Virtual Memory
October 14, 2008

Topics
- Address spaces
- Motivations for virtual memory
- Address translation
- Accelerating translation with TLBs
Announcements

**Autolab outage**
- The autolab machine was hacked on Saturday
  - not the autolab programs, but the underlying OS
  - rebuilt and brought back online Monday
- Should not block your progress on shelllab
  - all files needed were made available on class website (under docs link)
  - fish machines are working fine
- Re-submit tshlab, if you finished before the outage
  - as always, there was a time gap between last backup and the breakin
Byte-Oriented Memory Organization

- Programs Refer to Virtual Memory Addresses
  - Conceptually very large array of bytes
  - Actually implemented with hierarchy of different memory types
  - System provides address space private to particular “process”
    - Program being executed
    - Program can clobber its own data, but not that of others

- Compiler + Run-Time System Control Allocation
  - Where different program objects should be stored
  - All allocation within single virtual address space
Simple Addressing Modes

- **Normal**  (R)  Mem[Reg[R]]
  - Register R specifies memory address

  ```
  movl (%ecx),%eax
  ```

- **Displacement**  D(R)  Mem[Reg[R]+D]
  - Register R specifies start of memory region
  - Constant displacement D specifies offset

  ```
  movl 8(%ebp),%edx
  ```
How does everything fit?
- 32-bit addresses: ~4,000,000,000 (4 billion) bytes
- 64-bit addresses: ~16,000,000,000,000,000,000 (16 quintillion) bytes

How to decide which memory to use in your program?
- How about after a fork()?

What if another process stores data into your memory?
- How could you debug your program?
So, we add a level of indirection

One simple trick solves all three problems

- Each process gets its own private image of memory
  - appears to be a full-sized private memory range
- This fixes “how to choose” and “others shouldn’t mess w/yours”
  - surprisingly, it also fixes “making everything fit”
- Implementation: translate addresses transparently
  - add a mapping function
    - to map private addresses to physical addresses
    - do the mapping on every load or store

This mapping trick is the heart of *virtual memory*
Address Spaces

A *linear address space* is an ordered set of contiguous nonnegative integer addresses:

\{0, 1, 2, 3, \ldots \}\]

A *virtual address space* is a set of \( N = 2^n \) *virtual addresses*:

\{0, 1, 2, \ldots, N-1\}

A *physical address space* is a set of \( M = 2^m \) (for convenience) *physical addresses*:

\{0, 1, 2, \ldots, M-1\}

In a system based on virtual addressing, each byte of main memory has a physical address and a virtual address (or more)
A System Using Physical Addressing

Used by many embedded microcontrollers in devices like cars, elevators, and digital picture frames
A System Using Virtual Addressing

One of the great ideas in computer science

- used by all modern desktop and laptop microprocessors
Why Virtual Memory?

(1) VM allows efficient use of limited main memory (RAM)
   - Use RAM as a cache for the parts of a virtual address space
     - some non-cached parts stored on disk
     - some (unallocated) non-cached parts stored nowhere
     - Keep only active areas of virtual address space in memory
       - transfer data back and forth as needed

(2) VM simplifies memory management for programmers
   - Each process gets a full, private linear address space

(3) VM isolates address spaces
   - One process can’t interfere with another’s memory
     - because they operate in different address spaces
   - User process cannot access privileged information
     - different sections of address spaces have different permissions
(1) VM as a Tool for Caching

**Virtual memory** is an array of N contiguous bytes
- think of the array as being stored on disk

The contents of the array on disk are cached in **physical memory (DRAM cache)**

![Diagram showing virtual memory and physical memory relationship](image)
DRAM Cache Organization

DRAM cache organization driven by the enormous miss penalty

- DRAM is about 10x slower than SRAM
- Disk is about 100,000x slower than a DRAM
  - to get first byte, though fast for next byte

DRAM cache properties

- Large page (block) size (typically 4-8 KB)
- Fully associative
  - Any virtual page can be placed in any physical page
  - Requires a “large” mapping function – different from CPU caches
- Highly sophisticated replacement algorithms
  - Too complicated and open-ended to be implemented in hardware
- Write-back rather than write-through
Reminder: MMU checks the cache

One of the great ideas in computer science

- used by all modern desktop and laptop microprocessors
How? 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

![Diagram of page table and memory mapping](image)
Address Translation with a Page Table

Virtual page number (VPN) acts as index into the page table.

If valid=0 then page not in memory (page fault).

Page table base register (PTBR)

Virtual page number (VPN) → Virtual page offset (VPO) → Physical page number (PPN) → Physical page offset (PPO) → Physical address

The VPN acts as index into the page table.
Page Hits

A page hit is a reference to a VM word that is in physical (main) memory
Page Faults

A page fault is caused by a reference to a VM word that is not in physical (main) memory.

- Example: An instruction references a word contained in VP 3, a miss that triggers a page fault exception.

