Operating Systems

Address translation
(Segmentation and Paging)

Anton Burtsev
Two programs one memory

main() {
    ...
    yield()
}

main() {
    ...
    yield()
}

Save/restore
Two programs one memory

- How can we do this?

```
main() {
    ...
    yield()
}
```

```
main() {
    ...
    yield()
}
```
Relocation

• One way to achieve this is to relocate program at different addresses
  • Remember relocation (from linking and loading)
Relocate binaries to work at different addresses

Process 1 (ls)

Process text | Process data | Process heap | (free) | Process stack

Process 2 (ls)

Process text | Process data | Process heap | (free) | Process stack

Memory

Relocate to start at 0x110000
• One way to achieve this is to relocate program at different addresses
  • Remember relocation (from linking and loading)
  • This works! But not ideal.
• What is the problem?
• One way to achieve this is to relocate program at different addresses
  • Remember relocation (from linking and loading)
  • This works! But not ideal.
• What is the problem?
  • Isolation can be enforced in software
  • Software Fault Isolation (SFI)
    - Google NaCl (Chrome Sandbox)
    - WASM (Web Assembly, another sandbox standard)
Another way is to ask for hardware support
Segmentation
What are we aiming for?

- Illusion of a private address space
  - Identical copy of an address space in multiple programs
  - Simplifies software architecture
    - One program is not restricted by the memory layout of the others
Two processes, one memory?

Process 1 (ls)

Process 2 (ls)

Memory
Two processes, one memory?

- We want hardware to add base value to every address used in the program
Seems easy

- One problem
  - Where does this base address come from?
Seems easy

• One problem
  • Where does this base address come from?
  • Hardware can maintain a table of base addresses
    – One base for each process
  • Dedicate a special register to keep an index into that table
• One problem
  • Where does this base address come from?
  • Hardware can maintain a table of base addresses
    – One base for each process
  • Dedicate a special register to keep an index into that table
Segmentation: example

```
mov (%EBX), EAX  # mov value from the location pointed by EBX into EAX
EAX = 0x0
EBX = 0x300010
```

Physical Memory
Segmentation: address consists of two parts

- Segment register contains segment selector
- General registers contain offsets
- Intel calls this address: “logical address”
Segmenation: Global Descriptor Table

- GDT is an array of segment descriptors
- Each descriptor contains base and limit for the segment
- Plus access control flags
Location of GDT in physical memory is pointed by the GDT register
Segmentation: base + offset

- Segment register (0x1) chooses an entry in GDT
- This entry contains base of the segment (0x110000) and limit (size) of the segment (not shown)
• Physical address:
  • 0x410010 = 0x300010 (offset) + 0x110000 (base)
  • Intel calls this address “linear”
Each process has a private GDT

OS will switch between GDTs
New addressing mode: “logical addresses”
All addresses are logical address

- They consist of two parts
  - Segment selector (16 bit) + offset (32 bit)
- Segment selector (16 bit)
  - Is simply an index into an array (Descriptor Table)
  - That holds segment descriptors
    - Base and limit (size) for each segment
Elements of the descriptor table are segment descriptors

- Base address
  - 0 – 4 GB
- Limit (size)
  - 0 – 4 GB
- Access rights
  - Executable, readable, writable
  - Privilege level (0 - 3)
- Offsets into segments (x in our example) or "Effective addresses" are in registers.
• Logical addresses are translated into physical

  • Effective address + DescriptorTable[selector].Base
• Logical addresses are translated into physical
  • Effective address + DescriptorTable[selector].Base
• Logical addresses are translated into physical
• Effective address + DescriptorTable[selector].Base
• Logical addresses are translated into physical
• Effective address + DescriptorTable[selector].Base
- **Physical address** = 

  \[ \text{Effective address} + \text{DescriptorTable}[\text{selector}].\text{Base} \]

- Effective addresses (or offsets) are in registers
- **Selector is in a special register**
Segment registers

• Hold 16 bit segment selectors
  • Indexes into GDT

• Segments are associated with one of three types of storage
  • Code
  • Data
  • Stack
Segmented programming (not real)

```c
static int x = 1;
int y; // stack
if (x) {
    y = 1;
    printf ("Boo");
} else
    y = 0;
```

```c
ds:x = 1; // data
ss:y;     // stack
if (ds:x) {
    ss:y = 1;
    cs:printf(ds:"Boo");
} else
    ss:y = 0;
```
Programming model

- Segments for: code, data, stack, “extra”
  - A program can have up to 6 total segments
  - Segments identified by registers: cs, ds, ss, es, fs, gs

- Prefix all memory accesses with desired segment:
  - `mov eax, ds:0x80` (load offset 0x80 from data into eax)
  - `jmp cs:0xab8` (jump execution to code offset 0xab8)
  - `mov ss:0x40, ecx` (move ecx to stack offset 0x40)
Programming model, cont.

