18-600 Foundations of Computer Systems

Lecture 15: "Virtual Memory Concepts and Systems"

October 18, 2017
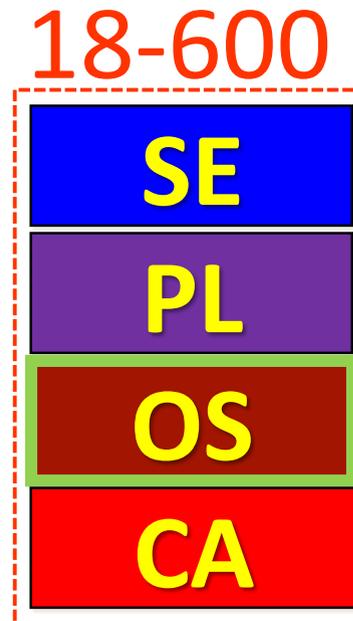
18-600

SE

PL

OS

CA

> Required Reading Assignment:
>   • **Chapter 9 of CS:APP (3rd edition) by Randy Bryant & Dave O'Hallaron.**

Electrical & Computer ENGINEERING

Carnegie Mellon University

Socrative Experiment (Continuing)

➢ Pittsburgh Students (18600PGH): https://api.socrative.com/rc/icJVVC

➢ Silicon Valley Students (18600SV): https://api.socrative.com/rc/iez85z

➢ Microphone/Speak out/Raise Hand: Still G-R-E-A-T!

➢ Socrative:

- Let's me open floor for electronic questions, putting questions into a visual queue so I don't miss any
- Let's me do flash polls, etc.
- Prevents cross-talk and organic discussions in more generalized forums from pulling coteries out of class discussion into parallel question space.
  - Keeps focus and reduces distraction while adding another vehicle for classroom interactivity.
- Won't allow more than 150 students per "room"
  - So, I created one room per campus
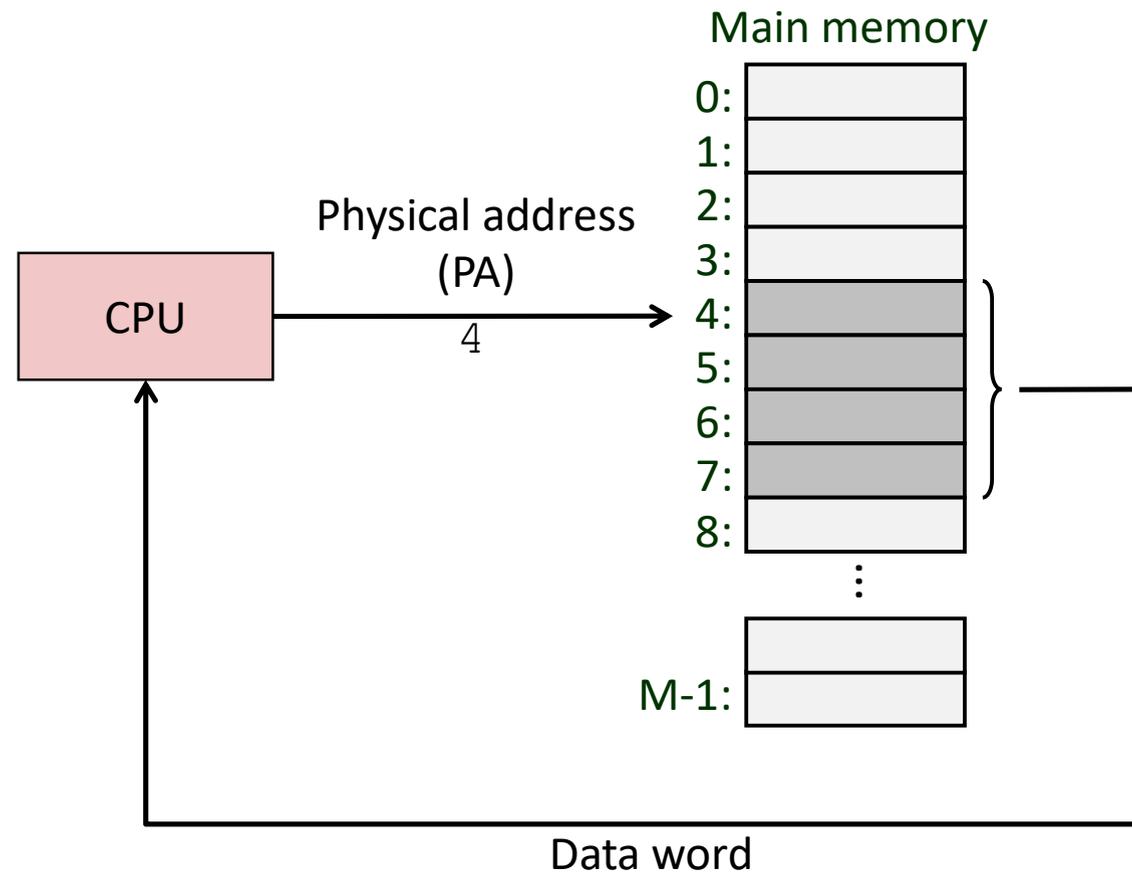 - May later try random assignment to a room, etc.

Carnegie Mellon University

18-600 Foundations of Computer Systems

Lecture 15:
"Virtual Memory Concepts and Systems"

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

Electrical & Computer ENGINEERING

A System Using Physical Addressing

Main memory

CPU

Physical address
(PA)

4

0:
1:
2:
3:
4:
5:
6:
7:
8:

⋮

M-1:

