CS315: Principles of Database Systems
Physical Design

Arnab Bhattacharya
arnabb@cse.iitk.ac.in

Computer Science and Engineering,
Indian Institute of Technology, Kanpur
http://web.cse.iitk.ac.in/~cs315/

2nd semester, 2013-14

Tue, Fri 1530-1700 at CS101

Physical storage media

- Cache (primary storage)
  - Fastest
  - Most costly
  - Volatile: contents vanish once power is off

Physical storage media

- Cache (primary storage)
  - Fastest
  - Most costly
  - Volatile: contents vanish once power is off
- Main memory (primary storage)
  - Fast
  - Not enough to hold a database
  - Volatile

Physical storage media

- Cache (primary storage)
  - Fastest
  - Most costly
  - Volatile: contents vanish once power is off
- Main memory (primary storage)
  - Fast
  - Not enough to hold a database
  - Volatile
- Flash memory (secondary storage, online storage)
  - Non-volatile
  - Read is quite fast
  - Write is slower due to erase
  - Supports a fixed number of write/erase cycles
  - Cheaper than main memory

Physical storage media (contd.)

- Magnetic disk (secondary storage, online storage)
  - Large
  - Direct-access: can read and write any location
  - Data needs to be brought to memory
  - Slower
  - Non-volatile

Physical storage media (contd.)

- Magnetic disk (secondary storage, online storage)
  - Large
  - Direct-access: can read and write any location
  - Data needs to be brought to memory
  - Slower
  - Non-volatile
- Optical storage (tertiary storage, offline storage)
  - CD, DVD, etc.
  - Non-volatile
  - Write-once, read-many
  - Slower
  - Re-writable also available

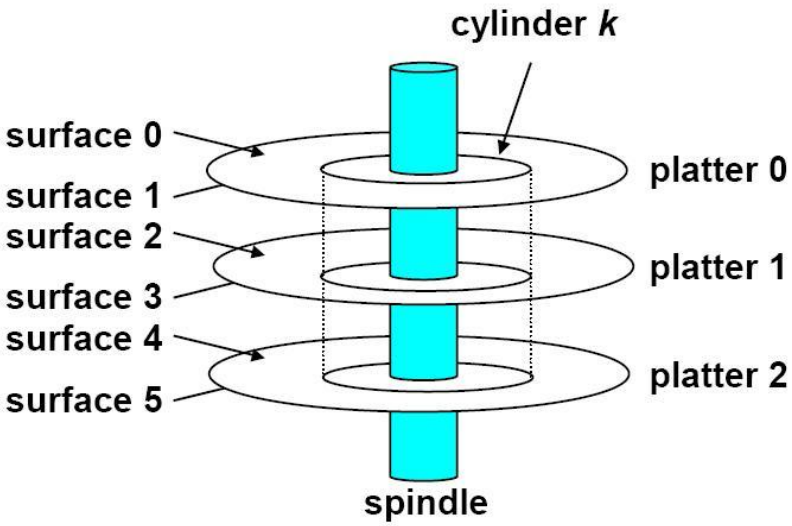
Physical storage media (contd.)

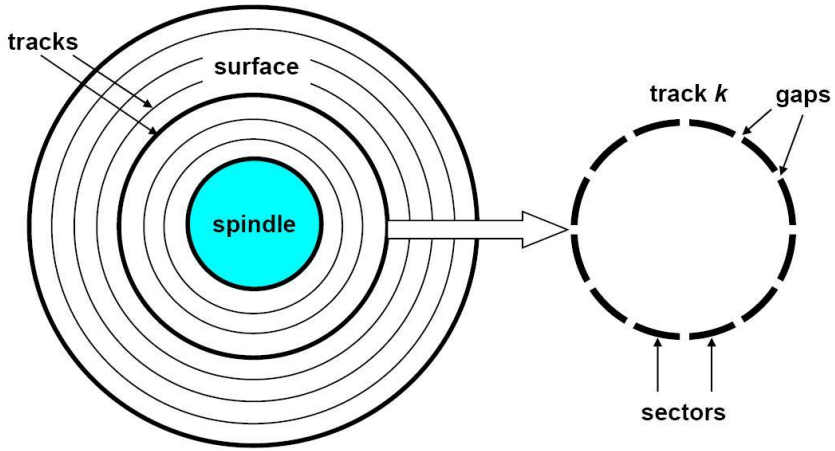
- Magnetic disk (secondary storage, online storage)
  - Large
  - Direct-access: can read and write any location
  - Data needs to be brought to memory
  - Slower
  - Non-volatile
- Optical storage (tertiary storage, offline storage)
  - CD, DVD, etc.
  - Non-volatile
  - Write-once, read-many
  - Slower
  - Re-writable also available
- Magnetic tape (tertiary storage, offline storage)
  - Sequential access
  - Much slower
  - Very high capacity
  - Much cheaper

