Virtual Memory

CS 351: Systems Programming
Michael Saelee <lee@iit.edu>
previously: SRAM ⇔ DRAM
next: DRAM ⇔ HDD, SSD, etc.
i.e., memory as a “cache” for disk
main goals:

1. maximize memory *throughput*
2. maximize memory *utilization*
3. provide *address space consistency* & *memory protection* to processes
\[ \text{throughput} = \# \text{ bytes per second} \]
- depends on access latencies (DRAM, HDD) and “hit rate”
utilization = fraction of allocated memory that contains “user” data (aka payload)

- vs. metadata and other overhead required for memory management
address space consistency → provide a uniform “view” of memory to each process
memory protection → prevent processes from directly accessing each other’s memory
“memory addresses”: what are they, really?
“physical” address: (byte) index into DRAM

CPU

address: \( N \)

Main Memory

\( N \)

data

(note: cache not shown)
```c
int glob = 0xDEADBEEE;

main() {
    fork();
    glob += 1;
}
```

(gdb) set detach-on-fork off
(gdb) break main
Breakpoint 1 at 0x400508: file memtest.c, line 7.
(gdb) run
Breakpoint 1, main () at memtest.c:7
7    fork();
(gdb) next
[New process 7450]
8    glob += 1;
(gdb) print &glob
$1 = (int *) 0x6008d4
(gdb) next
9 }
(gdb) print /x glob
$2 = 0xdeadbeef
(gdb) inferior 2
[Switching to inferior 2 [process 7450]
#0  0x000000310acac49d in __libc_fork ()
131       pid = ARCH_FORK ();
(gdb) finish
Run till exit from #0 in __libc_fork ()
8    glob += 1;
(gdb) print /x glob
$4 = 0xdeadbeee
(gdb) print &glob
$5 = (int *) 0x6008d4
```

**parent**

**child**
instructions executed by the CPU do not refer directly to *physical* addresses!
processes reference *virtual* addresses, the CPU relays virtual address requests to the *memory management unit* (MMU), which are *translated* to physical addresses
CPU → virtual address

MMU

address translation unit

physical address

disk address

Main Memory

“swap” space

(note: cache not shown)
essential problem: translate request for a virtual address → physical address

... this must be **FAST**, as *every* memory access from the CPU must be translated
both hardware/software are involved:

- **MMU (hw)** handles simple and fast operations (e.g., table lookups)

- **Kernel (sw)** handles complex tasks (e.g., eviction policy)
§ Virtual Memory Implementations
keep in mind goals:

1. maximize memory *throughput*
2. maximize memory *utilization*
3. provide *address space consistency* & *memory protection* to processes
1. simple relocation

- per-process relocation address is loaded by kernel on every context switch
1. simple relocation

- incorporate a *limit* register to provide memory protection
pros:
- simple & fast!
- provides protection
but: available memory for mapping depends on value of base address

i.e., address spaces are *not consistent!*
also: all of a process *below the address limit* must be loaded in memory

i.e., memory may be *vastly under-utilized*
2. segmentation

- partition virtual address space into multiple logical segments

- individually map them onto physical memory with relocation registers
virtual address has form $\text{seg#} : \text{offset}$
assert (offset ≤ L_2)

CPU

VA: seg#:offset

PA: offset + B_2

Main Memory

data

MMU

Segment Table

<table>
<thead>
<tr>
<th>Base</th>
<th>Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_0</td>
<td>L_0</td>
</tr>
<tr>
<td>B_1</td>
<td>L_1</td>
</tr>
<tr>
<td>B_2</td>
<td>L_2</td>
</tr>
<tr>
<td>B_3</td>
<td>L_3</td>
</tr>
</tbody>
</table>
- implemented as MMU registers
- part of kernel-maintained, per-process metadata (aka “process control block”)
- re-populated on each context switch
pros:

- still very fast
  - translation = register access & addition
- memory protection via limits
- segmented addresses improve consistency
simple relocation:

segmentation:
- variable segment sizes → memory fragmentation
- fragmentation potentially lowers utilization
- can fix through compaction, but expensive!
3. paging

- partition virtual and physical address spaces into *uniformly sized* pages

- only map pages onto physical memory that contain required data
- pages boundaries are *not aligned to segments*!

- simply aligned to multiples of page size
- minimum mapping granularity = page
- not all of a given segment need be mapped
new mapping problem:

- break a virtual address down into
  virtual page number & virtual page offset

- map VPN $\rightarrow$ physical page number
