Caches III
CSE 351 Spring 2017

Instructor:
Ruth Anderson

Teaching Assistants:
Dylan Johnson
Kevin Bi
Linxing Preston Jiang
Cody Ohlsen
Yufang Sun
Joshua Curtis
Administrivia

- Office Hours Changes – check calendar!!
- Homework 3, due TONIGHT (5/5)
- Midterm, Monday (5/8)
- Lab 3, due Thursday (5/11)
- Mid-Quarter Feedback Survey!
Question

- We have a cache of size 2 KiB with block size of 128 B. If our cache has 2 sets, what is its associativity?
  
  A. 2  
  B. 4  
  C. 8  
  D. 16

- If addresses are 16 bits wide, how wide is the Tag field?
Cache Read

1) Locate set
2) Check if any line in set is valid and has matching tag: hit
3) Locate data starting at offset

Address of byte in memory:

\[ E = \text{blocks/lines per set} \]

\[ S = \# \text{sets} = 2^s \]

\[ v \text{ tag} \]

\[ 0 \ 1 \ 2 \ \ldots \ \ldots \ K-1 \]

\[ \text{data begins at this offset} \]

\[ \text{valid bit} \]

\[ K = \text{bytes per block} \]
Example: Direct-Mapped Cache \((E = 1)\)

Direct-mapped: One line per set
Block Size \(K = 8\) B

\[ S = 2^s \text{ sets} \]

Address of int:

\[
\begin{array}{c}
\text{bits} \\
0...1 \\
100
\end{array}
\]

find set
Example: Direct-Mapped Cache ($E = 1$)

Direct-mapped: One line per set
Block Size $K = 8$ B

![Diagram of Direct-Mapped Cache]

- Address of `int`:
  - `t` bits: 0...01 100
- Block offset
- Valid? + Match?: yes = hit

Diagram showing the cache structure with tags and block offset.
Example: Direct-Mapped Cache ($E = 1$)

Direct-mapped: One line per set
Block Size $K = 8$ B

Address of `int`:

- `t` bits: 0...1 100

No match? Then old line gets evicted and replaced

This is why we want alignment!
Example: Set-Associative Cache \((E = 2)\)

2-way: Two lines per set
Block Size \(K = 8\) B

Address of `short int`:

```
τ bits  0...01  100
```

find set
Example: Set-Associative Cache ($E = 2$)

2-way: Two lines per set
Block Size $K = 8$ B

Valid? +
Match: yes = hit

Compare both

Address of `short int`:

Block offset
Example: Set-Associative Cache \((E = 2)\)

2-way: Two lines per set

Block Size \(K = 8\) B

Address of \text{short int}:

\begin{center}
\begin{tabular}{c}
\text{bits} \\
0...01 \quad 100
\end{tabular}
\end{center}

\text{valid?} + \text{match: yes} = \text{hit}

\begin{center}
\begin{tabular}{cccccccc}
\text{v} & \text{tag} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7
\end{tabular}
\end{center}

\begin{center}
\begin{tabular}{cccccccc}
\text{v} & \text{tag} & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7
\end{tabular}
\end{center}

\begin{center}
\begin{tabular}{c}
\text{block offset}
\end{tabular}
\end{center}

\text{short int} (2 B) is here

No match?

- One line in set is selected for eviction and replacement
- Replacement policies: random, least recently used (LRU), ...

}\[
\begin{array}{c}
\text{compare both}
\end{array}\]
Types of Cache Misses: 3 C’s!

- **Compulsory** (cold) miss
  - Occurs on first access to a block

- **Conflict** miss
  - Conflict misses occur when the cache is large enough, but multiple data objects all map to the same slot
    - *e.g.* referencing blocks 0, 8, 0, 8, ... could miss every time
  - Direct-mapped caches have more conflict misses than *E*-way set-associative (where *E* > 1)

- **Capacity** miss
  - Occurs when the set of active cache blocks (the *working set*) is larger than the cache (just won’t fit, even if cache was *fully-associative*)
  - **Note:** *Fully-associative* only has Compulsory and Capacity misses
Core i7: Associativity

Processor package

Core 0
- Regs
- L1 d-cache
- L1 i-cache
- L2 unified cache
- L3 unified cache (shared by all cores)

Core 3
- Regs
- L1 d-cache
- L1 i-cache
- L2 unified cache

Block/line size:
- 64 bytes for all

L1 i-cache and d-cache:
- 32 KiB, 8-way, Access: 4 cycles

L2 unified cache:
- 256 KiB, 8-way, Access: 11 cycles

L3 unified cache:
- 8 MiB, 16-way, Access: 30-40 cycles

slower, but more likely to hit
Making memory accesses fast!

