IA32/Linux Virtual Memory Architecture
Basic Execution Environment

Application Programming Registers

General-purpose registers

<table>
<thead>
<tr>
<th>Register</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAX</td>
<td>31:0</td>
</tr>
<tr>
<td>EBX</td>
<td>31:0</td>
</tr>
<tr>
<td>ECX</td>
<td>31:0</td>
</tr>
<tr>
<td>EDX</td>
<td>31:0</td>
</tr>
<tr>
<td>EBP</td>
<td>31:0</td>
</tr>
<tr>
<td>ESI</td>
<td>31:0</td>
</tr>
<tr>
<td>EDI</td>
<td>31:0</td>
</tr>
<tr>
<td>ESP</td>
<td>31:0</td>
</tr>
<tr>
<td>EAX</td>
<td>31:0</td>
</tr>
<tr>
<td>EBX</td>
<td>31:0</td>
</tr>
<tr>
<td>ECX</td>
<td>31:0</td>
</tr>
<tr>
<td>EDX</td>
<td>31:0</td>
</tr>
<tr>
<td>EBP</td>
<td>31:0</td>
</tr>
<tr>
<td>ESI</td>
<td>31:0</td>
</tr>
<tr>
<td>EDI</td>
<td>31:0</td>
</tr>
<tr>
<td>ESP</td>
<td>31:0</td>
</tr>
</tbody>
</table>

Segment registers

<table>
<thead>
<tr>
<th>Segment</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>15:0</td>
</tr>
<tr>
<td>DS</td>
<td>15:0</td>
</tr>
<tr>
<td>SS</td>
<td>15:0</td>
</tr>
<tr>
<td>ES</td>
<td>15:0</td>
</tr>
<tr>
<td>FS</td>
<td>15:0</td>
</tr>
<tr>
<td>GS</td>
<td>15:0</td>
</tr>
</tbody>
</table>

Control registers

<table>
<thead>
<tr>
<th>Control</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>CR0</td>
<td>31:0</td>
</tr>
<tr>
<td>CR1</td>
<td>31:0</td>
</tr>
<tr>
<td>CR2</td>
<td>31:0</td>
</tr>
<tr>
<td>CR3</td>
<td>31:0</td>
</tr>
<tr>
<td>CR4</td>
<td>31:0</td>
</tr>
</tbody>
</table>

System Table Registers

<table>
<thead>
<tr>
<th>Register</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDTR</td>
<td>47:16</td>
</tr>
<tr>
<td>IDTR</td>
<td>47:16</td>
</tr>
</tbody>
</table>

System Segment Registers

<table>
<thead>
<tr>
<th>Register</th>
<th>Bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>TR</td>
<td>15:0</td>
</tr>
<tr>
<td>LDTR</td>
<td>15:0</td>
</tr>
</tbody>
</table>
Segmented memory model

- Memory appears to a program as a group of independent address space called segments.
- A program must issue a logical address, which consists of a segment selector and an offset.
- Up to 16,383 segments of different sizes and types
  - Each segment can be as large as $2^{32}$ bytes.
- No way to disable segmentation.
- The use of paging is optional.
IA32 VM Architecture (2)

- **Logical address (far pointer)**
  - User’s view, segmented
  
  - Segment selector (16 bits)
  - Offset (32 bits)

- **Linear address**
  - 32-bit, flat

- **Physical address**
  - 32-bit, flat
  - Pentium Pro and later processors support an extension of the physical address space to $2^{36}$ bytes.
  - Invoked with the physical address extension (PAE) flag located in CR4 register.
IA32 VM Architecture (3)
Segmentation (1)

Basic flat model

- The OS and applications have access to a continuous, unsegmented address space.
- All segment descriptors have the same base address value of 0 and the same segment limit of 4GB.
## Segmentation (2)

### Protected flat model

- Segment limits are set to include only the range of addresses for which physical memory actually exists.
- May have multiple segments, but all overlay each other and start at address 0 in the linear address space.
### Segmentation (3)

#### Multisegment model

- Each program (or task) is given its own table of segment descriptors and its own segments.
- The segments can be completely private to their assigned programs or shared among programs.
Segmentation (4)

- **Segment registers**
  - Hold 16-bit segment selectors.
    - A segment selector is a special pointer that identifies a segment in memory.
    - To access a particular segment, the segment selector for that segment must be present in the appropriate segment register.
  - **Use of segment registers**
    - CS: for code segment
    - DS, ES, FS, and GS: for data segments (up to 4 segments simultaneously)
    - SS: for stack segment
  - FS and GS registers were introduced with the 80386 family of processors.
Segmentation (5)

