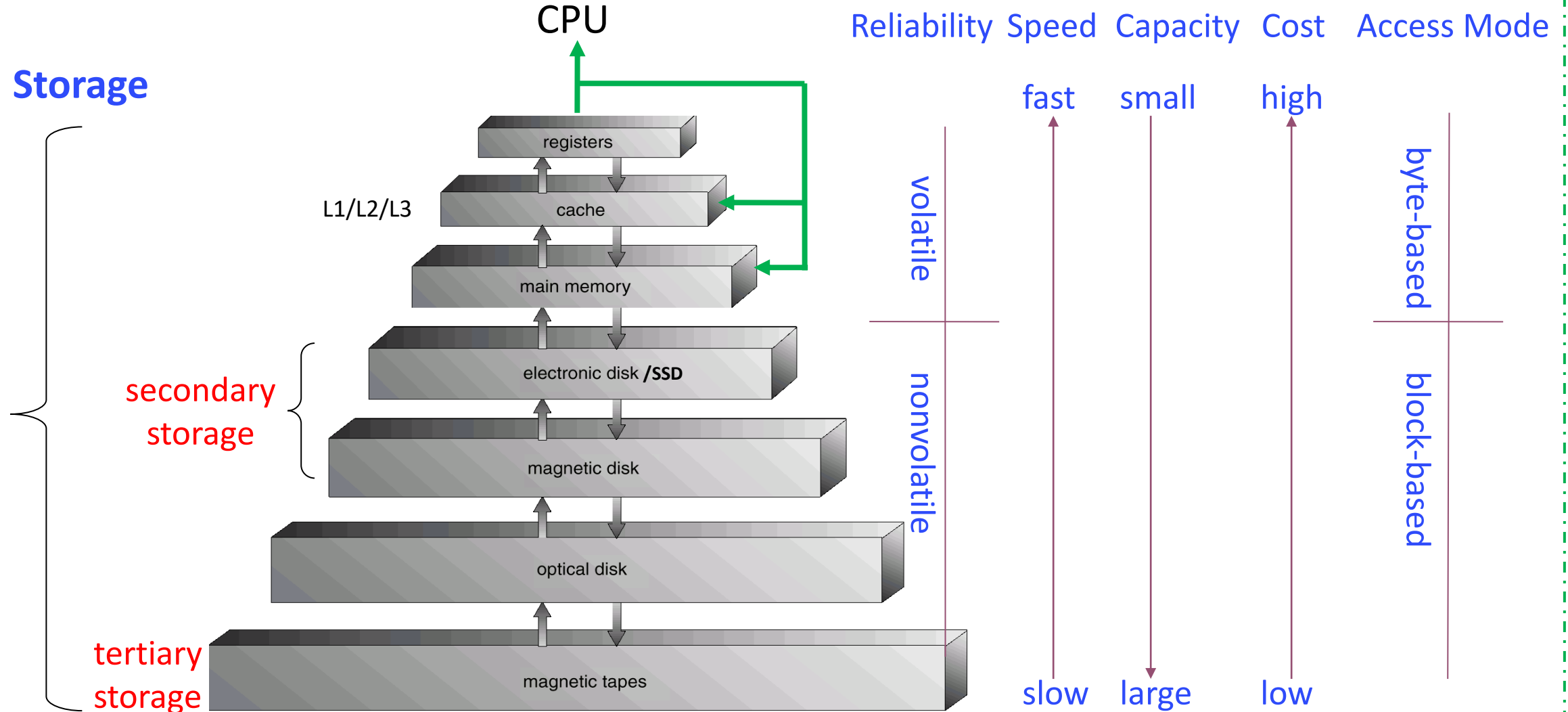
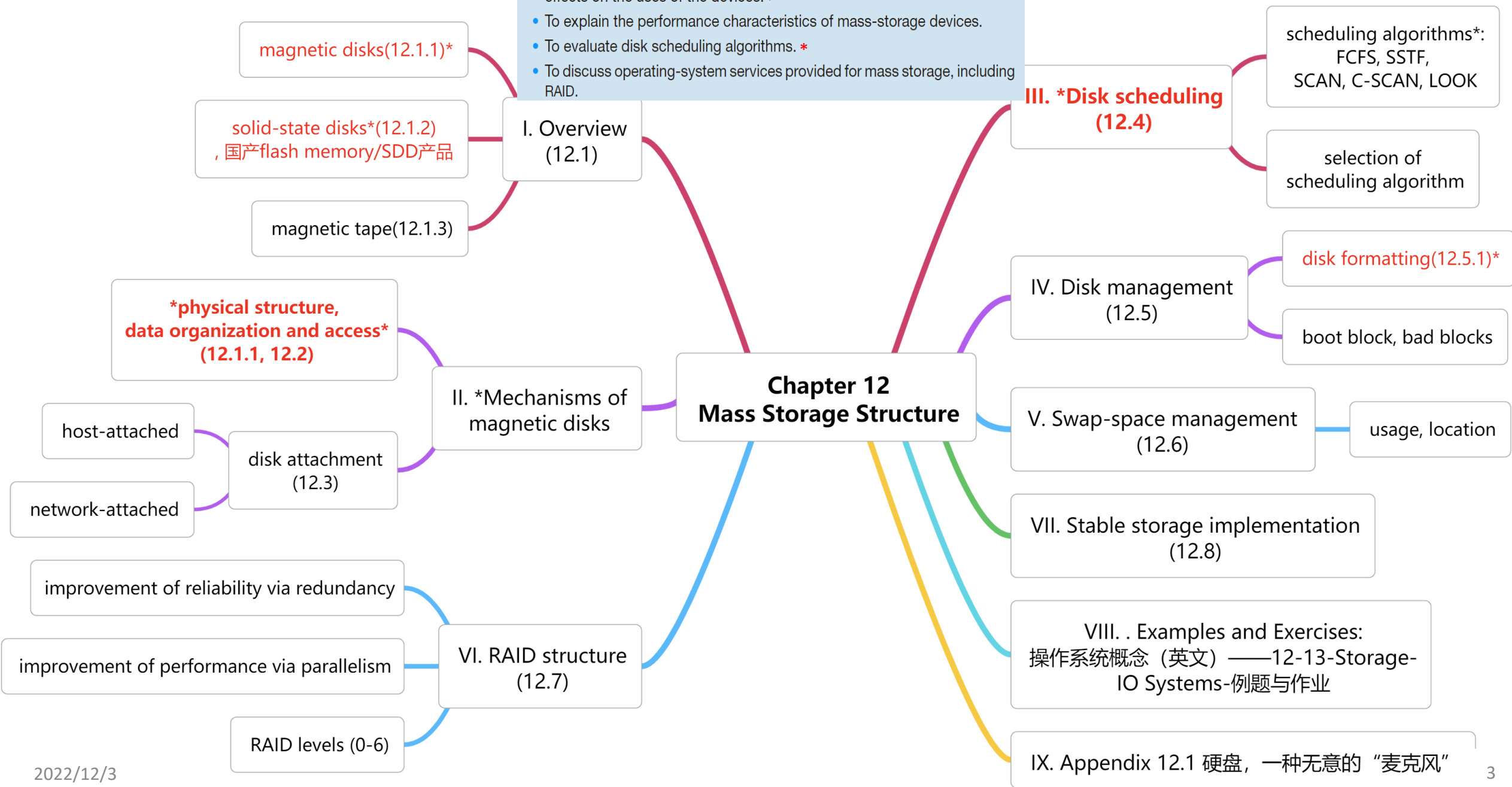


Fig.1.4 Storage-device hierarchy

- To describe the physical structure of secondary storage devices and its effects on the uses of the devices. *
- To explain the performance characteristics of mass-storage devices.
- To evaluate disk scheduling algorithms. *
- To discuss operating-system services provided for mass storage, including RAID.



12.1.1 Overview/12.2 Disk Structure/12.3 Disk Attachment

- HDD: 12.1.1 Magnetic disk/12.2 Disk structure/12.3 Disk attachment
 - ▣ structure, data organization, data access
 - ▣ addressing(寻址)
 - ▣ performance measures
 - ▣ attachment
- 12.1.2 Solid state disk (SSD)
- 12.1.3 Magnetic tape

HDD: Disk structure

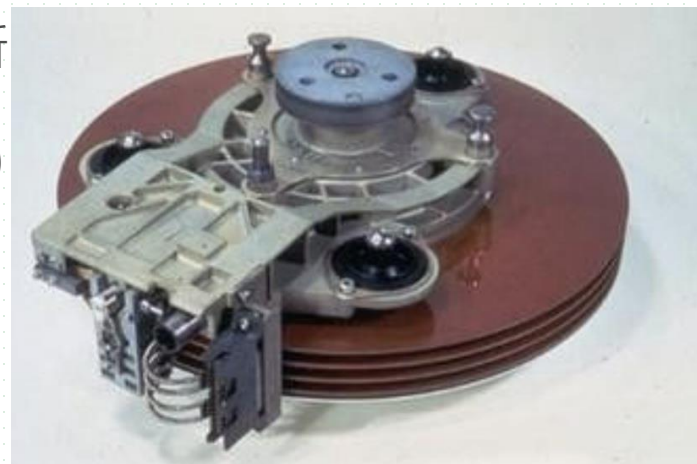
■ Hard Disk Drive, HDD

- 硬盘驱动器，机械硬盘
- information stored in magnetic disk(磁盘)
- consists of
 - disk platter/magnetic disk (盘片)
 - disk controller (I/O registers as I/O ports)

1956年，世界上第一块硬盘IBM 350 RAMAC (Random Access Method of Accounting and Control)，容量5M