- Cache basics
- Principle of locality
- Memory hierarchies
- Cache organization
  - Direct-mapped (sets; index + tag)
  - Associativity (ways)
  - Replacement policy
  - Handling writes
- Program optimizations that consider caches
What about writes?

- Multiple copies of data exist:
  - L1, L2, possibly L3, main memory

- What to do on a write-hit?
  - **Write-through**: write immediately to next level
  - **Write-back**: defer write to next level until line is evicted (replaced)
    - Must track which cache lines have been modified ("dirty bit")

- What to do on a write-miss?
  - **Write-allocate**: ("fetch on write") load into cache, update line in cache
    - Good if more writes or reads to the location follow
  - **No-write-allocate**: ("write around") just write immediately to memory

- Typical caches:
  - Write-back + Write-allocate, usually
  - Write-through + No-write-allocate, occasionally
Write-back, write-allocate example

Contents of memory stored at address G

tag (there is only one set in this tiny cache, so the tag is the entire block address!)

In this example we are sort of ignoring block offsets. Here a block holds 2 bytes (16 bits, 4 hex digits).

Normally a block would be much bigger and thus there would be multiple items per block. While only one item in that block would be written at a time, the entire line would be brought into cache.
Write-back, write-allocate example

```
mov 0xFACE, F
```

![Diagram showing cache and memory states after write-back and write-allocate]

- **Cache**: 
  - G: 0xBEEF
  - Dirty bit

- **Memory**: 
  - F: 0xCAFE
  - G: 0xBEEF
Write-back, write-allocate example

\[ \text{mov } 0\timesFACE, F \]

Step 1: Bring F into cache
Write-back, write-allocate example

mov 0xFACE, F

Step 2: Write 0xFACE to cache only and set dirty bit
Write-back, write-allocate example

mov 0xFACE, F  mov 0xFEED, F

Write hit!
Write 0xFEED to cache only

dirty bit
Write-back, write-allocate example

```
mov 0xFACE, F
mov 0xFEED, F
mov G, %rax
```
Write-back, write-allocate example

1. Write F back to memory since it is dirty
2. Bring G into the cache so we can copy it into %rax

mov 0xFACE, F  
mov 0xFEED, F  
mov G, %rax
Question

Which of the following cache statements is FALSE?

A. We can reduce compulsory misses by decreasing our block size
B. We can reduce conflict misses by increasing associativity
C. A write-back cache will save time for code with good temporal locality on writes
D. A write-through cache will always match data with the memory hierarchy level below it
Optimizations for the Memory Hierarchy

- Write code that has locality!
  - **Spatial**: access data contiguously
  - **Temporal**: make sure access to the same data is not too far apart in time

- How can you achieve locality?
  - Adjust memory accesses in *code* (software) to improve miss rate (MR)
    - Requires knowledge of *both* how caches work as well as your system’s parameters
  - Proper choice of algorithm
  - Loop transformations
Example: Matrix Multiplication

\[ c_{ij} = \sum_{k=1}^{n} a_{ik} \cdot b_{kj} \]
Matrices in Memory

- How do cache blocks fit into this scheme?
  - Row major matrix in memory:
    - COLUMN of matrix (blue) is spread among cache blocks shown in red
Naïve Matrix Multiply

# move along rows of A
for (i = 0; i < n; i++)
    # move along columns of B
    for (j = 0; j < n; j++)
        # EACH k loop reads row of A, col of B
        # Also read & write c(i,j) n times
        for (k = 0; k < n; k++)
            c[i*n+j] += a[i*n+k] * b[k*n+j];

\[
\begin{array}{c}
C(i,j) \\
\end{array} = \frac{\begin{array}{c}
C(i,j) \\
\end{array}}{\begin{array}{c}
A(i,:) \\
\end{array}} + \frac{\begin{array}{c}
\end{array}}{\begin{array}{c}
B(:,j) \\
\end{array}} \times \frac{\begin{array}{c}
\end{array}}{\begin{array}{c}
\end{array}}
\]
Cache Miss Analysis (Naïve)

- **Scenario Parameters:**
  - Square matrix \((n \times n)\), elements are doubles
  - Cache block size \(K = 64\) B = 8 doubles
  - Cache size \(C \ll n\) (much smaller than \(n\))

- **Each iteration:**
  \[ \frac{n}{8} + n = \frac{9n}{8} \text{ misses} \]

Ignoring matrix \(C\)
Cache Miss Analysis (Naïve)

- **Scenario Parameters:**
  - Square matrix \((n \times n)\), elements are doubles
  - Cache block size \(K = 64\) B = 8 doubles
  - Cache size \(C \ll n\) (much smaller than \(n\))

