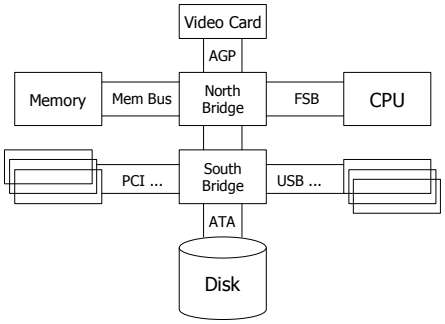
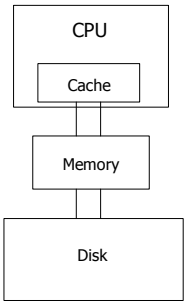
CS522 Advanced Database Systems
Disk Access

Chengyu Sun
California State University, Los Angeles

Computer Architecture



Storage Devices



Capacity??

Volatility??

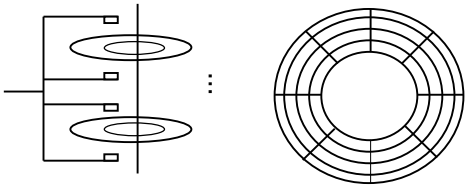
Access time??

What's the "typical" size of a database??

Understand Disk Access

- Data organization
- External algorithms
- Query optimization

Disk



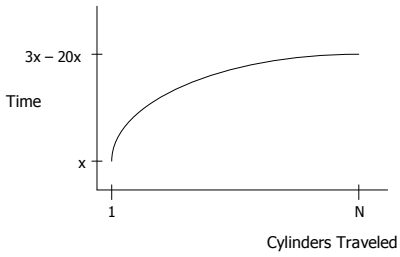
Disk Terminology

- Platter, surface, head
- Track, cylinder
- Sector, gap
- Block

Disk Access Time

- Disk Access Time
  - Seek time
  - Rotational latency
  - Transfer time
  - Other

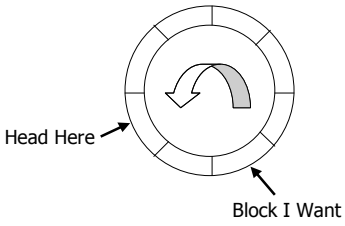
Seek Time



Average Seek Time

$$S = \frac{\sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \text{SEEKTIME } (i \rightarrow j)}{N(N-1)}$$

Average Rotational Latency



Transfer Time

- Average Transfer rate t
- Transfer time = Block size / t

Other Costs

- CPU time to issue I/O
- Contention for bus
- Contention for disk controller

Megatron 747 – Characteristics

- 8 double-surfaced platters
- 2^{14}=16,384 tracks per surface
- 2^{7}=128 sectors on average, 10% gap
- 2^{12}=4096 bytes per sector
- 16k block
- Seek time
  - 1ms to start and stop
  - 1ms for every 1000 cylinders

Megatron 747 – Calculations

- Capacity?
- Access time for one block?
  - Min
  - Max
  - Average

Write and Modify

- Write
  - slightly slower than Read
  - with Verify?
- Modify
  - Read block
  - Modify in memory
  - Write block

WD Caviar SE 250G

Interface	ATA100
Seek Time	
Average R/W (ms)	8.9/10.9
Track-to-Track R/W (ms)	2.0
Full Stroke R/W (ms)	21.0
Spindle Speed (RPM)	7,200
Average Rotational Latency (ms)	4.2
Buffer to Host Transfer Rates	
Mode 5 Ultra ATA (MB/s)	100
Mode 4 Ultra ATA (MB/s)	66.6
Mode 4 PIO (MB/s)	16.6
Cache (MB)	2

Maxtor Atlas 15K II

Byte per Sector	512 - 520
Interface	SCSI Ultra320
Seek Time	
Average R/W (ms)	3.0/3.4
Track-to-Track R/W (ms)	0.3/0.5
Full Stroke R/W (ms)	8.0/9.0
Spindle Speed (RPM)	15,000
Average Rotational Latency (ms)	2
Transfer Rates	
SCSI Maximum Burst (MB/s)	320
SCSI Maximum Host (MB/s)	270
Maximum Sustained (MB/s)	98
Cache (MB)	8

Lessons Learned So Far

- Disk access cost components and estimation
- Disk access vs. in-memory operations
  - several orders of magnitude difference
- Sequential vs. random access
  - 5 ~ 20 times difference

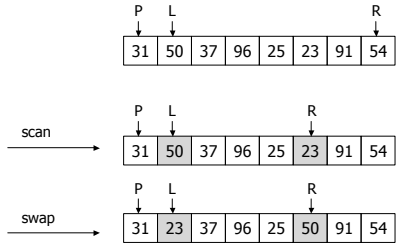
What We Can Do About It

- Algorithm design
  - Use as few disk accesses as possible
  - Access as sequentially as possible
- Disk optimizations

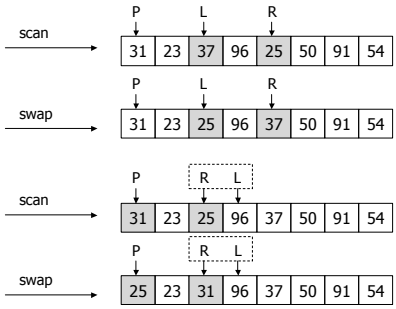
Example - Sorting

- Used quite extensively in DBMS operations
  - ORDER BY
  - DISTINCT
  - UNION, INTERSECTION, DIFFERENCE
  - Join

Quick Sort ...

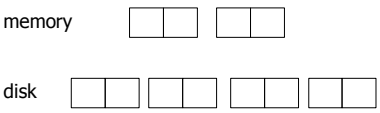


... Quick Sort



Disk I/O Model

- 2 numbers per page (block)
- 2 pages for data buffer
- O(logN) stack space



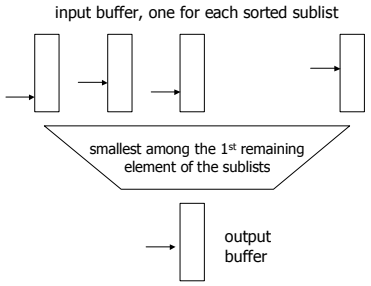
I/O Performance

- Quick Sort vs. Merge Sort

Two Phase Multiway Merge Sort (TPMMS)

- Phase 1: fill memory buffer, in-memory sort
- Phase 2: merge *sorted sublists*, one block from each

TPMMS Merge Phase



Disk Optimizations

- Buffering
- And??

Disk Failures

- Partial → total
- Intermittent → permanent

Coping with Disk Failures

- Detection
  - Checksum
- Correction
  - Redundancy

Recovery from Disk Crashes

- RAID 1, 2, 3, ... , ∞ !
  - Read performance
  - Write performance
  - Disk utilization
  - Number of recoverable simultaneous disks failures

Reading and Exercises

- Read Chapter 11
- Exercises
  - 11.3.2, 11.4.2, 11.5.2, 11.6.2, 11.7.2