CSCI-UA.0201

Computer Systems Organization

Memory Hierarchy

Mohamed Zahran (aka Z)
mzahran@cs.nyu.edu
http://www.mzahran.com
Programmer’s Wish List

Memory

- Private
- Infinitely large
- Infinitely fast
- Non-volatile
- Inexpensive

Programs are getting bigger faster than memories.
Example Memory Hierarchy

- **Regs**: CPU registers hold words retrieved from the L1 cache.
- **L1 cache (SRAM)**: L1 cache holds cache lines retrieved from the L1 cache.
- **L2 cache (SRAM)**: L2 cache holds cache lines retrieved from the L2 cache.
- **L3 cache (SRAM)**: L3 cache holds cache lines retrieved from L3 cache.
- **Main memory (DRAM)**: Main memory holds disk blocks retrieved from main memory.
- **Local secondary storage (local disks)**: Local disks hold files retrieved from disks on remote servers.
- **Remote secondary storage (e.g., Web servers)**: Remote secondary storage holds files retrieved from remote servers.

Levels:
- L0: Smaller, faster, and costlier (per byte) storage devices
- L1: Larger, slower, and cheaper (per byte) storage devices
- L2: Smaller, faster, and costlier (per byte) storage devices
- L3: Larger, slower, and cheaper (per byte) storage devices
## SRAM vs DRAM Summary

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SRAM</td>
<td>4 or 6</td>
<td>1X</td>
<td>No</td>
<td>Maybe</td>
<td>100x</td>
<td>Cache memories</td>
</tr>
<tr>
<td>DRAM</td>
<td>1</td>
<td>10X</td>
<td>Yes</td>
<td>Yes</td>
<td>1X</td>
<td>Main memories, frame buffers</td>
</tr>
</tbody>
</table>
Question: Who Cares About the Memory Hierarchy?

“Moore’s Law”

Processor-Memory Performance Gap:
(grows 50% / year)

CPU-DRAM Gap

µProc
60%/yr.

DRAM
7%/yr.
Disk
Funny facts:
• It took 51 years to reach 1TB and 2 years to reach 2TB!
• IBM introduced the first hard disk drive to break the 1 GB barrier in 1980.

Hard Disks

- spinning platter of special material
- mechanical arm with read/write head must be close to the platter to read/write data
- data is stored magnetically
- storage capacity is commonly between 100GB - 3TB
- disks are random access meaning data can be read/written anywhere on the disk

The disk surface spins at a fixed rotational rate

The read/write head is attached to the end of the arm and flies over the disk surface on a thin cushion of air

By moving radially, the arm can position the read/write head over any track
What’s Inside A Disk Drive?

- Spindle
- Arm
- Actuator
- Platters
- Electronics (including a processor and memory!)
- SCSI connector

Image courtesy of Seagate Technology
Disk Drives

- To access data:
  - seek time: position head over the proper track
  - rotational latency: wait for desired sector
  - transfer time: grab the data (one or more sectors)
A Conventional Hard Disk Structure
Hard Disk Architecture

- **Surface** = group of tracks
- **Track** = group of sectors
- **Sector** = group of bytes
- **Cylinder**: several tracks on corresponding surfaces
Disk Sectors and Access

• Each sector records
  - Sector ID
  - Data (512 bytes, 4096 bytes proposed)
  - Error correcting code (ECC)
    - Used to hide defects and recording errors
  - Synchronization fields and gaps

• Access to a sector involves
  - Queuing delay if other accesses are pending
  - Seek: move the heads
  - Rotational latency
  - Data transfer
  - Controller overhead
Disks: Other Issues

- Average seek and rotation times are helped by locality.
- Disk performance improves about 10%/year
- Capacity increases about 60%/year
- Example of disk controllers:
  - SCSI, ATA, SATA
Flash Storage

- Nonvolatile semiconductor storage
  - 100× - 1000× faster than disk
  - Smaller, lower power, more robust
  - But more $/GB (between disk and DRAM)
Flash Types

• **NOR flash**: bit cell like a NOR gate
  - Random read/write access
  - Used for instruction memory in embedded systems

• **NAND flash**: bit cell like a NAND gate
  - Denser (bits/area), but block-at-a-time access
  - Cheaper per GB
  - Used for USB keys, media storage, ...

