# Storage and File Structure

Dr. Shrutilipi Bhattacharjee, Assistant Professor, Dept. of IT, NIT Karnataka, India

# Classification of Physical Storage Media

- Speed with which data can be accessed
- Cost per unit of data
- Reliability
  - Data loss on power failure or system crash
  - Physical failure of the storage device
- Can differentiate storage into:
  - **Volatile storage:** Loses contents when power is switched off
  - **Non-volatile storage**:
    - Contents persist even when power is switched off
    - Includes secondary and tertiary storage, as well as batter-backed up main-memory

# Physical Storage Media

- **Cache**
  - Fastest and most costly form of storage
  - Volatile
  - Managed by the computer system hardware

- **Main memory**
  - Fast access (10s to 100s of nanoseconds; 1 nanosecond = $10^{-9}$ seconds)
  - Generally too small (or too expensive) to store the entire database
    - Capacities of up to a few gigabytes widely used currently
    - Capacities have gone up and per-byte costs have decreased steadily and rapidly (roughly factor of 2 every 2 to 3 years)

- **Volatile**
  - Contents of main memory are usually lost if a power failure or system crash occurs

# Physical Storage Media

- **Flash memory**
  - Data survives power failure
  - Data can be written at a location only once, but location can be erased and written to again
    - Can support only a limited number (10K – 1M) of write/erase cycles
    - Erasing of memory has to be done to an entire bank of memory
  - Reads are roughly as fast as main memory
  - But writes are slow (few microseconds), erase is slower
  - Widely used in embedded devices such as digital cameras, phones, and USB keys

# Physical Storage Media

- **Magnetic-disk**
  - Data is stored on spinning disk, and read/written magnetically
  - Primary medium for the long-term storage of data
    - Typically stores entire database
  - Data must be moved from disk to main memory for access, and written back for storage
    - Much slower access than main memory
  - **Direct-access**
    - Possible to read data on disk in any order, unlike magnetic tape
  - Capacities range up to roughly 16~32 TB
    - Much larger capacity and cost/byte than main memory/flash memory
    - Growing constantly and rapidly with technology improvements (factor of 2 to 3 every 2 years)
  - Survives power failures and system crashes
    - Disk failure can destroy data, but is rare

# Physical Storage Media

- **Optical storage**
  - Non-volatile, data is read optically from a spinning disk using a laser
  - CD-ROM (640 MB) and DVD (4.7 to 17 GB) most popular forms
  - Blu-ray disks: 27 GB to 54 GB
  - Write-one, read-many (WORM) optical disks used for archival storage (CD-R, DVD-R, DVD+R)
  - Multiple write versions also available (CD-RW, DVD-RW, DVD+RW, and DVD-RAM)
  - Reads and writes are slower than with magnetic disk
  - **Juke-box** systems, with large numbers of removable disks, a few drives, and a mechanism for automatic loading/unloading of disks available for storing large volumes of data

# Physical Storage Media

- **Tape storage**
  - Non-volatile, used primarily for backup (to recover from disk failure), and for archival data
  - **Sequential-access**
    - Much slower than disk
  - Very high capacity (40 to 300 TB tapes available)
  - Tape can be removed from drive ⇒ storage costs much cheaper than disk, but drives are expensive
  - Tape jukeboxes available for storing massive amounts of data
    - Hundreds of terabytes (1 terabyte = $10^{12}$ bytes) to even multiple **petabytes** (1 petabyte = $10^{15}$ bytes)