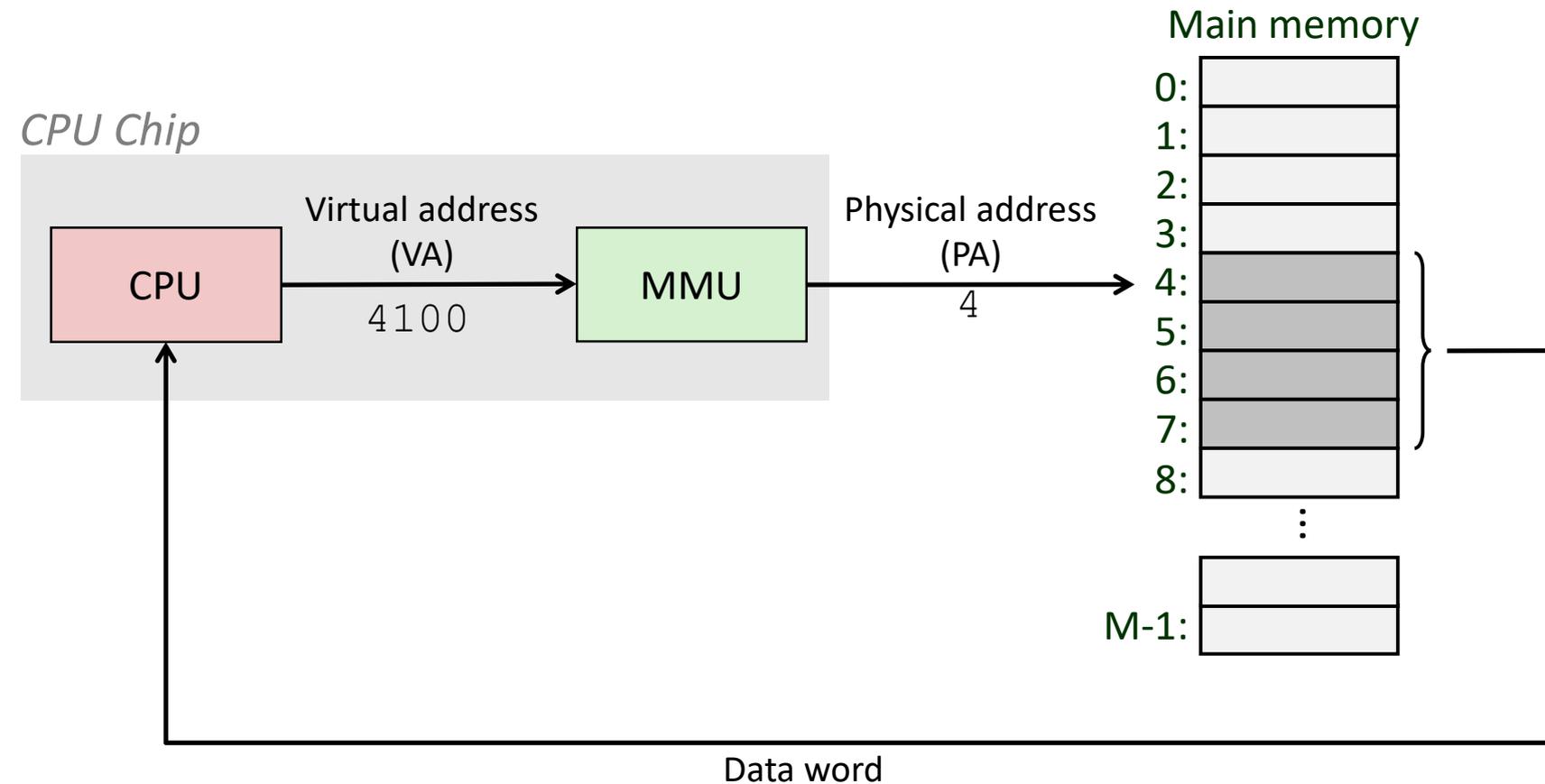Data word

- **Used in "simple" systems like embedded microcontrollers in devices like cars, elevators, and digital picture frames**

A System Using Virtual Addressing

- **Used in all modern servers, laptops, and smart phones**
- **One of the great ideas in computer science**

Address Spaces

- **Linear address space:** Ordered set of contiguous non-negative integer addresses:

    $$\{0, 1, 2, 3 \dots \}$$

- **Virtual address space:** Set of $N = 2^n$ virtual addresses

    $$\{0, 1, 2, 3, \dots, N-1\}$$

- **Physical address space:** Set of $M = 2^m$ physical addresses

    $$\{0, 1, 2, 3, \dots, M-1\}$$

Why Virtual Memory (VM)?

- **Uses main memory efficiently**
  - Use DRAM as a cache for parts of a virtual address space

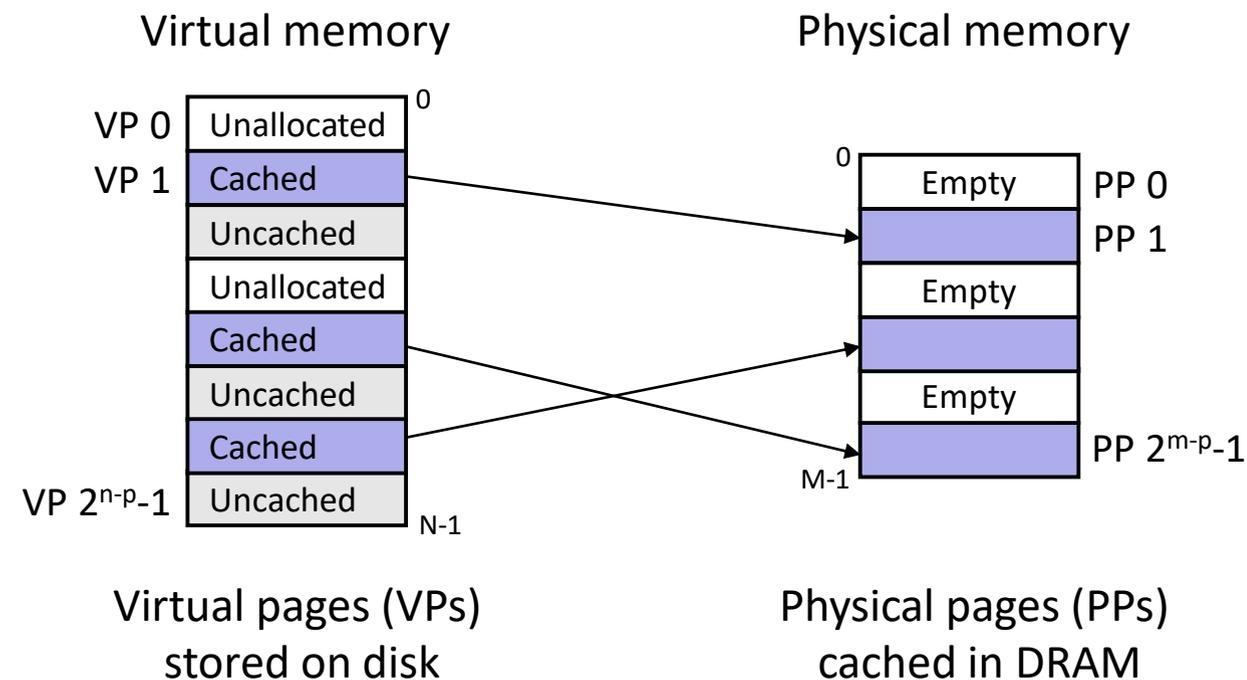- **Simplifies memory management**
  - Each process gets the same uniform linear address space

- **Isolates address spaces**
  - One process can't interfere with another's memory
  - User program cannot access privileged kernel information and code

VM as a Tool for Caching

- **Conceptually, *virtual memory* is an array of N contiguous bytes stored on disk.**

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

  - These cache blocks are called *pages* (size is $P = 2^p$ bytes)

Virtual memory   Physical memory

VP 0 | Unallocated | 0

VP 1 | Cached | → PP 0 Empty

| Uncached | PP 1

| Unallocated | Empty

| Cached |

| Uncached | Empty

| Cached | PP $2^{m-p}-1$

VP $2^{n-p}-1$ | Uncached | N-1   M-1

Virtual pages (VPs)
stored on disk

Physical pages (PPs)
cached in DRAM

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 KB, sometimes 4 MB
  - Fully associative
    - Any VP can be placed in any PP
    - Requires a "large" mapping function – different from cache memories
  - Highly sophisticated, expensive replacement algorithms
    - Too complicated and open-ended to be implemented in hardware
  - Write-back rather than write-through

