UMD DATA605 - Big Data Systems

Lesson 4.3: Data Storage

Instructor: Dr. GP Saggese, gsaggese@umd.edu

- Sources
  - Silberschatz et al. 2020, Chap 12, Physical Storage Systems
  - Silberschatz et al. 2020, Chap 13: Data Storage Structures

SCIENCE ACADEMY

- ***Storage***
  - Magnetic Disks / SSD
  - RAID
  - DB Internals

Storage Characteristics

- Storage media trade-offs:
  - Speed of access (e.g., 500-3,500MB/sec)
  - Cost per data unit (e.g., 50 USD/TB)
  - Medium reliability
- Volatile vs non-volatile storage
  - **Volatile**: loses contents when power switched off
  - **Non-volatile**: can survive failures and system crashes
- Sequential vs random access
  - **Sequential**: read the data contiguously

```
SELECT * FROM employee
```

- **Random**: read the data from anywhere at any time

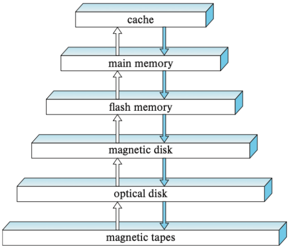
```
SELECT * FROM employee
WHERE name LIKE '\_\_a\_\_b'
```

- Need to know how data is stored in order to optimize access

Storage Hierarchy

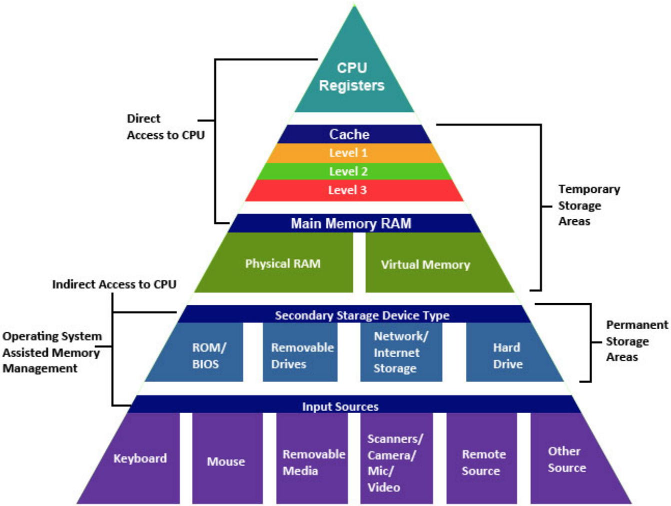
Organize storage by speed and cost

- Cache
  - Fastest, most costly
  - ~MBs on chip
  - DB developers consider cache effects
- Main memory
  - Up to 100s of GBs
  - Typically can't store entire DB
  - Volatile
- Flash memory / SSDs
  - More expensive than RAM, less than magnetic disk
  - Non-volatile, random access
- **Magnetic disk**
  - Long-term online storage
  - Non-volatile
- Optical disk (CD, Blue Ray)
  - Mainly read-only
- Magnetic tapes
  - Backup, archival data
  - Stored long-term, e.g., legal reasons
  - Sequential-access



- **Primary storage**: cache, main memory
- **Secondary (or online)**: flash memory, magnetic disk
- **Offline:** optical, magnetic tape

Storage Hierarchy



source: http://cse1.net/recaps/4-memory.html

SCIENCE ACADEMY

How Important Is Memory Hierarchy?