- This is cumbersome,
- Instead the idea is: infer code, data and stack segments from the instruction type:
  - Control-flow instructions use code segment (jump, call)
  - Stack management (push/pop) uses stack
  - Most loads/stores use data segment
- Extra segments (es, fs, gs) must be used explicitly
Code segment

• Code
  • CS register
  • EIP is an offset inside the segment stored in CS

• Can only be changed with
  • procedure calls,
  • interrupt handling, or
  • task switching
Data segment

- Data
  - DS, ES, FS, GS
  - 4 possible data segments can be used at the same time
Stack segment

- Stack
  - ss
- Can be loaded explicitly
  - OS can set up multiple stacks
  - Of course, only one is accessible at a time
Segmentation: what did we achieve

- Illusion of a private address space
  - Identical copy of an address space in multiple programs
    - We can implement `fork()`

- Isolation
  - Processes cannot access memory outside of their segments
Segmentation works for isolation, i.e., it does provide programs with illusion of private memory
Segmentation is ok... but
What if process needs more memory?

Diagram:

Process 1 (ls)
- `malloc() = x`
- `base_{p_1}`
- `x + base_{p_1}`

Process 2 (ls)
- `x`
- `base_{p_2}`
- `x + base_{p_2}`
What if process needs more memory?
You can move P2 in memory
Or even swap it out to disk
Problems with segments

• Segments are somewhat inconvenient
  • Relocating or swapping the entire process takes time

• Memory gets fragmented
  • There might be no space (gap) for the swapped out process to come in
  • Will have to swap out other processes
Paging
Paging idea

• Break up memory into 4096-byte chunks called pages
  • Modern hardware supports 2MB, 4MB, and 1GB pages
• Independently control mapping for each page of linear address space

• Compare with segmentation (single base + limit)
  • many more degrees of freedom
How can we build this translation mechanism?
Paging: naive approach: translation array

Linear address 0x410010
- Remember it’s result of logical to linear translation (aka segmentation)
  - \(0x410010 = 0x300010\) (offset) + \(0x110000\) (base)
Paging: naive approach: translation array

Linear address 0x410010

- 0x410010 = 0x300010 (offset) + 0x110000 (base)

Remember it’s result of logical to linear translation (aka segmentation)
What is wrong?
What is wrong?

• We need 4 bytes to relocate each page
  • 20 bits for physical page number
  • 12 bits of access flags

• Therefore, we need array of 4 bytes x 1M entries
  - 4MBs
Paging: naive approach: translation array

Virtual Address Space (aka Virtual Memory)

0x410010 = 00 0000 0001 00 0001 0000 0000 0001 0000

page number

page number = 1040 or 0x410 or (0b 100 0001 0000)

1M (1,048,575)

V2P translation array

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

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

1040

1025

Physical Memory

CR3 = 0x2

- Each entry is 4 bytes
  -- 20 bits to represent page number + access control bits
- 1 page can contain 1024 entries
- We need 1024 pages to represent all possible 1M translations
Paging: array with size

- The size controls how many entries are required
But still what may go wrong?
Paging: array with size
Paging: array with size
Paging: array with chunks

0x410010 = \text{00 0000 0001|00 0001 00000000 0001 0000}

Table of array regions

V2P translation array

500,000
Paging: array with chunks

0x410010 = 00 0000 0001 00 0001 0000 0000 0001 0000

Table of array regions
- 0x8
- 0x9
- ...
- 0xF

V2P translation array

Physical memory
Paging: page table

0x410010 = 00 0000 0001 00 0001 0000 0000 0001 0000

Page Table Directory (aka Level 1)

