Review: Memory Hierarchy of a Modern Computer System

- Take advantage of the principle of locality to:
  - Present as much memory as in the cheapest technology
  - Provide access at speed offered by the fastest technology

Review: A Summary on Sources of Cache Misses

- Compulsory (cold start): first reference to a block
  - “Cold” fact of life: not a whole lot you can do about it
  - Note: When running “billions” of instruction, Compulsory Misses are insignificant
- Capacity:
  - Cache cannot contain all blocks access by the program
  - Solution: increase cache size
- Conflict (collision):
  - Multiple memory locations mapped to same cache location
  - Solutions: increase cache size, or increase associativity
- Two others:
  - Coherence (Invalidation): other process (e.g., I/O) updates memory
  - Policy: Due to non-optimal replacement policy

Review: Set Associative Cache

- N-way set associative: N entries per Cache Index
  - N direct mapped caches operates in parallel
- Example: Two-way set associative cache
  - Cache Index selects a “set” from the cache
  - Two tags in the set are compared to input in parallel
  - Data is selected based on the tag result
Review: Where does a Block Get Placed in a Cache?

- Example: Block 12 placed in 8 block cache

32-Block Address Space:

<table>
<thead>
<tr>
<th>Block no.</th>
<th>0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1</th>
</tr>
</thead>
</table>

- Direct mapped: block 12 can go only into block 4 (12 mod 8)

- Set associative: block 12 can go anywhere in set 0 (12 mod 4)

- Fully associative: block 12 can go anywhere

Which block should be replaced on a miss?

- Easy for Direct Mapped: Only one possibility
- Set Associative or Fully Associative:
  - Random
  - LRU (Least Recently Used)

<table>
<thead>
<tr>
<th>Size</th>
<th>2-way LRU</th>
<th>4-way LRU</th>
<th>8-way LRU</th>
</tr>
</thead>
<tbody>
<tr>
<td>16 KB</td>
<td>5.2% 5.7%</td>
<td>4.7% 5.3%</td>
<td>4.4% 5.0%</td>
</tr>
<tr>
<td>64 KB</td>
<td>1.9% 2.0%</td>
<td>1.5% 1.7%</td>
<td>1.4% 1.5%</td>
</tr>
<tr>
<td>256 KB</td>
<td>1.15% 1.17%</td>
<td>1.13% 1.13%</td>
<td>1.12% 1.12%</td>
</tr>
</tbody>
</table>

Goals for Today

- Finish discussion of Caching/TLBs
- Concept of Paging to Disk
- Page Faults and TLB Faults
- Precise Interrupts
- Page Replacement Policies

Note: Some slides and/or pictures in the following are adapted from slides ©2005 Silberschatz, Galvin, and Gagne. Many slides generated from my lecture notes by Kubiatowicz.
**Review: Cache performance**

- **Miss-oriented Approach to Memory Access:**
  \[ \text{CPUtime} = IC \times \left( CPI_{\text{duction}} + \frac{\text{MemAccess}}{\text{Inst}} \times \text{MissRate} \times \text{MissPenalty} \right) \times \text{CycleTime} \]
  
- Separating out Memory component entirely
  - AMAT = Average Memory Access Time
  \[ \text{AMAT} = \text{HitRate} \times \text{HitTime} + \text{MissRate} \times \text{MissTime} \]
  \[ = \text{HitTime} + \text{MissRate} \times \text{MissPenalty} \]
  \[ = \frac{\text{Frac}_{\text{hit}}}{\text{hit}} \times (\text{HitTime}_{\text{hit}} + \text{MissRate}_{\text{hit}} \times \text{MissPenalty}_{\text{hit}}) + \frac{\text{Frac}_{\text{data}}}{\text{data}} \times (\text{HitTime}_{\text{data}} + \text{MissRate}_{\text{data}} \times \text{MissPenalty}_{\text{data}}) \]

- **AMAT for Second-Level Cache**
  \[ \text{AMAT}_{1st} = \text{HitTime}_{1st} + \text{MissRate}_{1st} \times \text{MissPenalty}_{1st} \]
  \[ = \text{HitTime}_{1st} + \text{MissRate}_{1st} \times \text{AMAT}_{2nd} \]
  \[ = \text{HitTime}_{1st} + \text{MissRate}_{1st} \times (\text{HitTime}_{2nd} + \text{MissRate}_{2nd} \times \text{MissPenalty}_{2nd}) \]

---

**Caching Applied to Address Translation**

- **Question is one of page locality: does it exist?**
  - Instruction accesses spend a lot of time on the same page (since accesses sequential)
  - Stack accesses have definite locality of reference
  - Data accesses have less page locality, but still some...

