CS31: Introduction to Computer Systems

Week 9, Class 2
Caching
03/28/24

Dr. Sukrit Venkatagiri
Swarthmore College
Welcome!
<table>
<thead>
<tr>
<th>Wk</th>
<th>Lecture</th>
<th>Lab</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Intro to C</td>
<td>C Arrays, Sorting</td>
</tr>
<tr>
<td>2</td>
<td>Binary Representation, Arithmetic</td>
<td>Data Rep. &amp; Conversion</td>
</tr>
<tr>
<td>3</td>
<td>Digital Circuits</td>
<td>Circuit Design</td>
</tr>
<tr>
<td>4</td>
<td>ISAs &amp; Assembly Language</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Pointers and Memory</td>
<td>Pointers and Assembly</td>
</tr>
<tr>
<td>6</td>
<td>Functions and the Stack</td>
<td>Maze Lab</td>
</tr>
<tr>
<td>7</td>
<td>Arrays, Structures &amp; Pointers</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>Spring Break</strong></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Storage and Memory Hierarchy</td>
<td>Game of Life</td>
</tr>
<tr>
<td>9</td>
<td>Storage, Caching</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Caching, Operating System, Processing</td>
<td>Strings</td>
</tr>
<tr>
<td>11</td>
<td>Virtual Memory</td>
<td>Unix Shell</td>
</tr>
<tr>
<td>12</td>
<td>Parallel Applications, Threading</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Threading</td>
<td>pthreads Game of Life</td>
</tr>
<tr>
<td>14</td>
<td>Threading</td>
<td></td>
</tr>
</tbody>
</table>

**Where are we?**

- **C** (programming language)
- **x86 Assembly** (instruction set architecture)
- **Binary** (logic / bits)
- **CPU / memory** (logic / bits)
- **logic gates, circuits**
- **voltage**
Reading Quiz
How many lines of code can steal AI data from graphical processing units?

A. 5

B. 100

C. 10

D. 27
Removing data from a cache to make room for new data is known as...

A. Displacement

B. Eviction

C. Expulsion

D. Relocation
Which of these reflects how memory addresses are partitioned for caching?

A. Byte offset | Index | Tag

B. Tag | Index | Byte offset

C. Index | Tag | Byte offset

D. Byte offset | Tag | Index
Which two one-bit pieces of metadata would you expect to find on a direct-mapped cache line?

A. a *dirty* and a *valid* bit

B. a *dirty* and an *overflow* bit

C. an *overflow* and a *valid* bit
Abstraction Goal

• **Reality**: There is no one type of memory to rule them all!

• **Abstraction**: hide the complex/undesirable details of reality

• **Illusion**: We have the speed of SRAM, with the capacity of disk, at reasonable cost
The Memory Hierarchy

- **Local secondary storage (disk)**
  - Larger
  - Slower
  - Cheaper per byte

- **Remote secondary storage** (tapes, Web servers / Internet)
  - Slower than local disk to access
  - ~100 M cycles to access

- **Main memory (DRAM)**
  - ~100 cycles to access

- **Cache(s) (SRAM)**
  - ~10’s of cycles to access

- **Registers**
  - 1 cycle to access

**Access Times and Characteristics**

- CPU instructions can directly access registers.
- On-chip storage is faster and smaller per byte.
- Local secondary storage is cheaper per byte.
- Remote secondary storage is the slowest and most expensive.
Data Access Time over Years

Over time, gap widens between DRAM, disk, and CPU speeds.

- Disk seek time
- Flash SSD access time
- DRAM access time
- SRAM access time
- CPU cycle time
- Effective CPU cycle time

Really want to avoid going to disk for data

Want to avoid going to Main Memory for data

multicore
Memory: Units of Storage

- **Bit**
- **Nibble (4 Bits)**
- **Byte (8 Bits)**
- **Kilobyte (KB) - 1024 Bytes**
- **Megabyte (MB) - 1024 Kilobyte (KB)**
- **Gigabyte (GB) - 1024 Megabyte (MB)**
- **Terabyte (TB) - 1024 Gigabyte (GB)**
- **Petabyte (PB) - 1024 (TB)**, **Exabyte (EB) - 1024 (PB)**
- **Zettabyte (ZB) - 1024 (EB)**, **Yottabyte (YB) - 1024 (ZB)**
Recall

• A cache is a smaller, faster memory, that holds a subset of a larger (slower) memory

• We take advantage of locality to keep data in cache as often as we can!

• When accessing memory, we check cache to see if it has the data we’re looking for
Why we miss...

