Assignment and Exam Reminders

- **Lab 4**: Due March 6 (this Friday!)
  - Control flow and branch prediction

- **Lab 5**: Due March 22
  - Data cache

- **HW 4**: March 18
- **Exam**: March 20

- **Advice**: Finish the labs early
  - You have almost a month for Lab 5

- **Advice**: Manage your time well
Lab 3 Grade Distribution

- Average Grade: 63.17
- Median Grade: 69
- Standard Deviation: 37.19
- Maximum Grade: 100
- Minimum Grade: 34
Lab 3 Extra Credits

- Stay tuned!
Agenda for the Rest of 447

- The memory hierarchy
- Caches, caches, more caches
- Virtualizing the memory hierarchy
- Main memory: DRAM
- Main memory control, scheduling
- Memory latency tolerance techniques
- Non-volatile memory

- Multiprocessors
- Coherence and consistency
- Interconnection networks
- Multi-core issues
Readings for Today and Next Lecture

- Memory Hierarchy and Caches

Required
- Cache chapters from P&H: 5.1-5.3
- Memory/cache chapters from Hamacher+: 8.1-8.7

Required + Review:
How to Improve Cache Performance

- Three fundamental goals

- Reducing miss rate
  - Caveat: reducing miss rate can reduce performance if more costly-to-refetch blocks are evicted

- Reducing miss latency or miss cost

- Reducing hit latency or hit cost
Improving Basic Cache Performance

- Reducing miss rate
  - More associativity
  - Alternatives/enhancements to associativity
    - Victim caches, hashing, pseudo-associativity, skewed associativity
  - Better replacement/insertion policies
  - Software approaches

- Reducing miss latency/cost
  - Multi-level caches
  - Critical word first
  - Subblocking/sectoring
  - Better replacement/insertion policies
  - Non-blocking caches (multiple cache misses in parallel)
  - Multiple accesses per cycle
  - Software approaches
Cheap Ways of Reducing Conflict Misses

- Instead of building highly-associative caches:
  - Victim Caches
  - Hashed/randomized Index Functions
  - Pseudo Associativity
  - Skewed Associative Caches
  - ...

Victim Cache: Reducing Conflict Misses


- **Idea:** Use a small fully associative buffer (victim cache) to store evicted blocks
  - + Can avoid ping ponging of cache blocks mapped to the same set (if two cache blocks continuously accessed in nearby time conflict with each other)
  - -- Increases miss latency if accessed serially with L2; adds complexity
Hashing and Pseudo-Associativity

- **Hashing**: Use better “randomizing” index functions
  - + can reduce conflict misses
    - by distributing the accessed memory blocks more evenly to sets
    - Example of conflicting accesses: strided access pattern where stride value equals number of sets in cache
  - -- More complex to implement: can lengthen critical path

- **Pseudo-associativity (Poor Man’s associative cache)**
  - Serial lookup: On a miss, use a different index function and access cache again
  - Given a direct-mapped array with K cache blocks
    - Implement K/N sets
    - Given address Addr, sequentially look up: \( \{0, \text{Addr}[\log(K/N)-1: 0]\}, \{1, \text{Addr}[\log(K/N)-1: 0]\}, \ldots, \{N-1, \text{Addr}[\log(K/N)-1: 0]\} \)
Skewed Associative Caches

- Idea: Reduce conflict misses by using different index functions for each cache way

Skewed Associative Caches (I)

- Basic 2-way associative cache structure
Skewed Associative Caches (II)

- Skewed associative caches
  - Each bank has a different index function
Skewed Associative Caches (III)

- Idea: Reduce conflict misses by using different index functions for each cache way

- Benefit: indices are more randomized (memory blocks are better distributed across sets)
  - Less likely two blocks have same index
    - Reduced conflict misses

- Cost: additional latency of hash function

Software Approaches for Higher Hit Rate

- Restructuring data access patterns
- Restructuring data layout

- Loop interchange
- Data structure separation/merging
- Blocking
- ...