# Storage Hierarchy

**Primary storage**

Volatile

Very fast

**Secondary storage:**

*on-line storage*

Non-volatile

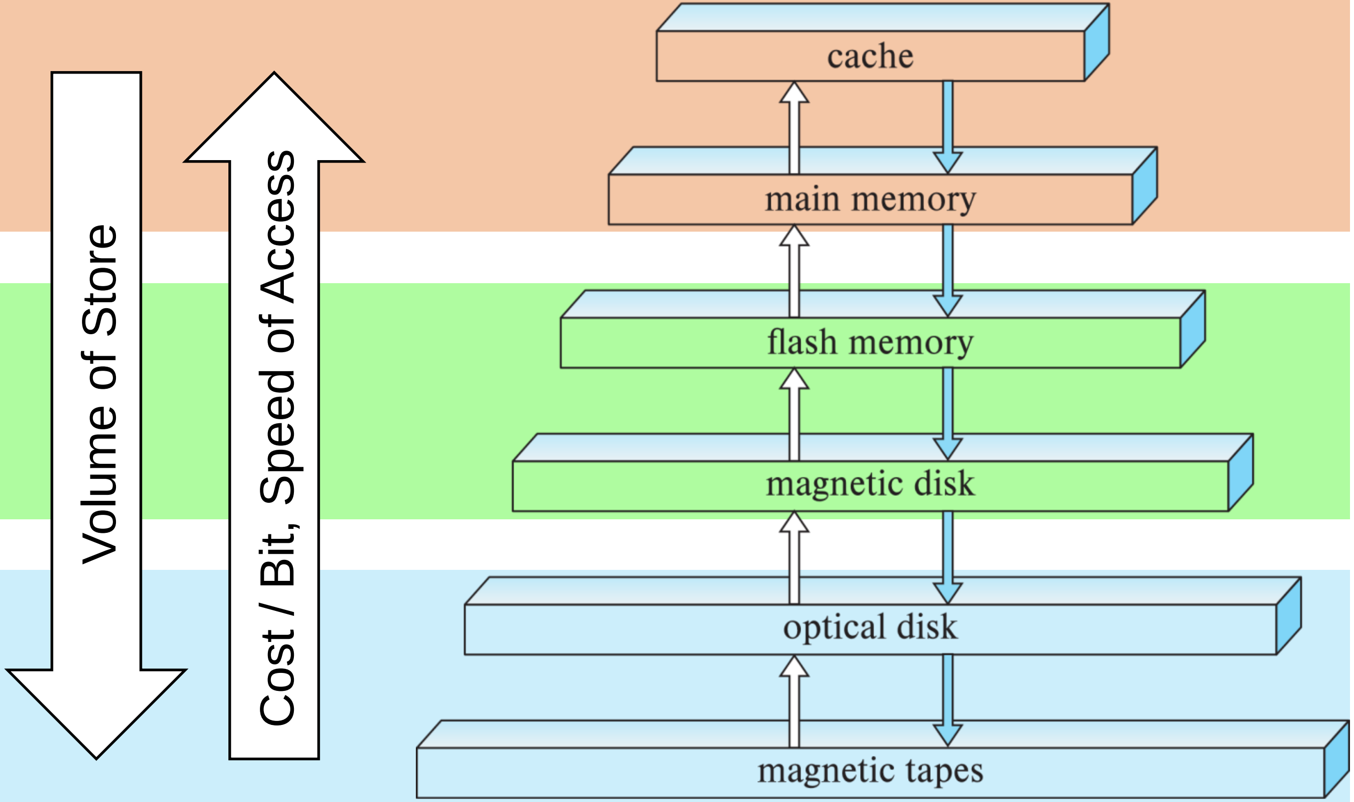
Moderately fast

**Tertiary storage:**

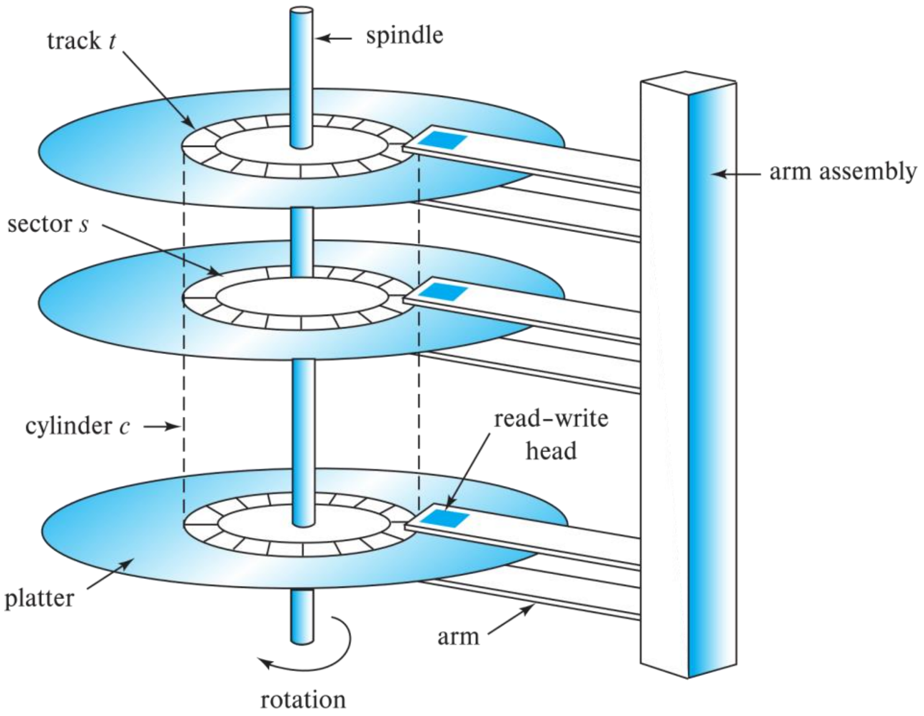
*off-line storage*

Non-volatile

Slow



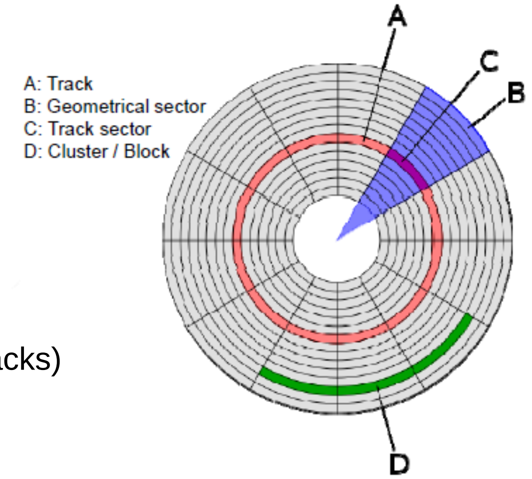
# Magnetic Hard Disk Mechanism

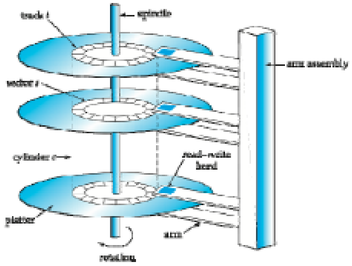


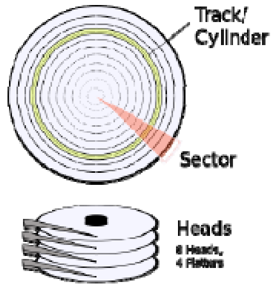

**NOTE: Diagram is schematic, and simplifies the structure of actual disk drives**

# Magnetic Disks

- **Read-write head**
  - Positioned very close to the platter surface (almost touching it)
  - Reads or writes magnetically encoded information
- Surface of platter divided into circular **tracks**
  - Over 50K-100K tracks per platter on typical hard disks
- Each track is divided into **sectors**
  - A sector is the smallest unit of data that can be read or written