• Compulsory (cold-start) miss:
  – First time we use data, load it into cache

• Capacity miss:
  – Cache is too small to store all the data we’re using

• Conflict miss:
  – To bring in new data to the cache, we evicted other data that we’re still using
Cache Design

• Lot’s of characteristics to consider:
  – Where should data be stored in the cache?
Cache Design

• Lot’s of **characteristics** to consider:
  – Where should data be stored in the cache?
  – What size data chunks should we store? (block size)
Cache Design

• Lot’s of characteristics to consider:
  – Where should data be stored in the cache?
  – What size data chunks should we store? (block size)

• Goals:
  – Maximize hit rate
  – Maximize (temporal & spatial) locality benefits
  – Reduce cost/complexity of design
Suppose the CPU asks for data, it’s not in cache. We need to move it into cache from memory. Where in the cache should it be allowed to go?

A. In exactly one place

B. In a few places

C. In most places, but not all

D. Anywhere in the cache
A larger *block size* (caching memory in larger chunks) is likely to exhibit... Why?

A. Better temporal locality

B. Better spatial locality

C. Fewer misses (better hit rate)

D. More misses (worse hit rate)

E. More than one of the above
Block Size Implications

- **Small blocks**
  - Room for more blocks
  - Fewer conflict misses

- **Large blocks**
  - Fewer trips to memory
  - Longer transfer time
  - Fewer cold-start misses
Trade-offs

• There is no single best design for all purposes!

• Common systems question: which point in the design space should we choose?

• Given a particular scenario:
  – Analyze needs
  – Choose design that fits the bill
Real CPUs

• **Goals**: general purpose processing
  – balance needs of many use cases
  – middle of the road: jack of all trades, master of none

• Some associativity, medium size blocks:
  – 8-way associative (memory in one of eight places)
  – 16 or 32 or 64-byte blocks
What should we use to determine whether or not data is in the cache?

A. The memory address of the data

B. The value of the data

C. The size of the data

D. Some other aspect of the data
Up Next:

- Direct-Mapped Cache
- Eviction
- Associativity
- Cache Conscious Programming

- **Office hours: today from 2-3:30pm**
Recall: Memory Reads

CPU places address A on the memory bus.

Load operation: `mov (Address A), %rax`
Recall: Memory Reads

Memory retrieves value and sends it across bus.

CPU reads value from the bus, and copies it into register rax, a copy also goes into the on-chip cache memory.
Memory Address Tells Us...

• Is the block containing the byte(s) you want already in the cache?

• If not, where should we put that block?
  – Do we need to kick out (“evict”) another block?

• Which byte(s) within the block do you want?
Memory Addresses

• Like everything else: series of bits (64)

• Keep in mind:
  – N bits gives us $2^N$ unique values.

• 64-bit address:
  – 1011000101110010110101000101011010110001011100101101010001010110

Divide into regions, each with distinct meaning.
Address Division

• First section: Tag
  – Of all the addresses that map to this location, which one is here?
  – Number of bits for this section is any bits left over after index and offset.

• Second section: Index
  – Which location(s) in the cache should we check for the data with this address?
  – Number of bits for this section depends on the number of cache locations.

• Third section: Offset
  – If we find a block of bytes in the cache (on a hit) which bytes within the block do we actually want?
  – Number of bits for this section depends on the block size – must be able to uniquely identify every byte in the block.
Direct-Mapped

• One place data can be

• Example: let’s assume some parameters:
  – 1024 cache locations (every block mapped to one)
  – Block size of 8 bytes
WS #1: Direct-Mapped

1024 cache locations (every block mapped to one)
Block size of 8 bytes

<table>
<thead>
<tr>
<th>Line</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</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>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1020</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1021</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1022</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1023</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Cache Metadata

• Valid bit: is the entry valid?
  – If set: data is correct, use it if we ‘hit’ in cache
  – If not set: ignore ‘hits’, the data is garbage

• Dirty bit: has the data been written?
  – Used by write-back caches
  – If set, need to update memory before eviction
Direct-Mapped

- Given:
  - 1024 cache locations
  - Block size of 8 bytes

- Address division:
  - Identify byte in block
    - How many bits? 3
  - Identify which row (line)
    - How many bits? 10
Direct-Mapped

- Address division:

<table>
<thead>
<tr>
<th>Tag (51 bits)</th>
<th>Index (10 bits)</th>
<th>Byte offset (3 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Index:
- Which line (row) should we check?
- Where could data be?

```
<table>
<thead>
<tr>
<th>Line</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</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>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1020</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1021</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1022</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1023</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
Direct-Mapped

• Address division:

<table>
<thead>
<tr>
<th>Tag (51 bits)</th>
<th>Index (10 bits)</th>
<th>Byte offset (3 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Index:
Which line (row) should we check?
Where could data be?

<table>
<thead>
<tr>
<th>Line</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</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>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1020</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1021</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1022</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1023</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Direct-Mapped

- **Address division:**

<table>
<thead>
<tr>
<th>Tag (51 bits)</th>
<th>Index (10 bits)</th>
<th>Byte offset (3 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4217</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

In parallel, check:

Tag:
Does the cache hold the data we’re looking for, or some other block?

Valid bit:
If entry is not valid, don’t trust garbage in that line (row).

If tag doesn’t match, or line is invalid, it’s a miss!
Direct-Mapped

• Address division:

<table>
<thead>
<tr>
<th>Tag (51 bits)</th>
<th>Index (10 bits)</th>
<th>Byte offset (3 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4217</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

Byte offset tells us which subset of block to retrieve.
Direct-Mapped

• Address division:

<table>
<thead>
<tr>
<th>Address</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0101</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0110</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0111</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1001</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1010</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1011</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1101</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1110</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1111</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Byte offset tells us which subset of block to retrieve.
### Input: Memory Address

<table>
<thead>
<tr>
<th>Tag</th>
<th>Index</th>
<th>Byte offset</th>
</tr>
</thead>
</table>

#### Table:

<table>
<thead>
<tr>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **0**: miss
- **1**: hit

**Select Byte(s)**

**Data**

**Diagram**
Direct-Mapped Example

• Suppose our addresses are 16 bits long.

• Our cache has 16 entries, block size of 16 bytes
  – 4 bits in address for the index
  – 4 bits in address for byte offset
  – Remaining bits (8): tag
Direct-Mapped Example

• Let’s say we access memory at address:
  – 0110101100110100

• Step 1:
  – Partition address into tag, index, offset
Direct-Mapped Example

• Let’s say we access memory at address:
  – 01101011 0011 0100

• Step 1:
  – Partition address into tag, index, offset
WS #2: Direct-Mapped Example

• Let’s say we access memory at address:
  – 01101011 0011 0100

• Step 2:
  – Use index to find line (row)
  – 0011 -> 3
Direct-Mapped Example

- Let’s say we access memory at address:
  - 01101011 0011 0100

- Step 2:
  - Use index to find line (row)
  - 0011 -> 3
Direct-Mapped Example

- Let’s say we access memory at address:
  - 01101011 0011 0100

- Note:
  - ANY address with 0011 (3) as the middle four index bits will map to this cache line.
  - e.g. 11111111 0011 0000

Use tag to store high-order bits. Let’s us determine which data is here! (many addresses map here)

So, which data is here?
Data from address 0110101100110100 OR 1111111001100000?
Direct-Mapped Example

• Let’s say we access memory at address:
  – 01101011 0011 0100

• Step 3:
  – Check the tag
  – Is it 01101011 (hit)?
  – Something else (miss)?
  – (Must also ensure valid)
Eviction

• If we don’t find what we’re looking for (miss), we need to bring in the data from memory.

• Make room by kicking something out.
  – If line to be evicted is dirty, write it to memory first

• Another important systems distinction:
  – Mechanism: An ability or feature of the system (What you can do)
  – Policy: Governs the decisions making for using the mechanism (What you should do)
Eviction

• For direct-mapped cache:
  – **Mechanism**: overwrite bits in cache line, updating
    • Valid bit
    • Tag
    • Data
  – **Policy**: not many options for a direct-mapped cache
    • Overwrite at the only location it could be!
Eviction: Direct-Mapped

• Address division:

<table>
<thead>
<tr>
<th>Tag (51 bits)</th>
<th>Index (10 bits)</th>
<th>Byte offset (3 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3941</td>
<td>1020</td>
<td></td>
</tr>
</tbody>
</table>

Find line:

Tag doesn’t match, bring in from memory.

If dirty, write back first!

<table>
<thead>
<tr>
<th>Line</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</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>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1020</td>
<td>1</td>
<td>0</td>
<td>1323</td>
<td>57883</td>
</tr>
<tr>
<td>1021</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1022</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1023</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Eviction: Direct-Mapped

• Address division:

<table>
<thead>
<tr>
<th>Tag (51 bits)</th>
<th>Index (10 bits)</th>
<th>Byte offset (3 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3941</td>
<td>1020</td>
<td></td>
</tr>
</tbody>
</table>

1. Send address to read main memory.

Main Memory
Eviction: Direct-Mapped

- Address division:

<table>
<thead>
<tr>
<th>Tag (51 bits)</th>
<th>Index (10 bits)</th>
<th>Byte offset (3 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3941</td>
<td>1020</td>
<td></td>
</tr>
</tbody>
</table>

1. Send address to read main memory.

2. Copy data from memory. Update tag.