- **Can we have a TLB hierarchy?**
  - Sure: multiple levels at different sizes/speeds

---

**What Actually Happens on a TLB Miss?**

- **Hardware traversed page tables:**
  - On TLB miss, hardware in MMU looks at current page table to fill TLB (may walk multiple levels)
    - If PTE valid, hardware fills TLB and processor never knows
    - If PTE marked as invalid, causes Page Fault, after which kernel decides what to do afterwards

- **Software traversed Page tables (like MIPS):**
  - On TLB miss, processor receives TLB fault
    - Kernel traverses page table to find PTE
      - If PTE valid, fills TLB and returns from fault
      - If PTE marked as invalid, internally calls Page Fault handler

- **Most chip sets provide hardware traversal:**
  - Modern operating systems tend to have more TLB faults since they use translation for many things
  - Examples:
    - shared segments
    - user-level portions of an operating system

---

**What happens on a Context Switch?**

- **Need to do something, since TLBs map virtual addresses to physical addresses**
  - Address Space just changed, so TLB entries no longer valid!

- **Options?**
  - Invalidate TLB: simple but might be expensive
    - What if switching frequently between processes?
  - Include ProcessID in TLB
    - This is an architectural solution: needs hardware

- **What if translation tables change?**
  - For example, to move page from memory to disk or vice versa...
  - Must invalidate TLB entry!
    - Otherwise, might think that page is still in memory!
Administrative

- Exam not graded yet
  - Will get solutions up early next week at latest
  - Will talk about Problem 3 in section on Friday
- Project 1 reports almost graded
  - Hoping to get them to you soon
- Project 2
  - Code due Tuesday (10/26)
  - Look at the lecture schedule to keep up with due dates!

Mobile devices are the future

- Android is the popular operating system from Google
  - For Mobile devices
    » Phones
    » Ebook Readers (i.e. B&N)
    » Tablets
  - Linux version 2.6.x
  - Java virtual machine and runtime system
  - Lots of media extensions
    » WebKit for browsing
    » Media Libraries
    » Cellular Networking
- Mobile Systems are the hottest new software stack
  - Ubiquitous Computing
  - Worldwide, more than 1 billion new cell phones purchased/year for last few years
    » Compare: worldwide number PCs purchased/year ~ 250M

What TLB organization makes sense?

- Needs to be really fast
  - Critical path of memory access
    » In simplest view: before the cache
    » Thus, this adds to access time (reducing cache speed)
  - Seems to argue for Direct Mapped or Low Associativity
- However, needs to have very few conflicts!
  - With TLB, the Miss Time extremely high!
  - This argues that cost of Conflict (Miss Penalty) is much higher than slightly increased cost of access (Hit Time)
- Thrashing: continuous conflicts between accesses
  - What if use low order bits of page as index into TLB?
    » First page of code, data, stack may map to same entry
    » Need 3-way associativity at least?
  - What if use high order bits as index?
    » TLB mostly unused for small programs

TLB organization: include protection

- How big does TLB actually have to be?
  - Usually small: 128-512 entries
  - Not very big, can support higher associativity
- TLB usually organized as fully-associative cache
  - Lookup is by Virtual Address
  - Returns Physical Address + other info
- Example for MIPS R3000:
  - What happens when fully-associative is too slow?
    - Put a small (4-16 entry) direct-mapped cache in front
    - Called a “TLB Slice”
- When does TLB lookup occur?
  - Before cache lookup?
  - In parallel with cache lookup?
Example: R3000 pipeline includes TLB “stages”

MIPS R3000 Pipeline

<table>
<thead>
<tr>
<th>Inst Fetch</th>
<th>Dcd/ Reg</th>
<th>ALU / E.A</th>
<th>Memory</th>
<th>Write Reg</th>
</tr>
</thead>
<tbody>
<tr>
<td>TLB</td>
<td>I-Cache</td>
<td>RF</td>
<td>Operation</td>
<td>WB</td>
</tr>
<tr>
<td>E.A</td>
<td>TLB</td>
<td>D-Cache</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TLB
64 entry, on-chip, fully associative, software TLB fault handler

Virtual Address Space

<table>
<thead>
<tr>
<th>ASID</th>
<th>V. Page Number</th>
<th>Offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>20</td>
<td>12</td>
</tr>
</tbody>
</table>

0x0 User segment (caching based on PT/TLB entry)
100 Kernel physical space, cached
101 Kernel physical space, uncached
11x Kernel virtual space

Allows context switching among 64 user processes without TLB flush

Combination
Segments and Paging!