1973年，第一款新型Winchester温彻斯特硬盘IBM 3340



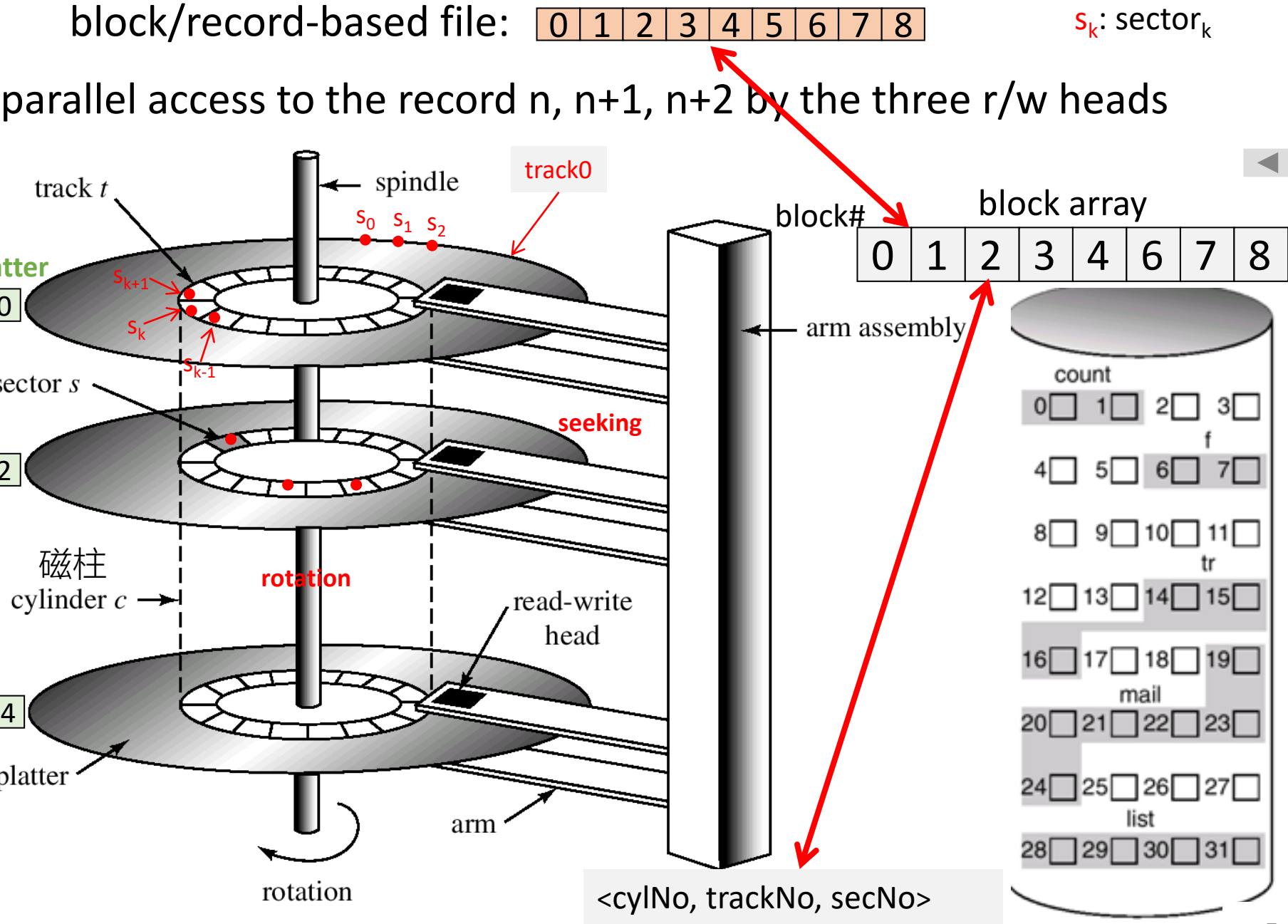
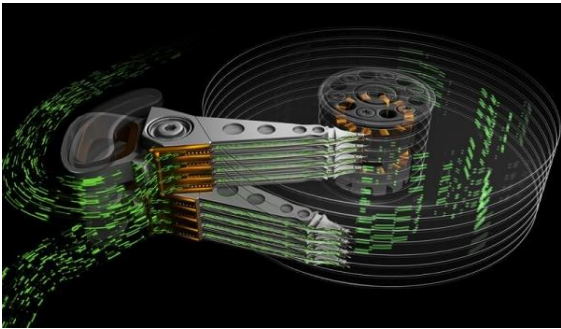
HDD: Disk structure

- HDD, or magnetic disks, provide the bulk of secondary storage for modern computer systems
 - ▣ each disk **platter** has a flat circular shape, with diameters ranging from 1.8 to 3.5 inches
 - the two surfaces of a platter are covered with a magnetic material, for storing information on the platter surfaces
 - ▣ a read– write **head** “flies” just above each surface of every platter, and are attached to a disk arm that moves all the heads as a unit
 - ▣ the surface of a platter is logically divided into circular **tracks**, which are subdivided into **sectors**
 - ▣ the set of tracks that are at one arm position makes up a **cylinder**
- When the disk is in use, a drive motor spins it at high speed
 - ▣ rotate 60 to 250 times per second, specified in terms of **rotations per minute (RPM)**
 - ▣ commonly, 5,400, 7,200, 10,000, and 15,000 RPM

Fig. 12.1

一维块地址block#


- 三维物理地址 on disk
- 柱面号cylNo
 - 磁道号trackNo
 - i.e. 盘面, 磁头
 - 扇区号secNo



HDD: Disk structure

- 为什么硬盘转速是3600、5400或7200这么奇怪的数字？7200转的硬盘一定比5400快吗？
 - ▣ 5400转/分钟=5400 RPM
- 借鉴交流电机设计
 - ▣ 美国交流电60Hz
 - ▣ 1分钟=60秒, $60\text{Hz} * 1\text{转/Hz} * 60/\text{分钟} = 3600\text{转/分钟}$
- $5400\text{ RPM} = 1.5 * 3600\text{RPM}$
- $7200\text{ RPM} = 2 * 3600\text{RPM}$

Addressing in HDD

- Disk drives are addressed as large one-dimensional arrays of logical blocks 
 - ▣ the block is the smallest unit of transfer
- The size of a logical block is usually 512 bytes, although some disks can be low-level formatted to have a different logical block size, such as 1,024 bytes
- The one-dimensional array of logical blocks is mapped onto the sectors of the disk sequentially
 - ▣ sector 0 is the first sector of the first track on the outermost cylinder
 - ▣ the mapping proceeds in order through that track, then through the rest of the tracks in that cylinder, and then through the rest of the cylinders from outermost to innermost.
- By this mapping, convert a logical block number **block#** into an old-style disk address that consists of
 - ▣ a cylinder number **cylNo**, a track number **trackNo** within that cylinder, and a sector number **secNo** within that track

Task Manager in Windows



Performance Measures of Disks

- **Positioning time , or random access time** – the time it takes from when a read or write request is issued to when data transfer begins, consisting of **seek time** and **rotational latency**
 - ▣ **seek time** – the time for the disk to move the disk arm to the cylinder containing the desired sector
 - 4 to 10 milliseconds on typical disks
 - on average, **average seek time is half the worst case seek time**

// 磁臂arm径向移动($1/2 * m$) 个磁道所需时间,
m: 盘片上磁道总数

Performance Measures of Disks

- ❑ **rotational latency** – the time for the desired sector to be accessed to appear under the disk head, i.e. to rotate to the head
 - 4 to 11 milliseconds on typical disks, depending on rotation rates RPM, e.g. 5400, 7200 RPM (rotations per minute)
 - the average latency is half of the above latency
// 磁盘旋转半圈所需时间
- Overall latency (seek time + rotational latency) is 5 to 20 msec depending on disk model
- **Data-transfer rate** – the rate at which data flow between the drive and the computer
 - ❑ i.e. the rate at which data can be retrieved from or stored to the disk
 - ❑ 25 to 200 MB per second max rate, lower for inner tracks ?
 - ❑ 进一步地，细分为内部速率、外部速率

■ 内部传输速率

- ❑ 磁头在盘片上读写数据速率，如50MB/s
- ❑ 也称为：（最大最小）持续传输速率
- ❑ 取决于seek time, latency

■ 外部传输速率

- ❑ 硬盘缓存cache（磁盘控制器中的I/O 寄存器）与内存间的传输速率
- ❑ 也称为：突发数据传输速率、接口速率
- ❑ 依磁盘接口类型而定，如SATA, m.2

