Announcements

- Previous lecture
  - Caches

- This lecture
  - Cache Performance.

- Tutorials happening this week & next
Recap: Memory Hierarchy Issues

- **Block size**: smallest unit that is managed at each level
  - E.g., 64B for cache lines, 4KB for memory pages

- **Block placement**: Where can a block be placed?
  - E.g., direct mapped, set associative, fully associative

- **Block identification**: How can a block be found?
  - E.g., hardware tag matching, OS page table

- **Block replacement**: Which block should be replaced?
  - E.g., Random, Least recently used (LRU), Not recently used (NRU)

- **Write strategy**: What happens on a write?
  - E.g., write-through, write-back, write-allocate

- **Inclusivity**: whether next lower level contains all the data found in the current level
  - Inclusive, exclusive
Cache Performance

Memory system and processor performance:

\[
\text{CPU time} = \text{IC} \times \text{CPI} \times \text{Clock time} \quad \longrightarrow \quad \text{CPU performance eqn.}
\]

\[
\text{CPI} = \text{CPI}_{ld/st} \times \frac{\text{IC}_{ld/st}}{\text{IC}} + \text{CPI}_{\text{others}} \times \frac{\text{IC}_{\text{others}}}{\text{IC}}
\]

\[
\text{CPI}_{ld/st} = \text{Average memory access time (AMAT)}
\]

\[
\text{AMAT} = \text{Hit time} + \text{Miss rate} \times \text{Miss penalty} \quad \longrightarrow \quad \text{Memory performance eqn.}
\]

Improving memory hierarchy performance:

- Decrease hit time
- Decrease miss rate
- Decrease miss penalty
Cache Performance – example problem

Assume we have a computer where the CPI is 1 when all memory accesses hit in the cache. Data accesses (ld/st) represent 50% of all instructions. If the miss penalty is 25 clocks and the miss rate is 2%, how much faster would the computer be if all instructions were cache hits?

[H&P 5th ed, B.1]
**Answer**  First compute the performance for the computer that always hits:

\[
\text{CPU execution time} = (\text{CPU clock cycles} + \text{Memory stall cycles}) \times \text{Clock cycle}
\]
\[
= (\text{IC} \times \text{CPI} + 0) \times \text{Clock cycle}
\]
\[
= \text{IC} \times 1.0 \times \text{Clock cycle}
\]

Now for the computer with the real cache, first we compute memory stall cycles:

\[
\text{Memory stall cycles} = \text{IC} \times \frac{\text{Memory accesses}}{\text{Instruction}} \times \text{Miss rate} \times \text{Miss penalty}
\]
\[
= \text{IC} \times (1 + 0.5) \times 0.02 \times 25
\]
\[
= \text{IC} \times 0.75
\]

where the middle term \((1 + 0.5)\) represents one instruction access and 0.5 data accesses per instruction. The total performance is thus

\[
\text{CPU execution time}_{\text{cache}} = (\text{IC} \times 1.0 + \text{IC} \times 0.75) \times \text{Clock cycle}
\]
\[
= 1.75 \times \text{IC} \times \text{Clock cycle}
\]

The performance ratio is the inverse of the execution times:

\[
\frac{\text{CPU execution time}_{\text{cache}}}{\text{CPU execution time}} = \frac{1.75 \times \text{IC} \times \text{Clock cycle}}{1.0 \times \text{IC} \times \text{Clock cycle}}
\]
\[
= 1.75
\]

The computer with no cache misses is 1.75 times faster.
Reducing Cache Miss Rates

Cache miss classification: the “three C’s”

- Compulsory misses (or cold misses): when a block is accessed for the first time
- Capacity misses: when a block is not in the cache because it was evicted because the cache was full
- Conflict misses: when a block is not in the cache because it was evicted because the cache set was full
  - Conflict misses only exist in direct-mapped or set-associative caches
  - In a fully associative cache, all non-compulsory misses are capacity misses
Cache Misses vs. Cache Size

- Miss rates are very small in practice (caching is effective!)
- Miss rates decrease significantly with cache size
  - Rule of thumb: miss rates change in proportion to $\sqrt{\text{cache size}}$
    e.g., 2x cache $\rightarrow \sqrt{2}$ fewer misses
- Miss rates decrease with set-associativity because of reduction in conflict misses
Reducing Cold Miss Rates