```
Line | V | D | Tag | Data (8 Bytes) |
-----|---|---|-----|----------------|
 0    |   |   |     |                |
 1    |   |   |     |                |
 2    |   |   |     |                |
 3    |   |   |     |                |
 4    |   |   |     |                |
...   |   |   |     |                |
1020  | 1 | 0 | 3941 | 92             |
1021  |   |   |     |                |
1022  |   |   |     |                |
1023  |   |   |     |                |
```
WS #3: Suppose we had 8-bit addresses, a cache with 8 lines, and a block size of 4 bytes.

- How many bits would we use for:
  - Tag?
  - Index?
  - Offset?
How would the cache change if we performed the following memory operations?

Read 01000100 (Value: 5)
Read 11100010 (Value: 17)
Write 01110000 (Value: 7)
Read 10101010 (Value: 12)
Write 01101100 (Value: 2)
How would the cache change if we performed the following memory operations?

- Read 01000100 (Value: 5)
- Read 11100010 (Value: 17)
- Write 01110000 (Value: 7)
- Read 10101010 (Value: 12)
- Write 01101100 (Value: 2)
How would the cache change if we performed the following memory operations?

<table>
<thead>
<tr>
<th>Line</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (4 Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>111</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>011 010</td>
<td>9 5</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>101</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>001</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>011</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>111</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>101</td>
<td>32</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>110</td>
<td>3</td>
</tr>
</tbody>
</table>

No change necessary.

Read 01000100 (Value: 5)
Read 11100010 (Value: 17)
Write 01110000 (Value: 7)
Read 10101010 (Value: 12)
Write 01101100 (Value: 2)
How would the cache change if we performed the following memory operations?

- Read 01000100 (Value: 5)
- Read 11100010 (Value: 17)
- Write 01110000 (Value: 7)
- Read 10101010 (Value: 12)
- Write 01101100 (Value: 2)
How would the cache change if we performed the following memory operations?

Read 01000100 (Value: 5)
Read 11100010 (Value: 17)
Write 01110000 (Value: 7)
Read 10101010 (Value: 12)
Write 01101100 (Value: 2)

Note: tag happened to match, but line was invalid.
How would the cache change if we performed the following memory operations?

Read 01000100 (Value: 5)
Read 11100010 (Value: 17)
Write 01110000 (Value: 7)
Read 10101010 (Value: 12)
Write 01101100 (Value: 2)

1. Write dirty line to memory.
2. Load new value, set it to 2, mark it dirty (write).
Associativity

• Problem: suppose we’re only using a small amount of data (e.g., two integers from different cache blocks)

• Bad luck: (both) blocks map to same cache line
  – Constantly evicting one another
  – Rest of cache is going unused!

• **Associativity**: allow a set blocks to be stored at the same index. Goal: reduce conflict misses.
Comparison

**Direct-mapped**
- Tag tells you if you found the correct data.
- Offset specifies which byte within block.
- Middle bits (index) tell you which 1 line to check.
- (+) Low complexity, fast.
- (-) Conflict misses.

**N-way set associative**
- Tag tells you if you found the correct data.
- Offset specifies which byte within block.
- Middle bits (set) tell you which N lines to check.
- (+) Fewer conflict misses.
- (-) More complex, slower, consumes more power.
Comparison: 1024 Lines
(For the same cache size, in bytes.)

**Direct-mapped**
- 1024 indices (10 bits)

**2-way set associative**
- 512 sets (9 bits)
  - Tag slightly (1 bit) larger.

<table>
<thead>
<tr>
<th>Set #</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></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>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>508</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>509</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>510</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>511</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## 2-Way Set Associative

<table>
<thead>
<tr>
<th>Tag (52 bits)</th>
<th>Set (9 bits)</th>
<th>Byte offset (3 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3941</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

### Bit Distribution

- **Tag (52 bits)**
- **Set (9 bits)**
- **Byte offset (3 bits)**

### Table

<table>
<thead>
<tr>
<th>Set #</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></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>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4063</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>508</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>509</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>510</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>511</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Same capacity as previous example: 1024 rows with 1 entry vs. 512 rows with 2 entries
### 2-Way Set Associative

<table>
<thead>
<tr>
<th>Tag (52 bits)</th>
<th>Set (9 bits)</th>
<th>Byte offset (3 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3941</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Set #</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></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>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4063</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>508</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>509</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>510</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>511</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Check all locations in the set, in parallel.
## 2-Way Set Associative

<table>
<thead>
<tr>
<th>Tag (52 bits)</th>
<th>Set (9 bits)</th>
<th>Byte offset (3 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3941</td>
<td>4</td>
<td></td>
</tr>
</tbody>
</table>

### Multiplexer
- Select correct value.
4-Way Set Associative Cache

Clearly, more complexity here!
Eviction

• Mechanism is the same…
  – Overwrite bits in cache line: update tag, valid, data

• Policy: choose which line in the set to evict
  – Pick a random line in set?
  – Choose an invalid line first
  – Choose the least recently used block
    • Has exhibited the least locality, kick it out!

Common combo in practice.
Least Recently Used (LRU)

- **Intuition**: if it hasn’t been used in a while, we have no reason to believe it will be used soon

- Need extra state to keep track of LRU info

<table>
<thead>
<tr>
<th>Set #</th>
<th>LRU</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (8 Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4063</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>0 3941</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Least Recently Used (LRU)

• **Intuition**: if it hasn’t been used in a while, we have no reason to believe it will be used soon

• Need extra state to keep track of LRU info

• For perfect LRU info:
  • 2-way: 1 bit
  • 4-way: 8 bits
  • N-way: \( N \times \log_2 N \) bits

Another reason why associativity often maxes out at 8 or 16.

These are metadata bits, not “useful” program data storage.

(Approximations make it not quite as bad.)
How would the cache change if we performed the following memory operations? (2-way set)

Read 01000100 (Value: 5)
Read 11100010 (Value: 17)
Write 01100100 (Value: 7)
Read 01000110 (Value: 5)
Write 01100000 (Value: 2)

<table>
<thead>
<tr>
<th>Set #</th>
<th>LRU</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (4 Bytes)</th>
<th>V</th>
<th>D</th>
<th>Tag</th>
<th>Data (4 Bytes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>111</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>001</td>
<td>17</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>111</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>010</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

LRU of 0 means the left line in the set was least recently used. 1 means the right line was used least recently.
Cache Conscious Programming

• Knowing about caching and designing code around it can significantly effect performance

(ex) 2D array accesses

Algorithmically, both $O(N \times M)$.

Is one faster than the other?