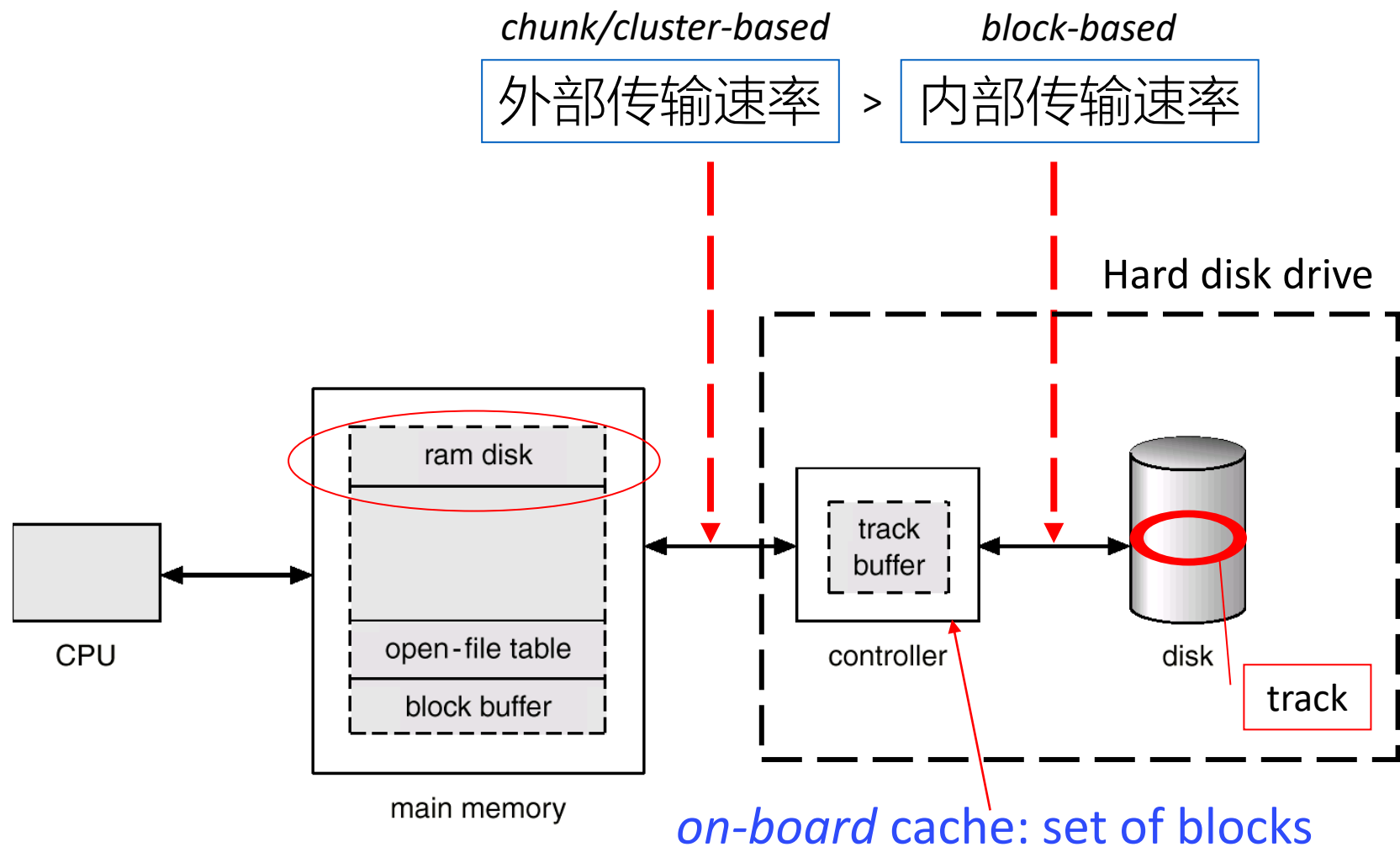


Fig. 12.11.0 on-board cache in device controller

Seagate(希捷) SV35 Series

■ 面向视频监控的高性能、大容量硬盘

硬盘和内存
间cache/buffer

外部传输速率

内部传输速率

平均寻道
时间 8.5ms

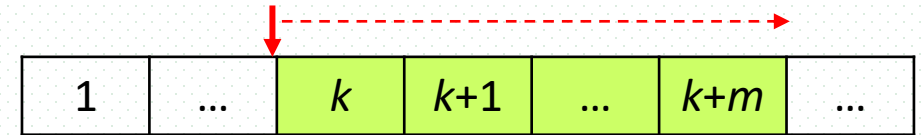
规格	2TB ¹	1TB ¹	500GB ¹
型号	ST2000VX002	ST31000526SV	ST3500411SV
接口选项	SATA 6Gb/秒 NCQ	SATA 6Gb/秒 NCQ	SATA 6Gb/秒 NCQ
性能			
转速 (RPM)	5900	7200	7200
多段缓存 (MB)	64	32	16
支持 SATA 数据传输率 (Gb/秒)	6.0/3.0/1.5	6.0/3.0/1.5	6.0/3.0/1.5
开机启动时间 (秒)	<17	<10	<10
持续数据传输率, 顺序写入 (MB/秒)	144	140	140
配置/结构			
磁头/磁盘	6/3	4/2	2/1
字节数/扇区	4096	512	512
电压			
电压容差 (包括噪声电压) 5V	±5%	±5%	±5%
电压容差 (包括噪声电压) 12V	±10%	±10%	±10%
可靠性/数据完整性			
接触启停周期	—	50,000	50,000
加载/卸载周期 (25°C, 50% 湿度)	300,000	—	—
不可恢复读错误/被读数据 (位), 最大	1/10 ¹⁴	1/10 ¹⁴	1/10 ¹⁴
年返修率 (AFR)	<1%	<1%	<1%
MTBF (小时)	>1M	>1M	>1M
开机小时数	8760	8760	8760

Performance Measures of Disks

- **Disk block** is a logical unit for storage allocation and retrieval
 - ▣ 4 to 16 kilobytes typically
 - smaller blocks: more transfers from disk
 - larger blocks: more space wasted due to partially filled blocks (碎片fragment)

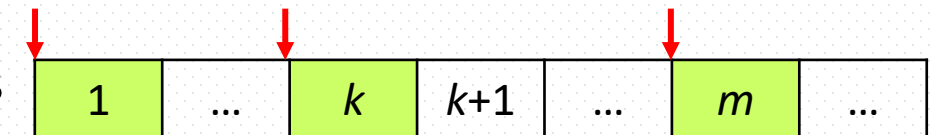
- **Sequential access pattern(顺序访问)**

- ▣ successive requests are for successive disk blocks
- ▣ disk seek required only for first block, e.g. #k block



- **Random access pattern(随机访问)**


- ▣ successive requests are for blocks that can be anywhere on disk
- ▣ each access requires a seek
- ▣ transfer rates are low since a lot of time is wasted in seeks



- **I/O operations per second (IOPS)**

- ▣ number of random block reads that a disk can support per second
- ▣ 50 to 200 IOPS on current generation magnetic disks

Performance Measures of Disks

- **Mean time to failure (MTTF, 平均无故障时间)** – the average time the disk is expected to run continuously without any failure.
 - typically 3 to 5 years 
 - probability of failure of new disks is quite low, corresponding to a “theoretical MTTF” of 500,000 to 1,200,000 hours for a new disk
 - e.g., an MTTF of 1,200,000 hours for a new disk means that given 1000 relatively new disks, on an average one will fail every 1200 hours
 - MTTF decreases as disk ages

Example 1

- Average I/O time =
 - ▣ average access time + (amount to transfer / transfer rate) + controller overhead
- For example to transfer a 4KB block on a 7200 RPM disk with a 5ms average seek time, 1Gb/sec transfer rate with a 0.1ms controller overhead
- The average access time is
 - ▣ average seek time + average rotation latency = 5ms + 4.17ms = 9.17ms
- Transfer time = amount to transfer / transfer rate =
$$4\text{KB} / 1\text{Gb/s} * 8\text{Gb} / \text{GB} * 1\text{GB} / 1024^2\text{KB} = 32 / (1024^2) = 0.031 \text{ ms}$$
- Average I/O time for 4KB block
$$= 9.17\text{ms} + 0.031\text{ms} + 0.1\text{ms} = 9.301\text{ms}$$

Spindle [rpm]	Average latency [ms]
4200	7.14
5400	5.56
7200	4.17
10000	3
15000	2

Example 2

21. 某磁盘的转速为 10 000 转/分，平均寻道时间是 6 ms，磁盘传输速率是 20 MB/s，磁盘控制器延迟为 0.2 ms，读取一个 4 KB 的扇区所需的平均时间约为
- A. 9 ms B. 9.4 ms C. 12 ms D. 12.4 ms

磁盘读操作步骤及各步骤时间

- step1. 平均地，磁头径向寻道时间，6ms
- step2. 盘片旋转，将目标扇区移动到磁头下——扇区查询/定位：
盘片旋转一圈耗时 $60 \times 1000 \text{ms} / 10000 = 6 \text{ms}$ ，
平均地，将读写头移到目标扇区所在位置，磁盘需要旋转半圈，故扇区查询时间为3ms
- step3. 磁头将4kB扇区数据从盘片读到I/O data-in寄存器时间（或者：将4KB数据从磁盘扇区读到内存buffer？），
 $4 \text{KB} \div \text{磁盘（内部）传输速率} 20 \text{MB/s} = 0.2 \text{ms}$
- step4. 磁盘控制器延时（driver读写I/O端口），0.2ms
- 读取4KB扇区的平均时间： $6 + 3 + 0.2 + 0.2 = 9.4 \text{ms}$

Disk Attachment

- Computers access disk storage in two ways, or the drive attached to computer
 - ▣ via local I/O ports, also known as host- attached storage
 - commonly used in small systems, e.g. PC
 - ▣ via a remote host in a distributed file system, referred to as
 - network-attached storage, storage-area network
- In host- attached mode, busses vary, including
 - ▣ IDE, ATA, SATA, USB, Fibre Channel, SCSI, SAS, Firewire
 - ▣ host controller in computer uses bus to talk to disk controller built into drive or storage array

Disk Attachment

- A network-attached storage (NAS) device is a special-purpose storage system that is accessed remotely over a data network
 - ▣ clients access network-attached storage via a **remote-procedure-call** interface such as NFS for UNIX systems or CIFS for Windows machines

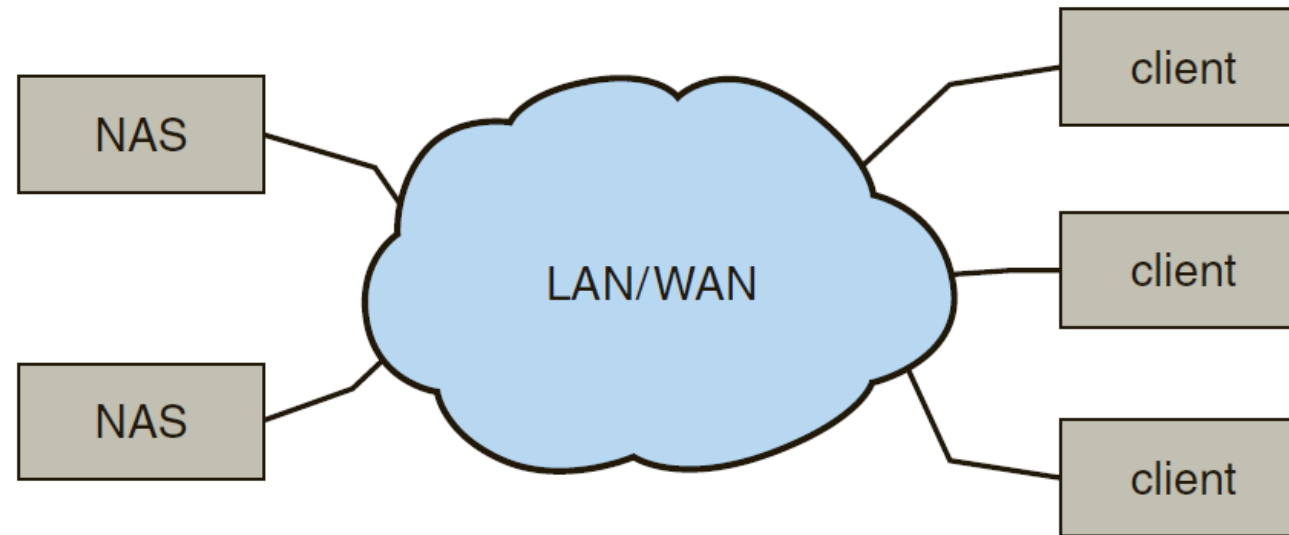


Figure 12.2 Network-attached storage.