Technique 1: Large block size
- Principle of spatial locality $\rightarrow$ other data in the block likely to be used soon
- Reduce cold miss rate
- May increase conflict and capacity miss rate for the same cache size (fewer blocks in cache)
- Increase miss penalty because more data has to be brought in each time
- Uses more memory bandwidth
Cache Misses vs. Block Size

- Small caches are very sensitive to block size
- Very large blocks (> 128B) never beneficial
- 64B is a sweet spot → common choice in today’s processors

H&P
Fig. 5.16
Reducing Cold Miss Rates

Technique 2: Prefetching

- Idea: bring into the cache ahead of time data or instructions that are likely to be used soon
- Can reduce cold misses (also capacity misses)
- Uses more memory bandwidth
- Does not typically increase miss penalty (prefetch is generally handled after main cache access is completed)
- May increase conflict and capacity miss rates by displacing useful blocks (cache pollution)
  - Can use a prefetch buffer to avoid polluting the cache
Prefetching

- Hardware prefetching: hardware automatically prefetches cache blocks on a cache miss
  - No need for extra prefetching instructions in the program
  - Effective for regular accesses, such as instructions
  - E.g., next blocks prefetching, stride prefetching

- Software prefetching: compiler inserts instructions at proper places in the code to trigger prefetches
  - Requires ISA support (nonbinding prefetch instruction)
  - Adds instructions to compute the prefetching addresses and to perform the prefetch itself (prefetch overhead)
  - E.g., data prefetching in loops, linked list prefetching
Software Prefetching

- E.g., prefetching in loops: Brings the next required block, two iterations ahead of time (assuming each element of x is 4-bytes long and the block has 64 bytes).

```c
for (i=0; i<=999; i++) {
    if (i%16 == 0)
        prefetch(x[i+32]);
    x[i] = x[i] + s;
}
```

- E.g., linked-list prefetching: Brings the next object in the list

```c
while (student) {
    student->mark = rand();
    prefetch(student->next);
    student = student->next;
}
```
Reducing Conflict Miss Rates

Technique 3: High associativity caches
- More options for block placement → fewer conflicts
- Reduce conflict miss rate
- May increase hit access time because tag match takes longer
- May increase miss penalty because replacement policy is more involved
**Cache Misses vs. Associativity**

- Small caches are very sensitive to associativity.
- In all cases more associativity decreases miss rate, but little difference between 4-way and fully associative.
Reducing Miss Rates

Technique 4: Compiler optimizations

- E.g., merging arrays: may improve spatial locality if the fields are used together for the same index

```c
int val[size];
int key[size];
struct valkey
{
    int val;
    int key;
};
Struct valkey merged_array[size];
```

- E.g., loop fusion: improves temporal locality

```c
for (i=0; i<1000; i++)
for (i=0; i<1000; i++)
    B[i] = B[i]+A[i];
```

```c
for (i=0; i<1000; i++)
    B[i] = B[i]+A[i];
```
Reducing Miss Rates

- E.g., blocking: change row-major and column-major array distributions to block distribution to improve spatial and temporal locality

```c
for (i=0; i<5; i++)
    for (j=0; j<5; j++) {
        r=0;
        for (k=0; k<5; k++) {
            r=r+y[i][k]*z[k][j];
            x[i][j]=r;
        }
    }
```

// matrix multiplication

```
i=0; j=0; 0<k<5
i=0; j=1; 0<k<5 ...
i=1; j=0; 0<k<5
```

Poor spatial and temporal locality

Poor temporal locality
Reducing Conflict Miss Rates – Loop Blocking or Tiling

\[
\text{for (jj = 0; jj < 5; jj = jj+2)} \\
\quad \text{for (kk = 0; kk < 5; kk = kk+2)} \\
\quad \text{for (i = 0; i < 5; i++)} \\
\quad \quad \text{for (j = jj; j < min(jj+2-1,5); j++)} \\
\quad \quad \quad \{ r = 0; \} \\
\quad \quad \quad \quad \text{for (k = kk; k < min(kk+2-1,5); k++)} \\
\quad \quad \quad \quad \quad r = r + y[i][k]*z[k][j]; \\
\quad \quad \quad \quad x[i][j]= x[i][j] + r; \\
\}
\]

