Mass-Storage Systems

Chapter12: Secondary Storage Structure

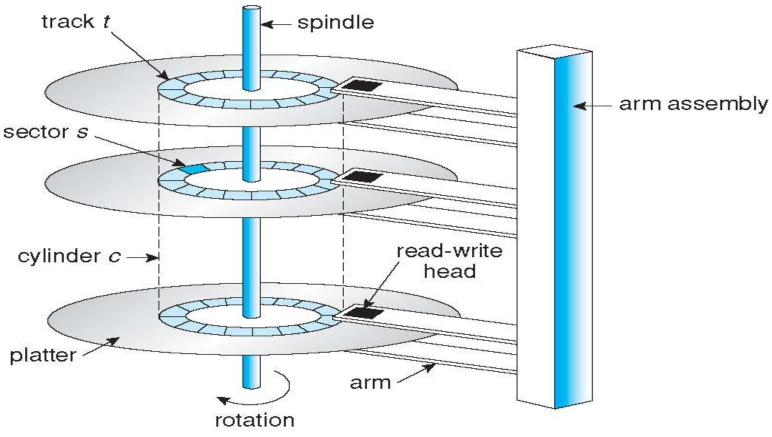
- Overview of Mass Storage Structure
- Disk Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management

Objectives

- Describe the physical structure of secondary and tertiary storage devices and the resulting effects on the uses of the devices
- Explain the performance characteristics of mass-storage devices

Overview of Mass Storage Structure

Magnetic disks



- Magnetic disk provide the bulk of secondary storage for modern computer systems.
- Each **disk platter** has a flat circular shape, like a CD. Common platter diameters range from 1.8 to 5.25 inches. The two surfaces of a platter are covered with a magnetic material.
- Information is stored by recording it magnetically on the platters.
- A **read -write head** moves above each surface of every platter.
- The heads 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.**

There may be thousands of concentric cylinders in a disk drive, and each track may contain hundreds of sectors. The storage capacity of common disk drives is measured in gigabytes.

- When the disk is in use, a drive motor spins it at high speed. Most drives rotate 60 to 200 times per second.
- Disk speed has two parts :
  a. The **transfer rate** is the rate at which data flow between the drive and the computer.
  b. The **positioning time** also called random access time consists of the time necessary to move the disk arm to the desired cylinder, called the seek time.
- **Rotational latency** time necessary for the desired sector to rotate to the disk head**.**
- Disk platter has magnetic surface which may be damaged sometimes by the head called **head crash.**

Magnetic tape

- Was early secondary-storage medium
- Relatively permanent and holds large quantities of data
- Access time is slow compared to main memory and secondary storage.
- Random access ~1000 times slower than disk
- Mainly used for backup, storage of infrequently-used data and as a transfer medium between systems.

Disk Structure

- Modern disk drives are addressed as large one-dimensional arrays of logical blocks, where the logical 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 using this mapping, a logical block number can be converted into an old-style disk address that consists of a cylinder number, a track number within that cylinder, and a sector number within that track.

Disadvantages :

1. Most disks have some defective sectors, but the mapping hides this by substituting spare sectors

from elsewhere on the disk.

2. The number of sectors per track is not constant on some drives.

- The number of sectors per track has been increasing as disk technology improves .
- The outer zone of a disk usually has several hundred sectors per track.
- Time taken to access sector or track in outer zone is more compared to time taken to access sector or track in inner zone.
- Number of cylinders per disk varies.

Disk Attachment

- Host-attached storage is storage accessed through local I/0 ports. These ports use several technologies.
- Desktop PC uses an I/0 bus architecture called IDE or ATA. This architecture supports a maximum of two drives per I/0 bus. Protocol that has simplified cabling is SATA.
- High-end workstations and servers generally use more sophisticated I/0 architectures, SCSI and fiber channel (FC).

SCSI is a bus architecture whose physical medium is a ribbon cable with a large number of conductors (typically 50 or 68).

