Water circuits

“If you think programming a computer is hard, just imagine what it would be if your bits were leaking all over the place.”

http://www.blikstein.com/paulo/projects/project_water.html

Review

• Pipeline challenge is hazards
  • Forwarding helps with many data hazards
  • Delayed branch helps with control hazard in 5 stage pipeline
  • Load delay slot / interlock necessary
• More aggressive performance:
  • Superscalar
  • Out-of-order execution
• Use caches to simulate fast large memory

Memory Hierarchy

As we move to deeper levels the latency goes up and price per bit goes down.

Motivation: Why We Use Caches (written $)

- 1989 first Intel CPU with cache on chip
- 1998 Pentium III has two levels of cache on chip

Cache Design

• How do we organize cache?
• Where does each memory address map to?
  (Remember that cache is subset of memory, so multiple memory addresses map to the same cache location.)
• How do we know which elements are in cache?
• How do we quickly locate them?

Direct-Mapped Cache (1/4)

• In a direct-mapped cache, each memory address is associated with one possible block within the cache
  • Therefore, we only need to look in a single location in the cache for the data if it exists in the cache
  • Block is the unit of transfer between cache and memory
Direct-Mapped Cache (2/4)

Cache Location 0 can be occupied by data from:
- Memory location 0, 4, 8, ...
- 4 blocks ⇒ ⇒ ⇒ ⇒ any memory location that is multiple of 4

What if we wanted a block to be bigger than one byte?

Direct-Mapped Cache (3/4)

When we ask for a byte, the system finds out the right block, and loads it all!
- How does it know right block?
- How do we select the byte?
- E.g., Mem address 11101?
- How does it know WHICH colored block it originated from?
- What do you do at baggage claim?

Direct-Mapped Cache (4/4)

What should go in the tag?
- Do we need the entire address?
- What do all these tags have in common?
- What did we do with the immediate when we were branch addressing, always count by bytes?
- Why not count by cache #?

It’s useful to draw memory with the same width as the block size

Issues with Direct-Mapped

- Since multiple memory addresses map to same cache index, how do we tell which one is in there?
- What if we have a block size > 1 byte?
- Answer: divide memory address into three fields

Tag to check Offset to select within correct block

Direct-Mapped Cache Terminology

- All fields are read as unsigned integers.
- Index: specifies the cache index (which “row”/block of the cache we should look in)
- Offset: once we’ve found correct block, specifies which byte within the block we want
- Tag: the remaining bits after offset and index are determined; these are used to distinguish between all the memory addresses that map to the same location

TIO Dan’s great cache mnemonic

AREA (cache size, B) = HEIGHT (# of blocks) * WIDTH (size of one block, B/block)

\[ 2^{(H+2W)} = 2^H \times 2^W \]
Direct-Mapped Cache Example (1/3)

- Suppose we have a 16KB of data in a direct-mapped cache with 4 word blocks.
- Determine the size of the tag, index, and offset fields if we're using a 32-bit architecture.
- Offset:
  - need to specify correct byte within a block
  - block contains 4 words
    \[ = 16 \text{ bytes} \]
    \[ = 2^4 \text{ bytes} \]
  - need 4 bits to specify correct byte

Direct-Mapped Cache Example (2/3)

- Index: (~index into an “array of blocks”)
  - need to specify correct block in cache
  - cache contains 16 KB = \(2^{14}\) bytes
  - block contains \(2^4\) bytes (4 words)
  - \# blocks/cache
    \[ = \frac{\text{bytes/cache}}{\text{bytes/block}} \]
    \[ = \frac{2^{14}}{2^4} \text{blocks/cache} = 2^{10} \text{blocks/cache} \]
  - need 10 bits to specify this many blocks

Direct-Mapped Cache Example (3/3)

- Tag: use remaining bits as tag
  - tag length = addr length – offset - index
    \[ = 32 - 4 - 10 \text{ bits} \]
    \[ = 18 \text{ bits} \]
  - so tag is leftmost 18 bits of memory address
- Why not full 32 bit address as tag?
  - All bytes within block need same address (4b)
  - Index must be same for every address within a block, so it’s redundant in tag check, thus can leave off to save memory (here 10 bits)