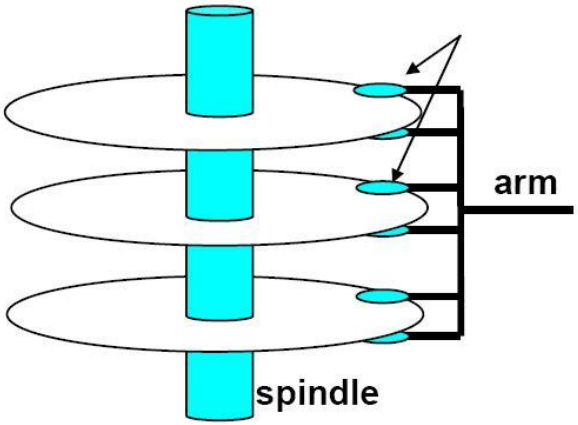
Disks

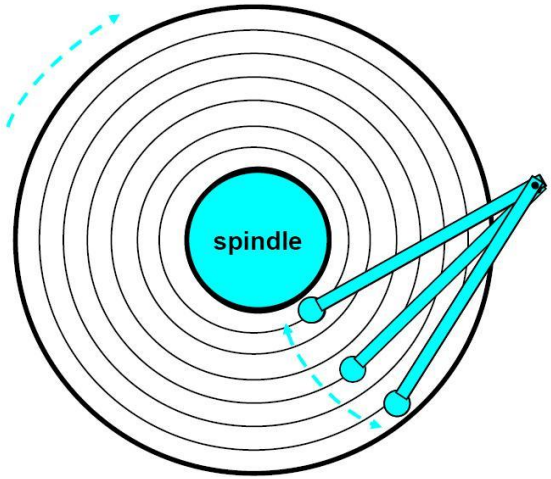
- Physically, disks consist of circular platters
- Both surfaces of a platter can be accessed
- Each surface contains concentric tracks
- Tracks are divided into sectors separated by gaps
- Aligned tracks form a cylinder

Disk access


cylinder k
surface 0
surface 1
surface 2
surface 3
surface 4
surface 5
platter 0
platter 1
platter 2
spindle


tracks
surface
track k
gaps
spindle
sectors


arm
spindle


spindle

Disk access time

- *Smallest* unit of information that can be read from or written to disk is a sector
- Block or page is a logical unit read from or written to by O/S
  - Block consists of a contiguous sequence of sectors

Disk access time

- *Smallest* unit of information that can be read from or written to disk is a sector
- Block or page is a logical unit read from or written to by O/S
  - Block consists of a contiguous sequence of sectors
- Access time T_{access}: Time to access a particular sector

$$T_{access} = T_{seek} + T_{rotation} + T_{transfer}$$

- Seek time T_{seek}: Time to position arm heads over cylinder containing the target sector
  - Typical seek time: 8 ms
- Rotational latency $T_{rotation}$: (Average) time to rotate r/w head to the first bit of the sector
  - $T_{rotation}$ = (1 / 2) $\times$ (1 / rpm) $\times$ (60 s / 1 min)
- Transfer time $T_{transfer}$: Time to read bits from the sector
  - $T_{transfer}$ = (1 / (#sectors / track)) $\times$ (60 / rpm)