V2P translation array

Physical memory

CR3 = 0x2
mov (%EBX), EAX  # mov value from the location pointed by EBX into EAX
EAX = 0
EBX = 20 983 809

20 983 809 = 00 0000 0101|00 0000 0011|0000 0000 0001

Virtual Address Space (or Memory) of the Process

page number

0 1 2

page number = 5123
or (0b1 0100 0000 0011)

Physical Memory

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

1M (1,048,575)
mov (%EBX), EAX  # mov value from the location pointed by EBX into EAX
EAX = 0
EBX = 20 983 809

20 983 809 = 00 0000 0101 00 0000 0011 0000 0000 0001

Virtual Address Space (or Memory) of the Process

CR3 = 0

page number = 5123
or (0b1 0100 0000 0011)

1M (1,048,575)
mov (%EBX), EAX  # mov value from the location pointed by EBX into EAX
EAX = 0
EBX = 20 983 809

20 983 809 = 00 0000 0101 00 0000 0011 0000 0000 0001

page number

Virtual Address Space (or Memory) of the Process

CR3 = 0

page number = 5123
or (0b1 0100 0000 0011)

Physical Memory

32 bits (4 bytes)

Level 1 (Page Table Directory)
mov (%EBX), EAX  # mov value from the location pointed by EBX into EAX
EAX = 0
EBX = 20 983 809

20 983 809 = 00 0000 010 [00 0000 0011] 0000 0000 0001

Virtual Address Space (or Memory) of the Process

CR3 = 0

Physical Memory

32 bits (4 bytes)

1023

Level 1 (Page Table Directory)

Level 2 (Page Table)

page number = 5123 or (0b1 0100 0000 0011)

1M (1,048,575)
mov (%EBX), EAX  # mov value from the location pointed by EBX into EAX
EAX = 0
EBX = 20 983 809

20 983 809 = 00 0000 0101|00 0000 0011|0000 0000 0001

Virtual Address Space (or Memory) of the Process

CR3 = 0

Virtual Address = 0 1 2
page number = 5123 or (0b1 0100 0000 0011)

Physical Memory

32 bits (4 bytes)

1M (1,048,575)

Level 1
(Page Table Directory)

Level 2
(Page Table)

Page

Level 1
(Page Table Directory)

Level 2
(Page Table)
mov (%EBX), EAX  # mov value from the location pointed by EBX into EAX
EAX = 0
EBX = 20 983 809

20 983 809 =

```
00 0000 0101 00 0000 0011 0000 0000 0001
```

Virtual Address Space (or Memory) of the Process

CR3 = 0

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

Physical Memory

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

page number = 5123 or (0b1 0100 0000 0011)

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

32 bits (4 bytes)

```
55
```

Level 1 (Page Table Directory)

```
Level 2 (Page Table)
```

Page

```
```

1M (1,048,575)

Result:

- EAX = 55
Page translation

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

| 12 10 |
| 4-KByte Page |
| Physical Address |

```

- Page Directory
- PDE with PS=0
- CR3
Page directory entry (PDE)

<table>
<thead>
<tr>
<th>31</th>
<th>30</th>
<th>29</th>
<th>28</th>
<th>27</th>
<th>26</th>
<th>25</th>
<th>24</th>
<th>23</th>
<th>22</th>
<th>21</th>
<th>20</th>
<th>19</th>
<th>18</th>
<th>17</th>
<th>16</th>
<th>15</th>
<th>14</th>
<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>Address of page table</td>
<td>Ignored</td>
<td>0</td>
<td>Ign</td>
<td>A</td>
<td>P</td>
<td>C</td>
<td>D</td>
<td>Pw</td>
<td>U</td>
<td>S</td>
<td>R</td>
<td>W</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 20 bit address of the page table
Page directory entry (PDE)

<table>
<thead>
<tr>
<th>31</th>
<th>30</th>
<th>29</th>
<th>28</th>
<th>27</th>
<th>26</th>
<th>25</th>
<th>24</th>
<th>23</th>
<th>22</th>
<th>21</th>
<th>20</th>
<th>19</th>
<th>18</th>
<th>17</th>
<th>16</th>
<th>15</th>
<th>14</th>
<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></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Address of page table</td>
<td>Ignored</td>
<td>0</td>
<td>I g n</td>
<td>A</td>
<td>P</td>
<td>C</td>
<td>D</td>
<td>P W</td>
<td>U</td>
<td>S</td>
<td>R</td>
<td>W</td>
<td>1</td>
<td>P D E: page table</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 20 bit address of the page table
- Wait... 20 bit address, but we need 32 bits
Page directory entry (PDE)

- 20 bit address of the page table
- Wait... 20 bit address, but we need 32 bits
  - Pages 4KB each, we need 1M to cover 4GB
  - Pages start at 4KB (page aligned boundary)
Page translation

Linear Address

31 22 21 12 11 0

Directory  Table  Offset

Page Directory

PDE with PS=0

CR3

Page Table

PTE

Physical Address

4-KByte Page
Page table entry (PTE)

- 20 bit address of the 4KB page
- Pages 4KB each, we need 1M to cover 4GB

| 31 | 30 | 29 | 28 | 27 | 26 | 25 | 24 | 23 | 22 | 21 | 20 | 19 | 18 | 17 | 16 | 15 | 14 | 13 | 12 | 11 | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 | 0 |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Address of 4KB page frame | Ignored | G | PAT | D | A | PC | D | PW | T | U | S | R | W | 1 | PTE: 4KB page |

- PTE: 4KB page
Benefit of page tables

... Compared to arrays?

- Page tables represent sparse address space more efficiently
  - An entire array has to be allocated upfront
  - But if the address space uses a handful of pages
  - Only page tables (Level 1 and 2 need to be allocated to describe translation)

- On a dense address space this benefit goes away
  - I'll assign a homework!
What about isolation?

- Two programs, one memory?
- Each process has its own page table
  - OS switches between them
P1 and P2 can't access each other memory
Compared to segments pages allow ...

- Emulate large virtual address space on a smaller physical memory
  - In our example we had only 12 physical pages
  - But the program can access all 1M pages in its 4GB address space
  - The OS will move other pages to disk
Compared to segments pages allow ...

- Share a region of memory across multiple programs
  - Well… segments allow this too
  - Communication (shared buffer of messages)
  - Shared libraries
Recap: complete address translation
Logical Address (or Far Pointer)

Segment Selector
Offset

Global Descriptor Table (GDT)
Segment Descriptor

Segment Base Address

Segment

Linear Address Space

Lin. Addr.

Page Directory

Page Table

Entry

Linear Address
Dir Table Offset

Page

Phy. Addr.

Physical Address Space

Segmentation
Paging
32bit x86 supports two page sizes

• 4KB pages
• 4MB pages
Page translation for 4MB pages

0x410010 = 00 0000 0001|00 0001 0000 0000 0001 0000

Page Table Directory (aka Level 1)

Physical memory

CR3 = 0x8192

4MB
Page translation for 4MB pages

Linear Address

31  22  21  0

Directory  Offset

Page Directory

PDE with PS=1

4-MByte Page

Physical Address

CR3

CR3

32

18

22
Page translation in 64bit mode

- Virtual addresses are 47 bits
- Physical addresses are 52 bits
Questions?
What pages are used for