Enabling Data Structure: Page Table

■ **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
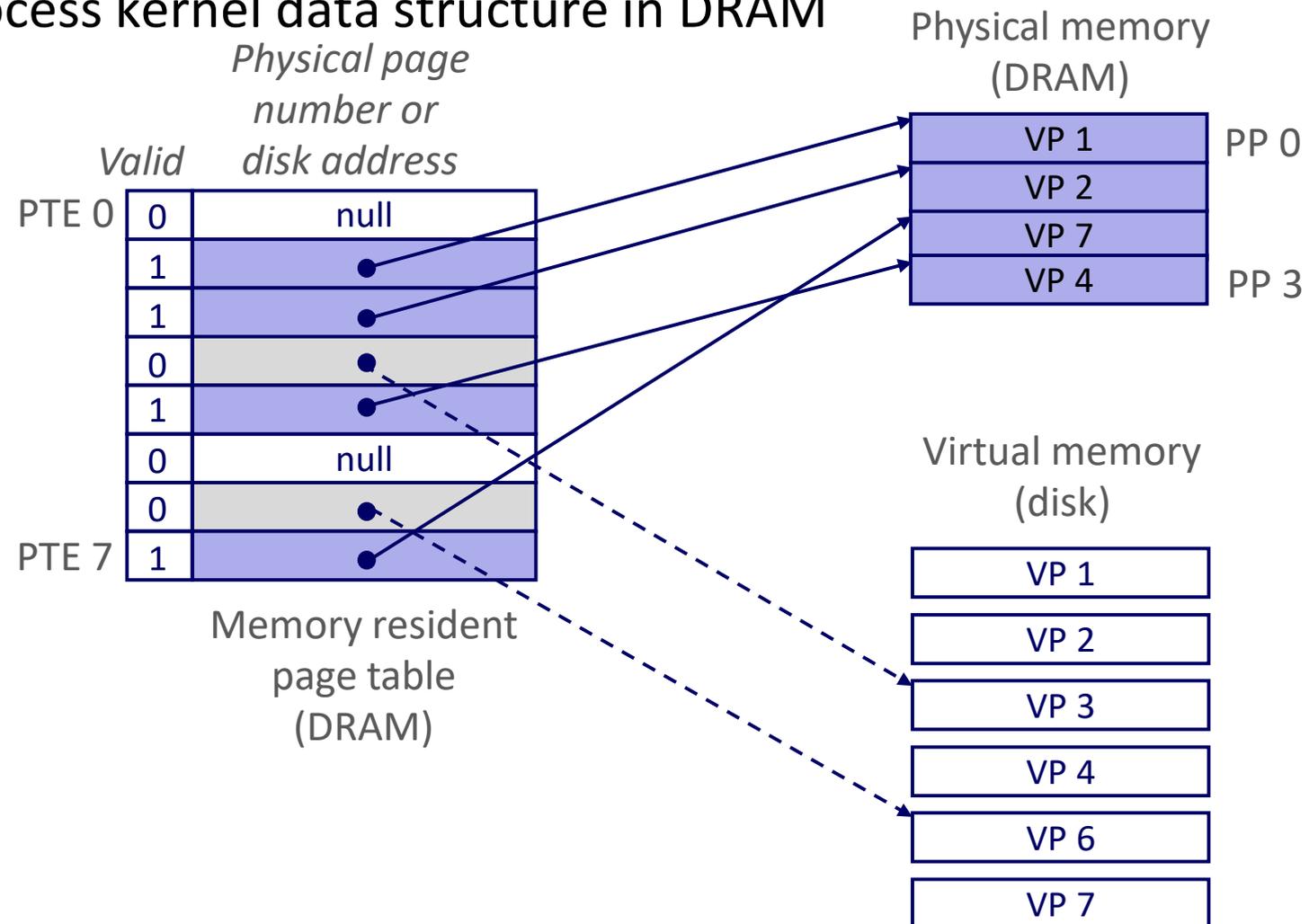
Page Hit

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

Carnegie Mellon University   11
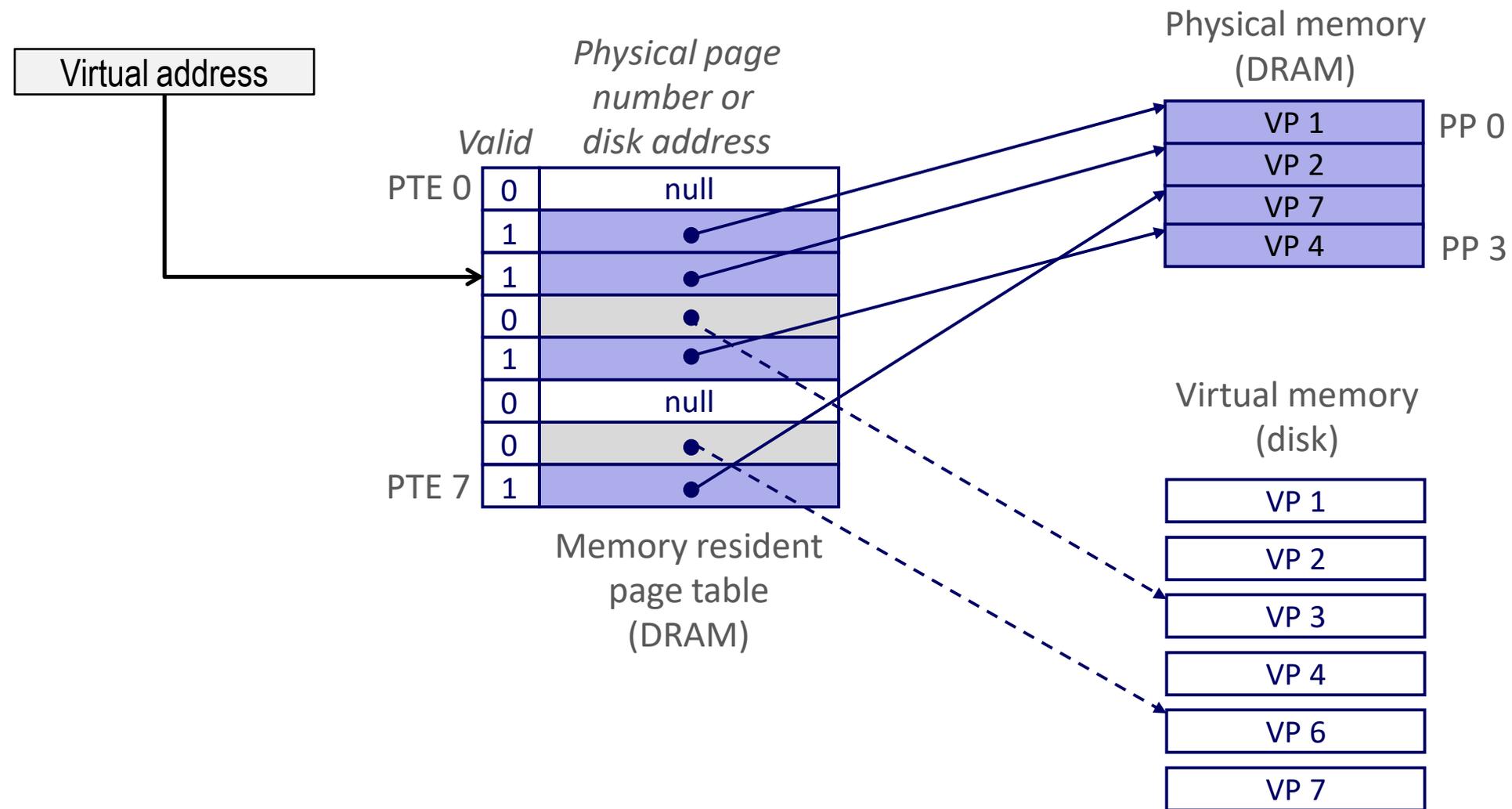
Page Fault

