CS 61C: Great Ideas in Computer Architecture

Direct-Mapped Caches, Set Associative Caches, Cache Performance

Instructor: Justin Hsia
Great Idea #3: Principle of Locality/ Memory Hierarchy
Extended Review of Last Lecture

• Why have caches?
  – Intermediate level between CPU and memory
  – In-between in size, cost, and speed

• Memory (hierarchy, organization, structures) set up to exploit *temporal* and *spatial locality*
  – *Temporal*: If accessed, will access again soon
  – *Spatial*: If accessed, will access others around it

• Caches hold a subset of memory (in *blocks*)
  – We are studying how they are designed for fast and efficient operation (lookup, access, storage)
Extended Review of Last Lecture

• Fully Associative Caches:
  – Every block can go in any slot
    • Use random or LRU replacement policy when cache full
  – Memory address breakdown (on request)
    • **Tag** field is identifier (which block is currently in slot)
    • **Offset** field indexes into block
  – *Each* cache slot holds block data, tag, valid bit, and dirty bit (dirty bit is only for *write-back*)
    • The whole cache maintains LRU bits
Extended Review of Last Lecture

• Cache read and write policies:
  – Affect consistency of data between cache and memory
  – Write-back vs. write-through
  – Write allocate vs. no-write allocate

• On memory access (read or write):
  1) Look at ALL cache slots in parallel
  2) If Valid bit is 0, then ignore
  3) If Valid bit is 1 and Tag matches, then use that data

• On write, set Dirty bit if write-back
Extended Review of Last Lecture

- Fully associative cache layout
  - 8-bit address space, 32-byte cache with 8-byte blocks
  - LRU replacement (2 bits), write-back and write allocate
  - Offset – 3 bits, Tag – 5 bits

- Each slot has 71 bits; cache has 4*71+2 = 286 bits
Agenda

• Direct-Mapped Caches
• Administrivia
• Set Associative Caches
• Cache Performance
Direct-Mapped Caches (1/3)

• Each memory block is mapped to exactly one slot in the cache (*direct-mapped*)
  – Every block has only one “home”
  – Use hash function to determine which slot

• Comparison with fully associative
  – Check just one slot for a block (faster!)
  – No replacement policy necessary
  – Access pattern may leave empty slots in cache
Direct-Mapped Caches (2/3)

- **Offset field** remains the same as before

- **Recall:** blocks consist of adjacent bytes
  - Do we want adjacent blocks to map to same slot?
  - **Index field:** Apply hash function to block address to determine *which slot* the block goes in
    - *(block address) modulo (# of blocks in the cache)*

- **Tag field** maintains same function (identifier), but is now shorter
TIO Address Breakdown

