Databases

- Data Models
  - Conceptual representation of the data
- Data Retrieval
  - How to ask questions of the database
  - How to answer those questions
- Data Storage
  - How/where to store data, how to access it
- Data Integrity
  - Manage crashes, concurrency
  - Manage semantic inconsistencies
Query Processing/Storage

Query Processing Engine

- Given a input user query, decide how to “execute” it
- Specify sequence of pages to be brought in memory
- Operate upon the tuples to produce results

Buffer Management

- Bringing pages from disk to memory
- Managing the limited memory

Space Management on Persistent Storage (e.g., Disks)

- Storage hierarchy
- How are relations mapped to files?
- How are tuples mapped to disk blocks?

Outline

- Storage hierarchy
- Disks
- RAID
- File Organization
- Etc....
Storage Hierarchy

- Tradeoffs between speed and cost of access
- Volatile vs nonvolatile
  - Volatile: Loses contents when power switched off
- 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'
- Why care?
  - Need to know how data is stored in order to optimize, to understand what’s going on

How important is this today?

- Trade-offs shifted drastically over last 10-15 years
  - Especially with fast network, SSDs, and high memories
  - However, the volume of data is also growing quite rapidly
- Some observations:
  - Cheaper to access another computer’s memory than accessing your own disk
  - Cache is playing more and more important role
  - Enough memory around that data often fits in memory of a single machine, or a cluster of machines
  - ”Disk” considerations less important
    - Still: Disks are where most of the data lives today
  - Similar reasoning/algorithms required though
Storage Hierarchy: Cache

- Cache
  - Super fast; volatile; Typically on chip
  - L1 vs L2 vs L3 caches ???
    - L1 about 64KB or so; L2 about 1MB; L3 8MB (on chip) to 256MB (off chip)
    - Huge L3 caches available now-a-days
  - Becoming more and more important to care about this
    - Cache misses are expensive
  - Similar tradeoffs as were seen between main memory and disks
  - Cache-coherency ??

source: http://cse1.net/recaps/4-memory.html
Main memory
- 10s or 100s of ns; volatile
- Pretty cheap and dropping: 1GB < 100$
- Main memory databases feasible now-a-days

Flash memory (EEPROM)
- Limited number of write/erase cycles
- Non-volatile, slower than main memory (especially writes)
- Examples?

Question
- How does what we discuss next change if we use flash memory only?
- Key issue: Random access as cheap as sequential access
Storage Hierarchy

- Magnetic Disk (Hard Drive)
  - Non-volatile
  - Sequential access much much faster than random access
  - Discuss in more detail later
- Optical Storage - CDs/DVDs; Jukeboxes
  - Used more as backups… Why?
  - Very slow to write (if possible at all)
- Tape storage
  - Backups; super-cheap; painful to access
  - IBM just released a secure tape drive storage solution

Storage…

- Primary
  - e.g. Main memory, cache; typically volatile, fast
- Secondary
  - e.g. Disks; Solid State Drives (SSD); non-volatile
- Tertiary
  - e.g. Tapes; Non-volatile, super cheap, slow
Jim Gray's Storage Latency Analogy:
How Far Away is the Data?

<table>
<thead>
<tr>
<th>Storage type</th>
<th>Access time</th>
<th>Relative access time</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 cache</td>
<td>0.5 ns</td>
<td>Blink of an eye</td>
</tr>
<tr>
<td>L2 cache</td>
<td>7 ns</td>
<td>4 seconds</td>
</tr>
<tr>
<td>1MB from RAM</td>
<td>0.25 ms</td>
<td>5 days</td>
</tr>
<tr>
<td>1MB from SSD</td>
<td>1 ms</td>
<td>23 days</td>
</tr>
<tr>
<td>HDD seek</td>
<td>10 ms</td>
<td>231 days</td>
</tr>
<tr>
<td>1MB from HDD</td>
<td>20 ms</td>
<td>1.25 years</td>
</tr>
</tbody>
</table>

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

- Storage hierarchy
- Disks
- RAID
- File Organization
- Etc....

1956
IBM RAMAC
24” platters
100,000 characters each
5 million characters
1979
SEAGATE
5MB

1998
SEAGATE
47GB

2006
Western Digital
500GB
Weight (max. g): 600g

Latest:

Single hard drive:
Seagate Barracuda 7200.10 SATA
750 GB
7200 rpm
weight: 720g
Uses “perpendicular recording”

Microdrives

IBM 1 GB

Toshiba 80GB
“Typical” Values

- **Diameter**: 1 inch → 15 inches
- **Cylinders**: 100 → 2000
- **Surfaces**: 1 or 2
- **(Tracks/cyl)**: 2 (floppies) → 30
- **Sector Size**: 512B → 50K
- **Capacity**: 360 KB to 2TB (as of Feb 2010)
- **Rotations per minute (rpm)**: 5400 to 15000