- ***Page fault:*** **reference to VM word that is not in physical memory (DRAM cache miss)**
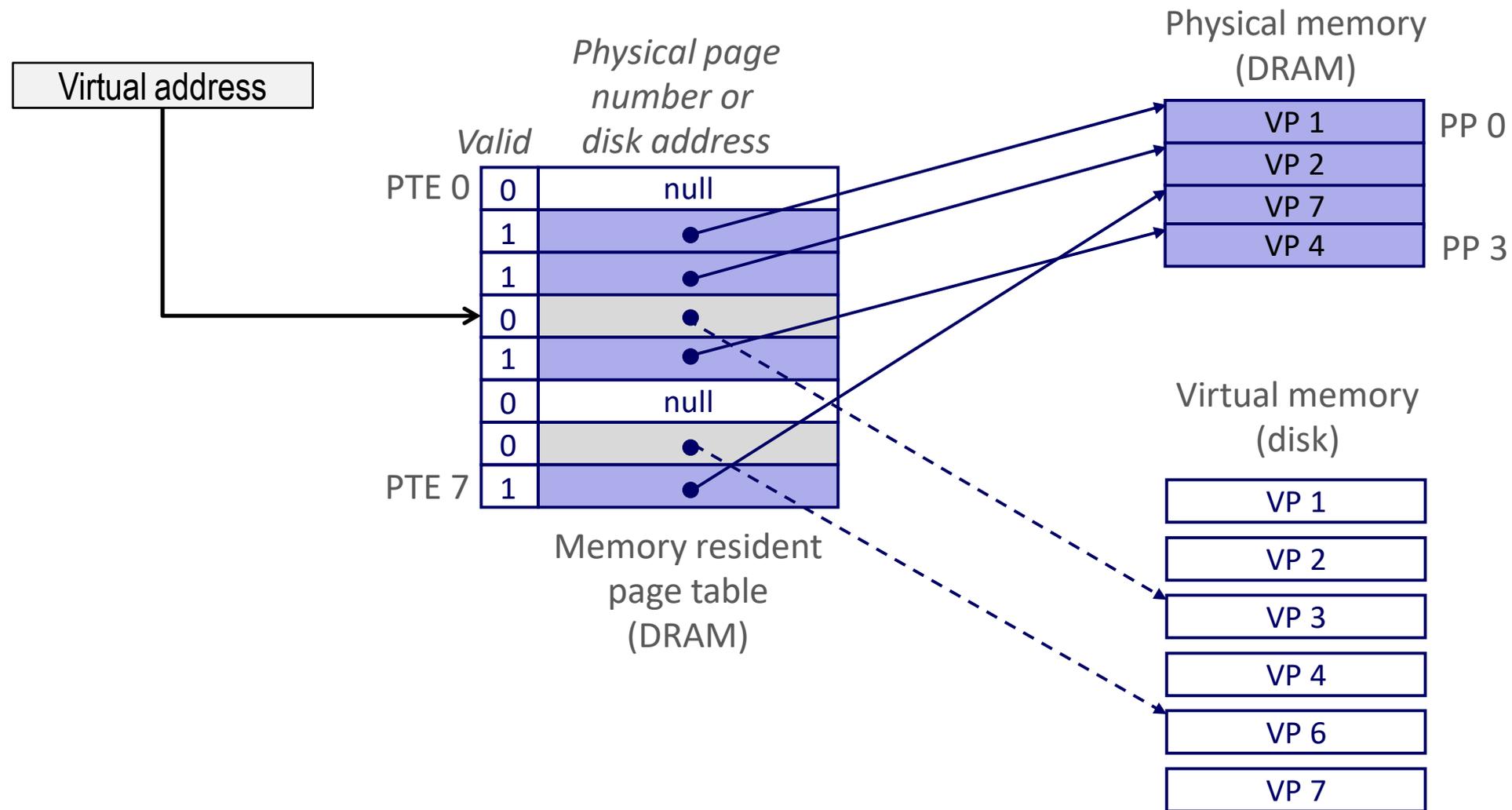
Handling Page Fault
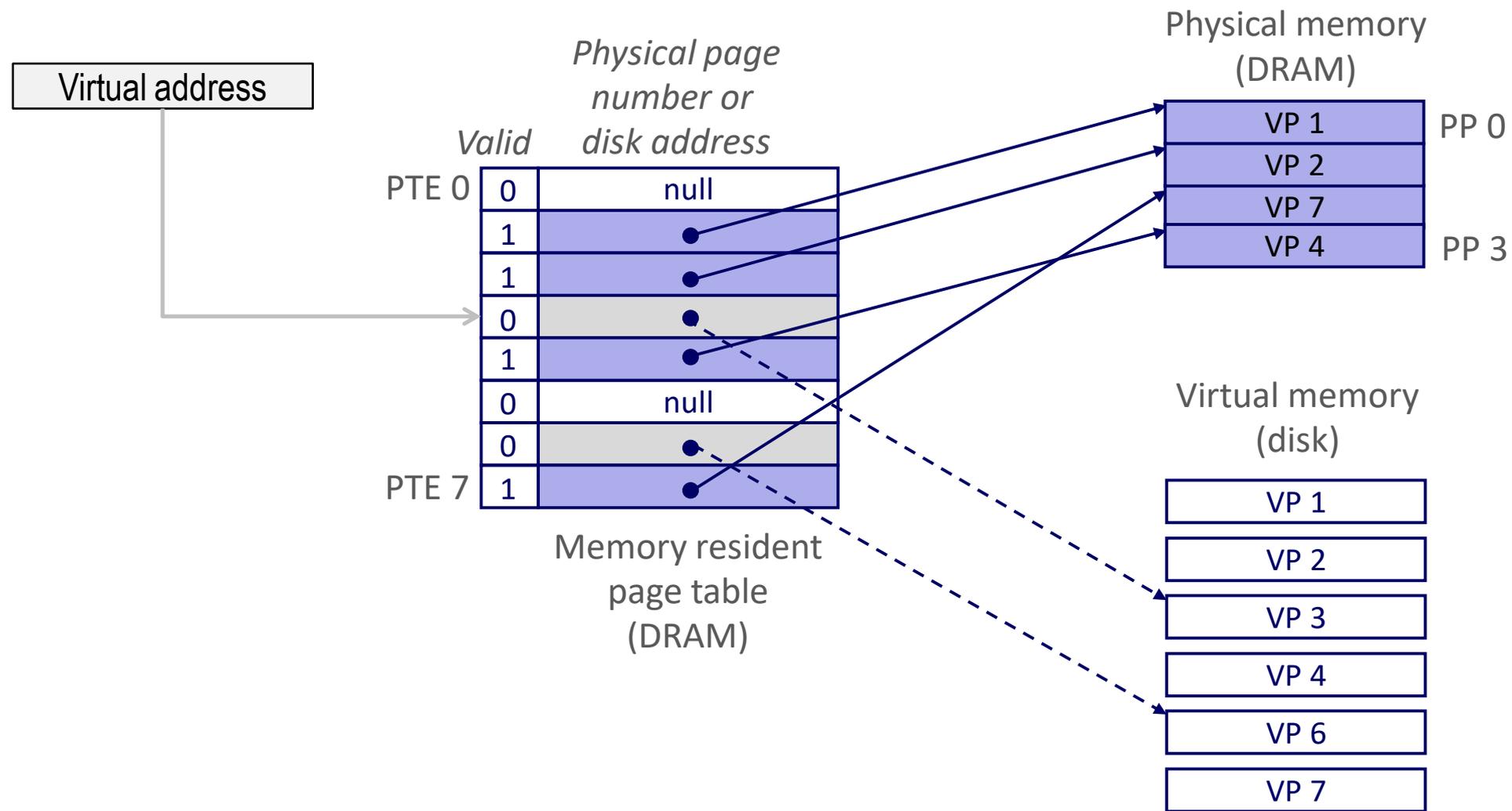
- Page miss causes page fault (an exception)