  - Sector size typically 512 bytes
  - Typical sectors per track: 500 to 1000 (on inner tracks) to 1000 to 2000 (on outer tracks)
- To read/write a sector
  - Disk arm swings to position head on right track
  - Platter spins continually; data is read/written as sector passes under head
- Head-disk assemblies
  - Multiple disk platters on a single spindle (1 to 5 usually)
  - One head per platter, mounted on a common arm
- **Cylinder** *i* consists of $i^{th}$ track of all the platters



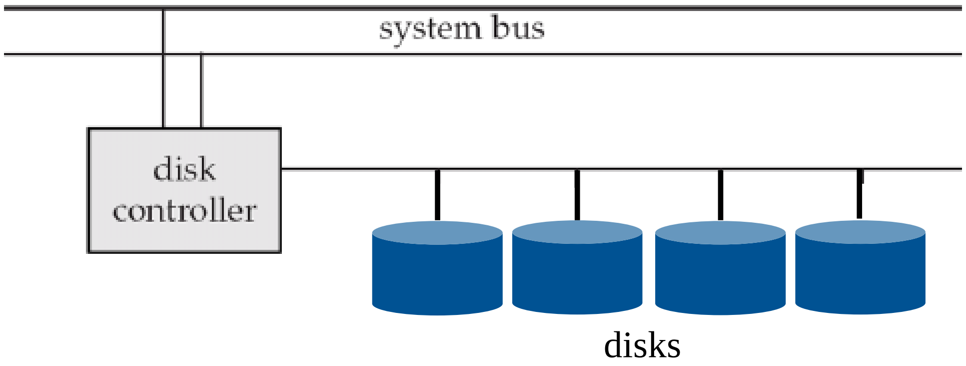


# Magnetic Disks

- Earlier generation disks were susceptible to head-crashes
  - Surface of earlier generation disks had metal-oxide coatings which would disintegrate on head crash and damage all data on disk
  - Current generation disks are less susceptible to such disastrous failures, although individual sectors may get corrupted