Disk Attachment

- A storage-area network (SAN) is a private network (using storage protocols rather than networking protocols) connecting servers and storage units
 - ▣ the power of a SAN lies in its flexibility. Multiple hosts and multiple storage arrays can attach to the same SAN, and storage can be dynamically allocated to hosts.

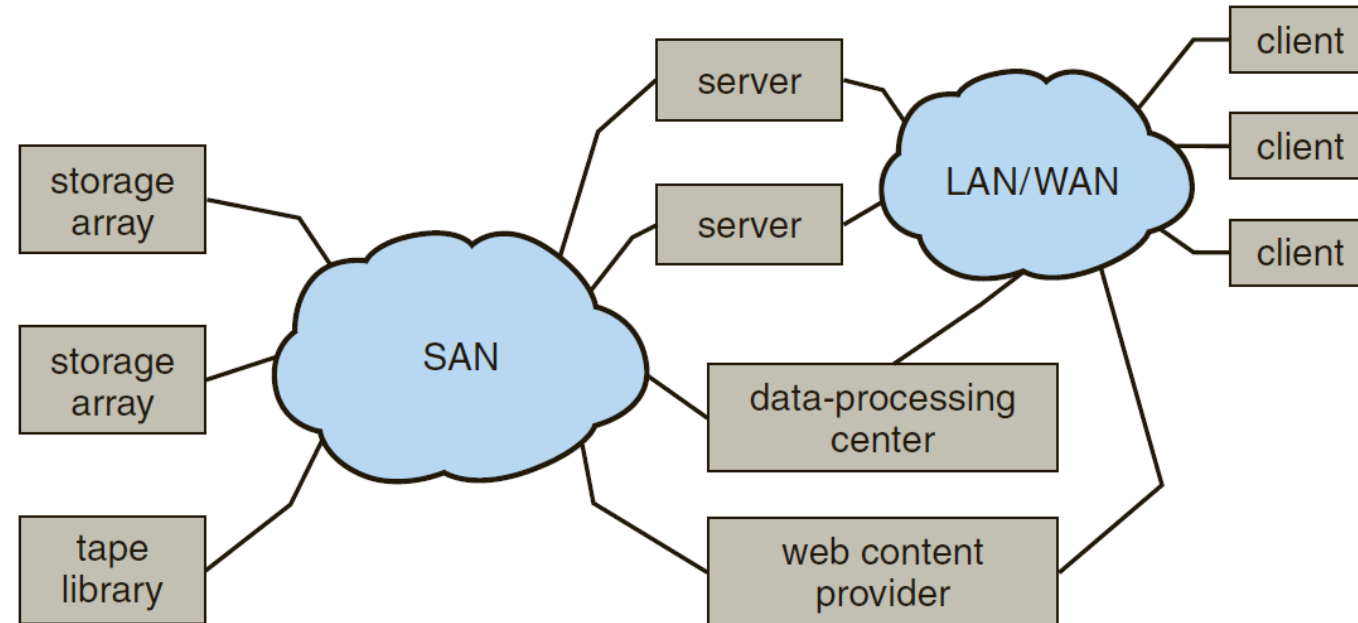


Figure 12.3 Storage-area network.

Solid-State Disks/SSD

- SSD, Solid State Disk, storing information in flash memory (闪存), two categories

- ▣ NOR flash
- ▣ NAND flash

- **NAND flash**

- ▣ used widely for storage, cheaper than NOR flash
- ▣ requires page-at-a-time read (page: 512 bytes to 4 KB)

//读操作：以page为单位读

- 20 to 100 microseconds (微秒) for a page read
- not much difference between sequential and random read

//顺序读、随机读的速度差别不大

- ▣ page can only be written once
 - must be **erased** to allow rewrite 【先擦后写】

//写操作：当向 a page中写入数据时，如果page中已有数据，需要先整体擦除page内数据，再向page内写入数据；可擦除次数有限(10w-100w)，严重影响使用寿命

SSD优缺点

■ 优点（与disk相比）

- 内部不存在机械运动部件，抗震性好，噪音低
- 低容量硬盘（< 1T）体积小，重量轻
- 相对固定的读取时间
 - 寻址时间与数据存储位置无关，磁盘碎片不会影响读取时间
 - 寻道时间为零；最小速度和最快速度差异很小，可获得更高的平均速度

■ 缺点

- 成本较高，容量不大，写入寿命有限
- 写入速度有限，相比传统硬盘无明显优势，易受到写入碎片的影响
 - 和顶级机械硬盘相比有一定差距
- 数据损坏后难以恢复，几乎不可能在失效、破碎或者被击穿的芯片中找回数据

微软试图在2023年之前放弃机械硬盘

<https://www.tomshardware.com/news/microsofts-reportedly-trying-to-kill-hdd-boot-drives-for-windows-11-pcs-by-2023>

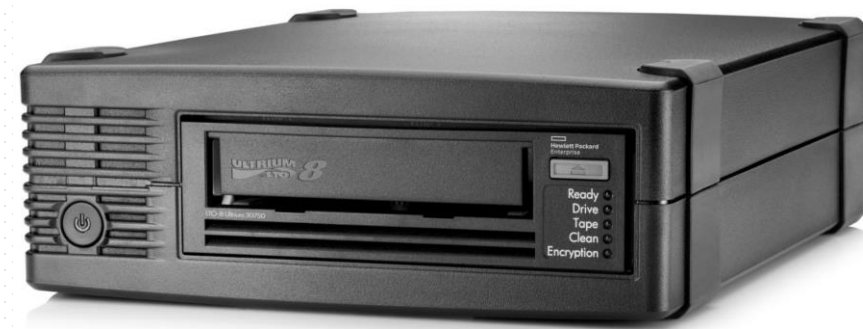
数据存储行业分析公司Trendfocus最近发布的一份高管简报报告称，一些OEM透露，微软正在推动他们放弃HDD作为预装Windows 11 PC的主要存储设备，转而使用SSD，目前的切换截止日期为2023年。

微软最新的硬件配置要求为 Windows 11 配备“64GB或更大的存储设备”，因此SSD并不是标准安装的最低要求。但是微软规定 DirectStorage 和 Windows Subsystem for Android这两个功能需要SSD，但您不必使用这些功能。

Magnetic Tape

- Was early secondary-storage medium
 - ▣ evolved from open spools (卷轴式) to cartridges (筒式)
- Relatively permanent and holds large quantities of data
- Sequential access, and access time is slow
 - ▣ random access ~1000 times slower than disk
- Mainly used for backup, storage of infrequently-used data, transfer medium between systems
- Kept in spool and wound or rewound past read-write head
- Once data under head, transfer rates comparable to disk
 - ▣ 140MB/sec and greater
- 200GB to 1.5TB typical storage
- Common technologies are LTO-{3,4,5} and T10000

1952年,
世界上第
一台磁带
机, IBM726



12.4 Disk Scheduling

- The operating system is responsible for guaranteeing disk drives having *a fast access time* and *disk bandwidth*
- Access time has two major components
 - seek time, $\text{seek time} \propto \text{seek distance}$
 - rotational latency
- For *several access requests*, disk bandwidth is the total number of bytes transferred, divided by the total time between the first request for service and the completion of the last transfer