Carnegie Mellon University
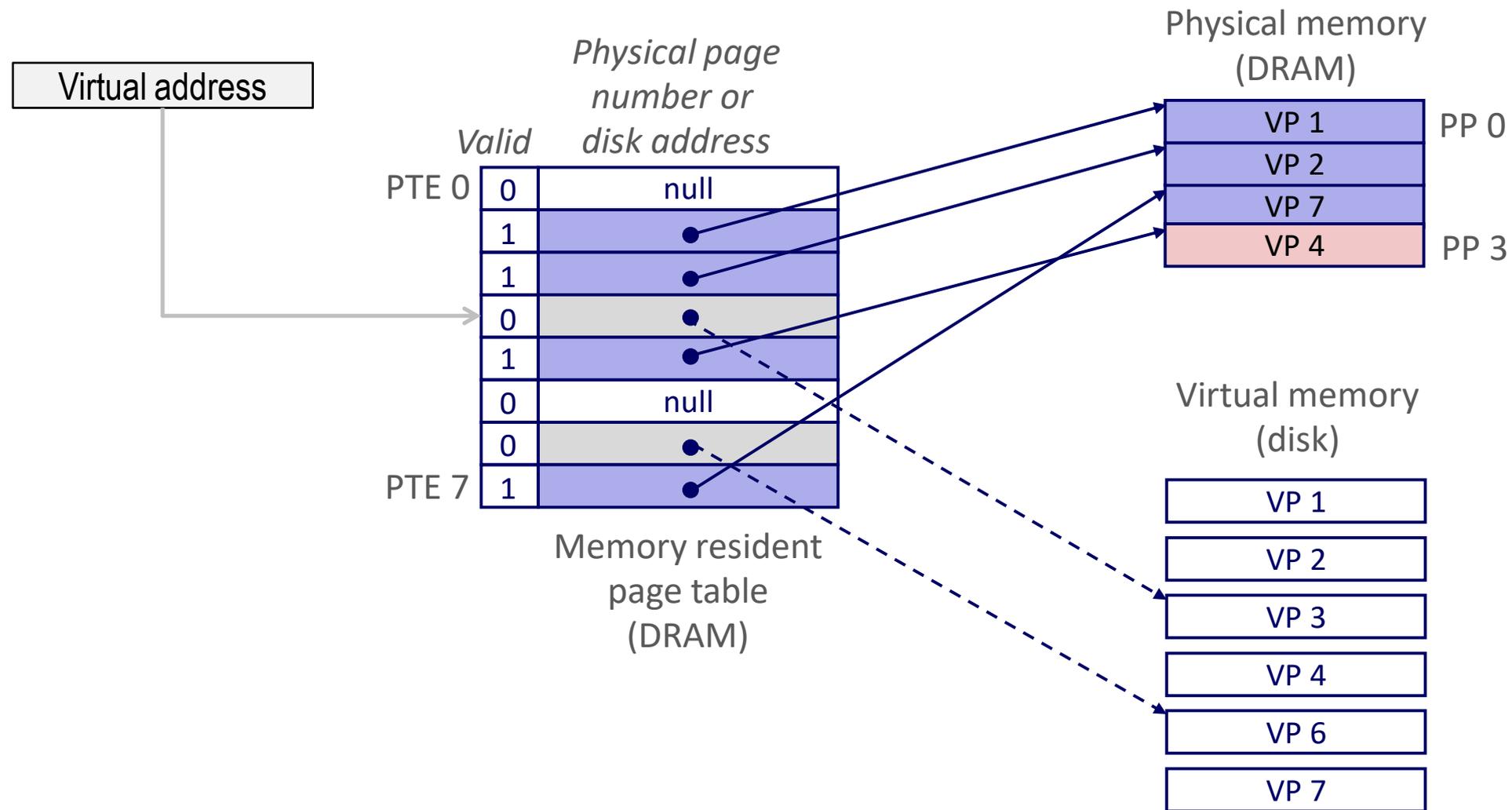
Handling Page Fault

- 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)