- **Disk controller** interfaces between the computer system and the disk drive hardware
  - Accepts high-level commands to read or write a sector
  - Initiates actions such as moving the disk arm to the right track and actually reading or writing the data
  - Computes and attaches **checksums** to each sector to verify that data is read back correctly
    - If data is corrupted, with very high probability stored checksum won't match recomputed checksum
  - Ensures successful writing by reading back sector after writing it
  - Performs **remapping of bad sectors**

# Disk Subsystem




- Multiple disks connected to a computer system through a controller
  - Controllers functionality (checksum, bad sector remapping) often carried out by individual disks; reduces load on controller
- Disk interface standards families
  - ATA (AT adaptor) range of standards
  - SATA (Serial ATA)
  - SCSI (Small Computer System Interconnect) range of standards
  - SAS (Serial Attached SCSI)
  - Several variants of each standard (different speeds and capabilities)

# Disk Subsystem

- Disks usually connected directly to computer system
- In **Storage Area Networks (SAN)**, a large number of disks are connected by a high-speed network to a number of servers
- In **Network Attached Storage (NAS)** networked storage provides a file system interface using networked file system protocol, instead of providing a disk system interface

# Performance Measures of Disks

- **Access time:** The time it takes from when a read or write request is issued to when data transfer begins and it consists of:
  - **Seek time:** Time it takes to reposition the arm over the correct track
    - Average seek time is 1/2 the worst case seek time
      - Would be 1/3 if all tracks had the same number of sectors, and we ignore the time to start and stop arm movement
    - 4 to 10 milliseconds on typical disks
  - **Rotational latency:** Time it takes for the sector to be accessed to appear under the head
    - Average latency is 1/2 of the above latency
    - 4 to 11 milliseconds on typical disks (5400 to 15000 r.p.m.)
- **Data-transfer rate:** 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
  - Multiple disks may share a controller, so rate that controller can handle is also important
    - E.g. SATA: 150 MB/sec, SATA-II 3Gb (300 MB/sec)
    - Ultra 320 SCSI: 320 MB/s, SAS (3 to 6 Gb/sec)
    - Fiber Channel (FC2Gb or 4Gb): 256 to 512 MB/s

# Performance Measures of Disks

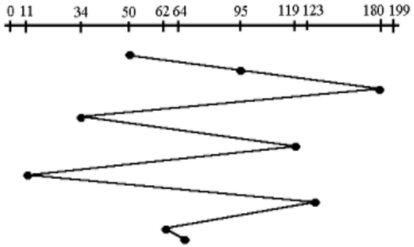
- **Mean time to failure (MTTF):** The average time the disk is expected to run continuously without any failure
  - Typically 3 to 5 years
  - Probability of failure of new disks is quite low, corresponding to a “theoretical MTTF” of 500,000 to 1,200,000 hours for a new disk
    - E.g., an MTTF of 1,200,000 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

# Optimization of Disk-Block Access

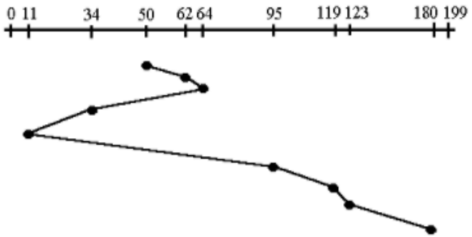
- **Block:** A contiguous sequence of sectors from a single track
  - Data is transferred between disk and main memory in blocks
  - Sizes range from 512 bytes to several kilobytes
    - Smaller blocks: More transfers from disk
    - Larger blocks: More space wasted due to partially filled blocks
    - Typical block sizes today range from 4 to 16 kilobytes