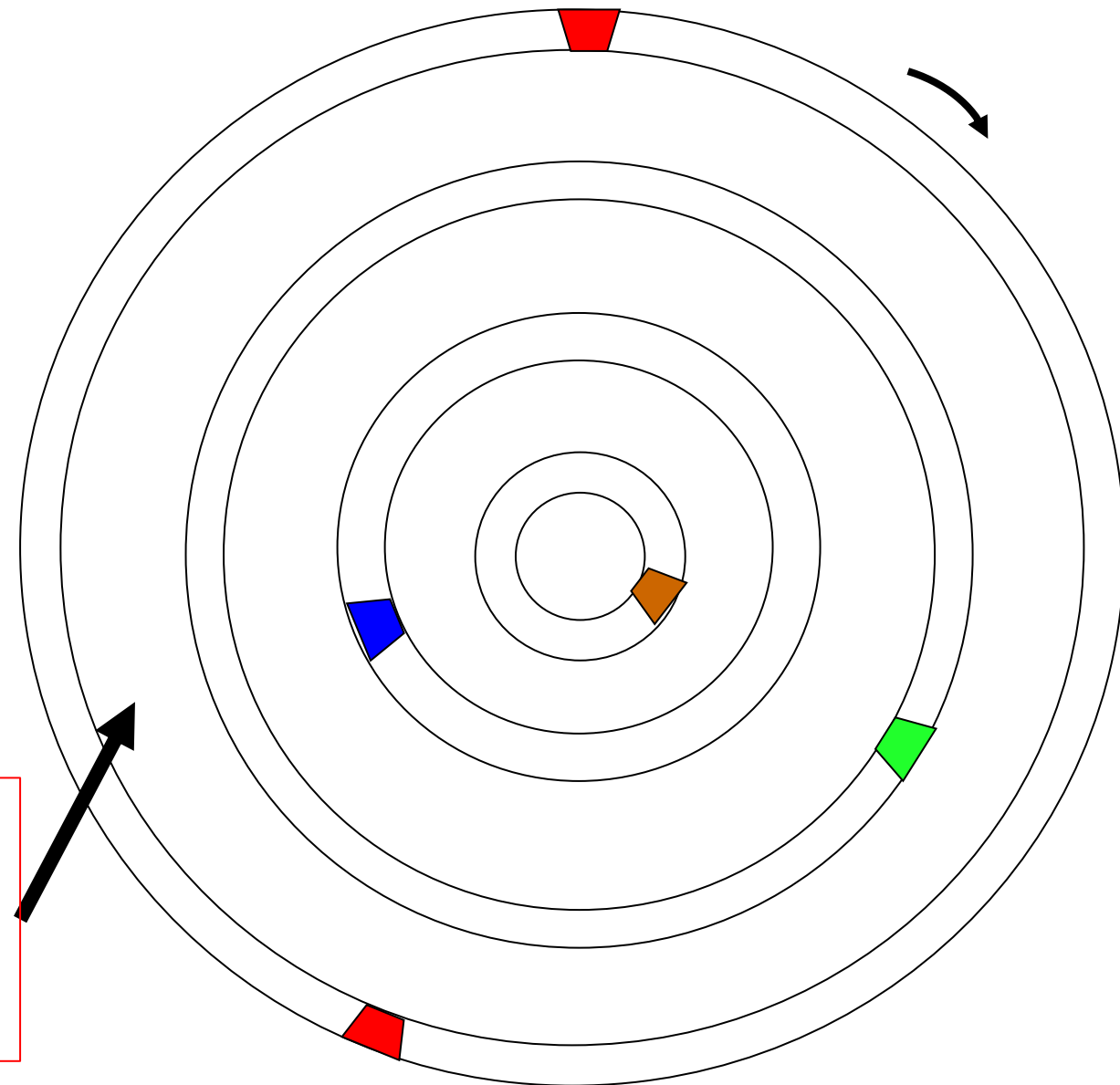
Disk Scheduling

- When a process needs I/O to or from the disk, it issues a system call to OS, specifying
 - ▣ whether this operator is input or output
 - ▣ what is the disk address for the data transfer is
 - ▣ what the memory address for the data transfer is
 - ▣ what the number of sectors to be transferred is
- When several I/O requests arrive at a disk drive and controller, the disk scheduling scheme selects one request to service, and the other requests are placed in the queue of pending requests for that drive
- Several algorithms exist to schedule the servicing of disk I/O requests

Disk Scheduling Queue

- A disk queue with requests for I/O to blocks on cylinder, i.e. a request cylinder queue
 - ▣ I/O request: 访问位于磁盘盘面不同磁道上的block
 - ▣ 磁盘（调度）队列 disk queue: I/O request需要访问的磁道号，即需要访问的数据block所在的磁道号
 - e.g. 98, 183, 37, 122, 14, 124, 65, 67
- Disk scheduling
 - ▣ arrange the disk arm to move to the cylinders in the queue, and on these cylinders the blocks to be input/output reside in

I/O request block#: 0, 30000, 8500, 130, 17500,
disk queue track/cylinder#: 0, 90, 30, 0, 60
块号block#→<柱面, 磁道, 扇区>



the disk head is now
at cylinder/track 10,
and moves to the
innermost of the
disk

Disk Scheduling

- We illustrate them with a request cylinder queue (0-199).

98, 183, 37, 122, 14, 124, 65, 67

Head pointer 53

- Scheduling algorithms
 - FCFS
 - SSTF
 - SCAN
 - C-SCAN
 - Look/C-Look

FCFS scheduling

- Fig. 12.4
 - ▣ illustration shows total head movement of **640** cylinders

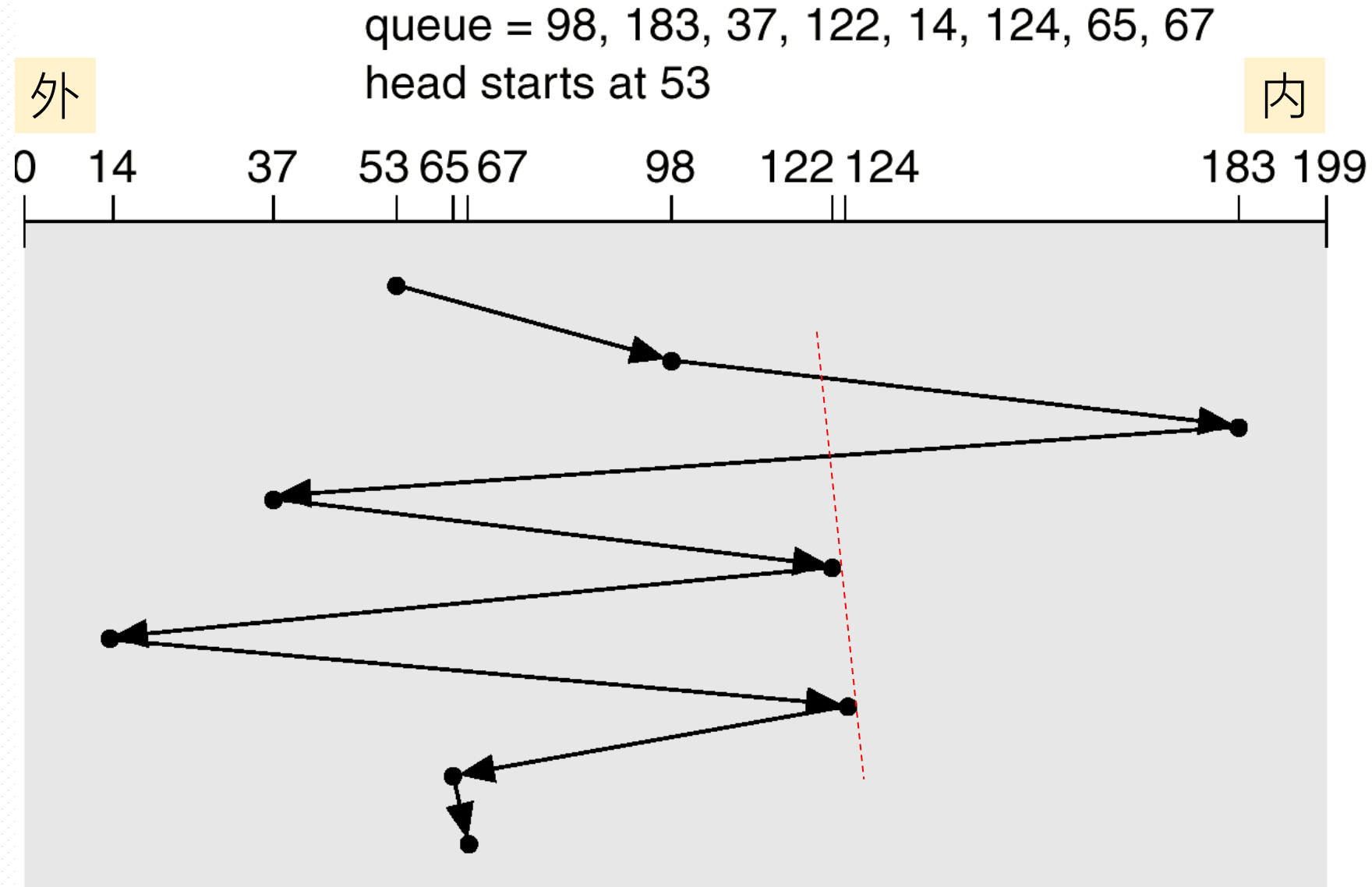


Fig. 12.4 FCFS disk scheduling

SSTF Scheduling

- Selects the request with the minimum seek time from the current head position
 - ❑ shortest-seek-time-first (最短寻道时间优先)
- SSTF scheduling is a form of SJF scheduling
 - ❑ may cause starvation of some requests
- In Fig.12.5, total head movement of **236** cylinders