- **Logical to linear address**
  - Examine the segment descriptor in GDT or LDT to check the access rights and the offset is within the limits.
  - Adds the segment base address from the segment descriptor to the offset to form a linear address.
## Segmentation (6)

### Segment selector

- **Index**
- **Table Indicator**:
  - 0 = GDT
  - 1 = LDT
- **Requested Privilege Level (RPL)**

### Segment registers

<table>
<thead>
<tr>
<th>Visible Part</th>
<th>Hidden Part</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment Selector</td>
<td>Base Address, Limit, Access Information</td>
</tr>
<tr>
<td>CS</td>
<td>SS</td>
</tr>
<tr>
<td>DS</td>
<td>ES</td>
</tr>
<tr>
<td>FS</td>
<td>GS</td>
</tr>
</tbody>
</table>
Segmentation (7)

- Segment descriptor tables
  - Each system must have one GDT (Global Descriptor Table), which may be used for all programs and tasks.
  - Optionally, one or more LDTs (Local Descriptor Tables) can be defined in a system segment.
  - GDT is not a segment, but a data structure in the linear address space pointed to by the GDTR register.
  - GDT must contain a segment descriptor for the LDT segment.
  - The first descriptor in GDT is not used.
  - The LDTR register caches the segment descriptor of the current LDT segment.
Segmentation (8)

- Global and local descriptor tables

[Diagram showing Global and Local Descriptor Tables]
Segmentation (9)

- Segment descriptor

```
+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+
| Base 31:24        | G                 | D/B               | AVL               | Seg. Limit 19:16  | P                 | D/P               | S                 | Type               | Base 23:16         |
|                   |                   |                   |                   |                   |                   |                   |                   |                   |                   |                   |
|                   | 31                | 24 23 22 21 20 19 | 16 15 14 13 12 11 | 8 7               | 0                 |                   |                   |                   | 4                 |
+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+
|                   |                   |                   |                   |                   |                   |                   |                   |                   |                   |
|                   |                   |                   |                   |                   |                   |                   |                   |                   | 31                |
|                   |                   |                   |                   |                   |                   |                   |                   |                   | 16 15             |
|                   |                   |                   |                   |                   |                   |                   |                   |                   | 0                 |
+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+
| Base Address 15:00|                   |                   |                   | Segment Limit 15:00|                   |                   |                   |                   | 0                 |
+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+-------------------+
```

L    — 64-bit code segment (IA-32e mode only)
AVL  — Available for use by system software
BASE — Segment base address
D/B  — Default operation size (0 = 16-bit segment; 1 = 32-bit segment)
DPL  — Descriptor privilege level
G    — Granularity
LIMIT — Segment Limit
P    — Segment present
S    — Descriptor type (0 = system; 1 = code or data)
TYPE — Segment type
Paging (1)

Paging support in IA-32
- Optional: enabled by PG flag of CR0 register
- Default page size: 4KB
  - PSE (page size extension) flag of CR4 enables 4MB page size
    (From Pentium)

36-bit physical addressing
- Pentium Pro and later processors support an extension of the physical address space to $2^{36}$ bytes.
  - Enabled by PAE (physical address extension) flag of CR4
  - With PAE enabled, 2MB page size is supported
- Pentium III introduced PSE-36 mechanism
  - Available when PSE-36 CPUID feature flag is set
  - Map up to 1024 4MB pages into 64GB physical address space
Paging (2)

- Linear to physical address (4KB)
  - The **physical** address of the current page directory is stored in the CR3 register (a.k.a. page directory base register or PDBR).

```
<table>
<thead>
<tr>
<th>Linear Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 22 21 12 11 0</td>
</tr>
<tr>
<td>Directory</td>
</tr>
</tbody>
</table>
```

```
CR3 (PDBR) ➔ Directory Entry ➔ Page Directory ➔ Page Table ➔ Page-Table Entry ➔ Physical Address
```

1024 PDE * 1024 PTE = $2^{20}$ Pages

*32 bits aligned onto a 4-KByte boundary.
Paging (3)

