Caching

(get it??)
Topics

- Caching
- Cache policies and implementations
- Performance impact
- General strategies
A cache is a small, fast memory that acts as a buffer or staging area for a larger, slower memory.

- Fundamental CS system design concept
- Data is transferred in blocks or lines
- Slower caches use larger block sizes
- Cache hit vs. cache miss
- Hit ratio: \# hits / \# memory accesses
Caches

- A cache always begins cold ("empty")
  - Every request will be a cold miss initially
- As the cache loads data, it is warmed up
  - This effect can cause performance measurement variation during experiments if not controlled for
- A working set is a collection of elements needed repeatedly for a particular computation
  - If the working set doesn't fit in cache, this is called a capacity miss
Cache implementations

- What data structure can we use to implement caches?
  - Need **FAST** lookups and containment checks
  - From CS 240: use a hash table!
  - Cache slot = "real address" % CACHE_SIZE

What if we wanted our cache to store blocks longer than a single byte?

What if multiple "real" addresses map to the same cache slot? (this is called a **conflict**)

![Diagram of cache and memory mapping](image)
Question

- Suppose we have a sixteen-element cache, with slots labeled starting at zero. Which slot would we use to store a cached version of a data element stored at address 0x4d6?
  - Reminder: cache slot = "real address" % CACHE_SIZE
  - Hint: powers of two make things easier in hex!
Cache implementations

- A cache line is a block or sequence of bytes that is moved between memory levels in a single operation.
- A cache set is a collection of one or more cache lines.
  - Each cache line contains a tag to identify the source address and a valid flag/bit indicating whether the value is up-to-date.
- Cache parameters (S, E, B, m):
  - \( S = \# \text{ of cache sets} = 2^s \)
    - \( s = \# \text{ of bits for set index} \)
  - \( E = \# \text{ of lines per cache set} \)
    - Level of associativity
  - \( B = \text{block (cache line) size} = 2^b \)
    - \( b = \# \text{ of bits for block offset} \)
  - \( m = \# \text{ of bits for memory address} \)
    - \( M = \text{size of memory in bytes} = 2^m \)
  - \( C = \text{total cache capacity} = S \times E \times B \)
  - \( t = \# \text{ of tag bits} = m - s - b \)
Cache implementations

- **Direct-mapped** \((E = 1)\) caches

---

**Diagram:**

- **Selected set**
  - \(t\) bits
  - \(s\) bits
  - \(b\) bits

- **Set 0**
  - Valid
  - Tag
  - Cache block

- **Set 1**
  - Valid
  - Tag
  - Cache block

- **Set \(S-1\)**
  - Valid
  - Tag
  - Cache block

- **Legend:**
  - \(E = 1\) line per set

---

**Notes:**

1. The valid bit must be set
2. The tag bits in the cache line must match the tag bits in the address
3. If (1) and (2), then cache hit, and block offset selects starting byte.
Cache implementations

- **Set-associative** \((1 < E < C/B)\) caches

**Diagram:**

- Set 0:
  - Valid
  - Tag
  - Cache block

- Set 1:
  - Valid
  - Tag
  - Cache block

- Set \(S-1\):
  - Valid
  - Tag
  - Cache block

**Legend:**

- \(t\) bits
- \(s\) bits
- \(b\) bits
- Tag
- Set index
- Block offset

**Notes:**

- \(E=2\) lines per set
- "Two-way set associative"

**Instructions:**

1. The valid bit must be set
2. The tag bits in one of the cache lines must match the tag bits in the address
3. If (1) and (2), then cache hit, and block offset selects starting byte
Cache implementations

- **Fully-associative** \((E = C/B)\) caches

The entire cache is one set, so by default set 0 is always selected.

Set 0:

\[
\text{Valid} \quad \text{Tag} \quad \text{Cache block}
\]

\[
\text{Valid} \quad \text{Tag} \quad \text{Cache block}
\]

\[
\vdots
\]

\[
\text{Valid} \quad \text{Tag} \quad \text{Cache block}
\]

\[
E = C/B \text{ lines in the one and only set}
\]

1. The valid bit must be set.
2. The tag bits in one of the cache lines must match the tag bits in the address.
3. If (1) and (2), then cache hit, and block offset selects starting byte.
Cache implementations

- Why use the middle bits for the set index?
  - Contiguous memory blocks should map to different cache sets

![Diagram](image)
Cache misses ("Three C’s")

• **Compulsory / cold miss**
  - First cache miss due to an “empty” cache

• **Conflict miss**
  - Cache miss due to multiple lines in working set mapping to the same cache line
  - Repeated conflict misses for the same cache lines or blocks is called **thrashing**

• **Capacity miss**
  - Working set is too large to fit in cache
Cache architecture

- **Example:** Intel Core i7
- **Per-core:**
  - Registers
  - L1 d-cache and i-cache
    - Data and instructions
  - L2 unified cache
- **Shared:**
  - L3 unified cache
  - Main memory
Cache policies

- If a cache set is full, a cache miss in that set requires lines to be replaced or evicted.