Carnegie Mellon University
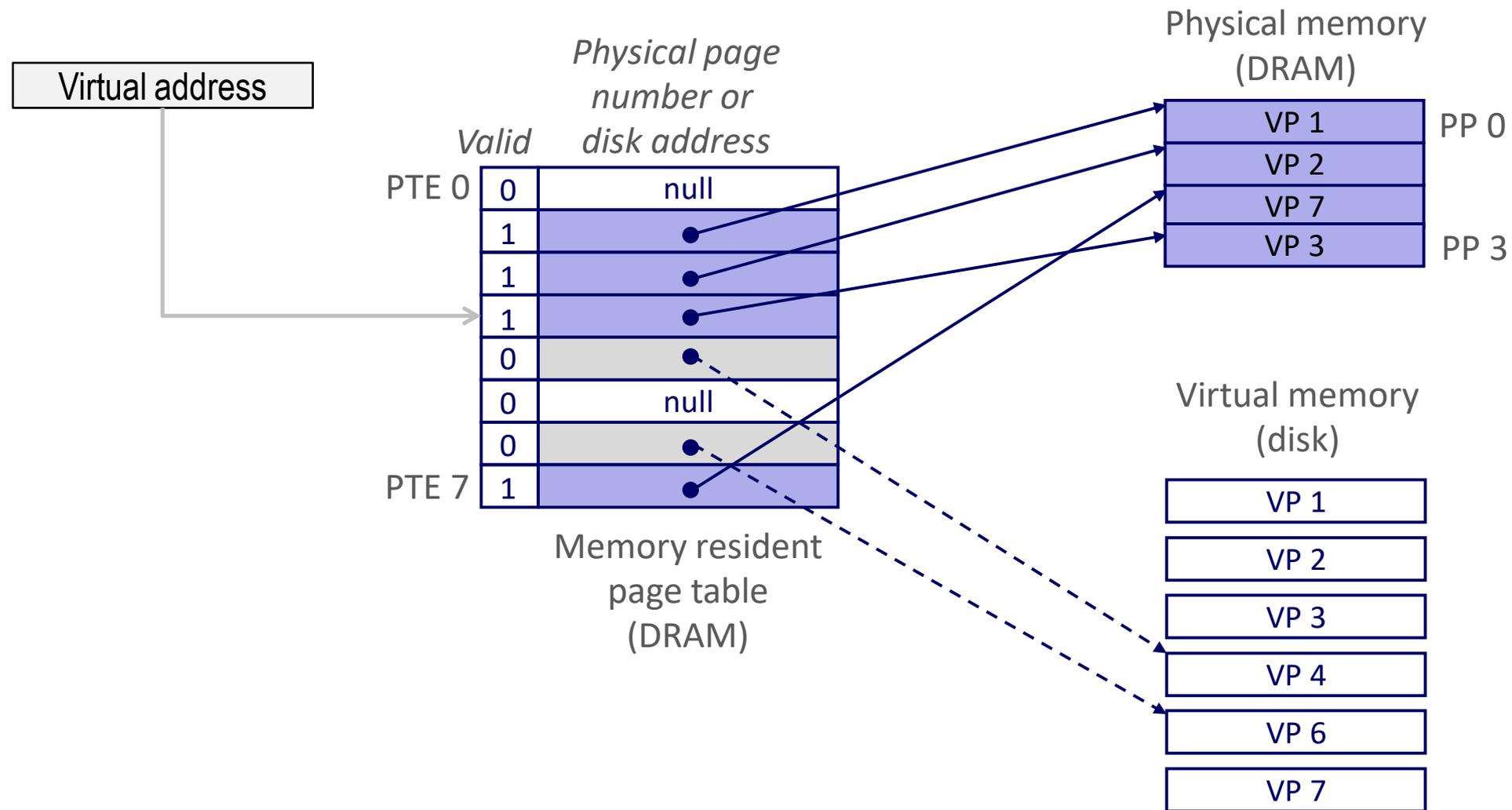
Handling Page Fault

- 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!

Virtual address

Physical page number or disk address

Valid

PTE 0

0	null
1	
1	
1	
0	
0	null
0	
1	

PTE 7

Memory resident
page table
(DRAM)

Physical memory
(DRAM)

VP 1
VP 2
VP 7
VP 3

PP 0

PP 3

Virtual memory
(disk)

VP 1
VP 2
VP 3
VP 4
VP 6
VP 7

Key point: Waiting until the miss to copy the page to
DRAM is known as *demand paging*

Carnegie Mellon University
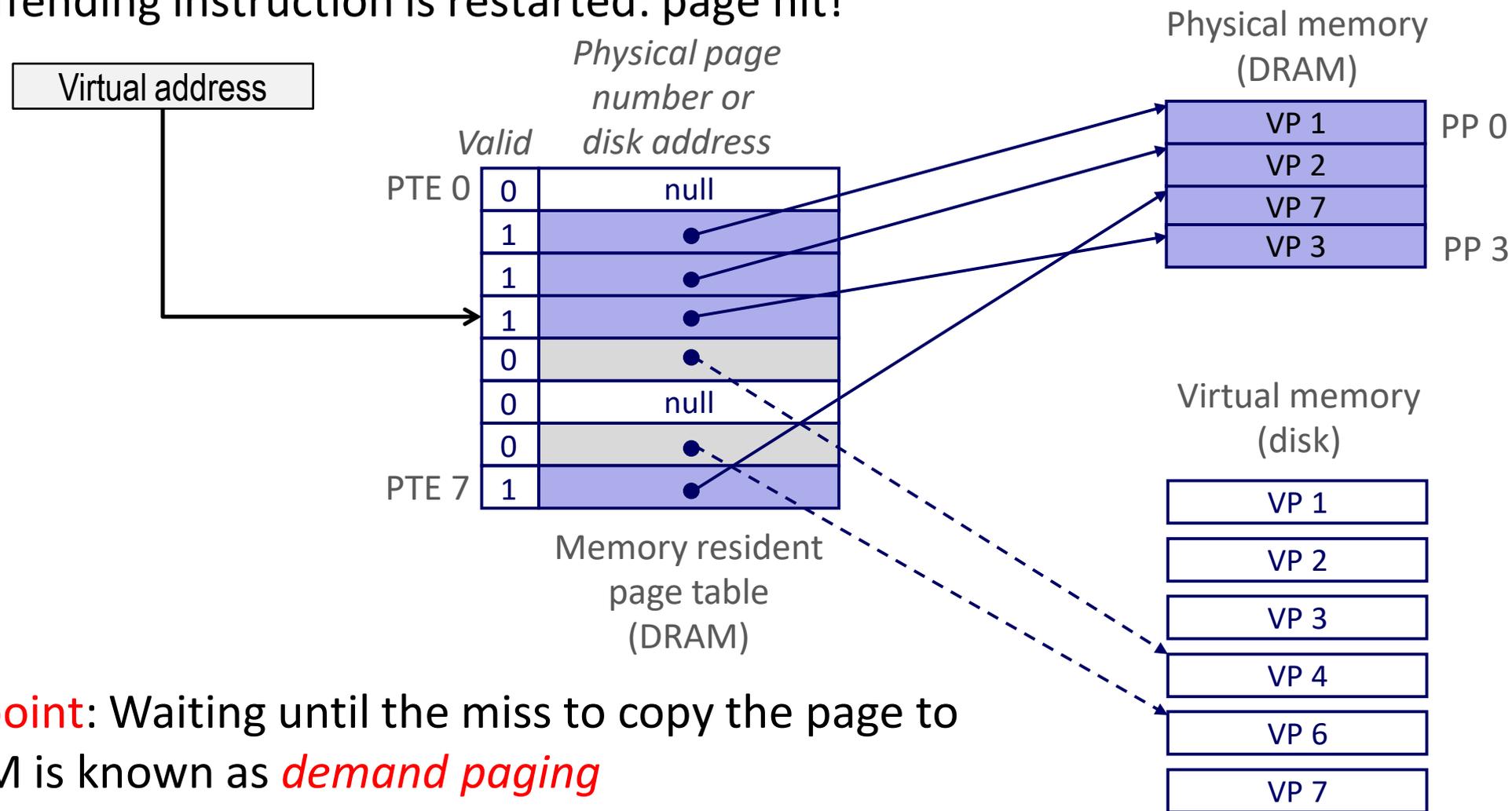
Allocating Pages