Reducing translation time further

• As described, TLB lookup is in serial with cache lookup:

Virtual Address

\[ \text{Virtual Address} = \text{V page no.} + \text{offset} \]

\[ \begin{align*}
\text{TLB Lookup} \\
\text{V Access} &\quad \text{Rights} & \quad \text{PA} \\
\end{align*} \]

Physical Address

\[ \text{Physical Address} = \text{P page no.} + \text{offset} \]

• Machines with TLBs go one step further: they overlap TLB lookup with cache access.
  - Works because offset available early

Overlapping TLB & Cache Access

• Here is how this might work with a 4K cache:

4K Cache

\[ \text{index} = 20, \quad \text{offset} = 10, \quad \text{disp} = 2 \]

Hit/ Miss

\[ \text{FN Data} \]

\[ \text{FN} \]

\[ \text{FN Hit/ Miss} \]

\[ \text{assoc lookup} \]

\[ 32 \]

\[ 1 \text{K} \]

\[ 4 \text{ bytes} \]

\[ 20 \]

\[ 3 \]

\[ \text{Hit/ Miss} \]

\[ \text{Hit/ Miss} \]

\[ \text{Hit/ Miss} \]

Overlapping TLB & Cache Access

• Here is how this might work with a 4K cache:

4K Cache

\[ \text{index} = 20, \quad \text{offset} = 10, \quad \text{disp} = 2 \]

Hit/ Miss

\[ \text{FN Data} \]

\[ \text{FN} \]

\[ \text{FN Hit/ Miss} \]

\[ \text{assoc lookup} \]

\[ 32 \]

\[ 1 \text{K} \]

\[ 4 \text{ bytes} \]

\[ 20 \]

\[ 3 \]

\[ \text{Hit/ Miss} \]

\[ \text{Hit/ Miss} \]

\[ \text{Hit/ Miss} \]

What if cache size is increased to 8KB?

- Overlap not complete
- Need to do something else. See CS152/252

Another option: Virtual Caches

- Tags in cache are virtual addresses
- Translation only happens on cache misses

Demand Paging

• Modern programs require a lot of physical memory
  - Memory per system growing faster than 25%-30%/year
  - But they don’t use all their memory all of the time
  - 90-10 rule: programs spend 90% of their time in 10% of their code
  - Wasteful to require all of user’s code to be in memory

• Solution: use main memory as cache for disk
Illusion of Infinite Memory

- Disk is larger than physical memory
  - In-use virtual memory can be bigger than physical memory
  - Combined memory of running processes much larger than physical memory
  - More programs fit into memory, allowing more concurrency
- Principle: Transparent Level of Indirection (page table)
  - Supports flexible placement of physical data
  - Data could be on disk or somewhere across network
  - Variable location of data transparent to user program
  - Performance issue, not correctness issue

Demand Paging is Caching

- Since Demand Paging is Caching, must ask:
  - What is block size?
    » 1 page
  - What is organization of this cache (i.e. direct-mapped, set-associative, fully-associative)?
    » Fully associative: arbitrary virtual→physical mapping
  - How do we find a page in the cache when it's time?
    » First check TLB, then page-table traversal
  - What is page replacement policy? (i.e. LRU, Random...)
    » This requires more explanation... (kinda LRU)
  - What happens on a miss?
    » Go to lower level to fill miss (i.e. disk)
  - What happens on a write? (write-through, write back)
    » Definitely write-back. Need dirty bit!

Review: What is in a PTE?