16
Restructuring Data Access Patterns (I)

- Idea: Restructure data layout or data access patterns
- Example: If column-major
  - $x[i+1,j]$ follows $x[i,j]$ in memory
  - $x[i,j+1]$ is far away from $x[i,j]$

<table>
<thead>
<tr>
<th>Poor code</th>
<th>Better code</th>
</tr>
</thead>
<tbody>
<tr>
<td>for $i = 1$, rows</td>
<td></td>
</tr>
<tr>
<td>for $j = 1$, columns</td>
<td></td>
</tr>
<tr>
<td>sum = sum + $x[i,j]$</td>
<td>sum = sum + $x[i,j]$</td>
</tr>
</tbody>
</table>

- This is called loop interchange
- Other optimizations can also increase hit rate
  - Loop fusion, array merging, ...
- What if multiple arrays? Unknown array size at compile time?
Restructuring Data Access Patterns (II)

- **Blocking**
  - Divide loops operating on arrays into computation chunks so that each chunk can hold its data in the cache
  - Avoids cache conflicts between different chunks of computation
  - Essentially: Divide the working set so that each piece fits in the cache

- But, there are still self-conflicts in a block
  1. there can be conflicts among different arrays
  2. array sizes may be unknown at compile/programming time
Restructuring Data Layout (I)

- Pointer based traversal (e.g., of a linked list)
- Assume a huge linked list (1M nodes) and unique keys
- Why does the code on the left have poor cache hit rate?
  - “Other fields” occupy most of the cache line even though rarely accessed!

```c
struct Node {
    struct Node* node;
    int key;
    char [256] name;
    char [256] school;
}

while (node) {
    if (node->key == input-key) {
        // access other fields of node
    }
    node = node->next;
}
```
Restructuring Data Layout (II)

- **Idea:** separate frequently-used fields of a data structure and pack them into a separate data structure

- **Who should do this?**
  - Programmer
  - Compiler
  - Profiling vs. dynamic
  - Hardware?
  - Who can determine what is frequently used?

```c
struct Node {
    struct Node* node;
    int key;
    struct Node-data* node-data;
}

struct Node-data {
    char [256] name;
    char [256] school;
}

while (node) {
    if (node->key == input-key) {
        // access node->node-data
    }
    node = node->next;
}
```
Improving Basic Cache Performance

- Reducing miss rate
  - More associativity
  - Alternatives/enhancements to associativity
    - Victim caches, hashing, pseudo-associativity, skewed associativity
  - Better replacement/insertion policies
  - Software approaches

- Reducing miss latency/cost
  - Multi-level caches
  - Critical word first
  - Subblocking/sectoring
  - Better replacement/insertion policies
  - Non-blocking caches (multiple cache misses in parallel)
  - Multiple accesses per cycle
  - Software approaches
Miss Latency/Cost

- What is miss latency or miss cost affected by?
  - Where does the miss get serviced from?
    - Local vs. remote memory
    - What level of cache in the hierarchy?
    - Row hit versus row miss
    - Queueing delays in the memory controller and the interconnect
    - ...
  - How much does the miss stall the processor?
    - Is it overlapped with other latencies?
    - Is the data immediately needed?
    - ...

Memory Level Parallelism (MLP) means generating and servicing multiple memory accesses in parallel [Glew’ 98].

Several techniques to improve MLP (e.g., out-of-order execution).

MLP varies. Some misses are isolated and some parallel. How does this affect cache replacement?
Traditional Cache Replacement Policies

- Traditional cache replacement policies try to reduce miss count

- **Implicit assumption**: Reducing miss count reduces memory-related stall time

- Misses with varying cost/MLP *breaks* this assumption!

- Eliminating an isolated miss helps performance more than eliminating a parallel miss

- Eliminating a higher-latency miss could help performance more than eliminating a lower-latency miss
An Example

Misses to blocks P1, P2, P3, P4 can be parallel
Misses to blocks S1, S2, and S3 are isolated

Two replacement algorithms:
1. Minimizes miss count (Belady’s OPT)
2. Reduces isolated miss (MLP-Aware)

For a fully associative cache containing 4 blocks
Fewest Misses ≠ Best Performance

Belady’s OPT replacement

MLP-Aware replacement

Saved cycles
MLP-Aware Cache Replacement

- How do we incorporate MLP into replacement decisions?

  - Required reading for this week
Enabling Multiple Outstanding Misses
Handling Multiple Outstanding Accesses

- **Question:** If the processor can generate multiple cache accesses, can the later accesses be handled while a previous miss is outstanding?