- **Each iteration:**
  - \(\frac{n}{8} + n = \frac{9n}{8}\) misses

- Afterwards **in cache:**
  - (schematic)
Linear Algebra to the Rescue (1)

- Can get the same result of a matrix multiplication by splitting the matrices into smaller submatrices (matrix “blocks”)

- For example, multiply two 4×4 matrices:

\[
A = \begin{bmatrix}
a_{11} & a_{12} & a_{13} & a_{14} \\
a_{21} & a_{22} & a_{23} & a_{24} \\
a_{31} & a_{32} & a_{33} & a_{34} \\
a_{41} & a_{42} & a_{43} & a_{44}
\end{bmatrix} = \begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix}, \text{ with } B \text{ defined similarly.}
\]

\[
AB = \begin{bmatrix}
(A_{11}B_{11} + A_{12}B_{21}) & (A_{11}B_{12} + A_{12}B_{22}) \\
(A_{21}B_{11} + A_{22}B_{21}) & (A_{21}B_{12} + A_{22}B_{22})
\end{bmatrix}
\]
Linear Algebra to the Rescue (2)

Matrices of size \( n \times n \), split into 4 blocks of size \( r \) (\( n=4r \))

\[
C_{22} = A_{21}B_{12} + A_{22}B_{22} + A_{23}B_{32} + A_{24}B_{42} = \sum_k A_{2k}B_{k2}
\]

- Multiplication operates on small “block” matrices
  - Choose size so that they fit in the cache!
  - This technique called “cache blocking”
Blocked Matrix Multiply

- Blocked version of the naïve algorithm:

```c
# move by rxr BLOCKS now
for (i = 0; i < n; i += r)
    for (j = 0; j < n; j += r)
        for (k = 0; k < n; k += r)
            # block matrix multiplication
            for (ib = i; ib < i+r; ib++)
                for (jb = j; jb < j+r; jb++)
                    for (kb = k; kb < k+r; kb++)
                        c[ib*n+jb] += a[ib*n+kb]*b[kb*n+jb];
```

- \( r \) = block matrix size (assume \( r \) divides \( n \) evenly)
Cache Miss Analysis (Blocked)

- **Scenario Parameters:**
  - Cache block size $K = 64 \text{ B} = 8$ doubles
  - Cache size $C \ll n$ (much smaller than $n$)
  - Three blocks $(r \times r)$ fit into cache: $3r^2 < C$

- **Each block iteration:**
  - $r^2/8$ misses per block
  - $2n/r \times r^2/8 = nr/4$

Ignoring matrix $C$
Cache Miss Analysis (Blocked)

- **Scenario Parameters:**
  - Cache block size $K = 64$ B = 8 doubles
  - Cache size $C \ll n$ (much smaller than $n$)
  - Three blocks $(r \times r)$ fit into cache: $3r^2 < C$

- **Each block iteration:**
  - $r^2/8$ misses per block
  - $2n/r \times r^2/8 = nr/4$

- Afterwards in cache (schematic)
Matrix Multiply Visualization

- Here $n = 100$, $C = 32$ KiB, $r = 30$

Naïve:

- $\approx 1,020,000$ cache misses

Blocked:

- $\approx 90,000$ cache misses
Cache-Friendly Code

- Programmer can optimize for cache performance
  - How data structures are organized
  - How data are accessed
    - Nested loop structure
    - Blocking is a general technique

- All systems favor “cache-friendly code”
  - Getting absolute optimum performance is very platform specific
    - Cache size, cache block size, associativity, etc.
  - Can get most of the advantage with generic code
    - Keep working set reasonably small (temporal locality)
    - Use small strides (spatial locality)
    - Focus on inner loop code
The Memory Mountain

Core i7 Haswell
2.1 GHz
32 KB L1 d-cache
256 KB L2 cache
8 MB L3 cache
64 B block size

Aggressive prefetching

Slopes of spatial locality

Ridges of temporal locality

Read throughput (MB/s)

Stride (x8 bytes)

Size (bytes)
Learning About Your Machine

Linux:
- `lscpu`
- `ls /sys/devices/system/cpu/cpu0/cache/index0/`
  - Ex: `cat /sys/devices/system/cpu/cpu0/cache/index*/size`
- `cat /proc/cpuinfo | grep cache | sort | uniq`

Windows:
- `wmic memcache get <query>`  (all values in KB)
  - Ex: `wmic memcache get MaxCacheSize`

Modern processor specs:  [http://www.7-cpu.com/](http://www.7-cpu.com/)