• Memory address fields:

```
<table>
<thead>
<tr>
<th>31</th>
<th>30</th>
<th>29</th>
<th>\ldots</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tag</td>
<td>Index</td>
<td>Offset</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

- **Tag** field: T bits
- **Index** field: I bits
- **Offset** field: O bits

• Meaning of the field sizes:
  - O bits \(\leftrightarrow 2^O \text{ bytes/block} = 2^{O-2} \text{ words/block} \)
  - I bits \(\leftrightarrow 2^I \text{ slots in cache} = \text{cache size / block size} \)
  - T bits = A – I – O, where A = # of address bits (A = 32 here)
Direct-Mapped Caches (3/3)

• What’s actually in the cache?
  – Block of data \( (8 \times K = 8 \times 2^O \) bits)  
  – Tag field of address as identifier (\( T \) bits)  
  – Valid bit (1 bit)  
  – Dirty bit (1 bit if write-back)  
  – No replacement management bits!

• Total bits in cache = \# slots \times (8\times K + T + 1 + 1)  
  \[ = 2^I \times (8\times 2^O + T + 1 + 1) \text{ bits} \]
DM Cache Example (1/5)

• Cache parameters:
  – Direct-mapped, address space of 64B, block size of 1 word, cache size of 4 words, write-through

• TIO Breakdown:
  – 1 word = 4 bytes, so $O = \log_2(4) = 2$
  – Cache size / block size = 4, so $I = \log_2(4) = 2$
  – $A = \log_2(64) = 6$ bits, so $T = 6 - 2 - 2 = 2$

• Bits in cache = $2^2 \times (8 \times 2^2 + 2 + 1) = 140$ bits
DM Cache Example (2/5)

• Cache parameters:
  – Direct-mapped, address space of 64B, block size of 1 word, cache size of 4 words, write-through
  – Offset – 2 bits, Index – 2 bits, Tag – 2 bits

<table>
<thead>
<tr>
<th>Index</th>
<th>V</th>
<th>Tag</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>X</td>
<td>XX</td>
<td>0x??</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0x??</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0x??</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0x??</td>
</tr>
<tr>
<td>01</td>
<td>X</td>
<td>XX</td>
<td>0x??</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0x??</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0x??</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0x??</td>
</tr>
<tr>
<td>10</td>
<td>X</td>
<td>XX</td>
<td>0x??</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0x??</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0x??</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0x??</td>
</tr>
<tr>
<td>11</td>
<td>X</td>
<td>XX</td>
<td>0x??</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0x??</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0x??</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0x??</td>
</tr>
</tbody>
</table>

• 35 bits per index/slot, 140 bits to implement
DM Cache Example (3/5)

Which blocks map to each row of the cache?
(see colors)

On a memory request:
(let’s say $001011_{\text{two}}$)

1) Take Index field (10)
2) Check if Valid bit is true in that row of cache
3) If valid, then check if Tag matches

Cache slots exactly match the Index field

Main Memory shown in blocks, so offset bits not shown (x’s)
DM Cache Example (4/5)

- Consider the sequence of memory address accesses

Starting with a cold cache:

\[
\begin{array}{ccccccc}
0 & 2 & 4 & 8 & 20 & 16 & 0 & 2
\end{array}
\]

\begin{itemize}
\item **0 miss**
\item **2 hit**
\item **4 miss**
\item **8 miss**
\end{itemize}
DM Cache Example (5/5)

- Consider the sequence of memory address accesses

Starting with a cold cache: 0 2 4 8 20 16 0 2

- 20 miss

- 16 miss

- 0 miss

- 2 hit

- 8 requests, 6 misses – last slot was never used!
Worst-Case for Direct-Mapped

- Cold DM $\$ that holds 4 1-word blocks
- Consider the memory accesses: 0, 16, 0, 16,...

<table>
<thead>
<tr>
<th>0 Miss</th>
<th>16 Miss</th>
<th>0 Miss</th>
</tr>
</thead>
<tbody>
<tr>
<td>00 M[0-3]</td>
<td>00 M[0-3]</td>
<td>01 M[16-19]</td>
</tr>
</tbody>
</table>

- HR of 0%
  - Ping pong effect: alternating requests that map into the same cache slot
- Does fully associative have this problem?
Comparison So Far

• Fully associative
  – Block can go into *any* slot
  – Must check ALL cache slots on request ("slow")
  – TO breakdown (i.e. \( I = 0 \) bits)
  – “Worst case” still fills cache (more efficient)

• Direct-mapped
  – Block goes into *one specific* slot (set by Index field)
  – Only check ONE cache slot on request ("fast")
  – TIO breakdown
  – “Worst case” may only use 1 slot (less efficient)
Agenda

• Direct-Mapped Caches
• Administrivia
• Set Associative Caches
• Cache Performance
Administrivia

• Proj1 due Sunday
  – Shaun extra OH Saturday 4-7pm