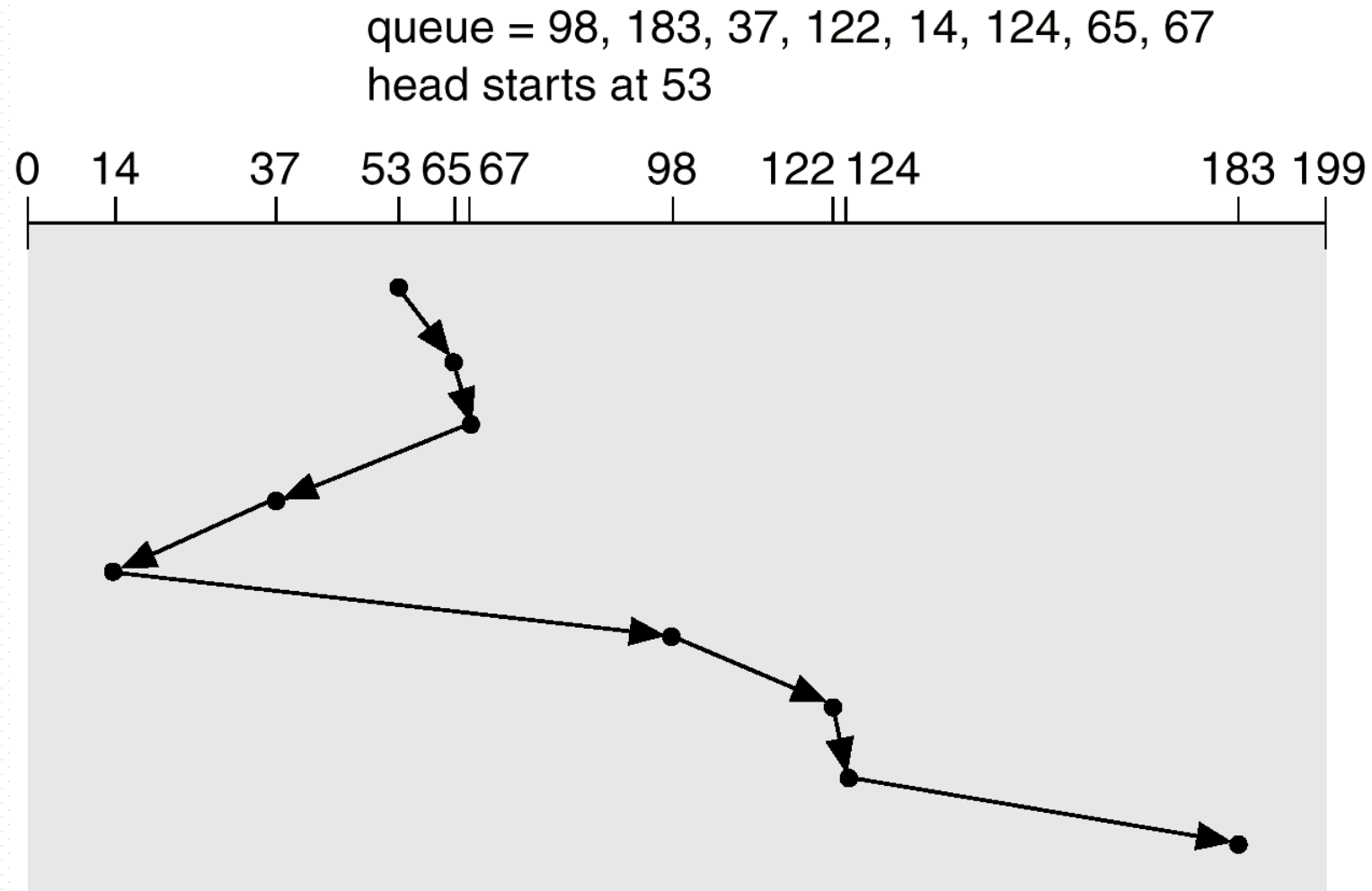


Fig. 12.5 SSTF disk scheduling

SCAN Scheduling

- The disk arm starts at one end of the disk (*outermost*, *innermost?*), 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

- 减少磁臂反复换向运动

- Also called the *elevator algorithm*

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53

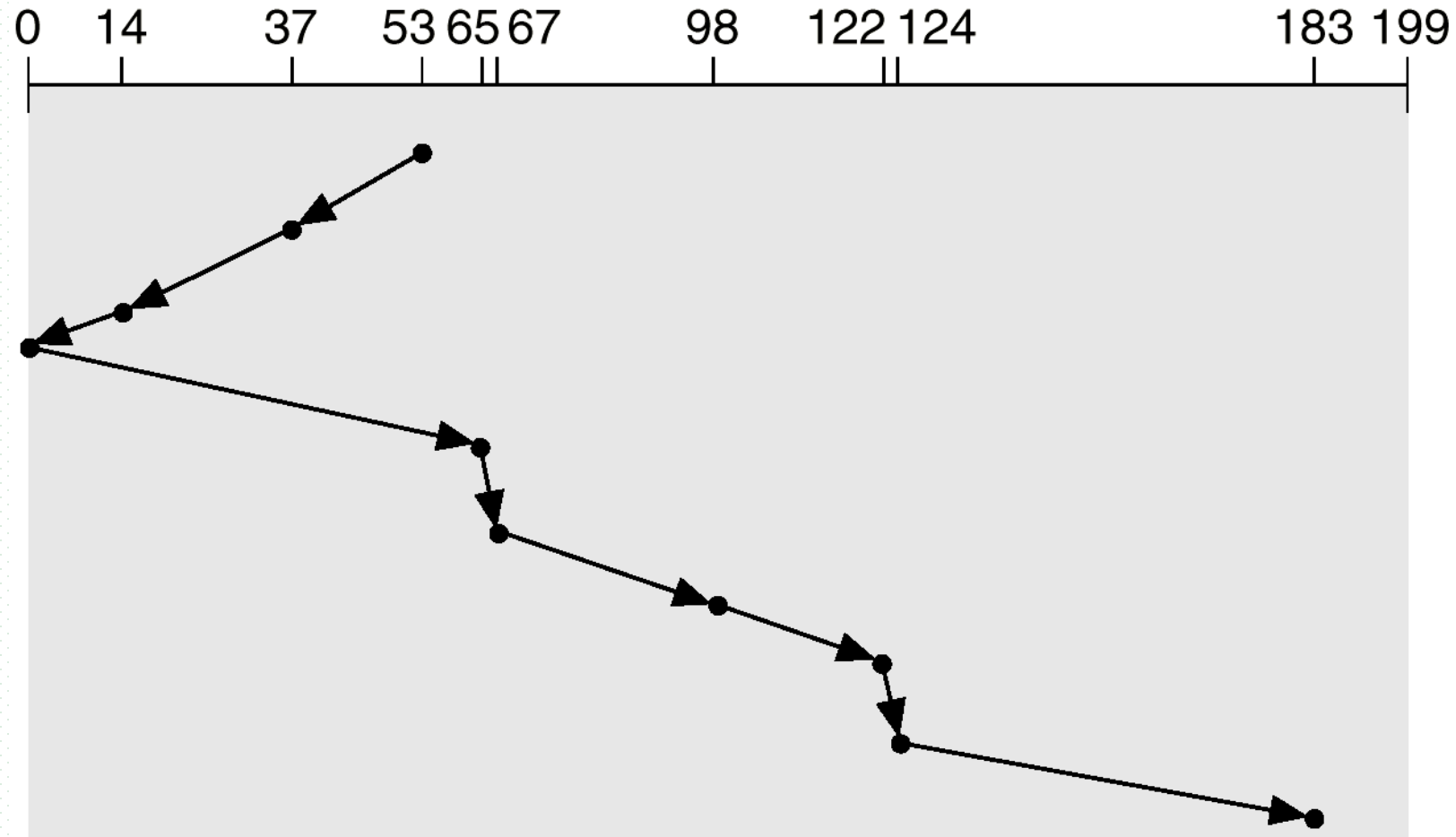


Fig. 12.6 SCAN disk scheduling

C-SCAN

- 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.
- Provides a more uniform wait time than SCAN

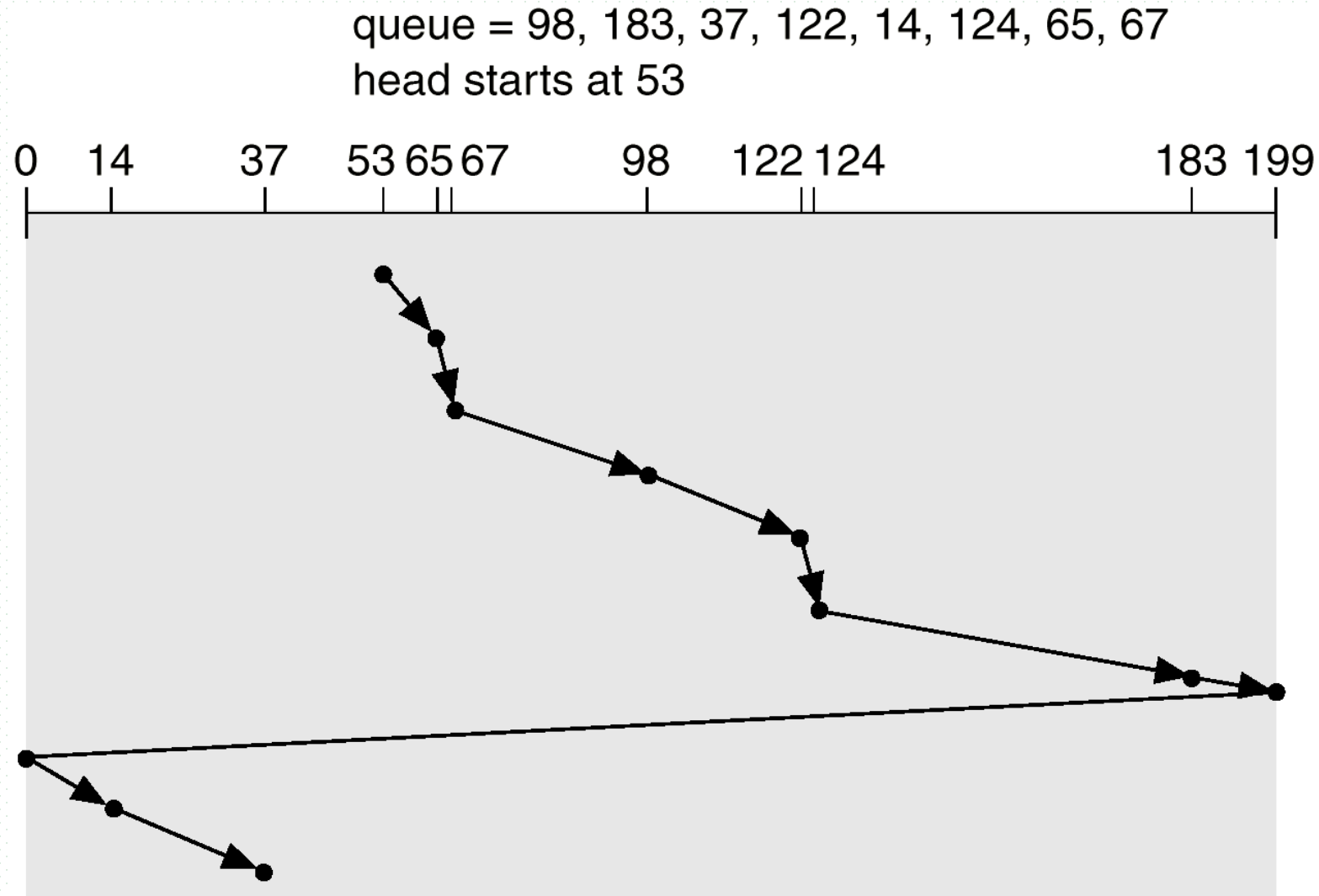


Fig. 12.7 C-SCAN disk scheduling

Look/C-Look

- Arm only goes **as far as the final request** in each direction, then reverses direction immediately, without first going all the way to the end of the disk
- Version of SCAN and C-SCAN following this patterns are called Look scheduling and C-Look scheduling
- Fig.12.8

