Lecture Topics

- Today: Memory Management
  (Stallings, chapter 7)
- Next: continued

Announcements

- Project #3 (due 2/16)
- Project #4 (due 2/23)
Memory Hierarchy

- Trade-offs among types of storage
  - Faster access time, greater cost per bit
  - Greater capacity, smaller cost per bit
  - Greater capacity, slower access speed

- Moving down the hierarchy:
  - Slower access time, decreasing cost per bit
  - Increasing capacity
  - Decreasing frequency of access by CPU
Locality of Reference

- Memory references for both instructions and data values tend to cluster over time.

- Example: once a loop is entered, there is frequent access to a small set of instructions.

- Hence: once an instruction is referenced, it is likely that the instruction (and nearby instructions) will be referenced again in the near future.

Types of Locality

- Temporal locality: same address referenced repeatedly in the near-term future
  - instructions: loops, functions
  - data: variables

- Spatial locality: nearby addresses referenced in the near-term future
  - instructions: sequential execution
  - data: arrays, similar data structures
Example: C/C++ Program

From www.cse.msu.edu/~cse410/Examples/example01

```c
int sum = 0;

int main()
{
    for( int i = 1; i <= 6; i++)
    {
        sum = sum + i;
    }
}
```

Equivalent ARM Assembly Language

```assembly
.global main
.text
main:   push    {lr}
        mov     r0, #1
loop:   cmp     r0, #6
        bgt     end
        ldr     r2, =sum
        ldr     r1, [r2]
        add     r1, r1, r0
        str     r1, [r2]
        add     r0, r0, #1
        b       loop
end:    pop     {lr}
        mov     pc, lr
```

Equivalent ARM Machine Language

```
.global main
.text
0000 E52DE004 main:   push    {lr}
0004 E3A00001    mov     r0, #1
0008 E3500006    loop:   cmp     r0, #6
000c CA000005    bgt     end
0010 E59F2018    ldr     r2, =sum
0014 E5921000    ldr     r1, [r2]
0018 E0811000    add     r1, r1, r0
001c E5821000    str     r1, [r2]
0020 E2800001    add     r0, r0, #1
0024 EAFFFFF7    b       loop
0028 E49DE004    end:    pop     {lr}
002c E1A0F00E    mov     pc, lr
```

Execution Trace

```
Time  PC        IR
-----  --------  ------
  0 00010800 E52DE004 * main:   push    {lr}
  1 00010804 E3A00001    mov     r0, #1
  2 00010808 E3500006    loop:   cmp     r0, #6
  3 0001080c CA000005    bgt     end
  4 00010810 E59F2018 *   ldr     r2, =sum
  5 00010814 E5921000 *   ldr     r1, [r2]
  6 00010818 E0811000    add     r1, r1, r0
  7 0001081c E5821000 *   str     r1, [r2]
  8 00010820 E2800001    add     r0, r0, #1
  9 00010824 EAFFFFF7    b       loop
 10 00010808 E3500006    loop:   cmp     r0, #6
 11 0001080c CA000005    bgt     end
```

....
Exploiting Locality

- RAM (primary storage) is slow compared to CPU registers (by a factor of about 200):
  - 0.5 ns to access registers
  - 100 ns to access RAM

- Exploit locality of reference by keeping a subset of the instructions and data values in high-speed storage (with mechanism to change the subset of instructions and data values when necessary)

Cache Memory

- Cache: fast (and thus small and expensive)
- Main memory: slow (and thus larger and cheaper)
- Processor first checks cache for requested word
- If not found in cache, a block of memory containing the word is moved to the cache
The Hit Ratio

- Hit ratio: fraction of accesses where item is in cache
- T1: access time for fast memory
- T2: access time for slow memory
- T2 >> T1

When hit ratio is close to 1.0, average access time is close to T1

Average Memory Access Time

Consider a two-level memory hierarchy, where M1 is faster than M2. The average memory access time can be calculated using:

\[
AMAT = H \times T1 + (1-H) \times (T1 + T2) = T1 + (1-H) \times T2
\]

H = hit ratio (fraction of references found in M1)
T1 = access time for M1
T2 = access time for M2
Single Level of Cache

Example

- Processor configuration:
  - cache access time is 1 clock cycle (1 ns)
  - cache miss penalty is 100 clock cycles

- If the requested item is in cache, then it can be accessed in one clock cycle (no delay)

- If the requested item is not in cache, then the processor has to stall until the item can be fetched from RAM
Example (continued)

- For a particular instruction sequence, the hit rate is 97%. What is the average memory access time?

\[
\text{AMAT} = \text{time for a hit} + \text{miss rate} \times \text{miss penalty} \\
= 1 \text{ clock cycle} + 0.03 \times 100 \text{ clock cycles} \\
= 4 \text{ clock cycles (or 4 ns)}
\]