Accessing Data

- **Accessing a sector**
  - Time to *seek* to the track (seek time)
    - average 4 to 10ms
  - + Waiting for the sector to get under the head (rotational latency)
    - average 4 to 11ms
  - + Time to transfer the data (transfer time)
    - very low
  - About 10ms per access
    - So if randomly accessed blocks, can only do 100 block transfers
    - 100 x 512bytes = 50 KB/s
- **Data transfer rates**
  - Rate at which data can be transferred (w/o any seeks)
    - 30-50MB/s to up to 200MB/s (Compare to above)
    - Seeks are bad!
Seagate Barracuda: 1TB

- Heads 8, Disks 4
- Bytes per sector: 512 bytes
- Default cylinders: 16,383
- Defaults sectors per track: 63
- Defaults read/write heads: 16
- Spindle speed: 7200 rpm
- Internal data transfer rate: 1287 Mbits/sec max
- Average latency: 4.16msec
- Track-to-track seek time: 1msec-1.2msec
- Average seek: 8.5-9.5msec
- We also care a lot about power now-a-days
  - Why?

Reliability

- Mean time to/between failure (MTTF/MTBF):
  - 57 to 136 years
- Consider:
  - 1000 new disks
  - 1,200,000 hours of MTTF each
  - On average, one will fail 1200 hours = 50 days!
Disk Controller

- Interface between the disk and the CPU
- Accepts the commands
- *checksums* to verify correctness
- Remaps bad sectors

---

Optimizing block accesses

- Typically sectors too small
- Block: A contiguous sequence of sectors
  - 512 bytes to several Kbytes
  - All data transfers done in units of blocks
- Scheduling of block access requests?
  - Considerations: *performance* and *fairness*
  - *Elevator algorithm*
Solid State Drives

- Essentially flash that emulates hard disk interfaces
- No seeks → Much better random reads performance
- Writes are slower, the number of writes at the same location limited
  - Must write an entire block at a time
- About a factor of 10 more expensive right now

- Will soon lead to perhaps the most radical hardware configuration change in a while

Outline

- Storage hierarchy
- Disks
- RAID
- File Organization
- Etc.…
RAID

- Redundant array of independent disks
- Goal:
  - Disks are very cheap
  - Failures are very costly
  - Use “extra” disks to ensure reliability
    - If one disk goes down, the data still survives
    - Also allows faster access to data
- Many raid “levels”
  - Different reliability and performance properties

RAID Levels

(a) No redundancy.

(b) Make a copy of the disks.
  - If one disk goes down, we have a copy.
  - Reads: Can go to either disk, so higher data rate possible.
  - Writes: Need to write to both disks.
RAID Levels

(c) Memory-style Error Correcting
    Keep extra bits around so we can reconstruct.
    Superceded by below.

(d) One disk contains “parity” for the main data disks.
    Can handle a single disk failure.
    Little overhead (only 25% in the above case).

RAID Level 5

- Distributed parity “blocks” instead of bits
- Subsumes Level 4
- Normal operation:
  - “Read” directly from the disk. Uses all 5 disks
  - “Write”: Need to read and update the parity block
    - To update 9 to 9’
      - read 9 and P2
      - compute $P2' = P2 \text{xor} 9 \text{xor} 9'$
      - write 9’ and P2’

<table>
<thead>
<tr>
<th>P0</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>P1</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>P2</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>13</td>
<td>14</td>
<td>P3</td>
<td>15</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>P4</td>
</tr>
</tbody>
</table>
RAID Level 5

- Failure operation (disk 3 has failed)
  - “Read block 0”: Read it directly from disk 2
  - “Read block 1” (which is on disk 3)
    - Read P0, 0, 2, 3 and compute $1 = P0 \oplus 0 \oplus 2 \oplus 3$
  - “Write”:
    - To update 9 to 9'
      - read 9 and P2
      - Oh… P2 is on disk 3
      - So no need to update it
      - Write 9'

| P0 | 0 | 2 | 3 |
|    | 4 | P1 | 6 | 7 |
|    | 8 | 9 | 10 | 11 |
|    | 12 | 13 | P3 | 15 |
|    | 16 | 17 | 19 | P4 |

(f) RAID 5: block-interleaved distributed parity

Choosing a RAID level

- Main choice between RAID 1 and RAID 5
- Level 1 better write performance than level 5
  - Level 5: 2 block reads and 2 block writes to write a single block
  - Level 1: only requires 2 block writes
  - Level 1 preferred for high update environments such as log disks
- Level 5 lower storage cost
  - Level 1 50% of disks used for redundancy
  - Level 5 is preferred for applications with low update rate, and large amounts of data