- **Goal:** Enable cache access when there is a pending miss

- **Goal:** Enable multiple misses in parallel
  - Memory-level parallelism (MLP)

- **Solution:** Non-blocking or lockup-free caches
Handling Multiple Outstanding Accesses

- **Idea:** Keep track of the status/data of misses that are being handled in Miss Status Handling Registers (MSHRs)

- A cache access checks MSHRs to see if a miss to the same block is already *pending*.
  - If pending, a new request is not generated
  - If pending and the needed data available, data forwarded to later load

- Requires buffering of outstanding miss requests
Miss Status Handling Register

- Also called “miss buffer”
- Keeps track of
  - Outstanding cache misses
  - Pending load/store accesses that refer to the missing cache block
- Fields of a single MSHR entry
  - Valid bit
  - Cache block address (to match incoming accesses)
  - Control/status bits (prefetch, issued to memory, which subblocks have arrived, etc)
  - Data for each subblock
  - For each pending load/store
    - Valid, type, data size, byte in block, destination register or store buffer entry address
## Miss Status Handling Register Entry

<table>
<thead>
<tr>
<th>Valid</th>
<th>Block Address</th>
<th>Issued</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1</th>
<th>3</th>
<th>5</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>Type</td>
<td>Block Offset</td>
<td>Destination</td>
</tr>
<tr>
<td>Valid</td>
<td>Type</td>
<td>Block Offset</td>
<td>Destination</td>
</tr>
<tr>
<td>Valid</td>
<td>Type</td>
<td>Block Offset</td>
<td>Destination</td>
</tr>
<tr>
<td>Valid</td>
<td>Type</td>
<td>Block Offset</td>
<td>Destination</td>
</tr>
</tbody>
</table>

- Load/store 0
- Load/store 1
- Load/store 2
- Load/store 3
MSHR Operation

- On a cache miss:
  - Search MSHRs for a pending access to the same block
    - Found: Allocate a load/store entry in the same MSHR entry
    - Not found: Allocate a new MSHR
    - No free entry: stall

- When a subblock returns from the next level in memory
  - Check which loads/stores waiting for it
    - Forward data to the load/store unit
    - Deallocate load/store entry in the MSHR entry
  - Write subblock in cache or MSHR
  - If last subblock, deallocate MSHR (after writing the block in cache)
Non-Blocking Cache Implementation

- When to access the MSHRs?
  - In parallel with the cache?
  - After cache access is complete?

- MSHRs need not be on the critical path of hit requests
  - Which one below is the common case?
    - Cache miss, MSHR hit
    - Cache hit
Enabling High Bandwidth Memories
Multiple Instructions per Cycle

- Can generate multiple cache/memory accesses per cycle
- How do we ensure the cache/memory can handle multiple accesses in the same clock cycle?

Solutions:
- true multi-porting
- virtual multi-porting (time sharing a port)
- multiple cache copies
- banking (interleaving)
Handling Multiple Accesses per Cycle (I)

- True multiporting
  - Each memory cell has multiple read or write ports
    + Truly concurrent accesses (no conflicts on read accesses)
    -- Expensive in terms of latency, power, area
  - What about read and write to the same location at the same time?
    - Peripheral logic needs to handle this
Peripheral Logic for True Multiporting
Peripheral Logic for True Multiporting
Virtual multiporting

- Time-share a single port
- Each access needs to be (significantly) shorter than clock cycle
- Used in Alpha 21264
- Is this scalable?
Handling Multiple Accesses per Cycle (III)

- **Multiple cache copies**
  - Stores update both caches
  - Loads proceed in parallel

- **Used in Alpha 21164**

- **Scalability?**
  - Store operations form a bottleneck
  - Area proportional to “ports”
Handling Multiple Accesses per Cycle (III)

- **Banking (Interleaving)**
  - Bits in address determines which bank an address maps to
    - Address space partitioned into separate banks
    - Which bits to use for “bank address”?
  - + No increase in data store area
  - -- Cannot satisfy multiple accesses to the same bank
  - -- Crossbar interconnect in input/output

- **Bank conflicts**
  - Two accesses are to the same bank
  - How can these be reduced?
    - Hardware? Software?
General Principle: Interleaving

- **Interleaving (banking)**
  - **Problem**: a single monolithic memory array takes long to access and does not enable multiple accesses in parallel
  - **Goal**: Reduce the latency of memory array access and enable multiple accesses in parallel
  - **Idea**: Divide the array into multiple banks that can be accessed independently (in the same cycle or in consecutive cycles)
    - Each bank is smaller than the entire memory storage
    - Accesses to different banks can be overlapped
  - **A Key Issue**: How do you map data to different banks? (i.e., how do you interleave data across banks?)
Further Readings on Caching and MLP


Multi-Core Issues in Caching
Caches in Multi-Core Systems

- Cache efficiency becomes even more important in a multi-core/multi-threaded system
  - Memory bandwidth is at premium
  - Cache space is a limited resource

- How do we design the caches in a multi-core system?