- What is in a Page Table Entry (or PTE)?
  - Pointer to next-level page table or to actual page
  - Permission bits: valid, read-only, read-write, write-only
- Example: Intel x86 architecture PTE:
  - Address same format previous slide (10, 10, 12-bit offset)
  - Intermediate page tables called "Directories"

<table>
<thead>
<tr>
<th>PTE Bits</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>Present (same as &quot;valid&quot; bit in other architectures)</td>
</tr>
<tr>
<td>W</td>
<td>Writeable</td>
</tr>
<tr>
<td>U</td>
<td>User accessible</td>
</tr>
<tr>
<td>PWT</td>
<td>Page write transparent: external cache write-through</td>
</tr>
<tr>
<td>PCD</td>
<td>Page cache disabled (page cannot be cached)</td>
</tr>
<tr>
<td>A</td>
<td>Accessed: page has been accessed recently</td>
</tr>
<tr>
<td>D</td>
<td>Dirty (PTE only); page has been modified recently</td>
</tr>
<tr>
<td>L</td>
<td>L=1⇒4MB page (directory only). Bottom 22 bits of virtual address serve as offset</td>
</tr>
</tbody>
</table>

Demand Paging Mechanisms

- PTE helps us implement demand paging
  - Valid ⇒ Page in memory, PTE points at physical page
  - Not Valid ⇒ Page not in memory; use info in PTE to find it on disk when necessary
- Suppose user references page with invalid PTE?
  - Memory Management Unit (MMU) traps to OS
    » Resulting trap is a "Page Fault"
    - What does OS do on a Page Fault?:
      » Choose an old page to replace
      » If old page modified ("D=1"), write contents back to disk
      » Change its PTE and any cached TLB to be invalid
      » Load new page into memory from disk
      » Update page table entry, invalidate TLB for new entry
      » Continue thread from original faulting location
    - TLB for new page will be loaded when thread continued!
    - While pulling pages off disk for one process, OS runs another process from ready queue
      » Suspended process sits on wait queue
Software-Loaded TLB

- MIPS/Nachos TLB is loaded by software
  - High TLB hit rate OK to trap to software to fill the TLB, even if slower
  - Simpler hardware and added flexibility: software can maintain translation tables in whatever convenient format

- How can a process run without access to page table?
  - Fast path (TLB hit with valid=1):
    » Translation to physical page done by hardware
  - Slow path (TLB hit with valid=0 or TLB miss)
    » Hardware receives a “TLB Fault”
    » What does OS do on a TLB Fault?
      » Traverse page table to find appropriate PTE
      » If valid=1, load page table entry into TLB, continue thread
      » If valid=0, perform “Page Fault” detailed previously
      » Continue thread

- Everything is transparent to the user process:
  - It doesn’t know about paging to/from disk
  - It doesn’t even know about software TLB handling

Consider weird things that can happen

- What if an instruction has side-effects?
  - Options:
    » Unwind side-effects (easy to restart)
    » Finish off side-effects (messy!)
  - Example 1: mov (sp)+, 10
    » What if page fault occurs when write to stack pointer?
    » Did sp get incremented before or after the page fault?
  - Example 2: strcpy (r1), (r2)
    » Source and destination overlap: can’t unwind in principle!
    » IBM S/370 and VAX solution: execute twice - once read-only

- What about “RISC” processors?
  - For instance delayed branches?
    » Example: bne somewhere
    » ld r1, (sp)
    » Precise exception state consists of two PCs: PC and nPC
  - Delayed exceptions:
    » Example: div r1, r2, r3
    » ld r1, (sp)
    » What if takes many cycles to discover divide by zero, but load has already caused page fault?

Precise Exceptions

- Precise \( \Rightarrow \) state of the machine is preserved as if program executed up to the offending instruction
  - All previous instructions completed
  - Offending instruction and all following instructions act as if they have not even started
  - Same system code will work on different implementations
  - Difficult in the presence of pipelining, out-of-order execution, ...
  - MIPS takes this position

- Imprecise \( \Rightarrow \) system software has to figure out what is where and put it all back together
  - Performance goals often lead designers to forsake precise interrupts
  - system software developers, user, markets etc. usually wish they had not done this

- Modern techniques for out-of-order execution and branch prediction help implement precise interrupts
Steps in Handling a Page Fault

Demand Paging Example

- Since Demand Paging like caching, can compute average access time ("Effective Access Time")
  - \[ \text{EAT} = \text{Hit Rate} \times \text{Hit Time} + \text{Miss Rate} \times \text{Miss Time} \]
  - \[ \text{EAT} = \text{Hit Time} + \text{Miss Rate} \times \text{Miss Penalty} \]

- Example:
  - Memory access time = 200 nanoseconds
  - Average page-fault service time = 8 milliseconds
  - Suppose \( p \) = Probability of miss, \( 1-p \) = Probability of hit
  - Then, we can compute EAT as follows:
    \[ \text{EAT} = 200\text{ns} + p \times 8 \text{ms} \]
    \[ = 200\text{ns} + p \times 8,000,000\text{ns} \]

- If one access out of 1,000 causes a page fault, then \( EAT = 8.2 \mu \text{s} \):
  - This is a slowdown by a factor of 40!

