Chapter 12: Secondary Storage Structure

Monday, November 12, 2007
12:38 PM

- HDD vocabs
  - Platter
  - Disk arm
  - Track
  - Sector
  - Cylinder
  - I/O bus
  - Disk controller
  - Host controller (called Host Bus Adaptor in Limoncelli)

- Transfer rate = seek time + rotational latency
  - Seek time = time to move disk arm
  - Rotation latency = time for the desired sector to come under disk arm

- Logical block
  - Smallest unit of transfer.
  - Typically of size 512 byte. But the size can be determined when we low-level format the disk.
  - Block numbers assigned sequentially
    - Starts with outer most track.
    - Proceeds inside track until all the blocks in the track are assigned.
    - Once a track is done, proceeds with other tracks on the same cylinder.
    - Once a cylinder is done, proceeds with the inner cylinder.

- Disk attachement
  - Host-attached storage or direct-attached storage (DAS)
    - HDD connects to CPU via a bus.
    - Example bus: SCSI, ATA, SATA, USB
  - Network-attached storage (NAS)
    - Basically, a server with HDD.
    - Clients access data via network file system protocol such as NFS or CIFS.
    - Not very efficient, because storage access shares bandwidth with other connections.
  - Storage area network (SAN)
    - A private network using storage protocol (such as iSCSI) instead of file server protocol.
    - Separates storage transfer from other transfer.

- Disk scheduling
  - Disk receives requests to read/write blocks. Conceptually, it receives a sequence of block numbers.
  - Algorithms to schedule visits to blocks
    - First come first serve (FCFS)
      - Inefficient.
    - Shortest seek time first (SSTF)
      - Improved performance over FCFS.
      - May cause starvation.
      - Not optimal
    - SCAN
      - Disk arm starts at one end, moves towards the other end, servicing requests that are found along the way.
      - Once disk arm reaches the end, it reverses the head movement direction.
    - C-SCAN
      - Modified Scan to make wait time uniform
      - Once disk arm reaches the end, it goes back to the beginning, not servicing any request that it finds along the way.

        - LOOK
            - The arm goes as far as the last request in one direction before reversing direction.
    - In most case, SSTF and LOOK are reasonable choices.
    - Different disk scheduling algorithms should be written as separate modules so that they are replaceable.

- RAID = Redundant Array of Inexpensive Disks
    - Used to provide redundancy and performance
        - Redundancy is provided by **mirroring** (copy data to separate piece of HDD)
        - Performance is provided by **striping** (store different bits or blocks in separate piece of HDD so that read and write can happen in parallel)
    - RAID Levels
        - RAID 0
            - Stiping at block level.
            - Gives performance but not redundancy.
        - RAID 1
            - Only mirroring.
            - Gives redundancy at the cost of write performance.
        - RAID 2
            - Memory-style error-correcting code (ECC)
            - Hamming code.
            - Say, 4-way striping on 4 HDDs with 3 HDDs containing Hamming code bits.
            - If one disk fails, the system can still read from the system.
        - RAID 3
            - Bit-interleaved parity organization.
            - Say, 4-way striping on 4 HDDs with 1 HDD dedicated to store parity bits.
            - Provides read performance.
            - However, write performance is quite bad, since the parity disk needs to be written every time, and, in order for the data to written to the parity disk, all disks must be read.
            - However, read is very fast, and it is popular with stream applications such as video or audio streaming.
        - RAID 4
            - Block-interleaved parity organization.
            - One disk store parity bits.
            - Single read slower than RAID 3, must large reads are much faster.
            - Still, small independent writes cannot be performed in parallel, and can be slower than RAID 3.
        - RAID 5
            - Block-interleaved distributed parity organization.
            - Parity bits distributed across all disks.
            - Avoid overuse of single parity disk.
            - RAID 4 and RAID 5 are the most common RAID types.
        - RAID 6
            - P+Q redundancy scheme.
            - Implement other error correcting codes such as Reed-Solomon codes.
            - Can safeguard against failure of more than one disks.
        - RAID 0+1
            - Disk is mirrored. Each mirror is striped.
            - Can tolerate one disk failure in one striped set. However, once a disk fail, then any disk in the other striped set is a single point failure.
        - RAID 1+0
            - Disk is striped. Each striped
            - The system can still run if no two disks from the same stripe fails.
    - To choose which level of RAID, consider the rebuild performance.
    - Raid 0+1 and 1+0 are used when both performance and reliability are important -- small data base.
    - Raid 5 is preferred when storing large data.
    - Raid is not a replacement for backup, neither it prevents software or user error.