Given page size $= 2^p$ bytes

VA:

\[
\begin{array}{|c|c|}
\hline
\text{virtual page number} & \text{virtual page offset} \\
\hline
\end{array}
\]

PA:

\[
\begin{array}{|c|c|}
\hline
\text{physical page number} & \text{physical page offset} \\
\hline
\end{array}
\]
VA: 

| virtual page number | virtual page offset |

PA: 

| physical page number | physical page offset |

address translation
translation structure: page table

if invalid, page is not mapped
page table entries (PTEs) typically contain additional metadata, e.g.:

- dirty (modified) bit
- access bits (shared or kernel-owned pages may be read-only or inaccessible)
e.g., 32-bit virtual address, 
4KB ($2^{12}$) pages, 
4-byte PTEs; 

- size of page table?
e.g., 32-bit virtual address, 4KB ($2^{12}$) pages, 4-byte PTEs;

- # pages = $2^{(32-12)} = 2^{20} = 1\text{M}$

- page table size = $1\text{M} \times 4$ bytes = $4\text{MB}$
4MB is much too large to fit in the MMU — insufficient registers and SRAM!

Page table resides in **main memory**
The translation process (aka *page table walk*) is performed by hardware (MMU).

The kernel must initially populate, then continue to manage a process’s page table.

The kernel also populates a *page table base register* on context switches.
translation: hit
translation: **miss**

1. VA: $N$
2. page table walk
3. page fault
4. transfer control to kernel
5. data transfer
6. PTE update
7. VA: $N$ (retry)
8. page table walk
9. PA: $N'$
10. data
kernel decides where to place page, and what to evict (if memory is full)

- e.g., using LRU replacement policy
this system enables \textbf{on-demand paging}
i.e., an active process need only be partly in memory (load rest from disk dynamically)
but if working set (of active processes) exceeds available memory, we may have swap thrashing
integration with caches?
Q: do caches use physical or virtual addresses for lookups?
Virtual address based Cache

Process A

Virtual Address Space

CPU

Cache

<table>
<thead>
<tr>
<th>Address</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>X</td>
</tr>
<tr>
<td>M</td>
<td>Y</td>
</tr>
<tr>
<td>N</td>
<td>Z</td>
</tr>
</tbody>
</table>

Process B

Virtual Address Space

ambiguous!
Physical address based Cache

Process A

Virtual Address Space

M

L

0

X

Process B

Virtual Address Space

N

M

0

Z

Y

CPU

Cache

<table>
<thead>
<tr>
<th>Address</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>X</td>
</tr>
<tr>
<td>Q</td>
<td>Y</td>
</tr>
<tr>
<td>R</td>
<td>Z</td>
</tr>
</tbody>
</table>

Physical Memory

X

Z

Y
Q: do caches use physical or virtual addresses for lookups?

A: caches typically use *physical* addresses
Main Memory

MMU (address translation unit)

Cache

CPU

data

page table walk

(hit) (miss)

(update)

VA

PA

process page table

%o*@$&##!!!
saved by hardware:

the *Translation Lookaside Buffer* (TLB) — a cache used solely for VPN→PPN lookups
only if TLB miss!

TLB + Page table
(exercise for reader: revise earlier translation diagrams!)
virtual address

n-1  virtual page number (VPN)  p  p-1  page offset

valid  tag  physical page number (PPN)

TLB Hit

physical address

valid  tag  data

Cache Hit

byte offset

Data

TLB

Cache
TLB mappings are *process specific* — requires flush & reload on context switch

- some architectures store PID (aka “virtual space” ID) in TLB
Familiar caching problem:

- TLB caches a few thousand mappings
- vs. *millions* of virtual pages per process!
we can improve TLB hit rate by reducing the number of pages …

by increasing the size of each page
compute # pages for 32-bit memory for:

- 1KB, 512KB, 4MB pages

- $2^{32} \div 2^{10} = 2^{22} = 4M$ pages
- $2^{32} \div 2^{19} = 2^{13} = 8K$ pages
- $2^{32} \div 2^{22} = 2^{10} = 1K$ pages

(not bad!)
lots of wasted space!
increasing page size results in increased internal fragmentation and lower utilization
i.e., TLB effectiveness needs to be balanced against memory utilization
so what about 64-bit systems?

$2^{64} = 16$ Exabyte address space

$\approx 4$ billion x 4GB
most modern implementations support a max of $2^{48}$ (256TB) addressable space
page table size?

- # pages  = $2^{48} \div 2^{12} = 2^{36}$
- PTE size  = 8 bytes (64 bits)
- PT size   = $2^{36} \times 8 = 2^{39}$ bytes
  = 512GB
512GB
(just for the virtual memory *mapping* structure)
(and we need *one per process*)
(these things aren’t going to fit in memory)
instead, use *multi-level* page tables:

- split an address translation into two (or more) separate table lookups
- unused parts of the table don’t need to be in memory!
“toy” memory system
- 8 bit addresses
- 32-byte pages

<table>
<thead>
<tr>
<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>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

VPN ─────────── page offset

Page Table

all 8 PTEs must be in memory at all times
“toy” memory system
- 8 bit addresses
- 32-byte pages

Page offset

Page "directory"
all unmapped; don’t need in memory!
“toy” memory system
- 8 bit addresses
- 32-byte pages
**Intel Architecture Memory Management**

http://www.intel.com/products/processor/manuals/

(Software Developer’s Manual Volume 3A)
In computer science, PROTECTED-MODE MEMORY MANAGEMENT involves managing memory in a way that allows for multiple processes to run concurrently. This is typically achieved through segment and paging mechanisms.

1. **Segmentation**: Each process is allocated a segment of memory, identified by its segment type and base address in the linear address space. The offset part of the logical address is added to the base address for addressing a byte within the segment. The base address plus the offset forms a linear address in the processor's linear address space.

2. **Paging**: If virtual memory is used, the linear address space is simulated with a small amount of physical memory (RAM and ROM) and some disk. This is done through the processor's paging mechanism, which uses page tables and page directories to translate logical addresses into physical addresses.

3. **Physical Address Space**: The physical address space is defined as the range of addresses that the processor can generate on its address bus. It is used to access physical memory directly, bypassing virtual memory.

The figure illustrates these concepts with a flowchart, showing how logical addresses are translated into physical addresses through segmentation and paging mechanisms.
Access checks can be used to protect not only against referencing an address outside the limit of a segment, but also against performing disallowed operations in certain segments. For example, since code segments are designated as read-only segments, hardware can be used to prevent writes into code segments. The access rights information created for segments can also be used to set up protection rings or levels.

Protection levels can be used to protect operating-system procedures from unauthorized access by application programs.

3.2.4 Segmentation in IA-32e Mode

In IA-32e mode of Intel 64 architecture, the effects of segmentation depend on whether the processor is running in compatibility mode or 64-bit mode. In compatibility mode, segmentation functions just as it does using legacy 16-bit or 32-bit protected mode semantics.

### Segment descriptors

- **CS**: Code Segment
- **SS**: Stack Segment
- **DS**: Data Segment
- **ES**: Extended Segment
- **FS**: Additional Segment
- **GS**: General Segment
PROTECTED-MODE MEMORY MANAGEMENT

FFFF_FFF0H. RAM (DRAM) is placed at the bottom of the address space because the initial base address for the DS data segment after reset initialization is 0.

3.2.2 Protected Flat Model

The protected flat model is similar to the basic flat model, except the segment limits are set to include only the range of addresses for which physical memory actually exists (see Figure 3-3). A general-protection exception (#GP) is then generated on any attempt to access nonexistent memory. This model provides a minimum level of hardware protection against some kinds of program bugs.

Figure 3-2. Flat Model

Figure 3-3. Protected Flat Model

Linear Address Space (or Physical Memory)

Segment Registers

CS
SS
DS
ES
FS
GS

Code- and Data-Segment Descriptors

Access
Limit
Base Address

Memory I/O
Stack
Not Present

Code
Not Present

Data and Stack

0

“Flat” model
Table 4-1 illustrates the key differences between the three paging modes.

Because they are used only if IA32_EFER.LME = 0, 32-bit paging and PAE paging is used only in legacy protected mode. Because legacy protected mode cannot produce linear addresses larger than 32 bits, 32-bit paging and PAE paging translate 32-bit linear addresses.

Because it is used only if IA32_EFER.LME = 1, IA-32e paging is used only in IA-32e mode. (In fact, it is the use of IA-32e paging that defines IA-32e mode.) IA-32e mode has two sub-modes:

- **Compatibility mode.** This mode uses only 32-bit linear addresses. IA-32e paging treats bits 47:32 of such an address as all 0.