- **Allocating a new page (VP 5) of virtual memory.**

Carnegie Mellon University
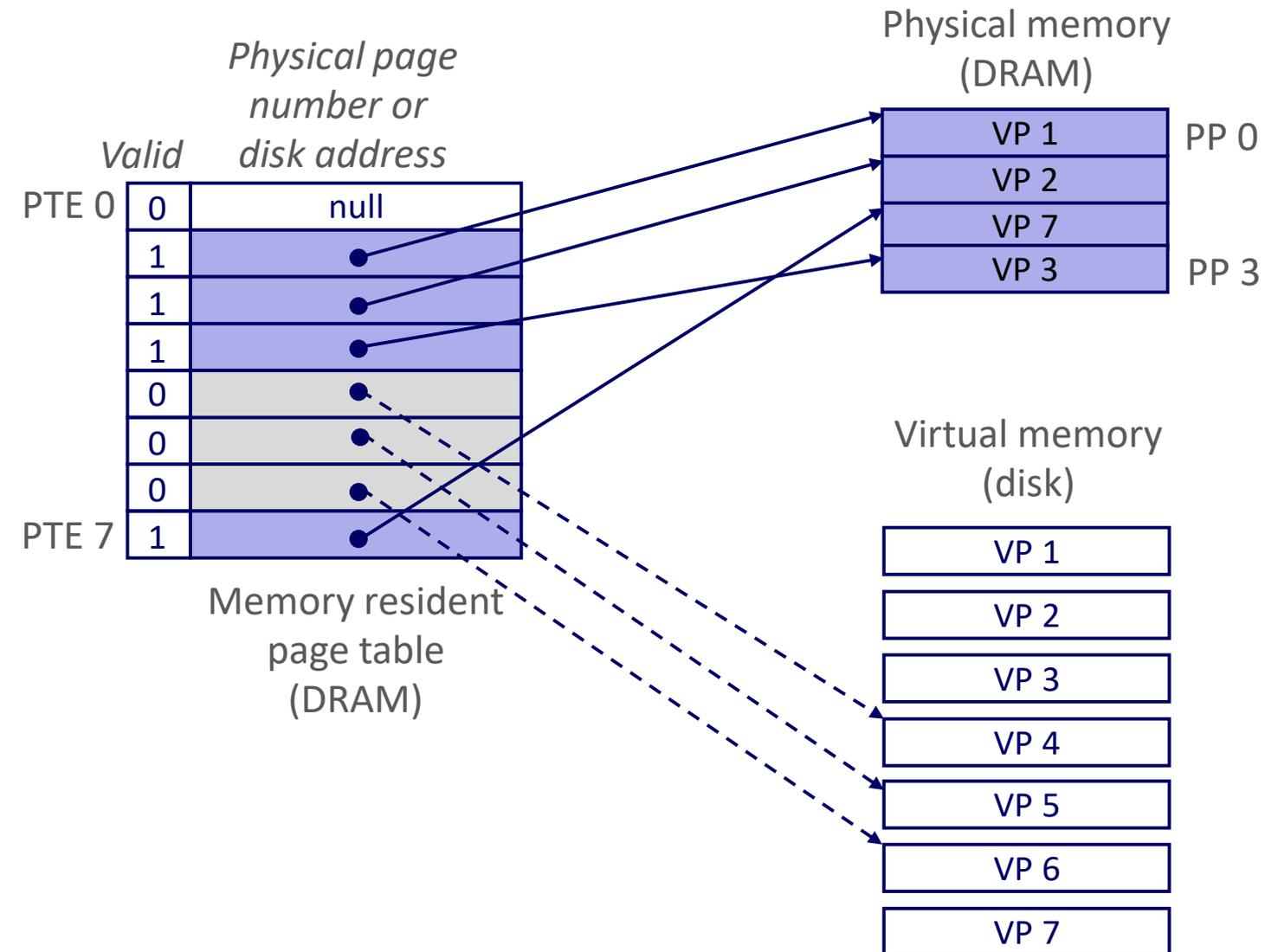
Locality to the Rescue Again!

- **Virtual memory seems terribly inefficient, but it 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

Carnegie Mellon University

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 can improve locality

Virtual Address Space for Process 1:

0

VP 1
VP 2
...

N-1

Address translation

0

PP 2

Physical Address Space (DRAM)

PP 6   **(e.g., read-only library code)**

PP 8

Virtual Address Space for Process 2:

0

VP 1
VP 2
...

N-1

...