# Optimization of Disk-Block Access

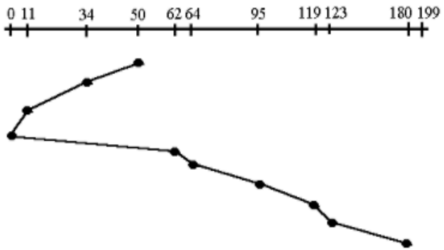
- **Disk-arm-scheduling** algorithms order pending accesses to tracks so that disk arm movement is minimized
- Example: *Queue 95, 180, 34, 119, 11, 123, 62, 64 with the Read-write head initially at the track 50 and the tail track being at 199*



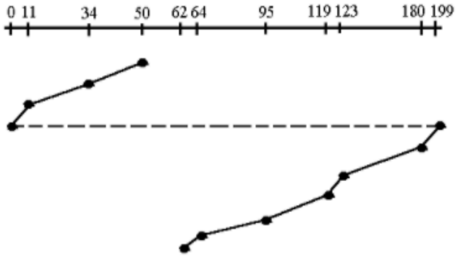

*First Come -First Serve (FCFS)*
*640 Tracks, Oscillations*



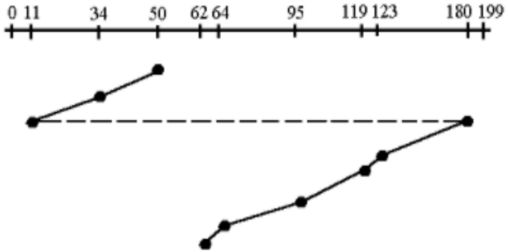

*Shortest Seek Time First (SSTF)*
*236 Tracks, Starvation*




*Elevator (SCAN)*
*230 Tracks*




*Circular Scan (C-SCAN)*
*187 Tracks*




*C-LOOK*
*157 Tracks, Optimal*

# Optimization of Disk-Block Access

- **File organization:** Optimize block access time by organizing the blocks to correspond to how data will be accessed
  - E.g. Store related information on the same or nearby cylinders
  - Files may get **fragmented** over time
    - E.g., if data is inserted to/deleted from the file
    - Or free blocks on disk are scattered, and newly created file has its blocks scattered over the disk
    - Sequential access to a fragmented file results in increased disk arm movement
  - Some systems have utilities to **defragment** the file system, in order to speed up file access

# Optimization of Disk-Block Access

- **Nonvolatile write buffers** speed up disk writes by writing blocks to a non-volatile RAM buffer immediately
  - Non-volatile RAM: Battery backed up RAM or flash memory
    - Even if power fails, the data is safe and will be written to disk when power returns
  - Controller then writes to disk whenever the disk has no other requests or request has been pending for some time
  - Database operations that require data to be safely stored before continuing can continue without waiting for data to be written to disk
  - *Writes can be reordered to minimize disk arm movement*

- **Log disk:** A disk devoted to writing a sequential log of block updates
  - Used exactly like nonvolatile RAM
    - Write to log disk is very fast since no seeks are required
    - No need for special hardware (NV-RAM)

- File systems typically reorder writes to disk to improve performance
  - **Journaling file systems** write data in safe order to NV-RAM or log disk
  - Reordering without journaling: Risk of corruption of file system data

# Flash Storage

- NOR flash vs NAND flash
- NAND flash
  - Used widely for storage, since it is much cheaper than NOR flash
  - Requires page-at-a-time read (page: 512 bytes to 4 KB)
  - Transfer rate around 20 MB/sec
  - **Solid state disks**: use multiple flash storage devices to provide higher transfer rate of 100 to 200 MB/sec
  - Erase is very slow (1 to 2 millisecs)
  - Erase block contains multiple pages
    - **Remapp**ing of logical page addresses to physical page addresses avoids waiting for erase
      - **Translation table** tracks mapping
        - Also stored in a label field of flash page
      - Remapping carried out by **flash translation layer**
    - After 100,000 to 1,000,000 erases, erase block becomes unreliable and cannot be used
      - **Wear leveling**

# RAID

- **RAID: Redundant Arrays of Independent Disks**
  - 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 chance that some disk out of a set of *N* disks will fail is much higher than the chance that a specific single disk will fail
  - E.g., a system with 100 disks, each with MTTF of 100,000 hours (approx. 11 years), will have a system MTTF of 1000 hours (approx. 41 days)
  - Techniques for using redundancy to avoid data loss are critical with large numbers of disks