• **Flash bits wears out after 1000’s of accesses**
  - Not suitable for direct RAM or disk replacement
  - Wear leveling: remap data to less used blocks
Typically:
• pages are 512–4KB in size
• a block consists of 32–128 pages
• A block wears out after roughly 100,000 repeated writes.
• Once a block wears out it can no longer be used.
Main Memory
(DRAM ... For now!)
DRAM

• packaged in memory modules that plug into expansion slots on the main system board (motherboard)

• Example package: 168-pin dual inline memory module (DIMM)
  – transfers data to and from the memory controller in 64-bit chunks
## Storage Trends

### SRAM

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$/MB</td>
<td>2,900</td>
<td>320</td>
<td>256</td>
<td>100</td>
<td>75</td>
<td>60</td>
<td>25</td>
<td>116</td>
</tr>
<tr>
<td>access (ns)</td>
<td>150</td>
<td>35</td>
<td>15</td>
<td>3</td>
<td>2</td>
<td>1.5</td>
<td>1.3</td>
<td>115</td>
</tr>
</tbody>
</table>

### DRAM

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$/MB</td>
<td>880</td>
<td>100</td>
<td>30</td>
<td>1</td>
<td>0.1</td>
<td>0.06</td>
<td>0.02</td>
<td>44,000</td>
</tr>
<tr>
<td>access (ns)</td>
<td>200</td>
<td>100</td>
<td>70</td>
<td>60</td>
<td>50</td>
<td>40</td>
<td>20</td>
<td>10</td>
</tr>
<tr>
<td>typical size (MB)</td>
<td>0.256</td>
<td>4</td>
<td>16</td>
<td>64</td>
<td>2,000</td>
<td>8,000</td>
<td>16,000</td>
<td>62,500</td>
</tr>
</tbody>
</table>

### Disk

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$/GB</td>
<td>100,000</td>
<td>8,000</td>
<td>300</td>
<td>10</td>
<td>5</td>
<td>0.3</td>
<td>0.03</td>
<td>3,333,333</td>
</tr>
<tr>
<td>access (ms)</td>
<td>75</td>
<td>28</td>
<td>10</td>
<td>8</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>typical size (GB)</td>
<td>0.01</td>
<td>0.16</td>
<td>1</td>
<td>20</td>
<td>160</td>
<td>1,500</td>
<td>3,000</td>
<td>300,000</td>
</tr>
</tbody>
</table>
Cache Memory
Von Neumann Bottleneck

Large gap between processor speed and memory speed
A...B...C of Cache
Cache Analogy

- Hungry! must eat!
  - Option 1: go to refrigerator
    - Found → eat!
    - Latency = 1 minute
  - Option 2: go to store
    - Found → purchase, take home, eat!
    - Latency = 20-30 minutes
  - Option 3: grow food!
    - Plant, wait ... wait ... wait ..., harvest, eat!
    - Latency = ~250,000 minutes (~ 6 months)
What Do We Gain?

Let $m =$ cache access time, $M:$ main memory access time
$p =$ probability that we find the data in the cache

$$\text{Average access time} = p \cdot m + (1-p)(m+M)$$
$$= m + (1-p) M$$

We need to increase $p$
Problem

Given the following:
Cache: 1 cycle access time
Main memory: 100 cycle access time

What is the average access time for 100 memory references if you measure that 90% of the cache accesses are hits
Class Problem

For application X, you measure that 40% of the operations access memory. The non-memory access operations take one cycle. You also measure that 90% of the memory references hit in the cache. Whenever a cache miss occurs, the processor is stalled for 20 cycles to transfer the block from memory into the cache. A cache hit takes one cycle. What is the average operations time?
Cache Organization

Fundamental Unit: CACHE BLOCK
Purpose: HOLD DATA

A Simple Cache
A More Complex Cache
(similar to having several caches)
A Single CACHE BLOCK (or LINE)
Parameter: BLOCK SIZE (or LINE SIZE)
A Single CACHE SET (equivalence class)
Parameter: ASSOCIATIVITY
Given 8 cache blocks ...

Direct Mapped

2-Way Set Associative

4-Way Set Associative

8-Way Set Associative
(Fully Associative, or Content-Addressable Memory)
Basic Cache Design

• Cache memory can copy data from any part of main memory
  - It has 2 parts:
    • The **TAG** (CAM) holds the memory address
    • The **BLOCK** (SRAM) holds the memory data

• Accessing the cache:
  - Compare the reference address with the tag
    • If they match, get the data from the cache block
    • If they don’t match, get the data from main memory
Direct Mapped Cache

Example: 32-bit address
Direct Mapped Cache
So...What is a cache?

- Small, fast storage used to improve average access time to slow memory.
- Exploits spatial and temporal locality
- In computer architecture, almost everything is a cache!
  - Registers a cache on variables
  - First-level cache a cache on second-level cache
  - Second-level cache a cache on memory
  - Memory a cache on disk (virtual memory)
  - etc…

![Diagram showing cache hierarchy with Proc/Regs at the top, followed by L1-Cache, L2-Cache, Memory, and Disk, Tape, etc. at the bottom. Slower (Cheaper, Bigger) moves downward, while Faster (More Expensive, Smaller) moves upward.]
Localities:
Why Cache Is a Good Idea?