Diagram:
- Virtual address
- Physical page number or disk address
- Valid
- Memory resident page table (DRAM)
- Physical memory (DRAM)
- Virtual memory (disk)
Handling a Page Fault

The kernel’s page fault handler selects VP 4 as the victim and replaces it with a copy of VP 3 from disk (demand paging).

- When the offending instruction restarts, it executes normally, without generating an exception.

---

The diagram shows the page table entries (PTEs) and the mapping between virtual and physical addresses. PTE 0 and PTE 7 are shown with their valid bits and physical page numbers or disk addresses.
Why does it work? Locality

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
(2) VM as a Tool for Memory Mgmt

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

![Diagram of virtual address space for processes 1 and 2, with address translation and physical address space (DRAM) for read-only library code)]
Simplifying Sharing and Allocation

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 – the program never knows

Sharing code and data among processes
- Map virtual pages to the same physical page (PP 7)

Virtual Address Space for Process 1:

Virtual Address Space for Process 2:

Address Translation

Physical Address Space (DRAM)
(e.g., read-only library code)
**IA32 Linux Memory Layout**

- **Stack**
  - Runtime stack (8MB limit)

- **Heap**
  - Dynamically allocated storage
  - When call `malloc()`, `calloc()`, `new()`

- **Data**
  - Statically allocated data
  - E.g., arrays & strings declared in code

- **Text**
  - Executable machine instructions
  - Read-only
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()` maps PTEs to the appropriate location in the executable binary file
- The `.text` and `.data` sections are copied, page by page, on demand by the virtual memory system.
(3) 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)

Page tables with permission bits

<table>
<thead>
<tr>
<th>VP 0:</th>
<th>SUP</th>
<th>READ</th>
<th>WRITE</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>PP 6</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>PP 4</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>PP 2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Process i:</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP 0:</td>
</tr>
<tr>
<td>VP 1:</td>
</tr>
<tr>
<td>VP 2:</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VP 0:</th>
<th>SUP</th>
<th>READ</th>
<th>WRITE</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>PP 9</td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>PP 6</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>PP 11</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Process j:</th>
</tr>
</thead>
<tbody>
<tr>
<td>VP 0:</td>
</tr>
<tr>
<td>VP 1:</td>
</tr>
<tr>
<td>VP 2:</td>
</tr>
</tbody>
</table>

Physical memory

- PP 0
- PP 2
- PP 4
- PP 6
- PP 9
- PP 11
Reminder: MMU checks the cache

One of the great ideas in computer science

- used by all modern desktop and laptop microprocessors
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
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
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 1-cycle 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
A TLB miss incurs an add’l memory access (the PTE)

- Fortunately, TLB misses are rare
Simple Memory System Example

Addressing

- 14-bit virtual addresses
- 12-bit physical address
- Page size = 64 bytes

```
  13  12  11  10  9   8   7   6   5   4   3   2   1   0

VPN                                VPO
  (Virtual Page Number)            (Virtual Page Offset)

  11  10  9   8   7   6   5   4   3   2   1   0

PPN                                PPO
  (Physical Page Number)           (Physical Page Offset)
```
## 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>
<th>VPN</th>
<th>PPN</th>
<th>Valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>28</td>
<td>1</td>
<td>08</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>01</td>
<td>–</td>
<td>0</td>
<td>09</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>02</td>
<td>33</td>
<td>1</td>
<td>0A</td>
<td>09</td>
<td>1</td>
</tr>
<tr>
<td>03</td>
<td>02</td>
<td>1</td>
<td>0B</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>04</td>
<td>–</td>
<td>0</td>
<td>0C</td>
<td>–</td>
<td>0</td>
</tr>
<tr>
<td>05</td>
<td>16</td>
<td>1</td>
<td>0D</td>
<td>2D</td>
<td>1</td>
</tr>
<tr>
<td>06</td>
<td>–</td>
<td>0</td>
<td>0E</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>07</td>
<td>–</td>
<td>0</td>
<td>0F</td>
<td>0D</td>
<td>1</td>
</tr>
</tbody>
</table>
Simple Memory System TLB

TLB

- 16 entries
- 4-way associative

### TLBT

### TLBI

### VPN

### VPO

<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>
Simple Memory System Cache

**Cache**
- 16 lines
- 4-byte line size
- Direct mapped

![Cache Diagram]

<table>
<thead>
<tr>
<th>Idx</th>
<th>Tag</th>
<th>Valid</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>

<table>
<thead>
<tr>
<th>Idx</th>
<th>Tag</th>
<th>Valid</th>
<th>B0</th>
<th>B1</th>
<th>B2</th>
<th>B3</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>24</td>
<td>1</td>
<td>3A</td>
<td>00</td>
<td>51</td>
<td>89</td>
</tr>
<tr>
<td>9</td>
<td>2D</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>A</td>
<td>2D</td>
<td>1</td>
<td>93</td>
<td>15</td>
<td>DA</td>
<td>3B</td>
</tr>
<tr>
<td>B</td>
<td>0B</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>C</td>
<td>12</td>
<td>0</td>
<td>—</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>
<td>15</td>
</tr>
<tr>
<td>E</td>
<td>13</td>
<td>1</td>
<td>83</td>
<td>77</td>
<td>1B</td>
<td>D3</td>
</tr>
<tr>
<td>F</td>
<td>14</td>
<td>0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
Address Translation Example
#1