- Many decisions
  - Shared vs. private caches
  - How to maximize performance of the entire system?
  - How to provide QoS to different threads in a shared cache?
  - Should cache management algorithms be aware of threads?
  - How should space be allocated to threads in a shared cache?
Private vs. Shared Caches

- **Private** cache: Cache belongs to one core (a shared block can be in multiple caches)
- **Shared** cache: Cache is shared by multiple cores
Resource Sharing Concept and Advantages

- Idea: Instead of dedicating a hardware resource to a hardware context, allow multiple contexts to use it
  - Example resources: functional units, pipeline, caches, buses, memory

- Why?
  - Resource sharing improves utilization/efficiency → throughput
    - When a resource is left idle by one thread, another thread can use it; no need to replicate shared data
  - Reduces communication latency
    - For example, shared data kept in the same cache in multithreaded processors
  - Compatible with the shared memory model
Resource Sharing Disadvantages

- Resource sharing results in contention for resources
  - When the resource is not idle, another thread cannot use it
  - If space is occupied by one thread, another thread needs to re-occupy it

- Sometimes reduces each or some thread’s performance
  - Thread performance can be worse than when it is run alone

- Eliminates performance isolation → inconsistent performance across runs
  - Thread performance depends on co-executing threads

- Uncontrolled (free-for-all) sharing degrades QoS
  - Causes unfairness, starvation

Need to efficiently and fairly utilize shared resources
Private vs. Shared Caches

- **Private** cache: Cache belongs to one core (a shared block can be in multiple caches)
- **Shared** cache: Cache is shared by multiple cores
Shared Caches Between Cores

- **Advantages:**
  - High effective capacity
  - *Dynamic partitioning* of available cache space
    - No fragmentation due to static partitioning
  - Easier to maintain coherence (a cache block is in a single location)
  - Shared data and locks do not ping pong between caches

- **Disadvantages**
  - Slower access
  - Cores incur *conflict misses due to other cores’ accesses*
    - Misses due to inter-core interference
    - Some cores can destroy the hit rate of other cores
  - Guaranteeing a minimum level of service (or fairness) to each core is harder (how much space, how much bandwidth?)
Shared Caches: How to Share?

- Free-for-all sharing
  - Placement/replacement policies are the same as a single core system (usually LRU or pseudo-LRU)
  - Not thread/application aware
  - An incoming block evicts a block regardless of which threads the blocks belong to

- Problems
  - Inefficient utilization of cache: LRU is not the best policy
  - A cache-unfriendly application can destroy the performance of a cache friendly application
  - Not all applications benefit equally from the same amount of cache: free-for-all might prioritize those that do not benefit
  - Reduced performance, reduced fairness
Example: Utility Based Shared Cache Partitioning

- **Goal:** Maximize system throughput
- **Observation:** Not all threads/applications benefit equally from caching → simple LRU replacement not good for system throughput
- **Idea:** Allocate more cache space to applications that obtain the most benefit from more space

- The high-level idea can be applied to other shared resources as well.

The Multi-Core System: A Shared Resource View
Need for QoS and Shared Resource Mgmt.

- Why is unpredictable performance (or lack of QoS) bad?
- Makes programmer’s life difficult
  - An optimized program can get low performance (and performance varies widely depending on co-runners)
- Causes discomfort to user
  - An important program can starve
  - Examples from shared software resources
- Makes system management difficult
  - How do we enforce a Service Level Agreement when hardware resources are sharing is uncontrollable?
Resource Sharing vs. Partitioning

- Sharing improves throughput
  - Better utilization of space

- Partitioning provides performance isolation (predictable performance)
  - Dedicated space

- Can we get the benefits of both?

- Idea: Design shared resources such that they are efficiently utilized, controllable and partitionable
  - No wasted resource + QoS mechanisms for threads
Shared Hardware Resources

- Memory subsystem (in both multithreaded and multi-core systems)
  - Non-private caches
  - Interconnects
  - Memory controllers, buses, banks

- I/O subsystem (in both multithreaded and multi-core systems)
  - I/O, DMA controllers
  - Ethernet controllers

- Processor (in multithreaded systems)
  - Pipeline resources
  - L1 caches