Caching Terminology

- When we try to read memory, 3 things can happen:
  1. cache hit:
     - cache block is valid and contains proper address, so read desired word
  2. cache miss:
     - nothing in cache in appropriate block, so fetch from memory
  3. cache miss, block replacement:
     - wrong data is in cache at appropriate block, so discard it and fetch desired data from memory (cache always copy)

Peer instruction

- Consider an address split into fields for cache access as follows:

<table>
<thead>
<tr>
<th>Address (hex)</th>
<th>Value of Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>00000010</td>
<td>a</td>
</tr>
<tr>
<td>00000014</td>
<td>b</td>
</tr>
<tr>
<td>00000018</td>
<td>c</td>
</tr>
<tr>
<td>0000001C</td>
<td>d</td>
</tr>
<tr>
<td>00000016</td>
<td>e</td>
</tr>
<tr>
<td>00000034</td>
<td>f</td>
</tr>
<tr>
<td>00000038</td>
<td>g</td>
</tr>
<tr>
<td>0000003C</td>
<td>h</td>
</tr>
<tr>
<td>0000008010</td>
<td>i</td>
</tr>
<tr>
<td>0000008014</td>
<td>k</td>
</tr>
<tr>
<td>000000801C</td>
<td>l</td>
</tr>
</tbody>
</table>

Accessing data in a direct mapped cache

- Ex.: 16KB of data, direct-mapped, 4 word blocks
  - Can you work out height, width, area?
  - Read 4 addresses
    1. \(0x00000014\)
    2. \(0x0000001C\)
    3. \(0x00000034\)
    4. \(0x000008014\)
  - Memory vals here:
### Accessing data in a direct mapped cache

- **4 Addresses:**
  - 0x00000014, 0x0000001C, 0x00000034, 0x00008014

- **4 Addresses divided (for convenience) into Tag, Index, Byte Offset fields**:
  - 0x000000000000000000 0000000001 0100
  - 0x000000000000000000 0000000001 1100
  - 0x000000000000000000 0000000011 0100
  - 0x000000000000000010 0000000001 0100

### 16 KB Direct Mapped Cache, 16B blocks

- **valid bit:** determines whether anything is stored in that row (when computer initially turned on, all entries invalid)

<table>
<thead>
<tr>
<th>Valid</th>
<th>Index</th>
<th>Tag</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1022</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>1023</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
</tbody>
</table>

### First Type of Cache Miss

- **“Three Cs” Model of Misses**
  - **1st C: Compulsory Misses**
    - occur when a program is first started
    - cache does not contain any of that program’s data yet, so misses are bound to occur
    - can’t be avoided easily, so won’t focus on these in this course

### 1. Read 0x00000014

<table>
<thead>
<tr>
<th>Valid</th>
<th>Index</th>
<th>Tag</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1022</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>1023</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
</tbody>
</table>

### So we read block 1 (0000000001)

<table>
<thead>
<tr>
<th>Valid</th>
<th>Index</th>
<th>Tag</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1022</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>1023</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
</tbody>
</table>

### No valid data

<table>
<thead>
<tr>
<th>Valid</th>
<th>Index</th>
<th>Tag</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1022</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
<tr>
<td>1023</td>
<td>0</td>
<td>0x0</td>
<td>0x0-3</td>
</tr>
</tbody>
</table>
So load that data into cache, setting tag, valid

<table>
<thead>
<tr>
<th>Valid</th>
<th>Tag</th>
<th>Index field</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0xc-f</td>
<td>0x8-b</td>
<td>0x4-7</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td>0x0-3</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>1022</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1023</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Read from cache at offset, return word b

<table>
<thead>
<tr>
<th>Valid</th>
<th>Tag</th>
<th>Index field</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0xc-f</td>
<td>0x8-b</td>
<td>0x4-7</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td>0x0-3</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>1022</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1023</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2. Read 0x00000001C = 0...00 0..01 1100