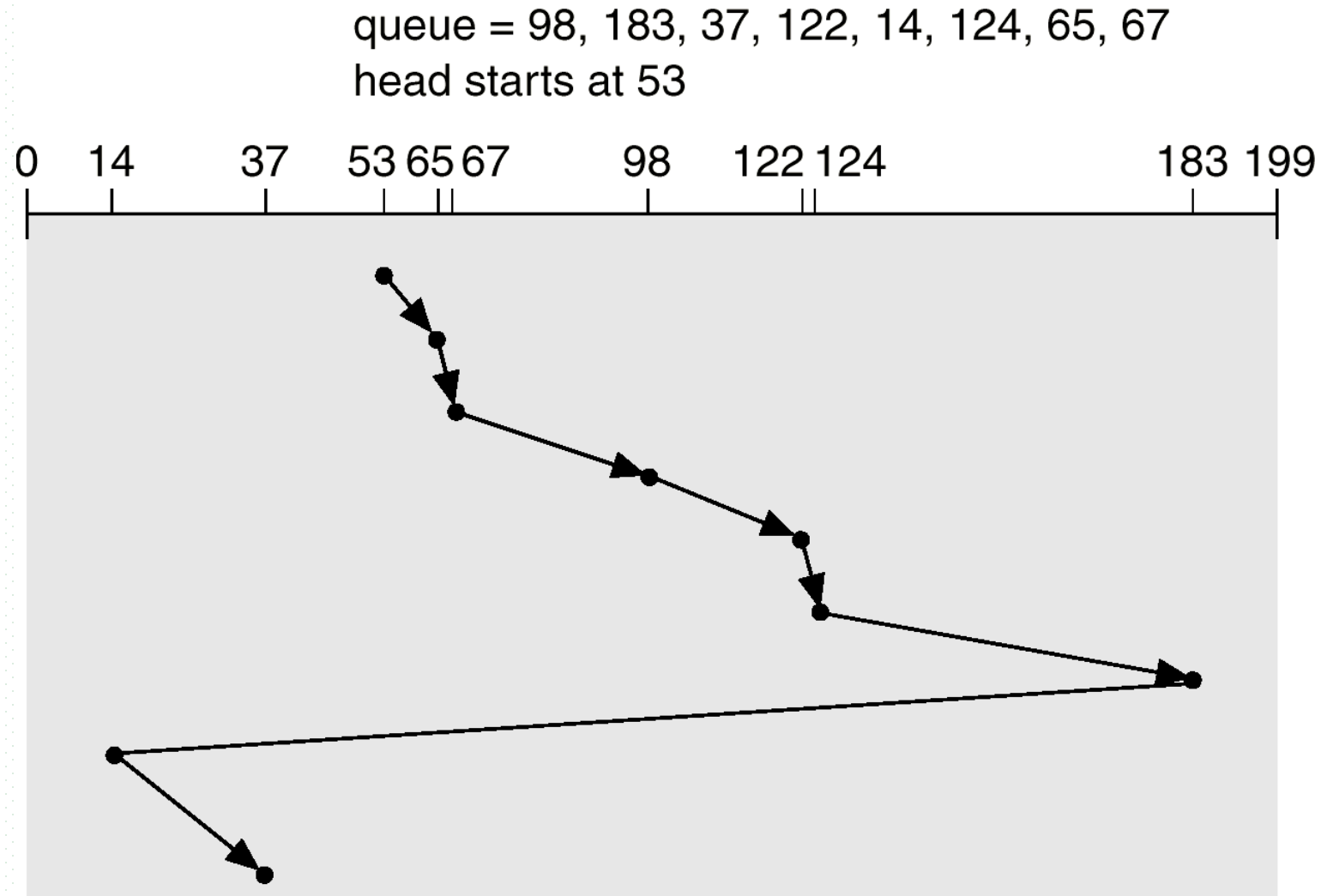


Fig.12.8 C-Look disk scheduling

Example 4

- Suppose that a disk drive has 200 cylinders, numbered 0 to 199. The queue of pending requests, in FIFO order, is
96,182,36,112,14,125,64,98
- Starting from the current head position, the SCAN algorithm is used to schedule the request queue, no matter which direction the disk head moves, the schedule algorithm results in a same total distance (in cylinders)
- Which position does the current head position? Why?
- Answer
 - 假设磁头所处的柱面编号为 X ，根据SCAN算法的调度策略，分两种情况讨论磁头所移动的总柱面数
 - 磁头向左移动，则磁头移动的总距离为： $X+182$
 - 磁头向右移动，则磁头移动的总距离为： $199-X+199-14$
 - 根据题意，有 $X+182=199-X+199-14$ ，解得 $X=101$
 - 即服务开始前磁头所处的位置是101

Selection of a Disk-Scheduling Algorithm

- SSTF is common and has a natural appeal
- SCAN and C-SCAN perform better for systems that place a heavy load on the disk.
- **Performance depends on the number and types of requests**
- Requests for disk service can be influenced by the file-allocation method
- 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
- Either **SSTF** or **LOOK** is a reasonable choice for the default algorithm

Disk Scheduling in SSD

- The disk-scheduling algorithms discussed in this section focus primarily on minimizing the amount of disk head movement in magnetic disk drives
- SSDs— which do not contain moving disk heads— commonly use a simple **FCFS** policy
 - e.g. the Linux Noop scheduler uses an FCFS policy but modifies it to merge adjacent requests
 - initial block access requests :
0, 10, 22, 46, 181, 150, 23, 182, 180, 199, 1
 - merged FCFS requests:
0, 1, 10, 22, 23, 46, 180, 181, 182, 150, 199
- The observed behavior of SSDs indicates that
 - the time required to service *reads* is uniform
//读取SSD中不同位置的数据块，读取时间基本一致
 - *write* service time is not uniform, due to the properties of flash memory

12.5 Disk Management

- Formatting
- Booting from disk
- Bad-block recovery

Formatting

- Initializing disks by formatting
 - ▣ physical formatting, or low-level formatting
 - ▣ logical formatting, or high-level formatting
- Low-level formatting, or physical formatting
 - ▣ dividing a disk into sectors that the disk controller can read and write
//划分硬盘的磁柱面、建立扇区、选择扇区间隔比
 - ▣ fills the disk with a special data structure for each sector, and the structure consists of
 - a header, a data area (usually 512 bytes in size), and a trailer
 - the header and trailer contain information used by the disk controller, such as a sector number and an error-correcting code (ECC) for read and write operations
- 出厂硬盘已经过低级格式化，一般无须用户再进行低级格式化；何时需要低级格式化？
 - ▣ 硬盘坏道太多，经常导致存取数据错误，甚至OS根本无法使用；
 - ▣ 硬盘上某些和低级格式化有关的参数被病毒等破坏，如硬盘间隔系数等，只能进行低级格式化重新建立这些参数

Formatting

- To use a disk to hold files, the operating system still needs to record its own data structures on the disk by two steps
 - ▣ step1. partition the disk into one or more groups of cylinders
 - 物理盘片划分为逻辑分区 (logical partition)
 - ▣ step2. Logical formatting, high-level formatting
 - makes a file system on the partitions that become volumes, stores the initial data structures onto the disk, including, e.g. maps of free and allocated space and an initial empty directory

Boot Block

- The bootstrap program is responsible for initializing all aspects of the system
 - ▣ from CPU registers to device controllers and the contents of main memory, and then starts the operating system
 - ▣ to do its job, the bootstrap program finds the operating system kernel on disk, loads that kernel into memory, and jumps to an initial address to begin the operating-system execution
- Most computers store a tiny **bootstrap loader program** in the boot ROM whose only job is to read into memory a full **bootstrap program** from disk
 - ▣ the full bootstrap program is stored in the “**boot blocks**” at a fixed location on the disk
 - ▣ a disk that has a boot partition is called a boot disk or system disk
- The full bootstrap program is able to load **the entire OS kernel** from a non-fixed location on disk and to start the operating system running

12.6 Swap-Space Management

■ Swap-space

- ❑ a part of disk space used as an extension of main memory to implement virtual memory, as a raw partition

