Virtual Memory: Systems

Instructors:
Alvin R. Lebeck

Slides provided by Randy Bryant and Dave O’Hallaron

Administrivia

- Work on Y86 simulator (no late submissions)
- HW #7 up, Due Monday April 25 (no late submissions)
- After VM, one main topic: storage & I/O
- One class on advanced topics
Today

- Review: VM as a tool for caching
- Review: VM as a tool for memory management
- Review: VM as a tool for memory protection
- Address translation
- Simple memory system example
- Case study: Core i7/Linux memory system

DRAM Cache Organization

- DRAM cache organization driven by the enormous miss penalty
  - DRAM is about \(10x\) slower than SRAM
  - Disk is about \(10,000x\) slower than DRAM

- Consequences
  - Large page (block) size: typically 4-8 KB, sometimes 4 MB
  - Fully associative
    - Any VP can be placed in any PP
    - Requires a “large” mapping function – different from CPU caches
  - Highly sophisticated, expensive replacement algorithms
    - Too complicated and open-ended to be implemented in hardware
  - Write-back rather than write-through
Page Tables

- A page table is an array of page table entries (PTEs) that maps virtual pages to physical pages.
  - Per-process kernel data structure in DRAM

Page Hit

- Page hit: reference to VM word that is in physical memory (DRAM cache hit)
**Page Fault**

- **Page fault**: reference to VM word that is not in physical memory (DRAM cache miss)

**Handling Page Fault**

- Page miss causes page fault (an exception)
Handling Page Fault

- Page miss causes page fault (an exception)
- Page fault handler selects a victim to be evicted (here VP 4)
Handling Page Fault

- Page miss causes page fault (an exception)
- Page fault handler selects a victim to be evicted (here VP 4)
- Offending instruction is restarted: page hit!

Locality to the Rescue Again!

- Virtual memory works because of locality

- At any point in time, programs tend to access a set of active virtual pages called the **working set**
  - Programs with better temporal locality will have smaller working sets

- If (working set size < main memory size)
  - Good performance for one process after compulsory misses

- If ( SUM(working set sizes) > main memory size )
  - **Thrashing:** Performance meltdown where pages are swapped (copied) in and out continuously
Today

- Review: VM as a tool for caching
- **Review: VM as a tool for memory management**
- Review: VM as a tool for memory protection
- Review: Address translation
- Core i7 case study
- Simple memory system example
- Case study: Core i7/Linux memory system

---

**VM as a Tool for Memory Management**

- **Key idea:** each process has its own virtual address space
  - It can view memory as a simple linear array
  - Mapping function scatters addresses through physical memory
    - Well chosen mappings simplify memory allocation and management

![Virtual Address Space Diagram](image)
VM as a Tool for Memory Management

- **Memory allocation**
  - Each virtual page can be mapped to any physical page
  - A virtual page can be stored in different physical pages at different times

- **Sharing code and data among processes**
  - Map virtual pages to the same physical page (here: PP 6)

Simplifying Linking and Loading

- **Linking**
  - Each program has similar virtual address space
  - Code, stack, and shared libraries always start at the same address

- **Loading**
  - `execve()` allocates virtual pages for `.text` and `.data` sections = creates PTEs marked as invalid
  - The `.text` and `.data` sections are copied, page by page, on demand by the virtual memory system
Today

- Review: VM as a tool for caching
- Review: VM as a tool for memory management
- Review: VM as a tool for memory protection
- Address translation
  - TLBs
  - VM & caches
- Core i7 case study
- Simple memory system example
- Case study: Core i7/Linux memory system

VM as a Tool for Memory Protection

- Extend PTEs with permission bits
- Page fault handler checks these before remapping
  - If violated, send process SIGSEGV (segmentation fault)
VM Address Translation

- Virtual Address Space
  - $V = \{0, 1, ..., N-1\}$

- Physical Address Space
  - $P = \{0, 1, ..., M-1\}$

- Address Translation
  - $MAP: V \rightarrow P \cup \{\emptyset\}$
  - For virtual address $a$:
    - $MAP(a) = a'$ if data at virtual address $a$ is at physical address $a'$ in $P$
    - $MAP(a) = \emptyset$ if data at virtual address $a$ is not in physical memory
      - Either invalid or stored on disk

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)
  - TLBI: TLB index
  - TLBT: TLB tag
  - VPO: Virtual page offset
  - VPN: Virtual page number