Index valid, Tag Matches, return d

<table>
<thead>
<tr>
<th>Valid</th>
<th>Tag</th>
<th>Index field</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0xc-f</td>
<td>0x8-b</td>
<td>0x4-7</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
<td>0x0-3</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td>...</td>
</tr>
<tr>
<td>1022</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1023</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. Read 0x00000034 = 0...0 0..011 0100

<table>
<thead>
<tr>
<th>Valid</th>
<th>Tag field</th>
<th>Index field</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0xc-f</td>
<td>0x8-b  0x4-7 0x0-3</td>
</tr>
</tbody>
</table>

Index | Tag | d | c | b | a |
--- | --- | --- | --- | --- | --- |
0    | 0   | 0   | 0   | 0   | 0   |
1    | 0   | 0   | 0   | 0   | 0   |
2    | 0   | 0   | 0   | 0   | 0   |
3    | 0   | 0   | 0   | 0   | 0   |
4    | 0   | 0   | 0   | 0   | 0   |
5    | 0   | 0   | 0   | 0   | 0   |
6    | 0   | 0   | 0   | 0   | 0   |
7    | 0   | 0   | 0   | 0   | 0   |

So read block 3

<table>
<thead>
<tr>
<th>Valid</th>
<th>Tag field</th>
<th>Index field</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0xc-f</td>
<td>0x8-b  0x4-7 0x0-3</td>
</tr>
</tbody>
</table>

Index | Tag | d | c | b | a |
--- | --- | --- | --- | --- | --- |
0    | 0   | 0   | 0   | 0   | 0   |
1    | 0   | 0   | 0   | 0   | 0   |
2    | 0   | 0   | 0   | 0   | 0   |
3    | 0   | 0   | 0   | 0   | 0   |
4    | 0   | 0   | 0   | 0   | 0   |
5    | 0   | 0   | 0   | 0   | 0   |
6    | 0   | 0   | 0   | 0   | 0   |
7    | 0   | 0   | 0   | 0   | 0   |

No valid data

<table>
<thead>
<tr>
<th>Valid</th>
<th>Tag field</th>
<th>Index field</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0xc-f</td>
<td>0x8-b  0x4-7 0x0-3</td>
</tr>
</tbody>
</table>

Index | Tag | d | c | b | a |
--- | --- | --- | --- | --- | --- |
0    | 0   | 0   | 0   | 0   | 0   |
1    | 0   | 0   | 0   | 0   | 0   |
2    | 0   | 0   | 0   | 0   | 0   |
3    | 0   | 0   | 0   | 0   | 0   |
4    | 0   | 0   | 0   | 0   | 0   |
5    | 0   | 0   | 0   | 0   | 0   |
6    | 0   | 0   | 0   | 0   | 0   |
7    | 0   | 0   | 0   | 0   | 0   |

Load that cache block, return word f

<table>
<thead>
<tr>
<th>Valid</th>
<th>Tag field</th>
<th>Index field</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0xc-f</td>
<td>0x8-b  0x4-7 0x0-3</td>
</tr>
</tbody>
</table>

Index | Tag | d | c | b | a |
--- | --- | --- | --- | --- | --- |
0    | 0   | 0   | 0   | 0   | 0   |
1    | 0   | 0   | 0   | 0   | 0   |
2    | 0   | 0   | 0   | 0   | 0   |
3    | 0   | 0   | 0   | 0   | 0   |
4    | 0   | 0   | 0   | 0   | 0   |
5    | 0   | 0   | 0   | 0   | 0   |
6    | 0   | 0   | 0   | 0   | 0   |
7    | 0   | 0   | 0   | 0   | 0   |

4. Read 0x00008014 = 0...10 0..001 0100

<table>
<thead>
<tr>
<th>Valid</th>
<th>Tag field</th>
<th>Index field</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0xc-f</td>
<td>0x8-b  0x4-7 0x0-3</td>
</tr>
</tbody>
</table>