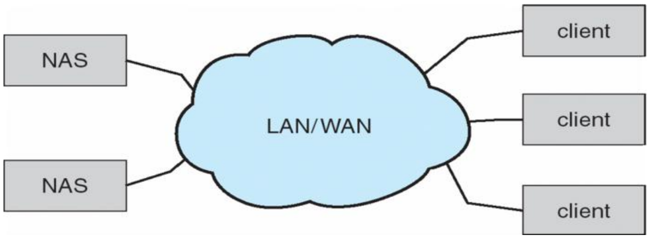
- The SCSI protocol supports a maximum of 16 devices per bus. The devices include one controller card in the host (the and up to 15 storage devices .
- A SCSI disk is a common SCSI target, but the protocol provides the ability to address up to 8 logical units in each SCSI target.

Fibre channel (FC) is a high-speed serial architecture that can operate over optical fibre or over a four-conductor copper cable.
It has two variants:

- One is a large switched fabric having a 24-bit address space. This variant is expected to dominate in the future and is the basis of Storage area networks.
- The other FC variant is an arbitrated loop that can address 126 devices (drives and controllers).

Network-Attached Storage

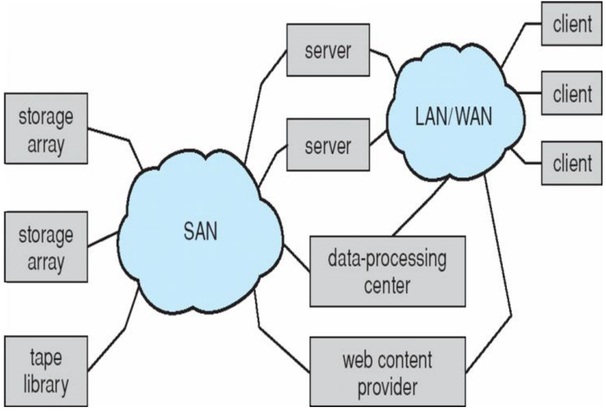
- A network-attached storage (NAS) device is a special-purpose storage system that is accessed remotely over a data network.
- Clients access network-attached storage via a remote-procedure-call interface such as NFS for UNIX systems or CIFS for Windows machines.
- The remote procedure calls(RPCs) are carried via TCP or UDP over an IP network- usually the same local-area network (LAN) that carries all data traffic to the clients.
- The network attached storage unit is usually implemented as a RAID array with software that implements the RPC interface.
- Network-attached storage (**NAS**) is storage made available over a network rather than over a local connection (such as a bus).



Disadvantage : One drawback of network-attached storage systems is that the storage I/O operations consume bandwidth on the data network, thereby increasing the latency(delay) of network communication.

Storage Area Network

- A storage-area network (SAN) is a private network 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.
- A SAN switch allows or prohibits access between the hosts and the storage. Example, if a host is running low on disk space, the SAN can be configured to allocate more storage to that host.
- SANs make it possible for clusters of servers to share the same storage and for storage arrays to include multiple direct host connections.
- SANs typically have more ports, and less expensive ports, than storage arrays.



Disk Scheduling

The operating system is responsible for using hardware efficiently — for the disk drives, this means having a fast access time and disk bandwidth.

Access time has two major components

- **Seek time** is the time for the disk are to move the heads to the cylinder containing the desired sector
- **Rotational latency** is the additional time waiting for the disk to rotate the desired sector to the disk head
- Minimize seek time