- Trade-offs shifted over last 10-15 years
- **Innovations:**
  - Fast network, SSDs, large memories
  - Data volume growing rapidly
- **Observations:**
  - Faster to access another computer's memory through network than your own disk
  - Cache plays a crucial role
  - In-memory DBs
    - Data often fits in memory of a machine cluster
  - Disk considerations less important
    - Disks still store most data today
- Algorithms depend on available technology

- Storage
  - ***Magnetic Disks / SSD***
  - RAID
  - DB Internals

Connecting disks to a server

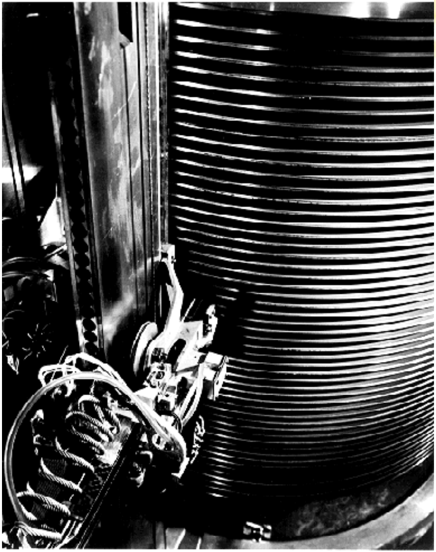
Connecting Disks to a Server

- **Disks** (magnetic and SSDs) connect to computer:
  - High-speed bus interconnections
  - High-speed network
- **High-speed interconnection**
  - Serial ATA (SATA)
  - Serial Attached SCSI (SAS)
  - NVMe (Non-volatile Memory Express)
- **High-speed networks**
  - Storage Area Network (SAN): ISCSI, Fiber Channel, InfiniBand
  - **Network Attached Storage (NAS)**
    - Provides file-system interface (e.g., NFS)
    - Cloud storage: Data stored in cloud, accessed via API, Object store, High latency

Magnetic Disks

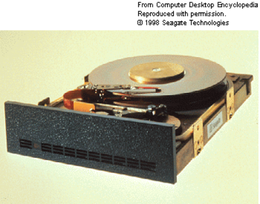
- *1956*
  - IBM RAMAC
  - 24" platters
  - 5 million characters

From Computer Desktop Encyclopedia
Reproduced with permission.
© 1996 International Business Machines Corporation
Unauthorized use not permitted.

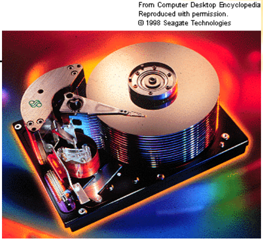


Magnetic Disks

- 1979
  - SEAGAT
  - 5MB



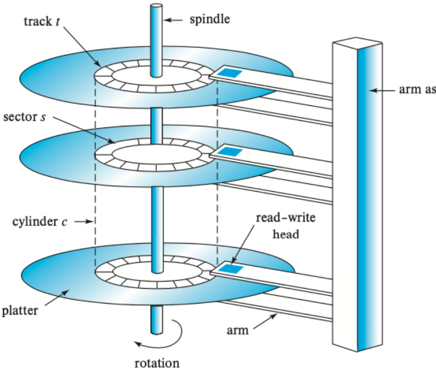
- 1998
  - SEAGAT
  - 47GB



- 2006
  - Western Digital
  - 500GB

Magnetic Disks: Components

- **Platters**
  - Rigid metal with magnetic material on both surfaces
  - Spins at 5400 or 9600 RPM
  - *Tracks* subdivided into *sectors* (smallest unit read/written, with a checksum)
- **Read-write heads**
  - Store information magnetically
  - Spinning creates a cushion maintaining heads a few microns from the surface
  - *Cylinder* is the i-th tracks of all platters (read/written together)
- **Arm**
  - Moves all heads along the disks
- **Disk controller**
  - Accepts commands to read/write a sector
  - Operates arm/heads
  - Remaps bad sectors to a different location



Magnetic Disks: Current Specs

- **Capacity**
  - 10 terabyte and more
- **Access time**
  - Time to start reading data
  - Seek time
    - Move arm across cylinders (2-20ms)
  - Rotational latency time
    - Wait for sector access (4-12ms)