Better temporal locality
Cache Performance II

- Memory system and processor performance:
  
  \[ \text{CPU time} = \text{IC} \times \text{CPI} \times \text{Clock time} \longrightarrow \text{CPU performance eqn.} \]

  \[ \text{Avg. mem. time} = \text{Hit time} + \text{Miss rate} \times \text{Miss penalty} \quad \longrightarrow \text{Memory performance eqn.} \]

- Improving memory hierarchy performance:
  - Decrease hit time
  - Decrease miss rate (block size, prefetching, associativity, compiler)
  - Decrease miss penalty
Reducing Cache Miss Penalty

Technique 1: Victim caches
- (Can also be considered to reduce miss rate)
- Very small cache used to capture evicted lines from cache
  - Targets conflict misses
- In case of cache miss the data may be found quickly in the victim cache
- Typically 8-32 entries, fully-associative
- Access victim cache in series or in parallel with main cache: trade-off?
Reducing Cache Miss Penalty

Technique 2: giving priority to reads over writes

- The value of a load is likely to be used soon (i.e., dependent instruction may stall), while writes are “fire-and-forget”
  - Insight: writes are “off the critical path” and their latency doesn’t usually matter. Thus, don’t stall for writes!

- Idea: place write misses in a write buffer, and let read misses overtake writes
  - Flush the writes from the write buffer when pipeline is idle or when buffer full

- Reads to the memory address of a pending write in the buffer now become hits in the buffer:

```
sw 512(r0), r3
...
lw r2, 512(0)
```

1. write miss goes into write buffer

<table>
<thead>
<tr>
<th>memory address</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>512</td>
<td>R[r3]</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

2. read hits in the write buffer and gets the value from the previous write
Reducing Cache Miss Penalty

Technique 3: early restart and critical word first

- On a read miss, processor needs just the requested word (or byte)
  - but processor must wait until the whole block is brought into the cache
- Early restart: as soon as the requested word arrives in the cache, send it to the processor
  - Meanwhile, continue reading the rest of the block into the cache

```assembly
lw r2, 3(0) 0x0003
```

Diagram:
- CPU
- L1 cache
- Main memory
- Cache block
- Tag
- Data
- 03
- 04 ... 05 ... 06 ...
Reducing Cache Miss Penalty

Technique 3: early restart and critical word first

- On a read miss, processor needs just the requested word (or byte)
  - but processor must wait until the whole block is brought into the cache
- Critical word first: get the requested word first from the memory and immediately send it to the processor
  - Meanwhile, continue reading the rest of the block into the cache

```assembly
lw r2, 3(0) 0x0003
```
Reducing Cache Miss Penalty

Technique 4: non-blocking (or lockup-free) caches

- Non-blocking caches: other memory instructions can overtake a cache miss instruction
  - Cache can service multiple hits while waiting on a miss: “hit under miss”
  - More aggressive: cache can service multiple hits while waiting on multiple misses: “miss under miss” or “hit under multiple misses”

- Cache and memory must be able to service multiple requests concurrently
  - Particularly valuable in dynamically scheduled (out-of-order) processors

- Must keep track of multiple outstanding memory operations
  - New hardware structure: Miss Status Handler Registers (MSHRs)
    - Address of a block being waited on
    - Capability to merge multiple requests to the same block
    - Destination register
Non-blocking Caches

- Significant improvement from small degree of outstanding memory operations
- Some applications benefit from large degrees
Reducing Cache Miss Penalty

Technique 5: second level caches (L2)
- Gap between main memory and L1 cache speeds is increasing
  - L2 makes main memory appear to be faster if it captures most of the L1 cache misses
    - L1 miss penalty becomes L2 hit access time if hit in L2
    - L1 miss penalty higher if miss in L2

- L2 considerations:
  - 256KB – 4MB capacity (last level of cache in smartphones & tablets)
  - ~10-20 cycles access time
  - Higher associativity (e.g., 8-16 ways) possible. Why?
  - Higher miss rate than L1. Why?