Seek time ≈ seek distance

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

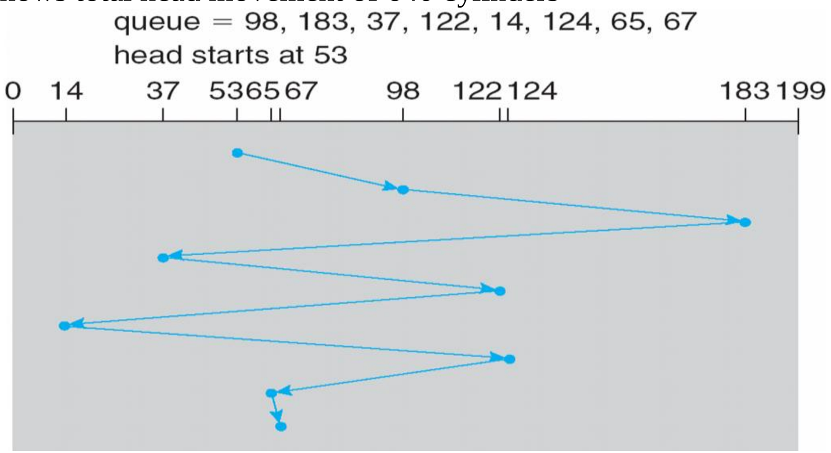
Several algorithms exist to schedule the servicing of disk I/O requests

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

98, 183, 37, 122, 14, 124, 65, 67
Head pointer 53

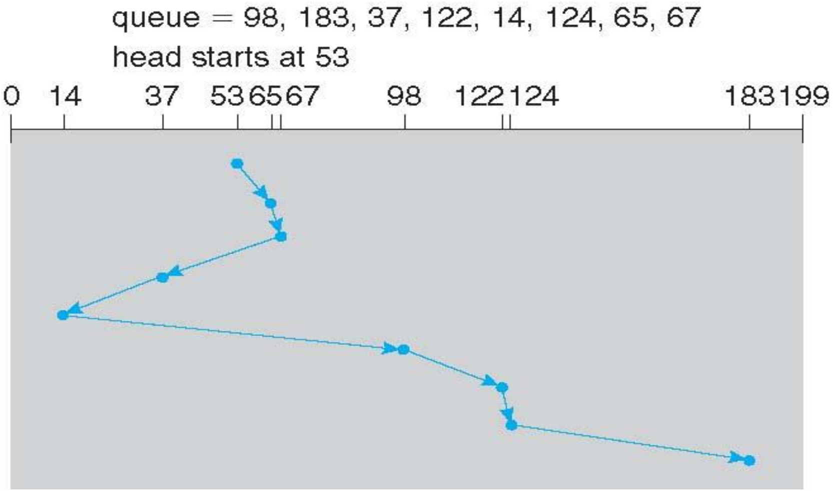
FCFS

- Illustration shows total head movement of 640 cylinders



SSTF

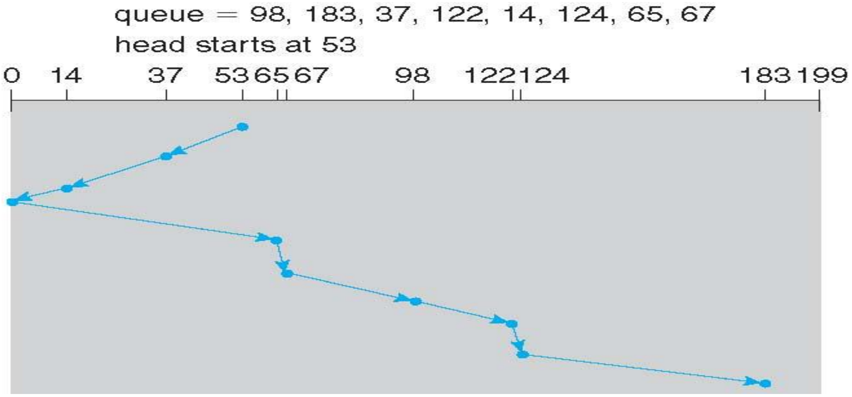
- Selects the request with the minimum seek time from the current head position
- SSTF scheduling is a form of SJF scheduling; may cause starvation of some requests
- Illustration shows total head movement of 236 cylinders