Index | Tag | d | c | b | a |
--- | --- | --- | --- | --- | --- |
0    | 0   | 0   | 0   | 0   | 0   |
1    | 0   | 0   | 0   | 0   | 0   |
2    | 0   | 0   | 0   | 0   | 0   |
3    | 0   | 0   | 0   | 0   | 0   |
4    | 0   | 0   | 0   | 0   | 0   |
5    | 0   | 0   | 0   | 0   | 0   |
6    | 0   | 0   | 0   | 0   | 0   |
7    | 0   | 0   | 0   | 0   | 0   |

So read Cache Block 1, Data is Valid

<table>
<thead>
<tr>
<th>Valid</th>
<th>Tag field</th>
<th>Index field</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0xc-f</td>
<td>0x8-b  0x4-7 0x0-3</td>
</tr>
</tbody>
</table>

Index | Tag | d | c | b | a |
--- | --- | --- | --- | --- | --- |
0    | 0   | 0   | 0   | 0   | 0   |
1    | 0   | 0   | 0   | 0   | 0   |
2    | 0   | 0   | 0   | 0   | 0   |
3    | 0   | 0   | 0   | 0   | 0   |
4    | 0   | 0   | 0   | 0   | 0   |
5    | 0   | 0   | 0   | 0   | 0   |
6    | 0   | 0   | 0   | 0   | 0   |
7    | 0   | 0   | 0   | 0   | 0   |
Cache Block 1 Tag does not match (0 != 2)

Miss, so replace block 1 with new data & tag

And return word J

Do an example yourself. What happens?

Answers

Administrivia
What to do on a write hit?

- **Write-through**
  - update the word in cache block and corresponding word in memory

- **Write-back**
  - update word in cache block
  - allow memory word to be “stale”

  ⇒ add ‘dirty’ bit to each block indicating that memory needs to be updated when block is replaced

  ⇒ OS flushes cache before I/O...

- Performance trade-offs?

---

Block Size Tradeoff (1/3)

- Benefits of Larger Block Size
  - **Spatial Locality**: if we access a given word, we’re likely to access other nearby words soon
  - Very applicable with Stored-Program Concept: if we execute a given instruction, it’s likely that we’ll execute the next few as well
  - Works nicely in sequential array accesses too

---

Block Size Tradeoff (2/3)

- Drawbacks of Larger Block Size
  - Larger block size means larger miss penalty
    - on a miss, takes longer time to load a new block from next level
  - If block size is too big relative to cache size, then there are too few blocks
    - Result: miss rate goes up

  - In general, minimize
    - **Average Memory Access Time (AMAT)**
      - Hit Time
      - Miss Penalty x Miss Rate

---

Block Size Tradeoff (3/3)

- **Hit Time** = time to find and retrieve data from current level cache
- **Miss Penalty** = average time to retrieve data on a current level miss (includes the possibility of misses on successive levels of memory hierarchy)
- **Hit Rate** = % of requests that are found in current level cache
- **Miss Rate** = 1 - Hit Rate

---

**Extreme Example: One Big Block**

- **Valid Bit**
- **Tag**
- **Cache Data**

  - Cache Size = 4 bytes  Block Size = 4 bytes
  - Only **ONE** entry (row) in the cache!

- If item accessed, likely accessed again soon
  - But unlikely will be accessed again immediately!

- The next access will likely to be a miss again
  - Continually loading data into the cache but discard data (force out) before use it again

- Nightmare for cache designer: **Ping Pong Effect**

---

Block Size Tradeoff Conclusions

- **Miss Penalty**
  - Exploits Spatial Locality
  - Fewer blocks: compromises temporal locality

- **Miss Rate**
  - Increased Miss Penalty & Miss Rate

- **Average Access Time**
  - Block Size
Second Type of Cache Miss

- **2nd C: Conflict Misses**
  - Miss that occurs because two distinct memory addresses map to the same cache location
  - Two blocks (which happen to map to the same location) can keep overwriting each other
  - Big problem in direct-mapped caches
  - How do we lessen the effect of these?

- **Dealing with Conflict Misses**
  - Solution 1: Make the cache size bigger
  - Fails at some point
  - Solution 2: Multiple distinct blocks can fit in the same cache Index?