- L3 caches are common on laptop/desktop/server processors
  - 30+ cycle access time
  - 2-20+ MB capacity
  - Very high associativity (16-32 ways)
Second Level Caches

- Memory subsystem performance:

  \[
  \text{Avg. mem. time} = \text{Hit time}_{L1} + \text{Miss rate}_{L1} \times \text{Miss penalty}_{L1}
  \]

  \[
  \text{Miss penalty}_{L1} = \text{Hit time}_{L2} + \text{Miss rate}_{L2} \times \text{Miss penalty}_{L2}
  \]

  \[
  \therefore \text{Avg. mem. time} = \text{Hit time}_{L1} + \text{Miss rate}_{L1} \times \left(\text{Hit time}_{L2} + \text{Miss rate}_{L2} \times \text{Miss penalty}_{L2}\right)
  \]

- Miss rates:
  - Local: the number of misses divided by the number of requests to the cache
    - E.g., Miss rate\textsubscript{L1} and Miss rate\textsubscript{L2} in the equations above
    - Usually not so small for lower level caches
  - Global: the number of misses divided by the total number of requests from the CPU
    - E.g, L2 global miss rate = Miss rate\textsubscript{L1} \times Miss rate\textsubscript{L2}
    - Represents the aggregate effectiveness of the cache hierarchy
Cache Misses vs. L2 size

- L2 caches must be much bigger than L1
- Local miss rates for L2 are larger than for L1 and are not a good measure of overall performance

H&P Fig. 5.10
Cache Performance II

- Memory system and processor performance:

  \[ \text{CPU time} = IC \times CPI \times \text{Clock time} \rightarrow \text{CPU performance eqn.} \]

  \[ \text{Avg. mem. time} = \text{Hit time} + \text{Miss rate} \times \text{Miss penalty} \rightarrow \text{Memory performance eqn.} \]

- Improving memory hierarchy performance:
  - Decrease hit time
  - Decrease miss rate (block size, prefetching, associativity, compiler)
  - Decrease miss penalty (victim caches, reads over writes, prioritize critical word, non-blocking caches, additional cache levels)
Reducing Cache Hit Time

Technique 1: small and simple caches

- Small caches are compact $\Rightarrow$ have short wire spans
  - Wires are slow

- Low associativity caches have few tags to compare against the requested data

- Direct mapped caches have only one tag to compare and comparison can be done in parallel with the fetch of the data
Reducing Cache Hit Time

Technique 2: **virtual-addressed caches**
- Programs use virtual addresses for data, while main memory uses physical addresses → addresses from processor must be translated at some point
- Idea: perform translation after the cache lookup

We will examine address translation in detail in the virtual memory lecture

![Diagram showing address translation]
# Cache Performance Techniques

<table>
<thead>
<tr>
<th>technique</th>
<th>miss rate</th>
<th>miss penalty</th>
<th>hit time</th>
<th>complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>large block size</td>
<td>🎁</td>
<td>🙁</td>
<td>🎁</td>
<td>🎁</td>
</tr>
<tr>
<td>high associativity</td>
<td>🎁</td>
<td></td>
<td>🙁</td>
<td>🙁</td>
</tr>
<tr>
<td>victim cache</td>
<td>🎁</td>
<td>🎁</td>
<td></td>
<td>🙁</td>
</tr>
<tr>
<td>hardware prefetch</td>
<td>🎁</td>
<td></td>
<td></td>
<td>🙁</td>
</tr>
<tr>
<td>compiler prefetch</td>
<td>🎁</td>
<td></td>
<td></td>
<td>🙁</td>
</tr>
<tr>
<td>compiler optimizations</td>
<td>🎁</td>
<td></td>
<td></td>
<td>🙁</td>
</tr>
<tr>
<td>prioritisation of reads</td>
<td>🎁</td>
<td></td>
<td></td>
<td>🙁</td>
</tr>
<tr>
<td>critical word first</td>
<td>🎁</td>
<td></td>
<td></td>
<td>🙁</td>
</tr>
<tr>
<td>nonblocking caches</td>
<td>🎁</td>
<td></td>
<td></td>
<td>🙁</td>
</tr>
<tr>
<td>L2 caches</td>
<td>🎁</td>
<td></td>
<td></td>
<td>🙁</td>
</tr>
<tr>
<td>small and simple caches</td>
<td>🙁</td>
<td>🎁</td>
<td>🎁</td>
<td>🎁</td>
</tr>
<tr>
<td>virtual-addressed caches</td>
<td>🎁</td>
<td></td>
<td></td>
<td>🙁</td>
</tr>
</tbody>
</table>