```plaintext
for(i=0; i < N; i++) {
    for(j=0; j < M; j++) {
        sum += arr[i][j];
    }
}
```
Cache Conscious Programming

• Knowing about caching and designing code around it can significantly effect performance

(ex) 2D array accesses

$$\text{for}(i=0; i < N; i++) \{$$
$$\quad \text{for}(j=0; j < M; j++) \{$$
$$\quad \quad \text{sum} += \text{arr}[i][j];$$
$$\quad \}$$
$$\}$$

$$\text{for}(j=0; j < M; j++) \{$$
$$\quad \text{for}(i=0; i < N; i++) \{$$
$$\quad \quad \text{sum} += \text{arr}[i][j];$$
$$\quad \}$$
$$\}$$

A. is faster.

B. is faster.

Algorithmically, both $O(N \times M)$.

Is one faster than the other?

C. Both would exhibit roughly equal performance.
Cache Conscious Programming

The first nested loop is more efficient if the cache block size is larger than a single array bucket (for arrays of basic C types, it will be).

```c
for(i=0; i < N; i++) {
    for(j=0; j< M; j++) {
        sum += arr[i][j];
    }
}
```

(ex) 1 miss every 4 buckets vs. 1 miss every bucket
Program Efficiency and Memory

• Be aware of how your program accesses data
  – Sequentially, in strides of size X, randomly, ...
  – How data is laid out in memory

• Will allow you to structure your code to run much more efficiently based on how it accesses its data

• Don’t go nuts...
  – Optimize the most important parts, ignore the rest
  – “Premature optimization is the root of all evil.” -Knuth
Amdahl’s Law

Idea: an optimization can improve total runtime at most by the fraction it contributes to total runtime.

If program takes 100 secs to run, and you optimize a portion of the code that accounts for 2% of the runtime, the best your optimization can do is improve the runtime by 2 secs.

Amdahl’s Law tells us to focus our optimization efforts on the code that matters:

Speed-up what is accounting for the largest portion of runtime to get the largest benefit. And, don’t waste time on the small stuff.
Up Next:

- Operating Systems, Processes
- Virtual Memory