Fully Associative Cache (1/3)

- Memory address fields:
  - Tag: same as before
  - Offset: same as before
  - Index: non-existent
- What does this mean?
  - No “rows”: any block can go anywhere in the cache
  - Must compare with all tags in entire cache to see if data is there

Fully Associative Cache (2/3)

- Fully Associative Cache (e.g., 32 B block)
  - Compare tags in parallel

<table>
<thead>
<tr>
<th>Cache Tag (27 bits long)</th>
<th>Byte Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>4 0</td>
</tr>
</tbody>
</table>

- Cache Tag | Valid | Cache Data
- B 31      | ... | B 1 B 0
- :         | 1   | :

Fully Associative Cache (3/3)

- Benefit of Fully Assoc Cache
  - No Conflict Misses (since data can go anywhere)
- Drawbacks of Fully Assoc Cache
  - Need hardware comparator for every single entry: If we have a 64KB of data in cache with 4B entries, we need 16K comparators: infeasible

Third Type of Cache Miss

- **Capacity Misses**
  - Miss that occurs because the cache has a limited size
  - Miss that would not occur if we increase the size of the cache
  - Sketchy definition, so just get the general idea
- This is the primary type of miss for Fully Associative caches.

N-Way Set Associative Cache (1/3)

- Memory address fields:
  - Tag: same as before
  - Offset: same as before
  - Index: points us to the correct “row” (called a set in this case)
- So what’s the difference?
  - Each set contains multiple blocks
  - Once we’ve found correct set, must compare with all tags in that set to find our data
**Associative Cache Example**

Here’s a simple 2 way set associative cache.

**N-Way Set Associative Cache (2/3)**

- **Basic Idea**
  - cache is direct-mapped w/respect to sets
  - each set is fully associative
  - basically N direct-mapped caches working in parallel: each has its own valid bit and data

- **Given memory address:**
  - Find correct set using Index value.
  - Compare Tag with all Tag values in the determined set.
  - If a match occurs, hit; otherwise a miss.
  - Finally, use the offset field as usual to find the desired data within the block.

**N-Way Set Associative Cache (3/3)**

- What’s so great about this?
  - even a 2-way set assoc cache avoids a lot of conflict misses
  - hardware cost isn’t that bad: only need N comparators

- In fact, for a cache with M blocks,
  - it’s Direct-Mapped if it’s 1-way set assoc
  - it’s Fully Assoc if it’s M-way set assoc
  - so these two are just special cases of the more general set associative design

**4-Way Set Associative Cache Circuit**

**Peer Instruction**

A. Mem hierarchies were invented before 1950. (UNIVAC I wasn’t delivered ‘til 1951)
B. If you know your computer’s cache size, you can often make your code run faster.
C. Memory hierarchies take advantage of spatial locality by keeping the most recent data items closer to the processor.

**Peer Instruction Answer**

A. “We are…forced to recognize the possibility of constructing a hierarchy of memories, each of which has greater capacity than the preceding but which is less accessible.” – von Neumann, 1946
B. Certainly! That’s call “tuning”
C. “Most Recent” items ⇒⇒⇒⇒ Temporal locality
And in Conclusion...

• We would like to have the capacity of disk at the speed of the processor: unfortunately this is not feasible.

• So we create a memory hierarchy:
  • each successively lower level contains “most used” data from next higher level
  • exploits temporal & spatial locality
  • do the common case fast, worry less about the exceptions (design principle of MIPS)

• Locality of reference is a Big Idea

And in Conclusion...

• Mechanism for transparent movement of data among levels of a storage hierarchy
  • set of address/value bindings
  • address → index to set of candidates
  • compare desired address with tag
  • service hit or miss
    • load new block and binding on miss

<table>
<thead>
<tr>
<th>address</th>
<th>tag</th>
<th>index</th>
<th>offset</th>
<th>Valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x0</td>
<td>0x0-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0x0-3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Index</th>
<th>Offset</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x0</td>
<td>0x0-3</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>