- **Page tables and directories**
  - **Page directory**
    - An array of 32-bit page-directory entries (PDEs) contained in a 4KB page (1024 PDEs/page).
  - **Page table**
    - An array of 32-bit page-table entries (PTEs) contained in a 4KB page (1024 PTEs/page).
    - Page tables are not used for 2MB or 4MB pages.
  - **Page**
    - Supports page sizes of 4KB, 2MB, and 4MB.
  - **Page-directory-pointer table**
    - An array of four 64-bit entries pointing to a page directory.
    - Only used when the physical address extension is enabled.
Paging (4)

- Linear to physical address (4MB, PSE enabled)
  - Both 4MB pages and page tables for 4KB pages can be accessed from the same page directory
  - Place OS kernel in 4MB pages to reduce TLB misses

*32 bits aligned onto a 4-KByte boundary.
Paging (5)

- Linear to physical address (4KB, PAE enabled)

```
Directory Pointer -> Directory Entry -> Page Directory -> Page-Directory-Pointer Table
                     |                                    | 32* CR3 (PDPTR)
                     |                                    | *32 bits aligned onto a 32-byte boundary
```

```
31 30 29 21 20 12 11 0

Directory Table Offset

Linear Address

32* 512 PDE * 512 PTE = 2^20 Pages
```

```
Page Table

Page-Table Entry

4-KByte Page

Physical Address

(=up to 40)
```
IA-32e paging mode in Intel64

- 48-bit virtual address $\rightarrow$ 52-bit physical address (4KB)
### Page directory entry (PDE)

#### Page-Directory Entry (4-KByte Page Table)

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>12</td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Page-Table Base Address**
- **Avail**: Available for system programmer’s use
- **G**, **P**, **S**: Global page (Ignored)
- **A**, **V**, **L**: Page size (0 indicates 4 KBytes)
- **P**: Available
- **C**: Accessed
- **D**: Cache disabled
- **W**: Write-through
- **T**: User/Supervisor
- **U**: Read/Write
- **R**, **W**: Present
### Paging (8)

#### Page table entry (PTE)

| 31 | 12  | 11  | 9  | 8  | 7  | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
|----|-----|-----|----|----|----|---|---|---|---|---|---|--|---|
|    |     |     |    |    |    |   |   |   |   |   |   |   |   |

- **Page Base Address**
- **Avail**
- **P**
- **G**
- **PAT**
- **D**
- **A**
- **PCD**
- **PWT**
- **USR**
- **R**
- **W**

| Available for system programmer’s use |
| Global Page |
| Page Table Attribute Index |
| Dirty |
| Accessed |
| Cache Disabled |
| Write-Through |
| User/Supervisor |
| Read/Write |
| Present |
### TLBs

- The P6 family and Pentium processors have separate TLBs for the data and instruction. (DTLB & ITLB)
- Separate TLBs for 4KB and 4MB page sizes
- All TLBs are automatically invalidated if the PDBR register is loaded.
  - by explicit MOV instruction
  - implicitly by executing a task switch
- A specific page-table entry in the TLB can be invalidated using INVLPG instruction.
- The page global enable (PGE) flag in CR4 and the global (G) flag of a PDE or PTE can be used to prevent frequently used pages from being automatically invalidated.
For more information, see

  - Volume 1: Basic Architecture
  - Volume 2: Instruction Set Reference
  - Volume 3: System Programming Guide

- Available at Intel’s web site:
Linux VM Architecture (1)

Virtual memory

0x00000000

PAGE_OFFSET = 0xC0000000

0xFFFFFFFF

3GB

1GB

0x00000000

0x3FFFFFFF

Physical memory

Kernel code

Kernel data

Page tables

Freelists, etc.

Available Page Frames

0xFFFFFFFF

0x3FFFFFFF

0x00000000

1GB

1GB
Linux VM Architecture (2)

Linux uses a virtual memory system with a linear address space. The address space is divided into two main segments: the kernel space and the user space.

- **Kernel Space**: Located from 0 to 3GB, it contains the BIOS, kernel text, and kernel data.
- **User Space**: Located from 3GB to the end of memory, it contains user applications and processes.

The kernel space is further divided into non-contiguous mappings, which include:
- BIOS
- kernel text
- kernel data

The user space is divided into three main regions:
- **kmalloc etc.**: Located from 0 to 4GB, it contains memory for user applications and processes.
- **vmalloc**: Located from 4GB to 890MB, it is used for allocating memory where possible.
- **pkmap and fixmap**: Located from 890MB to the end of memory, they are used for memory allocation and management.

The virtual address space is mapped to the physical address space using linear mapping. Use large pages where possible to improve performance and reduce memory fragmentation.