- Policies:
  - Random replacement
  - Least recently used
  - Least frequently used

- These policies require additional overhead
  - More important for lower levels of the memory hierarchy
Cache policies

• How should we handle writes to a cached value?
  – **Write-through**: immediately update to lower level
    • Typically used for higher levels of memory hierarchy
  – **Write-back**: defer update until replacement/eviction
    • Typically used for lower levels of memory hierarchy

• How should we handle write misses?
  – **Write-allocate**: load then update
    • Typically used for write-back caches
  – **No-write-allocate**: update without loading
    • Typically used for write-through caches
Performance impact

• Metrics
  – **Hit rate/ratio**: # hits / # memory accesses (1 – miss rate)
    • **Hit time**: delay in accessing data for a cache hit
  – **Miss rate/ratio**: # misses / # memory accesses
    • **Miss penalty**: delay in loading data for a cache miss
  – **Read throughput** (or "bandwidth"): the rate that a program reads data from a memory system

• General observations:
  – Larger cache = higher hit rate but higher hit time
  – Lower miss rates = higher read throughput
Core theme

- **Cache system design involves tradeoffs**
  - Larger caches ⇒ higher hit rate but higher hit time
    - Size vs. speed
  - Larger blocks ⇒ higher hit rate for programs with good spatial locality, but lower hit rate for others
    - Favor spatial vs. temporal locality
  - Higher associativity ⇒ lower chance of thrashing but expensive to implement w/ possibly increased hit time
    - Hit time vs. miss penalty
  - More writes ⇒ simpler to implement but lower performance
    - Write-through vs. write-back
Temporal locality

- Working set size vs. throughput

![Graph showing the relationship between working set size and throughput for different cache regions. The x-axis represents working set size in bytes, ranging from 128m to 16k, and the y-axis represents read throughput in MB/s, ranging from 0 to 14000. The graph indicates that higher is better.](image)
Spatial locality

- Stride vs. throughput

![Bar chart showing read throughput for different strides. Higher is better.](image)
Memory mountain (CS:APP)

- Stride and WSS vs. read throughput

[Diagram showing a 3D graph with axes labeled 'Stride (x8 bytes)', 'Size (bytes)', and 'Read throughput (MB/s)'. The graph illustrates varying read throughput with different combinations of stride and data size, highlighting the impact of cache levels (L1, L2, L3) on performance. Key features include:
  - L1, L2, L3 labels on the peaks.
  - 'Core i7 Haswell, 2.1 GHz, 32 KB L1 d-cache, 256 KB L2 cache, 8 MB L3 cache, 64 B block size.'
  - Arrows indicating slopes of spatial locality and ridges of temporal locality.
  - Higher read throughput is better.
Output of `lscpu`:

Architecture: x86_64
Byte Order: Little Endian
CPU(s): 24
Thread(s) per core: 2
Core(s) per socket: 6
Socket(s): 2
Vendor ID: Intel
Model name: Intel(R) Xeon(R) CPU E5-2640
CPU max MHz: 3000.0000
CPU min MHz: 1200.0000
L1d cache: 32K
L1i cache: 32K
L2 cache: 256K
L3 cache: 15360K
Output of `lscpu`:

Architecture: x86_64
Byte Order: Little Endian
CPU(s): 48
Thread(s) per core: 2
Core(s) per socket: 12
Socket(s): 2
Vendor ID: Intel
Model name: Intel(R) Xeon(R) CPU E5-2680
CPU max MHz: 3300.000
CPU min MHz: 1200.000
L1d cache: 32K
L1i cache: 32K
L2 cache: 256K
L3 cache: 30720K
Case study: matrix multiply

Figure 6.44  Six versions of matrix multiply. Each version is uniquely identified by the ordering of its loops.
Case study: matrix multiply

Lower is better
Optimization strategies

- Focus on the common cases
- Focus on the code regions that dominate runtime
- Focus on inner loops and minimize cache misses
- Favor repeated local accesses (temporal locality)
- Favor stride-1 access patterns (spatial locality)

**Tip:** You can use Valgrind to detect cache misses (look up a tool called `cachegrind`)
Next time

- **Virtual memory**: an OS-level memory cache
  - Bridge between module 4 (machine architectures) and module 5 (operating systems)
  - Module 4 unit test over the weekend (due Monday)