• HW4 released Friday, due next Sunday
• Midterm:
  – Please keep Sat 7/20 open just in case
  – Take old exams for practice
  – Doubled-sided sheet of handwritten notes
  – MIPS Green Sheet provided; no calculators
  – Will cover up through caches
Agenda

- Direct-Mapped Caches
- Administrivia
- Set Associative Caches
- Cache Performance
Set Associative Caches

• Compromise!
  – More flexible than DM, more structured than FA

• *N-way set-associative*: Divide $\$ \$ into sets, each of which consists of N slots
  – Memory block maps to a set determined by Index field and is placed in any of the N slots of that set
  – Call N the *associativity*
  – New hash function:
    (block address) modulo (# sets in the cache)
  – Replacement policy applies to every *set*
Effect of Associativity on TIO (1/2)

• Here we assume a cache of fixed size (C)
• **Offset**: # of bytes in a block (same as before)
• **Index**: Instead of pointing to a *slot*, now points to a *set*, so \( I = \log_2(C/K/N) \)
  - Fully associative (1 set): 0 Index bits!
  - Direct-mapped (\( N = 1 \)): max Index bits
  - Set associative: somewhere in-between
• **Tag**: Remaining identifier bits (\( T = A - I - O \))
Effect of Associativity on TIO (2/2)

• For a fixed-size cache, each increase by a factor of two in associativity doubles the number of blocks per set (i.e. the number of slots) and halves the number of sets – decreasing the size of the Index by 1 bit and increasing the size of the Tag by 1 bit.
Example: Eight-Block Cache Configs

- Total size of $ = \# \text{sets} \times \text{associativity}$
- For fixed $\#$ size, associativity ↑ means $\#$ sets ↓ and slots per set ↑
- With 8 blocks, an 8-way set associative $\#$ is same as a fully associative $\#$
Block Placement Schemes

• Place memory block 12 in a cache that holds 8 blocks

  - **Fully associative:** Can go in *any* of the slots (all 1 set)
  - **Direct-mapped:** Can only go in slot \(12 \mod 8 = 4\)
  - **2-way set associative:** Can go in either slot of set \(12 \mod 4 = 0\)
SA Cache Example (1/5)

• Cache parameters:
  – 2-way set associative, 6-bit addresses, 1-word blocks, 4-word cache, write-through

• How many sets?
  – \( C/K/N = 4/1/2 = 2 \) sets

• TIO Breakdown:
  – \( O = \log_2(4) = 2 \), \( I = \log_2(2) = 1 \), \( T = 6 - 1 - 2 = 3 \)

Memory Addresses: [XXX][X][XX]

Block address
SA Cache Example (2/5)

- **Cache parameters:**
  - 2-way set associative, 6-bit addresses, 1-word blocks, 4-word cache, write-through
  - Offset – 2 bits, Index – 1 bit, Tag – 3 bits

- 36 bits per slot, $36 \times 2 + 1 = 73$ bits per set, $2 \times 73 = 146$ bits to implement
SA Cache Example (3/5)

Cache:
Set Slot V Tag Data

Main Memory:
0000xx
0001xx
0010xx
0011xx
0100xx
0101xx
0110xx
0111xx
1000xx
1001xx
1010xx
1011xx
1100xx
1101xx
1110xx
1111xx

Each block maps into one set (either slot) (see colors)

On a memory request:
(let’s say 001011two)

1) Take Index field (0)
2) For EACH slot in set, check valid bit, then compare Tag

Set numbers exactly match the Index field

Main Memory shown in blocks, so offset bits not shown (x’s)
SA Cache Example (4/5)

- Consider the sequence of memory address accesses

  **Starting with a cold cache:**

  0  2  4  8  20  16  0  2

<table>
<thead>
<tr>
<th>0 miss</th>
<th>2 hit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>0 1 000</td>
<td>0x?? 0x?? 0x?? 0x??</td>
</tr>
<tr>
<td>0 1 000</td>
<td>0x?? 0x?? 0x?? 0x??</td>
</tr>
<tr>
<td>0 1 000</td>
<td>0x?? 0x?? 0x?? 0x??</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>4 miss</th>
<th>8 miss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>0 1 000</td>
<td>0x?? 0x?? 0x?? 0x??</td>
</tr>
<tr>
<td>0 1 000</td>
<td>0x?? 0x?? 0x?? 0x??</td>
</tr>
</tbody>
</table>

|        |       |
| 0 1 000 | 0x?? 0x?? 0x?? 0x?? |
SA Cache Example (5/5)

- Consider the sequence of memory address accesses
  
  Starting with a cold cache:

  0 2 4 8 20 16 0 2
  M H M M

  20 miss


  16 miss


  0 miss


  2 hit


- 8 requests, 6 misses
Worst Case for Set Associative

• Worst case for DM was repeating pattern of 2 into same cache slot (HR = 0/n)
  – Set associative for N > 1: HR = (n-2)/n
• Worst case for N-way SA with LRU?
  – Repeating pattern of at least N+1 that maps into same set
  – Back to HR = 0:
Question: What is the TIO breakdown for the following cache?

- 32-bit address space
- 32 KiB 4-way set associative cache
- 8 word blocks

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>I</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>21</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>(B)</td>