Example (continued)

- Assume the hit rate is 99% instead. What is the average memory access time?

\[
\text{AMAT} = \text{time for a hit} + \text{miss rate} \times \text{miss penalty} \\
= 1 \text{ clock cycle} + 0.01 \times 100 \text{ clock cycles} \\
= 2 \text{ clock cycles (or 2 ns)}
\]
Multiple Levels of Cache

Example

- Processor configuration:
  - Level 1 cache access time is 1 clock cycle (1 ns)
  - Level 2 cache access time is 10 clock cycles
  - RAM access time is 100 clock cycles

- Check L1 cache

- If not found, check L2 cache

- If not found, fetch from RAM
Example (continued)

- For a particular instruction sequence
  - L1 cache: 90% hit rate
  - L2 cache: 80% hit rate for remaining references

- Fraction of references found at each level?
  - Level 1: 90% of all references (0.9)
  - Level 2: 80% of remaining 10% (0.08)
  - RAM: 20% of remaining 10% (0.02)

Example (continued)

- For a particular instruction sequence
  - L1 cache: 90% hit rate
  - L2 cache: 80% hit rate for remaining references

- What is the average memory access time?

\[
AMAT = 1 \text{ clock cycle} + 0.10 \times 10 \text{ clock cycles} \\
+ 0.02 \times 100 \text{ clock cycles} \\
= 1 + 1 + 2 \text{ clock cycles} \\
= 4 \text{ clock cycles (4 ns)}
\]
Example (continued)

- For a particular instruction sequence
  - L1 cache: 95% hit rate
  - L2 cache: 80% hit rate for remaining references

- What is the average memory access time?
  \[
  \text{AMAT} = 1 \text{ clock cycle} + 0.05 \times 10 \text{ clock cycles} \\
  + 0.01 \times 100 \text{ clock cycles} \\
  = 1 + 0.5 + 1 \text{ clock cycles} \\
  = 2.5 \text{ clock cycles (2.5 ns)}
  \]

Cache and RAM Configuration

Unit of transfer between RAM and cache is one block
Each cache slot holds one block
RAM is viewed as being divided into fixed-size blocks
Read Operation

Load instruction: copy data from RAM to CPU

Check cache first – if desired item is already present in the cache, simply copy item from cache to CPU.

If desired item is not already present in the cache, copy a block (item and its neighbors) from RAM to the cache and copy the item to the CPU.

Diagram:

- Start
  - Receive address RA from CPU
  - In block containing RA in cache?
    - Yes: Fetch RA word and deliver to CPU
    - No: Access main memory for block containing RA
      - Allocate cache line for main memory block
      - Load main memory block into cache line
      - Deliver RA word to CPU
  - Done
Write Operation

Store instruction: copy data from CPU to RAM

Check cache first – if desired item is already present in the cache, simply copy item from CPU to cache

If desired item is not already present in the cache, copy item (and its neighbors) from RAM to the cache and copy the item from the CPU

Write Policies

After a store instruction, cache and RAM are inconsistent: contents of block in cache and RAM are different

Two strategies:
- Write through
- Write back
Write Policies

- **Write through**: whenever a cache block is changed, the block is written (copied) to RAM

- **Write back**: cache block is only written (copied) to RAM when the cache line is evicted (replaced)
  - multiple store instructions can occur before block has to be written to RAM
  - modified bit used to indicate that block has been changed (and must be written to RAM)

Cache Organizations
Direct Mapping

Mapping function:

\[ i = j \mod m \]

where

- \( i \) = cache line number
- \( j \) = main memory block number
- \( m \) = number of lines in the cache

Each block maps to exactly one cache line
Example Configuration

- Cache: 64 KB
- RAM: 16 MB (24-bit addresses)
- Block size: 4 bytes
- Cache is organized as $2^{14}$ lines, where each line holds 4 bytes
- RAM is viewed as 4M blocks of 4 bytes each

Example

- Address (24 bits) viewed as three fields:
  - Word: 2 bits to identify byte within word
  - Line: 14 bits to identify cache line
  - Tag: 8 bits (remaining bits)
Example (continued)

**Address:** 16339C

**Address in binary:**

```
000101100011001110011100
```

**Tag:** 00010110 (16)

**Line:** 00110011100111 (0CE7)

**Word:** 00 (0)

Example (continued)

<table>
<thead>
<tr>
<th>Cache line</th>
<th>Addresses of RAM blocks</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>000000, 010000, ..., FF0000</td>
</tr>
<tr>
<td>1</td>
<td>000004, 010004, ..., FF0004</td>
</tr>
<tr>
<td>2</td>
<td>000008, 010008, ..., FF0008</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>2^14 − 1</td>
<td>00FFFC, 01FFFC, ..., FFFFFC</td>
</tr>
</tbody>
</table>
Direct-Mapped Cache