- Components of the physical address (PA)
  - PPO: Physical page offset (same as VPO)
  - PPN: Physical page number
  - CO: Byte offset within cache line
  - CI: Cache index
  - CT: Cache tag
Address Translation With a Page Table

Address Translation: Page Hit

1) Processor sends virtual address to MMU
2-3) MMU fetches PTE from page table in memory
4) MMU sends physical address to cache/memory
5) Cache/memory sends data word to processor
**Address Translation: Page Fault**

1) Processor sends virtual address to MMU
2-3) MMU fetches PTE from page table in 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

---

**Integrating VM and Cache**

*VA: virtual address, PA: physical address, PTE: page table entry, PTEA = PTE address*
Speeding up Translation with a TLB

- Page table entries (PTEs) are cached in L1 like any other memory word
  - PTEs may be evicted by other data references
  - PTE hit still requires a small L1 delay

- **Solution:** *Translation Lookaside Buffer (TLB)*
  - Small hardware cache in MMU
  - Maps virtual page numbers to physical page numbers
  - Contains complete page table entries for small number of pages

---

**TLB Hit**

A TLB hit eliminates a memory access
TLB Miss

A TLB miss incurs an additional memory access (the PTE)
Fortunately, TLB misses are rare. Why?

Multi-Level Page Tables

- **Suppose:**
  - 4KB ($2^{12}$) page size, 48-bit address space, 8-byte PTE

- **Problem:**
  - Would need a 512 GB page table!
    - $2^{48} \times 2^{12} \times 2^{3} = 2^{39}$ bytes

- **Common solution:**
  - Multi-level page tables
  - Example: 2-level page table
    - Level 1 table: each PTE points to a page table (always memory resident)
    - Level 2 table: each PTE points to a page (paged in and out like any other data)
A Two-Level Page Table Hierarchy

Level 1 page table
- PTE 0
- PTE 1
- PTE 2 (null)
- PTE 3 (null)
- PTE 4 (null)
- PTE 5 (null)
- PTE 6 (null)
- PTE 7 (null)
- PTE 8
  - (1K - 9) null PTEs
- 1023 null PTEs
- PTE 1023

Level 2 page tables
- PTE 0
- ... (null)
- PTE 1023
- ... (null)
- PTE 0
- ... (null)
- PTE 1023

Virtual memory
- VP 0
- VP 1023
- VP 2047
- Gap
- 1023 unallocated pages
- VP 9215
- 6K unallocated VM pages
- 2K allocated VM pages for code and data
- 1023 unallocated pages
- 1 allocated VM page for the stack
- 1023 null PTEs

32 bit addresses, 4KB pages, 4-byte PTEs

Summary, thus far

- **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 programming
  - Simplifies protection by providing a convenient interpositioning point to check permissions
Review of 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)**
  - TLBI: TLB index
  - TLBT: TLB tag
  - VPO: Virtual page offset
  - VPN: Virtual page number

- **Components of the physical address (PA)**
  - PPO: Physical page offset (same as VPO)
  - PPN: Physical page number
  - CO: Byte offset within cache line
  - CI: Cache index
  - CT: Cache tag

Simple Memory System Example

- **Addressing**
  - 14-bit virtual addresses
  - 12-bit physical address
  - Page size = 64 bytes
Simple Memory System Page Table

Only show first 16 entries (out of 256)

<table>
<thead>
<tr>
<th>VPN</th>
<th>PPN</th>
<th>Valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>01</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>02</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>03</td>
<td>02</td>
<td>1</td>
</tr>
<tr>
<td>04</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>05</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>06</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>07</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>08</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>09</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>0A</td>
<td>09</td>
<td>1</td>
</tr>
<tr>
<td>0B</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>0C</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>0D</td>
<td>2D</td>
<td>1</td>
</tr>
<tr>
<td>0E</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>0F</td>
<td>0D</td>
<td>1</td>
</tr>
</tbody>
</table>

Simple Memory System TLB

- 16 entries
- 4-way associative

<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>
</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>
</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>
</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>
</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>
</tr>
</tbody>
</table>

Simple Memory System Cache

- 16 lines, 4-byte block size
- Physically addressed
- Direct mapped

Address Translation Example #1

Virtual Address: 0x03D4

Physical Address
Address Translation Example #2

Virtual Address: 0x0B8F

Physical Address

Address Translation Example #3

Virtual Address: 0x0020

Physical Address
Today

- Review: VM as a tool for caching
- Review: VM as a tool for memory management
- Review: VM as a tool for memory protection
- Review: Address translation
- Case study: Core i7/Linux memory system