SCAN

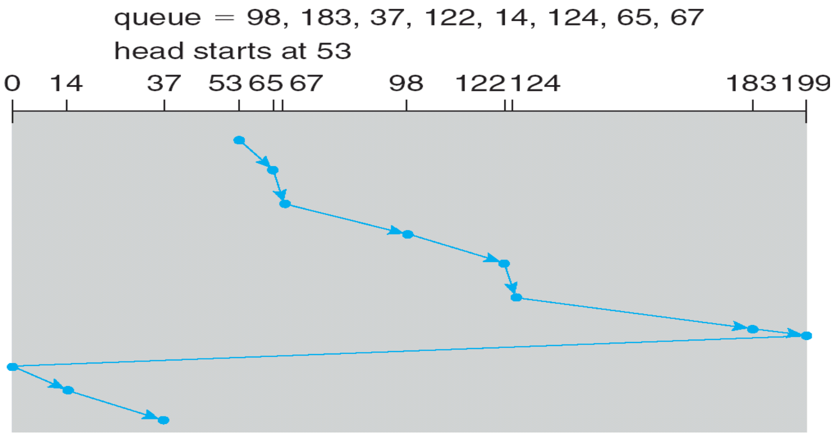
- The disk arm starts at one end of the disk, and moves toward the other end, servicing requests until it gets to the other end of the disk, where the head movement is reversed and servicing continues.
- SCAN algorithm sometimes called the elevator algorithm

- Illustration shows total head movement of 208 cylinders



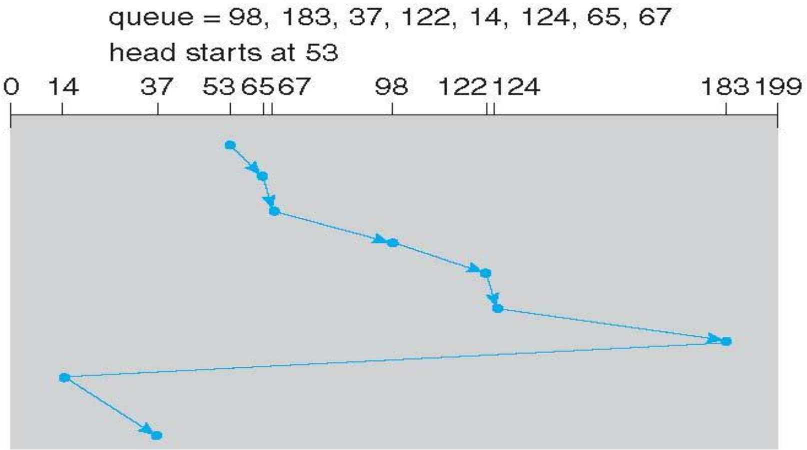
C_SCAN

- Provides a more uniform wait time than SCAN
- The head moves from one end of the disk to the other, servicing requests as it goes forward.
- When it reaches the other end, however, it immediately returns to the beginning of the disk, without servicing any requests on the return trip.
- Treats the cylinders as a circular list that wraps around from the last cylinder to the first one



C-LOOK

- Version of C-SCAN
- 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



Selecting a Disk-Scheduling Algorithm

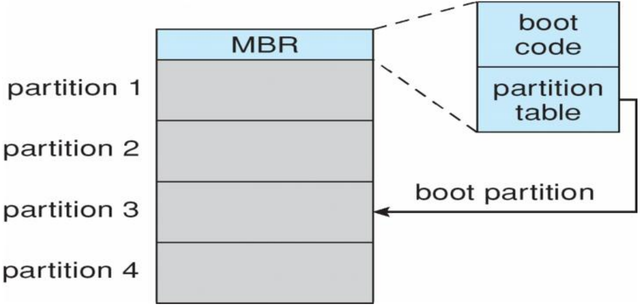
- SSTF is common and has a natural appeal.
- SCAN and C-SCAN perform better for systems that place a heavy load on the disk.
- Performance depends on the number and types of requests.
- Requests for disk service can be influenced by the file-allocation method.
- The disk-scheduling algorithm should be 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 Formatting