- **Spatial locality**: If block k is accessed, it is likely that block k+1 will be accessed.
- **Temporal locality**: If block k is accessed, it is likely that it will be accessed again.
Valid  Tag   0  1  \cdots  B-1
\vdots
Valid  Tag   0  1  \cdots  B-1
Set 0:
Valid  Tag   0  1  \cdots  B-1
\vdots
Set 1:
Valid  Tag   0  1  \cdots  B-1
\vdots
Set S-1:
Valid  Tag   0  1  \cdots  B-1

Cache size:  \[ C = B \times E \times S \]  data bytes

1 valid bit per line
t tag bits per line
\( B = 2^b \) bytes per cache block
\( E \) lines per set

\( S = 2^s \) sets
Problem

Show the breakdown of the address for the following cache configuration:

- 32 bit address
- 16K cache
- Direct-mapped cache
- 32-byte blocks

| tag | set index | block offset |
Problem

Show the breakdown of the address for the following cache configuration:

32 bit address
32K cache
4-way set associative cache
32-byte blocks

| tag | set index | block offset |
Cache

- **Associativity**: (DM, 2-way, 4-way,…FA)
- **Block size**
- **Replacement strategy**
  - (LRU, FIFO, LFU, RANDOM)

- **size**
Design Issues

- What to do in case of hit/miss?
- Block size
- Associativity
- Replacement algorithm
- Improving performance
Hits vs. Misses

• Read hits
  - this is what we want!

• Read misses
  - stall the CPU, fetch block from memory, deliver to cache, restart

• Write hits:
  - can replace data in cache and memory (write-through)
  - write the data only into the cache (write-back the cache later)

• Write misses:
  - read the entire block into the cache, then write the word
Improving Cache Performance

1. Reduce the miss rate,
2. Reduce the miss penalty,
3. Reduce power consumption (won’t be discussed here)
Reducing Misses

- **Classifying Misses: 3 Cs**
  - **Compulsory**—The first access to a block is not in the cache, so the block must be brought into the cache. Also called *cold start misses* or *first reference misses*. (Misses in even an Infinite Cache)
  - **Capacity**—If the cache cannot contain all the blocks needed during execution of a program, *capacity misses* will occur due to blocks being discarded and later retrieved.
  - **Conflict**—If block-placement strategy is set associative or direct mapped, *conflict misses* (in addition to compulsory & capacity misses) will occur because a block can be discarded and later retrieved if too many blocks map to its set. Also called *collision misses* or *interference misses*. 
How Can We Reduce Misses?

1) Change Block Size:

2) Change Associativity:

3) Increase Cache Size
Block Size

- Increasing the block size tends to decrease miss rate:
Decreasing miss ratio with associativity

One-way set associative
(direct mapped)

<table>
<thead>
<tr>
<th>Block</th>
<th>Tag</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Two-way set associative

<table>
<thead>
<tr>
<th>Set</th>
<th>Tag</th>
<th>Data</th>
<th>Tag</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Four-way set associative

<table>
<thead>
<tr>
<th>Set</th>
<th>Tag</th>
<th>Data</th>
<th>Tag</th>
<th>Data</th>
<th>Tag</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Eight-way set associative (fully associative)

<table>
<thead>
<tr>
<th>Tag</th>
<th>Data</th>
<th>Tag</th>
<th>Data</th>
<th>Tag</th>
<th>Data</th>
<th>Tag</th>
<th>Data</th>
<th>Tag</th>
<th>Data</th>
<th>Tag</th>
<th>Data</th>
<th>Tag</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Implementation of 4-way set associative
Effect of Associativity on Miss Rate

![Graph showing the effect of associativity on miss rate. The graph plots miss rate against associativity levels for different cache sizes. The x-axis represents one-way, two-way, four-way, and eight-way associativity, while the y-axis represents miss rate percentages from 0% to 15%. The cache sizes for each associativity level are 1 KB, 2 KB, 4 KB, 8 KB, 16 KB, 32 KB, 64 KB, and 128 KB.]
Reducing Miss Penalty

Write Policy 1:
Write-Through vs Write-Back

- **Write-through:** all writes update cache and underlying memory/cache
  - Can always discard cached data - most up-to-date data is in memory
  - Cache control bit: only a *valid* bit
- **Write-back:** all writes simply update cache
  - Can’t just discard cached data - may have to write it back to memory
  - Cache control bits: both *valid* and *dirty* bits

- **Other Advantages:**
  - **Write-through:**
    - memory (or other processors) always have latest data
    - Simpler management of cache
  - **Write-back:**
    - much lower bandwidth, since data often overwritten multiple times
    - Better tolerance to long-latency memory?
Reducing Miss Penalty
Write Policy 2:
Write Allocate vs Non-Allocate
(What happens on write-miss)

