Virtual Memory III
CSE 351 Autumn 2019

Instructor:
Justin Hsia

Teaching Assistants:
Andrew Hu
Antonio Castelli
Cosmo Wang
Diya Joy
Ivy Yu
Kaelin Laundry
Maurice Montag
Melissa Birchfield
Millicent Li
Suraj Jagadeesh

https://xkcd.com/648/
Administrivia

- Lab 4 due Friday (11/22)
- hw20 due Monday (11/25)
Address Translation: Page Hit

1) Processor sends *virtual* address to MMU (*memory management unit*)

2-3) MMU fetches PTE from page table in cache/memory
   (Uses PTBR to find beginning of page table for current process)

4) MMU sends *physical* address to cache/memory requesting data

5) Cache/memory sends data to processor

VA = Virtual Address  PTEA = Page Table Entry Address  PTE= Page Table Entry
PA = Physical Address  Data = Contents of memory stored at VA originally requested by CPU
Address Translation: Page Fault

1) Processor sends virtual address to MMU
2-3) MMU fetches PTE from page table in cache/memory
4) Valid bit is zero, so MMU triggers page fault exception
5) Handler identifies victim (and, if dirty, pages it out to disk)
6) Handler pages in new page and updates PTE in memory
7) Handler returns to original process, restarting faulting instruction
Hmm... Translation Sounds Slow

- The MMU accesses memory *twice*: once to get the PTE for translation, and then again for the actual memory request
  - The PTEs *may* be cached in L1 like any other memory word
    - But they may be evicted by other data references
    - And a hit in the L1 cache still requires 1-3 cycles

- *What can we do to make this faster?*
  - **Solution:** add another cache! 🎉
Speeding up Translation with a TLB

Translation "Lookaside Buffer" (TLB):

- Small hardware cache in MMU
  - Split VPN into TLB Tag and TLB Index based on # of sets in TLB
- Maps virtual page numbers to physical page numbers
- Stores *page table entries* for a small number of pages
  - Modern Intel processors have 128 or 256 entries in TLB
- Much faster than a page table lookup in cache/memory
**TLB Hit**
*(page has been used recently)*

- A TLB hit eliminates a memory access!
TLB Miss
(page has not been used recently)

A TLB miss incurs an additional memory access (the PTE)
- Fortunately, TLB misses are rare
Fetching Data on a Memory Read

1) Check TLB
   - **Input:** VPN, **Output:** PPN
   - **TLB Hit:** Fetch translation, return PPN
   - **TLB Miss:** Check page table (in memory)
     - **Page Table Hit:** Load page table entry into TLB
     - **Page Fault:** Fetch page from disk to memory, update corresponding page table entry, then load entry into TLB

2) Check cache
   - **Input:** physical address, **Output:** data
   - **Cache Hit:** Return data value to processor
   - **Cache Miss:** Fetch data value from memory, store it in cache, return it to processor

*these serve different purposes!*
Address Translation

Virtual Address

TLB Lookup

TLB Miss

Check the Page Table

($/Mem$)

TLB Hit

Protection Check Rwx

Protection Fault

Page not in Mem

Page Fault (OS loads page)

Find in Disk

Access Denied

Update TLB

Find in Mem

Access Permitted

Protection Fault

SIGSEGV

Physical Address

Check cache

Miss

Hit
Address Manipulation

request from CPU: $n$-bit virtual address

split to access TLB: TLB Tag TLB Index Page Offset

(on TLB miss) access PT: Virtual Page Number Page offset

$\textit{m}$-bit physical address:

split to access cache: Physical Page Number Page offset

Cache Tag Cache Index Offset

TRANSLATION
Context Switching Revisited

- What needs to happen when the CPU switches processes?
  - Registers:
    - Save state of old process, load state of new process
    - Including the Page Table Base Register (PTBR)
  - Memory:
    - Nothing to do! Pages for processes already exist in memory/disk and protected from each other
  - TLB:
    - **Invalidate** all entries in TLB – mapping is for old process’ VAs
  - Cache:
    - Can leave alone because storing based on PAs – good for shared data
Summary of Address Translation Symbols

- **Basic Parameters**
  - \( N = 2^n \)  Number of addresses in virtual address space
  - \( M = 2^m \)  Number of addresses in physical address space
  - \( P = 2^p \)  Page size (bytes)

- **Components of the virtual address (VA)**
  - \( \text{VPO} \) Virtual page offset
  - \( \text{VPN} \) Virtual page number
  - \( \text{TLBI} \) TLB index
  - \( \text{TLBT} \) TLB tag

- **Components of the physical address (PA)**
  - \( \text{PPO} \) Physical page offset (same as VPO)
  - \( \text{PPN} \) Physical page number
Simple Memory System Example (small)