- Before a disk can store data, it must be divided into sectors that the disk controller can read and write. This process is called physical formatting or Low-level formatting .
- Low-level formatting fills the disk with a special data structure for each sector. The data structure for a sector typically 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 (ECC).
- When the controller writes a sector of data during normal I/0, the ECC is updated with a value calculated from all the bytes in the data area.
- When the sector is read, the ECC is recalculated and compared with the stored value.
- If the stored and calculated numbers are different, this mismatch indicates that the data area of the sector has become corrupted and that the disk sector may be bad.
- The ECC is an *error-correcting* code because it contains enough information, if only a few bits of data have been corrupted, to enable the controller to identify which bits have changed and calculate what their correct values should be. It then reports a recoverable soft error.
- The controller automatically does the ECC processing whenever a sector is read or written.
- Before it can use a disk to hold files, the operating system still needs to record its own data structures on the disk. It does so in two steps :

- The first step is to partition the disk into one or more groups of cylinders. The operating system can treat each partition as though it were a separate disk. For instance, one partition can hold a copy of the operating system's executable code, while another holds user files.
- Second step is logical formatting or creation of a file system. In this step, the operating system stores the initial file-system data structures onto the disk. These data structures may include maps of free and allocated space (a FAT or inodes) and an initial empty directory.

Boot Block :

- When a computer is powered up or rebooted -it must have an initial program to run called *bootstrap* Program.
- It initializes all aspects of the system, from CPU registers to device controllers and the contents of main memory, and then starts the operating system.
- Operating-system kernel on disk, loads that kernel into memory, and jumps to an initial address to begin the operating-system execution.
- The bootstrap is stored in Read Only Memory (ROM).
- ROM needs no initialization and is at a fixed location that the processor can start executing when powered up or reset.
- ROM is read only, hence it cannot be infected by a computer virus.
- The problem is that changing this bootstrap code requires changing the ROM hardware chips.
- For this reason, most systems store a tiny bootstrap loader program in the boot ROM whose only job is to bring in a full bootstrap program from disk.
- The full bootstrap program can be changed easily: a new version is simply written onto the 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.
- 

Consider as an example the boot process in Windows 2000. The Windows 2000 system places its boot code in the first sector on the hard disk. Furthermore, Windows 2000 allows a hard disk to be divided into one or more partitions; one partition, identified as the boot partition, contains the operating system and device drivers. Booting begins in a Windows 2000 system by running code that is resident in the system's ROM memory. This code directs the system to read the boot code from the MBR. In addition to containing boot code, the MBR contains a table listing the partitions for the hard disk and a flag indicating which partition the system is to be booted from, as illustrated in above Figure. Once the system identifies the boot partition, it reads the first sector from that partition and continues with the remainder of the boot process, which includes loading the various subsystems and system services.

Bad Blocks

- Disks have moving parts and small tolerances, they are prone to failure.
- The disk needs to be replaced and its contents restored from backup media to the new disk.
- One or more sectors become defective.

Bad-block recovery

The controller maintains a list of bad blocks on the disk. The list is initialized during the low-level formatting at the factory and is updated over the life of the disk. Low-level formatting also sets aside spare sectors not visible to the operating system. The controller can be told to replace each bad sector logically with one of the spare sectors. This scheme is known as sector sparing or forwarding.
A typical bad-sector transaction might be as follows:

- The operating system tries to read logical block 87.
- The controller calculates the ECC and finds that the sector is bad. It reports this finding to the operating system.
- The next time the system is rebooted, a special command is run to tell the SCSI controller to replace the bad sector with a spare.
- After that, whenever the system requests logical block 87, the request is translated into the replacement sector's address by the controller.

Sector slipping :

Example: Suppose that logical block 17 becomes defective and the first available spare follows sector 202. Then, sector slipping remaps all the sectors front 17 to 202, moving them all down one spot. That is, sector 202 is copied into the spare, then sector 201 into 202, then 200 into 201, and so on, until sector 18 is copied into sector 19. Slipping the sectors in this way frees up the space of sector 18, so sector 17 can be mapped to it.