Virtual Address 0x03D4

<table>
<thead>
<tr>
<th>13</th>
<th>12</th>
<th>11</th>
<th>10</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>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

VPN 0xOF  TLBI 3  TLBT 0x03  TLB Hit? Y  Page Fault? NO  PPN  0x0D

Physical Address

| 11 | 10 | 9  | 8  | 7  | 6  | 5  | 4  | 3  | 2  | 1  | 0  |
|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 0  | 0  | 1  | 1  | 0  | 1  | 0  | 1  | 0  | 1  | 0  | 0  |

Offset 0  CI 0x5  CT 0x0D  Hit? Y  Byte 0x36
Address Translation Example #2

Virtual Address 0x0B8F

Virtual Address

<table>
<thead>
<tr>
<th>13</th>
<th>12</th>
<th>11</th>
<th>10</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>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

VPN ___ TLBI ___ TLBT ____ TLB Hit? __ Page Fault? __ PPN: ____

VPN: 0x2E TLBI: 2 TLBT: 0x0B TLB Hit?: NO Page Fault?: YES PPN: TBD

Physical Address

Physical Address

<table>
<thead>
<tr>
<th>11</th>
<th>10</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>
</table>

CT ___ CI ___ CO ___

Offset ___ CI___ CT ____ Hit? __ Byte: ____

PPN ___ PPO ___
Address Translation Example

#3

Virtual Address 0x0020

<table>
<thead>
<tr>
<th>Virtual Address</th>
<th>TLBT</th>
<th>TLBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>13 12 11 10 9 8 7 6 5 4 3 2 1 0</td>
<td>0 0 0 0 0 0 0 0 1 0 0 0 0 0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Virtual Address</th>
<th>VPN</th>
<th>VPO</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPN 0x00</td>
<td>TLBI 0</td>
<td>TLBT 0x00</td>
</tr>
<tr>
<td>TLB Hit? NO</td>
<td>Page Fault? NO</td>
<td>PPN 0x28</td>
</tr>
</tbody>
</table>

Physical Address

<table>
<thead>
<tr>
<th>Physical Address</th>
<th>CT</th>
<th>CI</th>
<th>CO</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 10 9 8 7 6 5 4 3 2 1 0</td>
<td>1 0 1 0 0 0 1 0 0 0 0 0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Physical Address</th>
<th>PPN</th>
<th>PPO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offset 0</td>
<td>CI 0x8</td>
<td>CT 0x28</td>
</tr>
<tr>
<td>Hit? NO</td>
<td>Byte: MEM</td>
<td></td>
</tr>
</tbody>
</table>
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 programming
- Simplifies protection by providing a convenient interpositioning point to check permissions
Allocating Virtual Pages

Example: Allocating new virtual page VP5

- Kernel allocates VP 5 on disk and points PTE 5 to it

Physical page number or disk address

Valid | PTE 0 | PTE 7
-------|-------|-------
0      | null  |
1      |
1      |
1      |
0      |
0      |
0      |
1      |

Physical memory (DRAM)

- PP 0
  - VP 1
  - VP 2
  - VP 7

- PP 3
  - VP 3

Virtual memory (disk)

- VP 1
- VP 2
- VP 3
- VP 4
- VP 5
- VP 6
- VP 7

Memory resident page table (DRAM)
Multi-Level Page Tables

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

Problem:
- Would need a 256 GB page table!
  - $2^{48} \times 2^{-12} \times 2^2 = 2^{38}$ bytes

Common solution
- Multi-level page tables
- Example: 2-level page table
  - Level 1 table: each PTE points to a page table (memory resident)
  - Level 2 table: Each PTE points to a page (paged in and out like other data)

- Level 1 table stays in memory
- Level 2 tables paged in and out
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

Level 2 page tables

Level 2 page tables

PTE 0
...
PTE 1023

Virtual memory

VP 0
...
VP 1023
VP 1024
...
VP 2047
Gap

2K allocated VM pages for code and data

6K unallocated VM pages

1023 unallocated pages

1023 unallocated pages

VP 9215

1 allocated VM page for the stack
Translating with a k-level Page Table

VIRTUAL ADDRESS

PHYSICAL ADDRESS

VPN 1  VPN 2  ...  VPN k  VPO

Level 1 page table

Level 2 page table

Level k page table

PPN

PPN

PPO
Servicing a Page Fault

(1) Processor signals disk controller
   - Read block of length \( P \) starting at disk address \( X \) and store starting at memory address \( Y \)

(2) Read occurs
   - Direct Memory Access (DMA)
   - Under control of I/O controller

(3) Controller signals completion
   - Interrupts processor
   - OS resumes suspended process