• Protect parts of the program
  • E.g., map code as read-only
    – Disable code modification attacks
    – Remember R/W bit in PTD/PTE entries!
  • E.g., map stack as non-executable
    – Protects from stack smashing attacks
    – Non-executable bit
More paging tricks

- Determine a working set of a program?
More paging tricks

• Determine a working set of a program?
  • Use “accessed” bit
More paging tricks

● Determine a working set of a program?
  ● Use “accessed” bit
● Iterative copy of a working set?
  ● Used for virtual machine migration
More paging tricks

• Determine a working set of a program?
  • Use “accessed” bit

• Iterative copy of a working set?
  • Used for virtual machine migration
  • Use “dirty” bit
More paging tricks

- Determine a working set of a program?
  - Use “accessed” bit
- Iterative copy of a working set?
  - Used for virtual machine migration
  - Use “dirty” bit
- Copy-on-write memory, e.g. lightweight fork()?
More paging tricks

- Determine a working set of a program?
  - Use “accessed” bit
- Iterative copy of a working set?
  - Used for virtual machine migration
  - Use “dirty” bit
TLB

- Walking page table is slow
- Each memory access is 240 (local) - 360 (one QPI hop away) cycles on modern hardware
- L3 cache access is 50 cycles
TLB

- CPU caches results of page table walks
  - In translation lookaside buffer (TLB)

<table>
<thead>
<tr>
<th>Virt</th>
<th>Phys</th>
</tr>
</thead>
<tbody>
<tr>
<td>0xf0231000</td>
<td>0x1000</td>
</tr>
<tr>
<td>0x00b31000</td>
<td>0x1f000</td>
</tr>
<tr>
<td>0xb0002000</td>
<td>0xc1000</td>
</tr>
</tbody>
</table>
TLB invalidation

- TLB is a cache (in CPU)
  - It is not coherent with memory
  - If page table entry is changes, TLB remains the same and is out of sync

<table>
<thead>
<tr>
<th>Virt</th>
<th>Phys</th>
</tr>
</thead>
<tbody>
<tr>
<td>0xf0231000</td>
<td>0x1000</td>
</tr>
<tr>
<td>0x00b31000</td>
<td>0x1f000</td>
</tr>
<tr>
<td>0xb0002000</td>
<td>0xc1000</td>
</tr>
</tbody>
</table>

Same Virt Addr.  No Change!!!
TLB invalidation

- After every page table update, OS needs to manually invalidate cached values
  - Flush TLB
    - Either one specific entry
    - Or entire TLB, e.g., when CR3 register is loaded
    - This happens when OS switches from one process to another
  - This is expensive
    - Refilling the TLB with new values takes time
Tagged TLBs

- Modern CPUs have “tagged TLBs”,
  - Each TLB entry has a “tag” – identifier of a process
  - No need to flush TLBs on context switch
- On Intel this mechanism is called
  - Process-Context Identifiers (PCIDs)

<table>
<thead>
<tr>
<th>Virt</th>
<th>Phys</th>
<th>Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>0xf0231000</td>
<td>0x1000</td>
<td>P1</td>
</tr>
<tr>
<td>0x00b31000</td>
<td>0x1f000</td>
<td>P2</td>
</tr>
<tr>
<td>0xb0002000</td>
<td>0xc1000</td>
<td>P1</td>
</tr>
</tbody>
</table>
When would you disable paging?
When would you disable paging?

• Imagine you're running a memcached
  • Key/value cache
• You serve 1024 byte values (typical) on 10Gbps connection
  • 1024 byte packets can leave every 835ns, or 1670 cycles (2GHz machine)
  • This is your target budget per packet
When would you disable paging?

- Now, to cover 32GB RAM with 4K pages
  - You need 64MB space
  - 64bit architecture, 4-level page tables (or 5-levels now)
- Page tables do not fit in L3 cache
  - Modern servers come with 32MB cache
- Every cache miss results in up to 4 cache misses due to page walk (remember 4-level page tables)
  - Each cache miss is 250 cycles

- Solution: 1GB pages
Back of the envelope

• If a page is 4K and an entry is 4 bytes, how many entries per page?
Back of the envelope

• If a page is 4K and an entry is 4 bytes, how many entries per page?
  • 1k
Back of the envelope

- If a page is 4K and an entry is 4 bytes, how many entries per page?
  - 1k
- How large of an address space can 1 page represent?
Back of the envelope

- If a page is 4K and an entry is 4 bytes, how many entries per page?
  - 1k
- How large of an address space can 1 page represent?
  - 1k entries * 1page/entry * 4K/page = 4MB
Back of the envelope