Typical disk parameters

- Average seek times from 4-10 ms
- Rotational speeds are 60, 90, 120, 250 revolutions per second, i.e., 3600, 5400, 7200, 15000 rpm respectively
- Sector sizes vary between 512 bytes and 1024 bytes
- 400 to 1000 sectors per track
- 20,000 to 50,000 tracks per surface
- 1 to 5 platters per disk

Typical disk parameters

- Average seek times from 4-10 ms
- Rotational speeds are 60, 90, 120, 250 revolutions per second, i.e., 3600, 5400, 7200, 15000 rpm respectively
- Sector sizes vary between 512 bytes and 1024 bytes
- 400 to 1000 sectors per track
- 20,000 to 50,000 tracks per surface
- 1 to 5 platters per disk
- Example
  - Rotational speed = 7200 rpm
  - Average seek time T_{seek} = 8 ms
  - Average #sectors / track = 400
  - $T_{rotation} = (1 / 2) \times (1 / 7200) \times 60 = 4.17$ ms
  - $T_{transfer} = (1 / 400) \times (1 / 7200) \times 60 = 0.02$ ms
  - $\therefore T_{access} = 8 + 4.17 + 0.02 = 12$ ms

Access times

- This disk access time is for random I/O

Access times

- This disk access time is for random I/O
- Once the first bit is read, the rest (sequential I/O) is almost free (only 0.02 ms)
- Ratio of random I/O to sequential I/O is, therefore, (12 / 0.02) = *600* times
- Data transfer rates or bulk transfer rates are calculated more precisely using gaps

Access times

- This disk access time is for random I/O
- Once the first bit is read, the rest (sequential I/O) is almost free (only 0.02 ms)
- Ratio of random I/O to sequential I/O is, therefore, (12 / 0.02) = *600* times
- Data transfer rates or bulk transfer rates are calculated more precisely using gaps
- Time to access a byte = 60 ns for DRAM (used for main memory) and 4 ns for SRAM (used for cache memory)
- Ratio of disk access time to memory access time = *40000* for SRAM and *2500* for DRAM

Access times

- This disk access time is for random I/O
- Once the first bit is read, the rest (sequential I/O) is almost free (only 0.02 ms)
- Ratio of random I/O to sequential I/O is, therefore, (12 / 0.02) = *600* times
- Data transfer rates or bulk transfer rates are calculated more precisely using gaps
- Time to access a byte = 60 ns for DRAM (used for main memory) and 4 ns for SRAM (used for cache memory)
- Ratio of disk access time to memory access time = *40000* for SRAM and *2500* for DRAM
- Disk access time is dominated by seek time and rotational latency
- Sequential access algorithms exploit the (almost) free access time of later bits heavily
- Most algorithms aim to avoid random I/Os

Optimization of disk block access

- *Disk arm scheduling*: schedule such that movement of disk arm head is minimized
  - Elevator algorithm: move arm in one direction, process all requests in that order, and then move arm back in reverse direction

Optimization of disk block access

- *Disk arm scheduling*: schedule such that movement of disk arm head is minimized
  - Elevator algorithm: move arm in one direction, process all requests in that order, and then move arm back in reverse direction
- *File organization*: organize blocks of a file to minimize random I/Os
  - Defragmention: put all blocks contiguously, and reduce fragmentation

Optimization of disk block access

- *Disk arm scheduling*: schedule such that movement of disk arm head is minimized
  - Elevator algorithm: move arm in one direction, process all requests in that order, and then move arm back in reverse direction
- *File organization*: organize blocks of a file to minimize random I/Os
  - Defragmention: put all blocks contiguously, and reduce fragmentation
- *Deferred writes*: Postpone and perform writes batchwise
  - Use non-volatile write buffers, e.g., flash memory
  - Maintain logs for correctness

Data redundancy and parallelism

- Redundancy improves reliability
- RAID: Redundant arrays of independent disks
- Uses mirroring or shadowing
  - Failure only if both fail
- Mean time to data loss depends on mean time to failure for each disk and mean time to repair