- What if want slowdown by less than 10%?
  - \[ 200\text{ns} \times 1.1 < EAT \Rightarrow p < 2.5 \times 10^{-6} \]
  - This is about 1 page fault in 400000!

What Factors Lead to Misses?

- Compulsory Misses:
  - Pages that have never been paged into memory before
  - How might we remove these misses?
    - Prefetching: loading them into memory before needed
    - Need to predict future somehow! More later.

- Capacity Misses:
  - Not enough memory. Must somehow increase size.
  - Can we do this?
    - One option: Increase amount of DRAM (not quick fix!)
    - Another option: If multiple processes in memory: adjust percentage of memory allocated to each one!

- Conflict Misses:
  - Technically, conflict misses don't exist in virtual memory, since it is a "fully-associative" cache

- Policy Misses:
  - Caused when pages were in memory, but kicked out prematurely because of the replacement policy
  - How to fix? Better replacement policy

Page Replacement Policies

- Why do we care about Replacement Policy?
  - Replacement is an issue with any cache
  - Particularly important with pages
    - The cost of being wrong is high: must go to disk
    - Must keep important pages in memory, not toss them out

- FIFO (First In, First Out)
  - Throw out oldest page. Be fair - let every page live in memory for same amount of time.
  - Bad, because throws out heavily used pages instead of infrequently used pages

- MIN (Minimum):
  - Replace page that won't be used for the longest time
  - Great, but can't really know future...
  - Makes good comparison case, however

- RANDOM:
  - Pick random page for every replacement
  - Typical solution for TLB's. Simple hardware
  - Pretty unpredictable - makes it hard to make real-time guarantees
Replacement Policies (Con’t)

- LRU (Least Recently Used):
  - Replace page that hasn’t been used for the longest time
  - Programs have locality, so if something not used for a while, unlikely to be used in the near future.
  - Seems like LRU should be a good approximation to MIN.

- How to implement LRU? Use a list!
  - On each use, remove page from list and place at head
  - LRU page is at tail

- Problems with this scheme for paging?
  - Need to know immediately when each page used so that can change position in list...
  - Many instructions for each hardware access

- In practice, people approximate LRU (more later)

Example: FIFO

- Suppose we have 3 page frames, 4 virtual pages, and following reference stream:
  - A B C A B D A D B C B
- Consider FIFO Page replacement:

Example: MIN

- Suppose we have the same reference stream:
  - A B C A B D A D B C B
- Consider MIN Page replacement:

When will LRU perform badly?

- Consider the following: A B C D A B C D A B C D
- LRU Performs as follows (same as FIFO here):

- Every reference is a page fault!
- MIN Does much better:

10/20/10 Kubiatowicz CS162 ©UCB Fall 2010 Lec 14.33

10/20/10 Kubiatowicz CS162 ©UCB Fall 2010 Lec 14.34

10/20/10 Kubiatowicz CS162 ©UCB Fall 2010 Lec 14.35

10/20/10 Kubiatowicz CS162 ©UCB Fall 2010 Lec 14.36
Summary

- TLB is cache on translations
  - Fully associative to reduce conflicts
  - Can be overlapped with cache access
- Demand Paging:
  - Treat memory as cache on disk
  - Cache miss ⇒ get page from disk
- Transparent Level of Indirection
  - User program is unaware of activities of OS behind scenes
  - Data can be moved without affecting application correctness
- Software-loaded TLB
  - Fast Path: handled in hardware (TLB hit with valid=1)
  - Slow Path: Trap to software to scan page table
- Precise Exception specifies a single instruction for which:
  - All previous instructions have completed (committed state)
  - No following instructions nor actual instruction have started
- Replacement policies
  - FIFO: Place pages on queue, replace page at end
  - MIN: replace page that will be used farthest in future
  - LRU: Replace page that hasn’t be used for the longest time