- **64-bit mode.** While this mode produces 64-bit linear addresses, the processor ensures that bits 63:47 of such an address are identical. IA-32e paging does not use bits 63:48 of such addresses.

<table>
<thead>
<tr>
<th>Paging Mode</th>
<th>CR0.PG</th>
<th>CR4.PAE</th>
<th>LME in IA32_EFER</th>
<th>Linear-Address Width</th>
<th>Physical-Address Width¹</th>
<th>Page Size(s)</th>
<th>Supports Execute-Disable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0</td>
<td>N/A</td>
<td>N/A</td>
<td>32</td>
<td>32</td>
<td>N/A</td>
<td>No</td>
</tr>
<tr>
<td>32-bit</td>
<td>1</td>
<td>0</td>
<td>0²</td>
<td>32</td>
<td>Up to 40³</td>
<td>4-KByte</td>
<td>No</td>
</tr>
<tr>
<td>PAE</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>32</td>
<td>Up to 52</td>
<td>4-KByte</td>
<td>Yes⁵</td>
</tr>
<tr>
<td>IA-32e</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>48</td>
<td>Up to 52</td>
<td>4-KByte</td>
<td>Yes⁵</td>
</tr>
</tbody>
</table>

1. The physical-address width is always bounded by MAXPHYADDR; see Section 4.1.4.
2. The processor ensures that IA32_EFER.LME must be 0 if CR0.PG = 1 and CR4.PAE = 0.
3. 32-bit paging supports physical-address widths of more than 32 bits only for 4-MByte pages and only if the PSE-36 mechanism is supported; see Section 4.1.4 and Section 4.3.
4. 4-MByte pages are used with 32-bit paging only if CR4.PSE = 1; see Section 4.3.
5. Execute-disable access rights are applied only if IA32_EFER.NXE = 1; see Section 4.6.
6. Not all processors that support IA-32e paging support 1-GByte pages; see Section 4.1.4.

### Paging modes

![IIT College of Science logo](image)
IA-32 paging (4KB pages)
IA-32 paging (4MB pages)
Figure 4-4 gives a summary of the formats of CR3 and the paging-structure entries with 32-bit paging. For the paging structure entries, it identifies separately the format of entries that map pages, those that reference other paging structures, and those that do neither because they are “not present”; bit 0 (P) and bit 7 (PS) are highlighted because they determine how such an entry is used..

<table>
<thead>
<tr>
<th>Address of page directory</th>
<th>Ignored</th>
<th></th>
<th></th>
<th>Ignored</th>
<th>CR3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bits 31:22 of address of 2MB page frame</td>
<td>Reserved (must be 0)</td>
<td>Ignored</td>
<td>G</td>
<td>1</td>
<td>D</td>
</tr>
<tr>
<td>Bits 39:32 of address</td>
<td>G</td>
<td>1</td>
<td>D</td>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>Address of page table</td>
<td>Ignored</td>
<td>0</td>
<td>Ignored</td>
<td>0</td>
<td>CR3</td>
</tr>
<tr>
<td>Address of 4KB page frame</td>
<td>Ignored</td>
<td>0</td>
<td>CR3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### PTE formats

<table>
<thead>
<tr>
<th>Address of page directory</th>
<th>Ignored</th>
<th></th>
<th></th>
<th>Ignored</th>
<th>CR3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bits 31:22 of address of 2MB page frame</td>
<td>Reserved (must be 0)</td>
<td>Ignored</td>
<td>G</td>
<td>1</td>
<td>D</td>
</tr>
<tr>
<td>Bits 39:32 of address</td>
<td>G</td>
<td>1</td>
<td>D</td>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>Address of page table</td>
<td>Ignored</td>
<td>0</td>
<td>Ignored</td>
<td>0</td>
<td>CR3</td>
</tr>
<tr>
<td>Address of 4KB page frame</td>
<td>Ignored</td>
<td>0</td>
<td>CR3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
IA-32e paging (4KB pages)
The following items describe the IA-32e paging process in more detail as well as how the page size is determined.

- A 4-KByte naturally aligned PML4 table is located at the physical address specified in bits 51:12 of CR3 (see Table 4-12). A PML4 table comprises 512 64-bit entries (PML4Es). A PML4E is selected using the physical address defined as follows:
  - Bits 51:12 are from CR3.
  - Bits 11:3 are bits 47:39 of the linear address.
  - Bits 2:0 are all 0.

Because a PML4E is identified using bits 47:39 of the linear address, it controls access to a 512-GByte region of the linear-address space.

- A 4-KByte naturally aligned page-directory-pointer table is located at the physical address specified in bits 51:12 of the PML4E (see Table 4-14). A page-directory-pointer table comprises 512 64-bit entries (PDPTEs). A PDPTE is selected using the physical address defined as follows:
  - Bits 51:12 are from the PML4E.

![IA-32e paging (1GB pages)](image_url)