- **Addressing**
  - 14-bit virtual addresses \( n = 14 \text{ bits} \) \( \iff \) \( N = 16 \text{ KiB} \) VA space
  - 12-bit physical address \( m = 12 \text{ bits} \) \( \iff \) \( M = 4 \text{ KiB} \) PA space
  - Page size = 64 bytes \( P = 64 \text{ B} \) \( \iff \) \( p = 6 \text{ bits} \)
**Simple Memory System: Page Table**

- Only showing first 16 entries (out of $2^8 = 256$)
  - **Note:** showing 2 hex digits for PPN even though only 6 bits
  - **Note:** other management bits not shown, but part of PTE

<table>
<thead>
<tr>
<th>VPN</th>
<th>PPN</th>
<th>Valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>02</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>–</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VPN</th>
<th>PPN</th>
<th>Valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>0x13</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td>09</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>2D</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>0D</td>
<td>1</td>
</tr>
</tbody>
</table>

$2^{n-p}$
Simple Memory System: TLB

- 16 entries total
- 4-way set associative

![TLB Diagram](image)

<table>
<thead>
<tr>
<th>Set</th>
<th>Tag</th>
<th>PPN</th>
<th>Valid</th>
<th>Tag</th>
<th>PPN</th>
<th>Valid</th>
<th>Tag</th>
<th>PPN</th>
<th>Valid</th>
<th>Tag</th>
<th>PPN</th>
<th>Valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>03</td>
<td>–</td>
<td>0</td>
<td>09</td>
<td>0D</td>
<td>1</td>
<td>00</td>
<td>–</td>
<td>0</td>
<td>07</td>
<td>02</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>03</td>
<td>2D</td>
<td>1</td>
<td>02</td>
<td>–</td>
<td>0</td>
<td>04</td>
<td>–</td>
<td>0</td>
<td>0A</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>02</td>
<td>–</td>
<td>0</td>
<td>08</td>
<td>–</td>
<td>0</td>
<td>06</td>
<td>–</td>
<td>0</td>
<td>03</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>07</td>
<td>–</td>
<td>0</td>
<td>03</td>
<td>0D</td>
<td>1</td>
<td>0A</td>
<td>34</td>
<td>1</td>
<td>02</td>
<td>–</td>
<td>0</td>
</tr>
</tbody>
</table>

Why does the TLB ignore the page offset? (address translation)
Simple Memory System: Cache

- Direct-mapped with $K = 4\ B$, $C/K = 16$
- Physically addressed

Note: It is just coincidence that the PPN is the same width as the cache Tag

```latex
\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|}
\hline
Index & Tag & Valid & B0 & B1 & B2 & B3 \\
\hline
0 & 19 & 1 & 99 & 11 & 23 & 11 \\
1 & 15 & 0 & - & - & - & - \\
2 & 1B & 1 & 00 & 02 & 04 & 08 \\
3 & 36 & 0 & - & - & - & - \\
4 & 32 & 1 & 43 & 6D & 8F & 09 \\
5 & 0D & 1 & 36 & 72 & F0 & 1D \\
6 & 31 & 0 & - & - & - & - \\
7 & 16 & 1 & 11 & C2 & DF & 03 \\
\hline
\end{tabular}
\end{table}
```

```latex
\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Index & Tag & Valid & B0 & B1 & B2 & B3 \\
\hline
8 & 24 & 1 & 3A & 00 & 51 & 89 \\
9 & 2D & 0 & - & - & - & - \\
A & 2D & 1 & 93 & 15 & DA & 3B \\
B & 0B & 0 & - & - & - & - \\
C & 12 & 0 & - & - & - & - \\
D & 16 & 1 & 04 & 96 & 34 & 15 \\
E & 13 & 1 & 83 & 77 & 1B & D3 \\
F & 14 & 0 & - & - & - & - \\
\hline
\end{tabular}
\end{table}
```
Current State of Memory System

Circled #s refer to Memory Request Example #

TLB:

<table>
<thead>
<tr>
<th>Set</th>
<th>Tag</th>
<th>PPN</th>
<th>V</th>
<th>Tag</th>
<th>PPN</th>
<th>V</th>
<th>Tag</th>
<th>PPN</th>
<th>V</th>
<th>Tag</th>
<th>PPN</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>03</td>
<td>–</td>
<td>0</td>
<td>09</td>
<td>0D</td>
<td>1</td>
<td>00</td>
<td>–</td>
<td>0</td>
<td>07</td>
<td>02</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>02</td>
<td>–</td>
<td>0</td>
<td>02</td>
<td>–</td>
<td>0</td>
<td>04</td>
<td>–</td>
<td>0</td>
<td>0A</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>02</td>
<td>2D</td>
<td>1</td>
<td>02</td>
<td>–</td>
<td>0</td>
<td>06</td>
<td>–</td>
<td>0</td>
<td>03</td>
<td>–</td>
<td>0X</td>
</tr>
<tr>
<td>3</td>
<td>07</td>
<td>–</td>
<td>0</td>
<td>03</td>
<td>0D</td>
<td>1</td>
<td>0A</td>
<td>34</td>
<td>1</td>
<td>02</td>
<td>–</td>
<td>0</td>
</tr>
</tbody>
</table>