**Key Points**:
- **PAGE_OFFSET**: Used for offsetting virtual addresses.
- **VMALLOC_START**: The start address for user memory allocation.
- **PKMAP_BASE**: The base address for memory management.
- **FIXADDR_START**: The start address for fixed memory.
- **high_memory**: Memory above 890MB.
- **low_memory**: Memory below 890MB.

---

SSE3044: Operating Systems | Fall 2013 | Jin-Soo Kim (jinsookim@skku.edu)
Segmentation: Minimal approach

- For better portability across machines

### GDT

<table>
<thead>
<tr>
<th>Address</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x00</td>
<td>NULL</td>
</tr>
<tr>
<td>0x08</td>
<td>(not used)</td>
</tr>
<tr>
<td>0x10</td>
<td>Kernel text from 0 (4GB)</td>
</tr>
<tr>
<td>0x18</td>
<td>Kernel data from 0 (4GB)</td>
</tr>
<tr>
<td>0x20</td>
<td>User text from 0 (4GB)</td>
</tr>
<tr>
<td>0x28</td>
<td>User data from 0 (4GB)</td>
</tr>
<tr>
<td>0xa0</td>
<td>Used for APM (4 entries)</td>
</tr>
<tr>
<td></td>
<td>Used for PNPBIOS (8 entries)</td>
</tr>
<tr>
<td></td>
<td>4 entries per CPU For TSS’s &amp; LDT’s</td>
</tr>
</tbody>
</table>

### Segment selectors

```
15 3 2 1 0
Index RPL
```

- Table Indicator:
  - 0 = GDT
  - 1 = LDT
- Requested Privilege Level (RPL)

<table>
<thead>
<tr>
<th>Selector</th>
<th>Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>__KERNEL_CS</td>
<td>0x10</td>
</tr>
<tr>
<td>__KERNEL_DS</td>
<td>0x18</td>
</tr>
<tr>
<td>__USER_CS</td>
<td>0x23</td>
</tr>
<tr>
<td>__USER_DS</td>
<td>0x2b</td>
</tr>
</tbody>
</table>
Paging: Four-level address translation

- The size of PUD and PMD is 1, if the physical address extension (PAE) flag is disabled.
Virtual memory areas (VMA)

- Nonoverlapping regions, each region representing a continuous, page-aligned subset of the virtual address space.
- Described by a single `vm_area_struct`
- VMAs are linked into a balanced binary tree to allow fast lookup of the region corresponding to any virtual address.
  - VMAs form a red-black tree.
Linux VM Architecture (6)

- **task_struct**
  - mm
  - mm_struct
    - map_count
    - pgd
    - mmap
    - mm_rb
  - page directory
    - PFN

- **mm_struct**
  - pgd

- **vm_area_struct**
  - vm_start
  - vm_end
  - vm_mm
  - vm_rb
  - vm_ops
  - vm_next

- **Virtual address space**
  - VM Area 1
  - VM Area 2
Linux VM Architecture (7)

- VMA example

```
[root@oz0 jinsoo]# cat /proc/1/maps
08048000-0804e000 r-xp 00000000 03:03 716858 /sbin/init
0804e000-0804f000 rw-p 00000000 03:03 716858 /sbin/init
0804f000-08053000 r-xp 00000000 00:00 0
40000000-40013000 r-xp 00000000 03:03 244332 /lib/ld-2.2.5.so
40013000-40014000 rw-p 00013000 03:03 244332 /lib/ld-2.2.5.so
40031000-40032000 rw-p 00000000 00:00 0
42000000-4212c000 r-xp 00000000 03:03 915244 /lib/i686/libc-2.2.5.so
4212c000-42131000 rw-p 0012c000 03:03 915244 /lib/i686/libc-2.2.5.so
42131000-42135000 rw-p 00000000 00:00 0
bffff000-c0000000 rwxp 00000000 00:00 0
[root@oz0 jinsoo]#
```

VMA permission offset device i-node mapped file
Linux VM Architecture (8)

- Page replacement: File pages

```
Not freeable

(INACTIVE LIST) head[New page] → tail[Evicted]

Not referenced

ACTIVE LIST head[2x Referenced] → tail[Referenced]

Referenced once or never

|Active list| / |Inactive list| = 1
```
Page replacement: Anonymous pages

\[ \text{Active list} / \text{Inactive list} = 1, \quad \text{if mem < 1GB} \]
\[ = \sqrt{10 \times \text{gb}}, \quad \text{otherwise} \]