• **Write allocate**: allocate new cache line in cache
  – Usually means that you have to do a “read miss” to fill in rest of the cache-line!
• **Write non-allocate** (or “write-around”):
  – Simply send write data through to underlying memory/cache - don’t allocate new cache line!
Decreasing miss penalty with multilevel caches

• Add a second (and third) level cache:
  - often primary cache is on the same chip as the processor
  - use SRAMs to add another cache above primary memory (DRAM)
  - miss penalty goes down if data is in 2nd level cache

• Using multilevel caches:
  - try and optimize the hit time on the 1st level cache
  - try and optimize the miss rate on the 2nd level cache
What about Replacement Algorithm?

- LRU: Least Recently Used
- LFU: Least Frequently Used
- FIFO: First-In First-Out
- Random
How to write cache friendly code?
Is The Following Code Cache Friendly?

```c
int sumvec(int v[N])
{
    int i, sum = 0;

    for (i = 0; i < N; i++)
        sum += v[i];

    return sum;
}
```
Matrix Multiplication Example

- **Description:**
  - Multiply $N \times N$ matrices
  - Matrix elements are doubles (8 bytes)
  - $O(N^3)$ total operations

```c
/* ijk */
for (i=0; i<n; i++) {
    for (j=0; j<n; j++) {
        sum = 0.0;
        for (k=0; k<n; k++)
            sum += a[i][k] * b[k][j];
        c[i][j] = sum;
    }
}
```

Variable `sum` held in register
Miss Rate Analysis for Matrix Multiply

• Assume:
  – Block size = 32B (big enough for four doubles)
  – Matrix dimension (N) is very large
  – Cache is not even big enough to hold multiple rows

• Analysis Method:
  – Look at access pattern of inner loop

\[
C_{i,j} = A_{i,k} \times B_{k,j}
\]
Layout of C Arrays in Memory

- C arrays allocated in row-major order
  - each row in contiguous memory locations
- Stepping through columns in one row:
  - \( \text{for } (i = 0; i < N; i++) \)
    \[ \text{sum } += a[0][i]; \]
  - accesses successive elements
  - if block size (B) > sizeof\( (a_{ij}) \) bytes, exploit spatial locality
- Stepping through rows in one column:
  - \( \text{for } (i = 0; i < n; i++) \)
    \[ \text{sum } += a[i][0]; \]
  - accesses distant elements
  - no spatial locality!
    - miss rate = 1 (i.e. 100%)
Matrix Multiplication (ijk)

```c
/* ijk */
for (i=0; i<n; i++)  {
    for (j=0; j<n; j++) {
        sum = 0.0;
        for (k=0; k<n; k++)
            sum += a[i][k] * b[k][j];
        c[i][j] = sum;
    }
}
```

Misses per inner loop iteration:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.25</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Matrix Multiplication (jik)

```c
/* jik */
for (j=0; j<n; j++) {
    for (i=0; i<n; i++) {
        sum = 0.0;
        for (k=0; k<n; k++)
            sum += a[i][k] * b[k][j];
        c[i][j] = sum
    }
}
```

Misses per inner loop iteration:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.25</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Matrix Multiplication (kij)

/* kij */
for (k=0; k<n; k++) {
    for (i=0; i<n; i++) {
        r = a[i][k];
        for (j=0; j<n; j++)
            c[i][j] += r * b[k][j];
    }
}

Inner loop:

Misses per inner loop iteration:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0</td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Matrix Multiplication (ikj)

```c
/* ikj */
for (i=0; i<n; i++) {
    for (k=0; k<n; k++) {
        r = a[i][k];
        for (j=0; j<n; j++)
            c[i][j] += r * b[k][j];
    }
}
```

Misses per inner loop iteration:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misses</td>
<td>0.0</td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>
Matrix Multiplication (jki)

```c
/* jki */
for (j=0; j<n; j++) {
    for (k=0; k<n; k++) {
        r = b[k][j];
        for (i=0; i<n; i++)
            c[i][j] += a[i][k] * r;
    }
}
```

Inner loop:

- \((*,k)\)
- \((k,j)\)
- \((*,j)\)

Misses per inner loop iteration:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misses</td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Matrix Multiplication (kji)

```c
/* kji */
for (k=0; k<n; k++) {
    for (j=0; j<n; j++) {
        r = b[k][j];
        for (i=0; i<n; i++)
            c[i][j] += a[i][k] * r;
    }
}
```

**Misses per inner loop iteration:**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Conclusions

• The computer system’s storage is organized as a hierarchy.
• The reason of this hierarchy is to try to get an memory that is very fast, cheap, and almost infinite.
• A good programmer must try to make the code cache friendly → make the common case cache friendly → locality