<td>19</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>(C)</td>
<td>19</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>(D)</td>
<td>17</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>
Get To Know Your Instructor
Agenda

• Direct-Mapped Caches
• Administrivia
• Set Associative Caches
• Cache Performance
Cache Performance

• Two things hurt the performance of a cache:
  – Miss rate and miss penalty
• Average Memory Access Time (AMAT): average time to access memory considering both hits and misses
  \[ \text{AMAT} = \text{Hit time} + \text{Miss rate} \times \text{Miss penalty} \]
  (abbreviated \( \text{AMAT} = \text{HT} + \text{MR} \times \text{MP} \))

• **Goal 1:** Examine how changing the different cache parameters affects our AMAT (Lec 12)
• **Goal 2:** Examine how to optimize your code for better cache performance (Lec 14, Proj 2)
AMAT Example

• **Processor specs:** 200 ps clock, MP of 50 clock cycles, MR of 0.02 misses/instruction, and HT of 1 clock cycle
  
  \[
  AMAT = 1 + 0.02 \times 50 = 2 \text{ clock cycles} = 400 \text{ ps}
  \]

• Which improvement would be best?
  
  – 190 ps clock
    
    380 ps
  
  – MP of 40 clock cycles
    
    360 ps
  
  – MR of 0.015 misses/instruction
    
    350 ps
Cache Parameter Example

• What is the potential impact of much larger cache on AMAT? (same block size)
  1) Increase HR
  2) Longer HT: smaller is faster
     – At some point, increase in hit time for a larger cache may overcome the improvement in hit rate, yielding a decrease in performance

• Effect on TIO? Bits in cache? Cost?
Effect of Cache Performance on CPI

- **Recall:** CPU Performance
  
  \[
  \text{CPU Time} = \text{Instructions} \times \text{CPI} \times \text{Clock Cycle Time} \tag{IC} \times \text{CC}
  \]

- Include memory accesses in CPI:
  
  \[
  \text{CPI}_{\text{stall}} = \text{CPI}_{\text{base}} + \text{Average Memory-stall Cycles}
  \]
  
  \[
  \text{CPU Time} = \text{IC} \times \text{CPI}_{\text{stall}} \times \text{CC}
  \]

- Simplified model for memory-stall cycles:
  
  \[
  \text{Memory-stall cycles} = \frac{\text{Accesses}}{\text{Instruction}} \times \text{MR} \times \text{MP}
  \]

  — Will discuss more complicated models next lecture
**CPI Example**

- **Processor specs:** $CPI_{\text{base}}$ of 2, a 100 cycle MP, 36% load/store instructions, and 2% I$ and 4% D$ MRs
  - How many times per instruction do we access the I$? The D$?
  - MP is assumed the same for both I$ and D$
  - Memory-stall cycles will be sum of stall cycles for both I$ and D$
CPI Example

- **Processor specs:** CPI\textsubscript{base} of 2, a 100 cycle MP, 36% load/store instructions, and 2% I$ and 4% D$ MRs
  - Memory-stall cycles
    \[
    \text{I$} = (100\% \times 2\% + 36\% \times 4\%) \times 100 = 3.44
    \]
    \[
    \text{D$} = (more \text{ than } 2 \times \text{ CPI}_{\text{base}}!)
    \]
  - CPI\textsubscript{stall} = 2 + 3.44 = 5.44

- What if the CPI\textsubscript{base} is reduced to 1?
- What if the D$ miss rate went up by 1%?
Impacts of Cache Performance

\[ CPI_{stall} = CPI_{base} + \text{Memory-stall Cycles} \]

- Relative penalty of cache performance increases as processor performance improves (faster clock rate and/or lower \( CPI_{base} \))
  - Relative contribution of \( CPI_{base} \) and memory-stall cycles to \( CPI_{stall} \)
  - Memory speed unlikely to improve as fast as processor cycle time

- What can we do to improve cache performance?
Sources of Cache Misses: The 3Cs

- **Compulsory**: (cold start or process migration, 1st reference)
  - First access to block impossible to avoid; Effect is small for long running programs

- **Capacity**:
  - Cache cannot contain all blocks accessed by the program

- **Conflict**: (collision)
  - Multiple memory locations mapped to the same cache location
The 3Cs: Design Solutions

• Compulsory:
  – Increase block size (increases MP; too large blocks could increase MR)

• Capacity:
  – Increase cache size (may increase HT)

• Conflict:
  – Increase cache size
  – Increase associativity (may increase HT)
Summary

• Set associativity determines flexibility of block placement
  – Fully associative: blocks can go anywhere
  – Direct-mapped: blocks go in one specific location
  – N-way: cache split into sets, each of which have \( n \) slots to place memory blocks

• Cache Performance
  – AMAT = HT + MR \times MP
  – CPU time = IC \times CPI_{\text{stall}} \times CC
    = IC \times (CPI_{\text{base}} + \text{Memory-stall cycles}) \times CC