Swap-Space Management

- Swapping in that setting occurs when the amount of physical memory reaches a critically low point and processes are moved from memory to swap space to free available memory.
- Low-level task of the operating system. Virtual memory uses disk space as an extension of main memory.
- Since disk access is much slower than memory access, using swap space significantly decreases system performance.
- The main goal for the design and implementation of swap space is to provide the best throughput for the virtual memory system.

Swap Space Use :

Swap space is used in various ways by different operating systems, depending on the memory-management algorithms in use. Systems that implement swapping may use swap space to hold an entire process image, including the code and data segments. The amount of swap space needed on a system can therefore vary from a few megabytes of disk space to gigabytes.

It is safer to overestimate than to underestimate the amount of swap space required, because if a system runs out of swap space it may be forced to abort processes or may crash entirely.

Overestimation wastes disk space that could otherwise be used for files, but it does no other harm. Some systems recommend the amount to be set aside for swap space.

Swap-Space Location

A swap space can reside in one of two places: it can be carved out of the normal file system, or it can be in a separate disk partition.

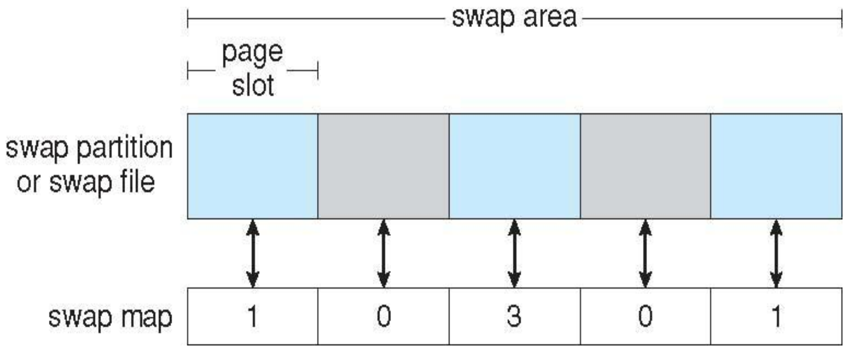
File System :

- If the swap space is a large file within the file system, normal file-system routines can be used to create it, name it and allocate its space. This approach, though easy to implement is inefficient.
- Navigating the directory structure and the disk allocation data structures takes time and extra disk accesses.
- External fragmentation can greatly increase swapping times by forcing multiple seeks during reading or writing of a process image.
- We can improve performance by caching the block location information in physical memory and by using special tools to allocate physically contiguous blocks for the swap file.

Disk Partition :

- Swap space can be created in a separate partition.
- A separate swap-space storage manager is used to allocate and deallocate the blocks from the raw partition.
- The manager uses algorithms optimized for speed rather than for storage efficiency, because swap space is accessed much more frequently than file systems (when it is used).
- Internal fragmentation may increase, but this trade-off is acceptable because the life of data in the swap space generally is much shorter than that of files in the file system. Since swap space is reinitialized at boot time, any fragmentation is short-lived.
- The raw-partition approach creates a fixed amount of swap space during disk partitioning. Adding more swap space requires either repartitioning the disk.

Data Structures for Swapping on Linux Systems



- In LINUX system swap space is only used for anonymous memory or for regions of memory shared by several processes.
- Linux allows one or more swap areas to be established. A swap area may be in either a swap file on a regular file system or a raw-swap-space partition.
- Each swap area consists of a series of 4-KB which are used to hold swapped pages.
- Associated with each swap area is a swap map - an array of integer counters, each corresponding to a page slot in the swap area.
- If the value of a counter is 0, the corresponding page slot is available. Values greater than 0 indicate that the page slot is occupied by a swapped page.
- The value of the counter indicates the number of mappings to the swapped page;
  For example, a value of 3 indicates that the swapped page is mapped to three different processes