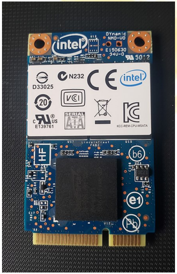
- **Data-transfer rate**
  - Transfer begins once data is reached
  - Transfer rate: 50-200MB/sec
  - Sector (disk block): logical unit of storage (4-16KB)
  - Sequential access: blocks on same or adjacent tracks
  - Random access: each request requires a seek
    - IOPS: number of random single block accesses per second (50-200 IOPS)
- **Reliability**
  - Mean time to failure (MTTF): average time system runs without failure
  - HDD lifespan: ~5 years

Accessing Data Speed

- **Random data transfer rates**
  - Time to read a random sector
  - It has 3 components
    - Seek time: Time to seek to the track (Average 4 to 10ms)
    - Rotational latency: Waiting for the sector to get under the head (Average 4 to 11ms)
    - Transfer time: Time to transfer the data (Very low)
  - About 10ms per access
    - Randomly accessed blocks: 100 block transfers (100/sec x 4 KB/block = 50 KB/s)
- **Serial data transfer rates**
  - Data transfer rate without seek
  - 30-50MB/s to 200MB/s
- **Seeks are bad!**

Solid State Disk (SSD)

- Mainstream around 2000s
- Like non-volatile RAM (NAND and NOR)
- **Capacity**
  - 250, 500 GBs (vs 1-10 TB for HDD)
- **Access time**
  - Latency for random access 1,000x smaller than HDD
    - E.g., 20-100 us (vs 10 ms HDDs)
  - Multiple random requests (e.g., 32) in parallel
  - 10,000 IOPS (vs 50/200 for HDDs)
  - Require reading an entire "page" of data (typically 4KB)
    - Equivalent to a block in magnetic disks



- **Data-transfer rate**
  - 1 GB/s (vs 200 MB/s HDD)
  - Typically limited by interface speed
  - Reads and writes ~500MB/s for SATA and 2-3 GB/s for NVMe
  - Lower power consumption than HDDs
  - Writing to SSD slower than reading (~2-3x)
    - Requires erasing all pages in the block
- **Reliability**
  - Limit to how many times a flash page can be erased (~1M times)
- Better than HDD from any point of view, but more expensive per GB

- Storage
    - Magnetic Disks / SSD
    - ***RAID***
    - DB Internals

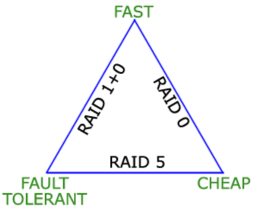
RAID

- RAID = Redundant Array of Independent Disks
- **Problem**
  - Storage capacity growing exponentially
  - Data-storage needs (web, DBs, multimedia) growing faster
  - Need many disks
  - MTTF between disk failures shrinking (e.g., days)
    - Single data copy leads to unacceptable data loss frequency
- **Observations**
  - Disks cheap
  - Failures costly
  - Use extra disks for reliability
    - Store data redundantly
    - Data survives disk failure
  - Bonus: faster data access
- **Goal**
  - Present a logical view of a large, reliable disk from many unreliable disks
  - Different RAID levels (reliability vs performance)

SCIENCE ACADEMY

Improve Reliability / Performance with RAID

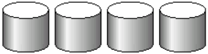
- **Reliability**
  - Use redundancy
    - Store data multiple times: E.g., mirroring
    - Reconstruct data if a disk fails
    - Increase MTTF
  - Assume independence of disk failure
    - Consider power failures and natural disasters
    - Aging disks increase failure probability
- **Performance**
  - Parallel access to multiple disks: E.g., mirroring, increase read requests
  - Stripe data across multiple disks: Increase transfer rate