Data redundancy and parallelism

- Redundancy improves reliability
- RAID: Redundant arrays of independent disks
- Uses mirroring or shadowing
  - Failure only if both fail
- Mean time to data loss depends on mean time to failure for each disk and mean time to repair
- Parallelism reduces mean response time
- Two main ways
- Bit-level striping: Each bit of a byte is stored in a separate disk
  - Time to re-form a byte increases
- Block-level striping: Each block of a file is stored in a separate disk
  - Certain tasks (e.g., searching) may need to combine only the results
  - Much faster

Data redundancy and parallelism

- Redundancy improves reliability
- RAID: Redundant arrays of independent disks
- Uses mirroring or shadowing
  - Failure only if both fail
- Mean time to data loss depends on mean time to failure for each disk and mean time to repair
- Parallelism reduces mean response time
- Two main ways
- Bit-level striping: Each bit of a byte is stored in a separate disk
  - Time to re-form a byte increases
- Block-level striping: Each block of a file is stored in a separate disk
  - Certain tasks (e.g., searching) may need to combine only the results
  - Much faster
- Job of buffer manager in memory to store disk blocks

Block organization

- Fixed-length records are stored contiguously
- When records are deleted, file gets fragmented
- Free lists are used to store address of next record
- Periodic re-organization to defragment

Block organization

- Fixed-length records are stored contiguously
- When records are deleted, file gets fragmented
- Free lists are used to store address of next record
- Periodic re-organization to defragment
- Variable-length records use slotted page structure
- Header of a slotted page (block) contains
  - Number of record entries
  - Address and size of each record
  - Address of start of free space
- Records are placed contiguously in the block

File organization

- Records in a file can be organized differently
- *Heap*: A record is placed anywhere where there is space
- *Sequential*: Records are placed sequentially in the order of the search key
- *Hashing*: Records are put in the block where they hash to

File organization

- Records in a file can be organized differently
- *Heap*: A record is placed anywhere where there is space
- *Sequential*: Records are placed sequentially in the order of the search key
- *Hashing*: Records are put in the block where they hash to
- Generally, a single relation is organized as a single file
- Sometimes, records from multiple relations are placed in the same file
- Suppose, two relations are joined frequently
- Records pertaining to the same value of the joined attrbute are stored in the same block
- This is called multi-table clustering file organization

Storage of special data

- Data dictionary or system catalog stores metadata (data about data)

Storage of special data

- Data dictionary or system catalog stores metadata (data about data)
- Information about relations
  - Name of relation
  - Name and type of attributes
  - Name and definition of views
  - Constraints

Storage of special data

- Data dictionary or system catalog stores metadata (data about data)
- Information about relations
  - Name of relation
  - Name and type of attributes
  - Name and definition of views
  - Constraints
- User information including password and access priviledge

Storage of special data

- Data dictionary or system catalog stores metadata (data about data)
- Information about relations
  - Name of relation
  - Name and type of attributes
  - Name and definition of views
  - Constraints
- User information including password and access priviledge
- Statistics about relations
  - Number of tuples

Storage of special data

- Data dictionary or system catalog stores metadata (data about data)
- Information about relations
  - Name of relation
  - Name and type of attributes
  - Name and definition of views
  - Constraints
- User information including password and access priviledge
- Statistics about relations
  - Number of tuples
- Organization of relations
  - Storage organization
  - Physical address

Storage of special data

- Data dictionary or system catalog stores metadata (data about data)
- Information about relations
  - Name of relation
  - Name and type of attributes
  - Name and definition of views
  - Constraints
- User information including password and access priviledge
- Statistics about relations
  - Number of tuples
- Organization of relations
  - Storage organization
  - Physical address
- Information about indices
  - Name of attribute and relation
  - Physical address of index

Storage of special data

- Data dictionary or system catalog stores metadata (data about data)
- Information about relations
  - Name of relation
  - Name and type of attributes
  - Name and definition of views
  - Constraints
- User information including password and access priviledge
- Statistics about relations
  - Number of tuples
- Organization of relations
  - Storage organization
  - Physical address
- Information about indices
  - Name of attribute and relation
  - Physical address of index
- Large objects need pointers and buffer management