Intel Core i7 Memory System
Review of 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)**
  - TLBI: TLB index
  - TLBT: TLB tag
  - VPO: Virtual page offset
  - VPN: Virtual page number

- **Components of the physical address (PA)**
  - PPO: Physical page offset (same as VPO)
  - PPN: Physical page number
  - CO: Byte offset within cache line
  - CI: Cache index
  - CT: Cache tag

---

### End-to-end Core i7 Address Translation

- CPU
- CR3
- VPN
- TLB
- PTE
- Page tables
- L1 d-cache
- L2, L3, and main memory
- Result
- Physical address (PA)

<table>
<thead>
<tr>
<th>CR3</th>
<th>PTE</th>
<th>Page tables</th>
<th>L1 d-cache</th>
<th>L2, L3, and main memory</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>40</td>
<td>12</td>
<td>40, 6, 6</td>
<td>32/64</td>
<td></td>
</tr>
</tbody>
</table>
### Core i7 Level 1-3 Page Table Entries

<table>
<thead>
<tr>
<th>63</th>
<th>62</th>
<th>52</th>
<th>51</th>
<th>12</th>
<th>11</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>XD</td>
<td>Unused</td>
<td>Page physical base address</td>
<td>Unused</td>
<td>G</td>
<td>PS</td>
<td>A</td>
<td>CD</td>
<td>WT</td>
<td>U/S</td>
<td>R/W</td>
<td>P=1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Available for OS (page table location on disk)  

P=0

Each entry references a 4K child page table

- **P:** Child page table present in physical memory (1) or not (0).
- **R/W:** Read-only or read-write access access permission for all reachable pages.
- **U/S:** User or supervisor (kernel) mode access permission for all reachable pages.
- **WT:** Write-through or write-back cache policy for the child page table.
- **CD:** Caching disabled (0) or enabled (1) for the child page table.
- **A:** Reference bit (set by MMU on reads and writes, cleared by software).
- **PS:** Page size either 4 KB or 4 MB (defined for Level 1 PTEs only).
- **G:** Global page (don’t evict from TLB on task switch)

**Page table physical base address:** 40 most significant bits of physical page table address (forces page tables to be 4KB aligned)

### Core i7 Level 4 Page Table Entries

<table>
<thead>
<tr>
<th>63</th>
<th>62</th>
<th>52</th>
<th>51</th>
<th>12</th>
<th>11</th>
<th>9</th>
<th>8</th>
<th>7</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>XD</td>
<td>Unused</td>
<td>Page physical base address</td>
<td>Unused</td>
<td>G</td>
<td>D</td>
<td>A</td>
<td>CD</td>
<td>WT</td>
<td>U/S</td>
<td>R/W</td>
<td>P=1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Available for OS (page location on disk)  

P=0

Each entry references a 4K child page

- **P:** Child page is present in memory (1) or not (0)
- **R/W:** Read-only or read-write access access permission for child page
- **U/S:** User or supervisor mode access
- **WT:** Write-through or write-back cache policy for this page
- **CD:** Cache disabled (1) or enabled (0)
- **A:** Reference bit (set by MMU on reads and writes, cleared by software)
- **D:** Dirty bit (set by MMU on writes, cleared by software)
- **G:** Global page (don’t evict from TLB on task switch)

**Page physical base address:** 40 most significant bits of physical page address (forces pages to be 4KB aligned)
Core i7 Page Table Translation

Cute Trick for Speeding Up L1 Access

- **Observation**
  - Bits that determine CI identical in virtual and physical address
  - Can index into cache while address translation taking place
  - Generally we hit in TLB, so PPN bits (CT bits) available next
  - “Virtually indexed, physically tagged”
  - Cache carefully sized to make this possible
Virtual Memory of a Linux Process

- **Process-specific data structs** (ptables, task and mm structs, kernel stack)
- **Kernel code and data**
- **User stack**
- **Memory mapped region for shared libraries**
- **Runtime heap (malloc)**
- **Initialized data (.data)**
- **Uninitialized data (.bss)**
- **Program text (.text)**

Linux Organizes VM as Collection of “Areas”

- **pgd**: Page global directory address
- **vm_prot**: Read/write permissions for this area
- **vm_flags**: Pages shared with other processes or private to this process
Linux Page Fault Handling

Virtualization (VMware)

- Boot and run multiple operating systems
  - Windows, macOS, Linux all at once
  - Not dual booting!
- Requires special support for efficiency
  - New privilege mode for hypervisor