Cache:

<table>
<thead>
<tr>
<th>Index</th>
<th>Tag</th>
<th>V</th>
<th>B0</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>19</td>
<td>1</td>
<td>99</td>
<td>11</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td>1</td>
<td>15</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>1B</td>
<td>1</td>
<td>00</td>
<td>02</td>
<td>04</td>
<td>08</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>1</td>
<td>43</td>
<td>6D</td>
<td>8F</td>
<td>09</td>
</tr>
<tr>
<td>5</td>
<td>0D</td>
<td>1</td>
<td>36</td>
<td>72</td>
<td>F0</td>
<td>1D</td>
</tr>
<tr>
<td>6</td>
<td>31</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>1</td>
<td>11</td>
<td>C2</td>
<td>DF</td>
<td>03</td>
</tr>
</tbody>
</table>

Page table (partial):

<table>
<thead>
<tr>
<th>VPN</th>
<th>PPN</th>
<th>V</th>
<th>Tag</th>
<th>PPN</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>28</td>
<td>1X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>–</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>33</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>02</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>–</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>–</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>–</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VPN</th>
<th>PPN</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td>0X</td>
</tr>
<tr>
<td>E</td>
<td></td>
<td>0X</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td>0D</td>
</tr>
</tbody>
</table>

Index

<table>
<thead>
<tr>
<th>Tag</th>
<th>V</th>
<th>B0</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>24</td>
<td>1X</td>
<td>3A</td>
<td>00</td>
<td>51</td>
</tr>
<tr>
<td>9</td>
<td>2D</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>A</td>
<td>2D</td>
<td>1X</td>
<td>93</td>
<td>15</td>
<td>DA</td>
</tr>
<tr>
<td>B</td>
<td>0B</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>C</td>
<td>12</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>D</td>
<td>16</td>
<td>1</td>
<td>04</td>
<td>96</td>
<td>34</td>
</tr>
<tr>
<td>E</td>
<td>13</td>
<td>1</td>
<td>83</td>
<td>77</td>
<td>1B</td>
</tr>
<tr>
<td>F</td>
<td>14</td>
<td>0</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Memory Request Example #1

- **Virtual Address:** \(0x03D4\)
  - TLBT: 11011111010100
  - TLBI: 1111011010100
  - VPN: 11111010100
  - VPO: 0000100

- **Physical Address:**
  - CT: 01101010010101
  - CI: 0101010101
  - CO: 0000000
  - PPN: 01010101
  - PPO: 0000000

- **Note:** It is just coincidence that the PPN is the same width as the cache Tag.

- **VPN:** \(0xF\)
  - Check this entry of the page table.

- **TLBT:** \(0x03\)
  - Look for this tag within TLB set.

- **TLBI:** 3
  - Check this set of the TLB.

- **TLB Hit?** Y
- **Page Fault?** N
- **PPN:** \(0x0D\)

- **CT:** \(0x0D\)
  - Look for this tag within cache set.

- **CI:** 5
  - Check this set of the cache.

- **CO:** 0
  - Which byte of cache block.

- **Cache Hit?** Y
- **Data (byte):** \(0x36\)
Memory Request Example #2

- **Virtual Address:** \(0x038E\)

  - TLBT: 13 12 11 10 9 8 7 6 5 4 3 2 1 0
  - TLBI: 0 0 0 0 1 1 1 0 0 0 1 1 1 1

  - VPN: 0x0E
  - TLBT: 0x03
  - TLBI: 2
  - TLB Hit?: N
  - Page Fault?: Y
  - PPN: n/a

- **Physical Address:**

  - CT: 11 10 9 8 7 6 5 4 3 2 1 0
  - CI: _____
  - CO: _____
  - Cache Hit?: ___
  - Data (byte): _______
Memory Request Example #3

Virtual Address: \(0x0020\)

Note: It is just coincidence that the PPN is the same width as the cache Tag

Physical Address:
Memory Request Example #4

- **Virtual Address:** \(0x036B\)

  - TLBT: 13 12 11 10 9 8 7 6 5 4 3 2 1 0
  - VPN: 0 0 0 0 1 1 0 1 1 0 1 0 1 1

  - VPO: 0x0 D
  - TLBT: 0x03
  - TLBI: 1
  - TLB Hit? \(\checkmark\)
  - Page Fault? \(\checkmark\)
  - PPN: 0x2D

- **Physical Address:**

  - CT: 11 10 9 8 7 6 5 4 3 2 1 0
  - PPN: 0x2D
  - CI: A
  - CO: 3
  - Cache Hit? \(\checkmark\)
  - Data (byte): 0x3B

**Note:** It is just coincidence that the PPN is the same width as the cache Tag.
Memory Overview

- `movl 0x8043ab, %rdi`
Page Table Reality

- Just one issue... the numbers don’t work out for the story so far!