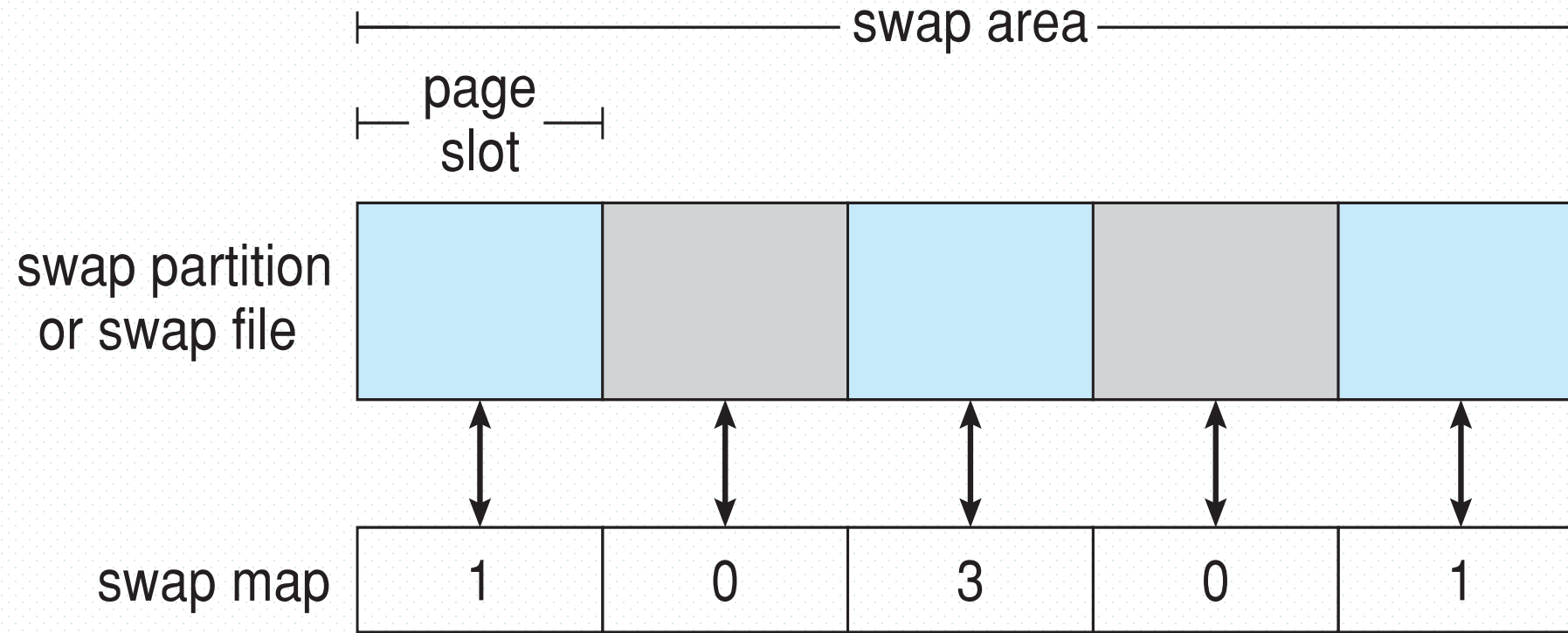
■ Swap-space location

- ❑ can be carved out of the normal file system
- ❑ more commonly, can be in a separate disk partition

■ Swap-space management

- ❑ BSD 4.3 allocates swap space when process starts; holds *text segment* (the program) and *data segment*
- ❑ kernel uses *swap maps* to track swap-space use

Data Structures for Swapping on Linux Systems



12.7 RAID Structure

- RAID
 - Redundant Arrays of Independent Disks, 独立磁盘冗余阵列
 - disk organization techniques that manage a large numbers of disks, providing a view of a single disk of
 - **high capacity** and **high speed** by using multiple disks in parallel,
 - **high reliability** by storing data redundantly, so that data can be recovered even if a disk fails
 - the hard **disk array** is governed by array management software and disk controller, which handles the error correction
 - RAID is generally used on network servers
 - //利用**磁盘阵列控制器**统一管理和控制一组磁盘驱动器, 组成速度快、可靠性高、性能价格比好的大容量磁盘系统, 填补快速CPU与慢速磁盘设备之间的间隙

Improvement of Reliability via Redundancy

- **Redundancy** – store extra information that can be used to rebuild information lost in a disk failure
- E.g., **Mirroring** (or **shadowing**)
 - ▣ duplicate every disk. Logical disk consists of two physical disks.
 - ▣ every write is carried out on both disks
 - Reads can take place from either disk
 - ▣ if one disk in a pair fails, data still available in the other
 - Data loss would occur only if a disk fails, and its mirror disk also fails before the system is repaired

Probability of combined event is very small
Except for dependent failure modes such as fire or building collapse or electrical power surges
- **Mean time to data loss** depends on mean time to failure, and **mean time to repair**
 - ▣ e.g., MTTF of 100,000 hours, mean time to repair of 10 hours gives mean time to data loss of 500×10^6 hours (or 57,000 years) for a mirrored pair of disks (ignoring dependent failure modes)

Improvement in Performance via Parallelism

- Two main goals of parallelism in a disk system:
 - ▣ 1. Load balance multiple small accesses to increase throughput
 - ▣ 2. Parallelize large accesses to reduce response time.
- Improve transfer rate by striping data across multiple disks
- **Bit-level striping** – split the bits of each byte across multiple disks
 - ▣ In an array of eight disks, write bit i of each byte to disk i .
 - ▣ Each access can read data at eight times the rate of a single disk.
 - ▣ But seek/access time worse than for a single disk
 - Bit level striping is not used much any more
- **Block-level striping** – with n disks, block i of a file goes to disk $(i \bmod n) + 1$
 - ▣ Requests for different blocks can run in parallel if the blocks reside on different disks
 - ▣ A request for a long sequence of blocks can utilize all disks in parallel

RAID Levels

- Schemes to provide redundancy at lower cost by using disk striping (磁盘分条) combined with parity bits
 - ▣ different RAID organizations, or RAID levels, have differing cost, performance and reliability characteristics
- **RAID Level 0:** Block striping; non-redundant.
 - ▣ used in high-performance applications where data loss is not critical.
- **RAID Level 1:** Mirrored disks with block striping
 - ▣ offers best write performance.
 - ▣ popular for applications such as storing log files in a database system



(a) RAID 0: nonredundant striping



(b) RAID 1: mirrored disks

RAID Levels

- **RAID Level 5: Block-Interleaved Distributed Parity** (块交错分布奇偶校验结构) ; partitions data and parity among all $N + 1$ disks, rather than storing data in N disks and parity in 1 disk.
 - ▣ E.g., with 5 disks, parity block for n th set of blocks is stored on disk $(n \bmod 5) + 1$, with the data blocks stored on the other 4 disks
 - ▣ Block writes occur in parallel if the blocks and their parity blocks are on different disks

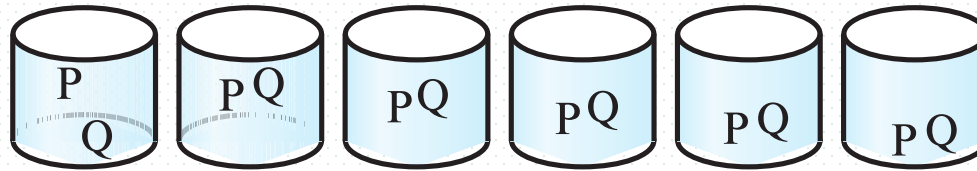


(c) RAID 5: block-interleaved distributed parity

P0	0	1	2	3
4	P1	5	6	7
8	9	P2	10	11
12	13	14	P3	15
16	17	18	19	P4

RAID Levels

- **RAID Level 6: P+Q Redundancy** scheme; similar to Level 5, but stores two error correction blocks (P, Q) instead of single parity block to guard against multiple disk failures.
 - ▣ Better reliability than Level 5 at a higher cost
 - Becoming more important as storage sizes increase



(d) RAID 6: P + Q redundancy

RAID Levels

- Other levels (**not used in practice**):
 - ▣ **RAID Level 2:** Memory-Style Error-Correcting-Codes (ECC) with bit striping.
 - ▣ **RAID Level 3:** Bit-Interleaved Parity
 - ▣ **RAID Level 4:** Block-Interleaved Parity; uses block-level striping, and keeps a parity block on a separate ***parity disk*** for corresponding blocks from N other disks.
 - RAID 5 is better than RAID 4, since with RAID 4 with random writes, parity disk gets much higher write load than other disks and becomes a bottleneck

- 高档主板可直接实现RAID功能

- RAID卡

- ▣ 实现RAID功能的板卡，由I/O处理器、硬盘控制器、硬盘连接器和缓存等构成
- ▣ 将多个磁盘/固态硬盘驱动器在逻辑上组织为一个磁盘驱动器，管理多个磁盘驱动器同时传输数据

LSI MegaRAID SAS 8708EM2 ¥ 2000



主要性能

RAID功能	RAID 0、1、5 和 6
接口 ⓘ	Serial SATA/SAS
内置接口	8
数据传输率 ⓘ	3Gb/s
最多连接设备	32个
插槽类型 ⓘ	PCI-E

练习题1

12.2 Suppose that a disk drive has 5000 cylinders, numbered 0 to 4999. The drive is currently serving a request at cylinder 143, and the previous request was at cylinder 125. The queue of pending requests, in FIFO order, is

86, 1470, 913, 1774, 948, 1509, 1022, 1750, 130

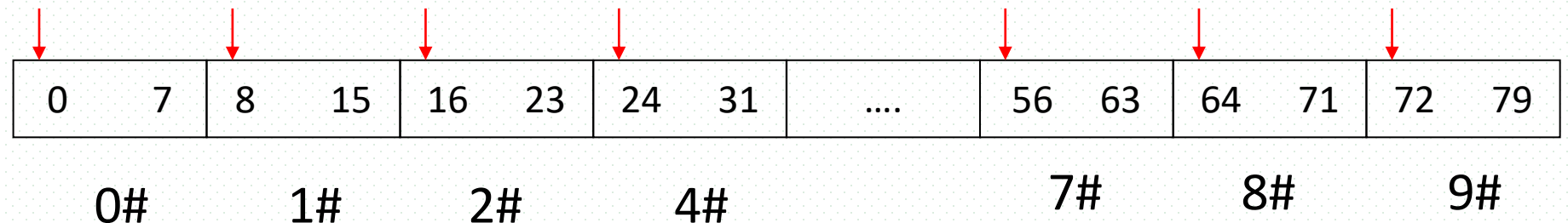
Starting from the current head position, what is the total distance (in cylinders) that the disk arm moves to satisfy all the pending requests, for each of the following disk-scheduling algorithms?

- a. FCFS
- b. SSTF
- c. SCAN
- d. LOOK
- e. C-SCAN

练习题2

- A file is made up of 128-byte fix-sized logical records and stored on the disk in the unit of the block that is of 1024 bytes. The size of the file is 10240 bytes
- Physical I/O operations transfer data on the disk into an OS buffer in main memory, in terms of 1024-byte block
- If a process issues *read* requests to read the file's records in the sequential access manner, what is the percentage of the *read* requests that will result in I/O operations?
 - the file is of $10240/128=80$ records
 - a block can hold $1024/128=8$ records
 - the file is stored in $10240/1024=10$ blocks
 - percentage = $10/80=1/8=0.125$

注意：每个记录对应一个I/O请求，每8个I/O请求导致1个实际I/O操作





Thanks for your
attention



北京邮电大学