- Originally a cost-effective alternative to large, expensive disks
  - I in RAID originally stood for "inexpensive''
  - Today RAIDs are used for their higher reliability and bandwidth
    - The "I" is interpreted as independent

# Improvement of Reliability via Redundancy

- **Redundancy:** Store extra information that can be used to rebuild information lost in a disk failure

- **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)

- **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

# Improvement of Reliability via Redundancy

- **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* mod *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

# Improvement of Reliability via Redundancy

- **Bit-Interleaved Parity:** A single parity bit is enough for error correction, not just detection, since
- We know which disk has failed
  - When writing data, corresponding parity bits must also be computed and written to a parity bit disk
  - To recover data in a damaged disk, compute XOR of bits from other disks (including parity bit disk)

- **Block-Interleaved Parity**: Uses block-level striping, and keeps a parity block on a separate disk for corresponding blocks from *N* other disks
  - When writing data block, corresponding block of parity bits must also be computed and written to parity disk
  - To find value of a damaged block, compute XOR of bits from corresponding blocks (including parity block) from other disks

# Choice of RAID Level

- Factors in choosing RAID level
  - Monetary cost
  - Performance: Number of I/O operations per second, and bandwidth during normal operation
  - Performance during failure
  - Performance during rebuild of failed disk
    - Including time taken to rebuild failed disk

- RAID 0 is used only when data safety is not important
  - E.g. data can be recovered quickly from other sources

- Level 2 and 4 never used since they are subsumed by 3 and 5

- Level 3 is not used anymore since bit-striping forces single block reads to access all disks, wasting disk arm movement, which block striping (level 5) avoids

- Level 6 is rarely used since levels 1 and 5 offer adequate safety for most applications

# Choice of RAID Level

- Level 1 provides much better write performance than level 5
  - Level 5 requires at least 2 block reads and 2 block writes to write a single block, whereas Level 1 only requires 2 block writes
  - Level 1 preferred for high update environments such as log disks

- Level 1 had higher storage cost than level 5
  - Disk drive capacities increasing rapidly (50%/year) whereas disk access times have decreased much less (x 3 in 10 years)
  - I/O requirements have increased greatly, e.g. for Web servers
  - When enough disks have been bought to satisfy required rate of I/O, they often have spare storage capacity
    - So there is often no extra monetary cost for Level 1!

- Level 5 is preferred for applications with low update rate, and large amounts of data

- Level 1 is preferred for all other applications

# Tertiary Storage: Optical Disks

- Compact disk-read only memory (CD-ROM)
  - Removable disks, 640 MB per disk
  - Seek time about 100 msec (optical read head is heavier and slower)
  - Higher latency (3000 RPM) and lower data-transfer rates (3-6 MB/s) compared to magnetic disks

- Digital Video Disk (DVD)
  - DVD-5 holds 4.7 GB , and DVD-9 holds 8.5 GB
  - DVD-10 and DVD-18 are double sided formats with capacities of 9.4 GB and 17 GB
  - Blu-ray DVD: 27 GB (54 GB for double sided disk)
  - Slow seek time, for same reasons as CD-ROM

- Record once versions (CD-R and DVD-R) are popular
  - Data can only be written once, and cannot be erased.
  - High capacity and long lifetime; used for archival storage
  - Multi-write versions (CD-RW, DVD-RW, DVD+RW and DVD-RAM) also available

# Tertiary Storage: Magnetic Tapes

- Hold large volumes of data and provide high transfer rates
- Few GB for DAT (Digital Audio Tape) format, 10-40 GB with DLT (Digital Linear Tape) format, 100 GB+ with Ultrium format, and 330 GB with Ampex helical scan format
- Transfer rates from few to 10s of MB/s

- Tapes are cheap, but cost of drives is very high

- Very slow access time in comparison to magnetic and optical disks
  - Limited to sequential access
  - Some formats (Accelis) provide faster seek (10s of seconds) at cost of lower capacity

- Used mainly for backup, for storage of infrequently used information, and as an off-line medium for
- transferring information from one system to another

- Tape jukeboxes used for very large capacity storage
  - Multiple petabyes ($10^{15}$ bytes)

# Next Lecture

## Storage and File Structure

# Thank you for your attention...

Any question?

**Contact:**
Department of Information Technology, NITK Surathkal, India
6th Floor, Room: 13
**Phone:** +91-9477678768
**E-mail:** shrutilipi@nitk.edu.in