- The problem is the page table for each process:
  - Suppose 64-bit VAs, 8 KiB pages, 8 GiB physical memory
  - How many page table entries is that?
    - 1 PTE for every virtual page
    - \(2^{n-p} = 2^{51}\) PTEs
    - \(n = 64\) bits, \(p = 13\) bits, \(m = 33\) bits
  - About how long is each PTE?
    - PPN width + management bits = 20 + 5 = 25 bits \(\approx 3\) bytes
    - \(m-p = (v,d,r,w,x)\)

- **Moral:** Cannot use this naïve implementation of the virtual → physical page mapping – it’s way too big
A Solution: Multi-level Page Tables

This is called a page walk

Virtual Address

Level 1 page table

Level 2 page table

Level k page table

TLB

VPN → PTE

VPN → PTE

VPN → PTE

Physical Address

Page table base register (PTBR)

VPN: 0 0 1 1 0 1 0 1

This is extra (non-testable) material

8 virtual pages: split VPN into 1-bit fields

arrays of pointers

n-1

p-1

m-1

0
Multi-level Page Tables

- A tree of depth $k$ where each node at depth $i$ has up to $2^j$ children if part $i$ of the VPN has $j$ bits
- Hardware for multi-level page tables inherently more complicated
  - But it’s a necessary complexity – 1-level does not fit
- Why it works: Most subtrees are not used at all, so they are never created and definitely aren’t in physical memory
  - Parts created can be evicted from cache/memory when not being used
  - Each node can have a size of ~1-100KB
- But now for a $k$-level page table, a TLB miss requires $k + 1$ cache/memory accesses
  - Fine so long as TLB misses are rare – motivates larger TLBs

This is extra (non-testable) material
Practice VM Question

- Our system has the following properties
  - 1 MiB of physical address space
  - 4 GiB of virtual address space
  - 32 KiB page size
  - 4-entry fully associative TLB with LRU replacement

a) Fill in the following blanks:

- \( 2^{17} \) Entries in a page table
- \( 2^{n-p} \) \( \leq \) # of virtual pages
- \( 17 \) TLBT bits
- \( 2^5 \) Max # of valid entries in a page table
- Min bit-width of PTBR
- Physical address of PT
- \( 2^{m-p} \) # of pages in physical memory
Practice VM Question

- One process uses a page-aligned square matrix `mat[]` of 32-bit integers in the code shown below:

  ```
  #define MAT_SIZE = 2048
  for(int i = 0; i < MAT_SIZE; i++)
    mat[i*(MAT_SIZE+1)] = i;
  ```

b) What is the largest stride (in bytes) between successive memory accesses (in the VA space)?

```
array index accessed
\hline
0 & \hline
0
2049
2*2049
:  
```

The stride is always 2049 ints, which is equal to `2049*4` bytes.
Practice VM Question

One process uses a page-aligned square matrix `mat[]` of 32-bit integers in the code shown below:

```c
#define MAT_SIZE = 2048
for(int i = 0; i < MAT_SIZE; i++)
    mat[i*(MAT_SIZE+1)] = i;
```

c) Assuming all of `mat[]` starts on disk, what are the following hit rates for the execution of the for-loop?

- 3/4 = 75% TLB Hit Rate
- 0% Page Table Hit Rate

**access pattern:** single write to index never revisit indices (always increasing) we access every row of matrix exactly once

each page holds \(2^{15}/2^{13} = 4\) rows of matrix

within each page: M T T H H

page size = 32 KiB = \(2^{15}\) B

\(\text{ints} = 2^{13}\) B
Virtual Memory Summary

- Programmer’s view of virtual memory
  - Each process has its own private linear address space
  - Cannot be corrupted by other processes

- System view of virtual memory
  - Uses memory efficiently by caching virtual memory pages
    - Efficient only because of locality
  - Simplifies memory management and sharing
  - Simplifies protection by providing permissions checking
Memory System Summary

- **Memory Caches (L1/L2/L3)**
  - Purely a speed-up technique
  - Behavior invisible to application programmer and (mostly) OS
  - Implemented totally in hardware

- **Virtual Memory**
  - Supports many OS-related functions
    - Process creation, task switching, protection
  - Operating System (software)
    - Allocates/shares physical memory among processes
    - Maintains high-level tables tracking memory type, source, sharing
    - Handles exceptions, fills in hardware-defined mapping tables
  - Hardware
    - Translates virtual addresses via mapping tables, enforcing permissions
    - Accelerates mapping via translation cache (TLB)