- If a page is 4K and an entry is 4 bytes, how many entries per page?
  - 1k
- How large of an address space can 1 page represent?
  - 1k entries * 1page/entry * 4K/page = 4MB
- How large can we get with a second level of translation?
If a page is 4K and an entry is 4 bytes, how many entries per page?
- 1k

How large of an address space can 1 page represent?
- 1k entries * 1page/entry * 4K/page = 4MB

How large can we get with a second level of translation?
- 1k tables/dir * 1k entries/table * 4k/page = 4 GB
- Nice that it works out that way!
### Segment descriptors

![Diagram of segment descriptors](image)

<table>
<thead>
<tr>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base 31:24</td>
<td>Segment base address</td>
</tr>
<tr>
<td>G</td>
<td>Granularity</td>
</tr>
<tr>
<td>D/B</td>
<td>Default operation size (0 = 16-bit segment; 1 = 32-bit segment)</td>
</tr>
<tr>
<td>AVL</td>
<td>Available for use by system software</td>
</tr>
<tr>
<td>Seg. Limit 19:16</td>
<td>Segment limit (16-bit mode)</td>
</tr>
<tr>
<td>P</td>
<td>Segment present</td>
</tr>
<tr>
<td>DPL</td>
<td>Descriptor privilege level</td>
</tr>
<tr>
<td>S</td>
<td>Descriptor type (0 = system; 1 = code or data)</td>
</tr>
<tr>
<td>Type</td>
<td>Segment type</td>
</tr>
<tr>
<td>Base 23:16</td>
<td>Segment base address (32-bit mode)</td>
</tr>
<tr>
<td>Base Address 15:00</td>
<td>Segment limit (32-bit mode)</td>
</tr>
</tbody>
</table>

- L  — 64-bit code segment (IA-32e mode only)
Page translation

Linear Address

31 22 21 12 11 0
Directory  Table  Offset

12 4-KByte Page

10
Page Directory

10
Page Table

20
Physical Address

20
PTE

32
CR3

PDE with PS=0
Page directory entry (PDE)

| 31  | 30  | 29  | 28  | 27  | 26  | 25  | 24  | 23  | 22  | 21  | 20  | 19  | 18  | 17  | 16  | 15  | 14  | 13  | 12  | 11  | 10  | 9   | 8   | 7   | 6   | 5   | 4   | 3   | 2   | 1   | 0   |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
|     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |     |

- Address of page table
  - Ignored
  - 0
  - Ignored
  - A
  - P
  - C
  - D
  - PW
  - U/S
  - R
  - W
  - PDE: page table

- 20 bit address of the page table
### Page directory entry (PDE)

| 31 | 30 | 29 | 28 | 27 | 26 | 25 | 24 | 23 | 22 | 21 | 20 | 19 | 18 | 17 | 16 | 15 | 14 | 13 | 12 | 11 | 10 | 9  | 8  | 7  | 6  | 5  | 4  | 3  | 2  | 1  | 0  |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| Address of page table | Ignored | 0 | Ign | A | P | C | D | P | W | U | S | R | W | 1 | PDE: page table |

- 20 bit address of the page table
- Wait... 20 bit address, but we need 32 bits
Page directory entry (PDE)

- 20 bit address of the page table
- Wait... 20 bit address, but we need 32 bits

- Pages 4KB each, we need 1M to cover 4GB
- Pages start at 4KB (page aligned boundary)
Page directory entry (PDE)

- Bit #1: R/W – writes allowed?
- But allowed where?
Page directory entry (PDE)

- Bit #1: R/W – writes allowed?
  - But allowed where?
  - One page directory entry controls 1024 Level 2 page tables
    - Each Level 2 maps 4KB page
  - So it's a region of 4KB x 1024 = 4MB
Page directory entry (PDE)

- Bit #2: U/S – user/supervisor
  - If 0 – user-mode access is not allowed
  - Allows protecting kernel memory from user-level applications
Page translation

Linear Address
31 22 21 12 11 0
Directory Table Offset

Page Directory
10
PDE with PS=0
32
CR3

10
Page Table

12
4-KByte Page

12
Physical Address

20
PTE
### Page table entry (PTE)

<table>
<thead>
<tr>
<th>Bit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>PTE: 4KB page</td>
</tr>
<tr>
<td>1</td>
<td>RW – writes allowed?</td>
</tr>
<tr>
<td>2</td>
<td>U/S – user/supervisor</td>
</tr>
<tr>
<td>3</td>
<td>Accessed</td>
</tr>
<tr>
<td>4</td>
<td>Dirty</td>
</tr>
<tr>
<td>5</td>
<td>4KB page frame address</td>
</tr>
</tbody>
</table>

- **20 bit address of the 4KB page**
  - Pages 4KB each, we need 1M to cover 4GB
- **Bit #1: RW** – writes allowed?
  - To a 4KB page
- **Bit #2: U/S** – user/supervisor
  - If 0 user-mode access is not allowed
- **Bit #5: A** – accessed
- **Bit #6: D** – dirty – software has written to this page
Questions?