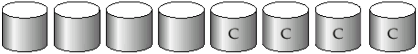
SCIENCE ACADEMY

RAID Levels

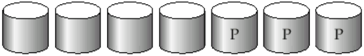
- **RAID 0: No redundancy**
  - Array of independent disks
  - Same access-time
  - Increased transfer rate
- **RAID 1: Mirroring**
  - Copy of disks
  - If one disk fails, you have a copy
  - Reads: higher data rate possible
  - Writes: write to both disks
- **RAID 2: Memory-style error correction**
  - Use extra bits to reconstruct
  - Superseded by RAID 5
- **RAID 3: Interleaved parity**
  - One disk contains parity for main data disks
  - Handle single disk failure
  - Little overhead (only 25%)
- **RAID 5: Block-interleaved distributed parity**
  - Distributed parity blocks



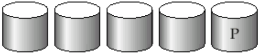
(a) RAID 0: nonredundant striping



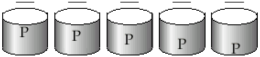
(b) RAID 1: mirrored disks



(c) RAID 2: memory-style error-correcting codes



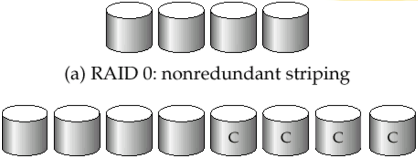
(d) RAID 3: bit-interleaved parity



(f) RAID 5: block-interleaved distributed parity

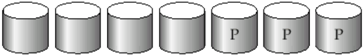
Choosing a RAID Level

- Main choice: RAID 1 vs. RAID 5
- **RAID 1 better write performance**
  - E.g., writing a single block
    - RAID 1: 2 block writes
    - RAID 5: 2 block reads, 2 block writes
  - Best for high update rate, small data (e.g., log disks)
- **RAID 5 lower storage cost**
  - RAID 1: 2x more disks
  - Best for low update rate, large data

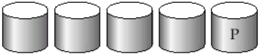


(a) RAID 0: nonredundant striping

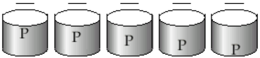
(b) RAID 1: mirrored disks



(c) RAID 2: memory-style error-correcting codes



(d) RAID 3: bit-interleaved parity



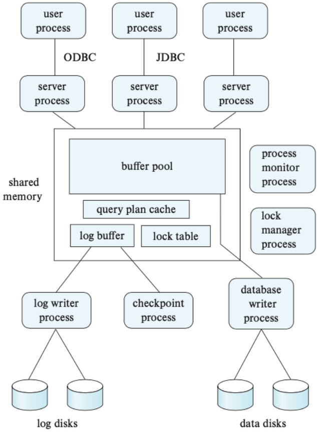
(f) RAID 5: block-interleaved distributed parity

SCIENCE ACADEMY

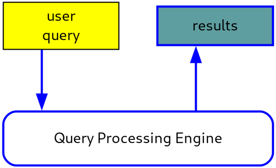
- Storage
  - Magnetic Disks / SSD
  - RAID
  - ***DB Internals***

(Centralized) DB Internals

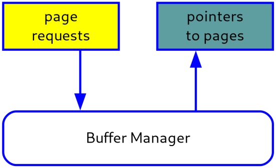
- User processes
  - Issue commands to DB
- Server processes
  - Receive commands, call DB code
- Process monitor process
  - Monitor DB processes
  - Recover from failures
- Lock manager process
  - Lock grant/release
  - Detect deadlocks
- Database writer process
  - Write modified buffer blocks to disk continuously
- Log writer process
  - Write log records to stable storage
- Checkpoint process
  - Perform periodic checkpoints
- Shared memory
  - Contain shared data
    - Buffer pool, Lock table, Log buffer, Caches (e.g., query plans)
  - Protect data with mutual exclusion locks



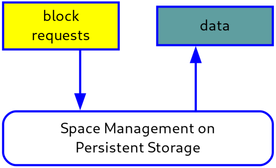
DB Internals



- **Query Processing Engine**
  - Execute user query
  - Specify page sequence for memory
  - Operate on tuples for results



- **Buffer Manager**
  - Transfer pages from disk to memory
  - Manage limited memory



- **Storage hierarchy**
  - Map tables to files
  - Map tuples to disk blocks