M-1
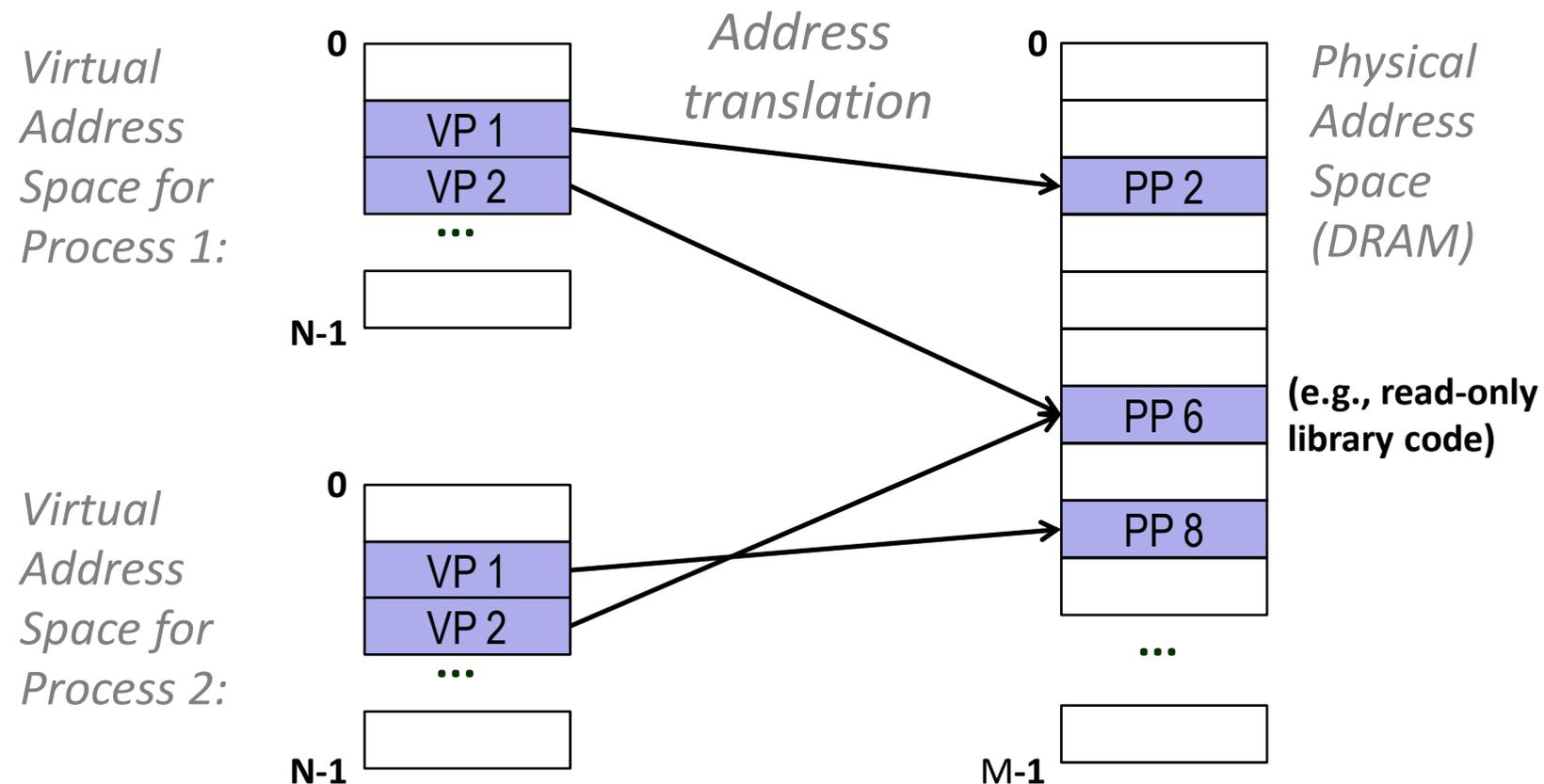
VM as a Tool for Memory Management

- **Simplifying 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)
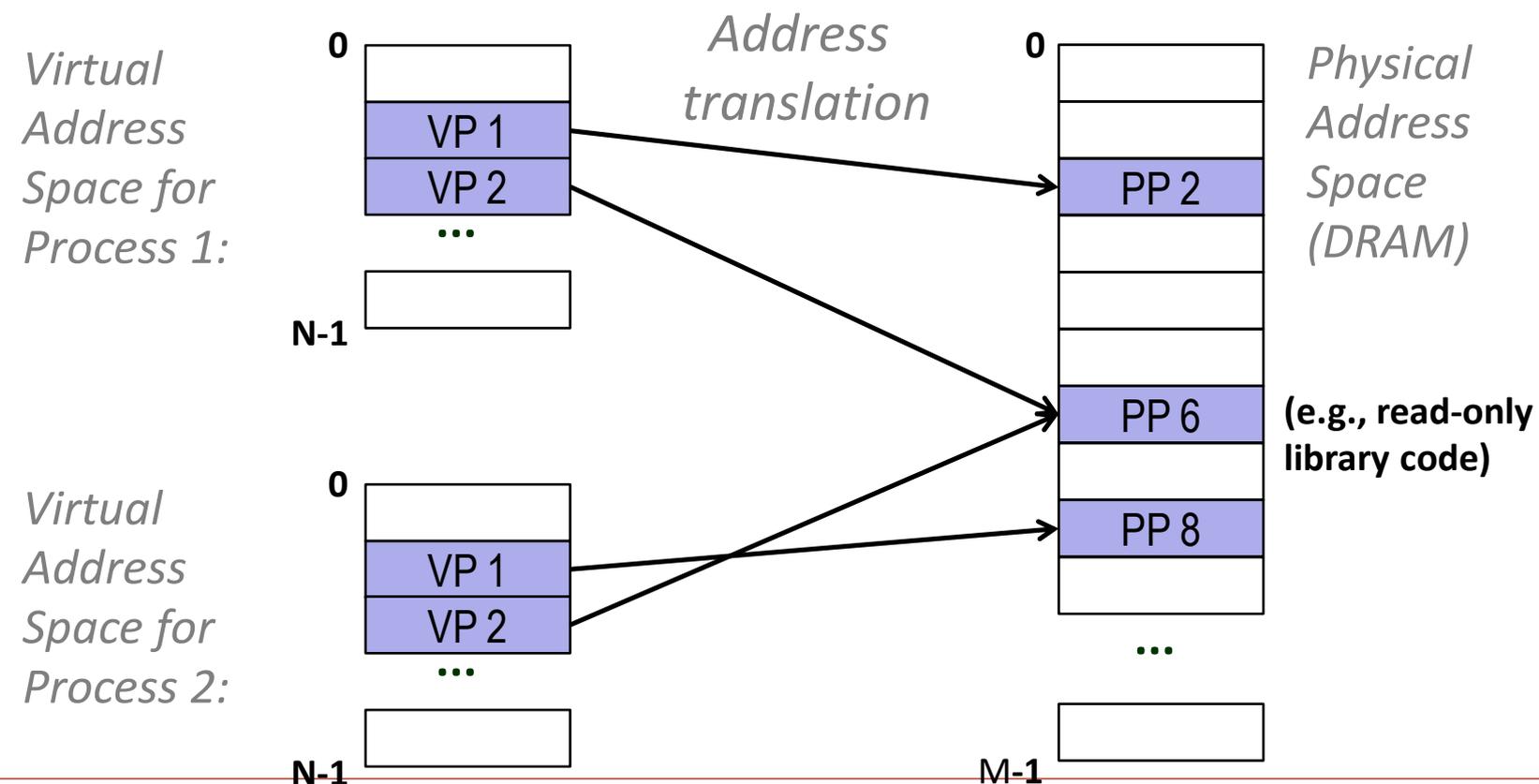
Simplifying Linking and Loading

Linking

- Each program has similar virtual address space

- Code, data, and heap always start at the same addresses.

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

Kernel virtual memory
User stack **(created at runtime)**
Memory-mapped region for **shared libraries**
Run-time heap **(created by malloc)**
Read/write segment **(.data,.bss)**
Read-only segment **(.init,.text,.rodata)**
Unused

Memory invisible to user code

%rsp (stack pointer)

brk

0x400000

0

Loaded from the executable file
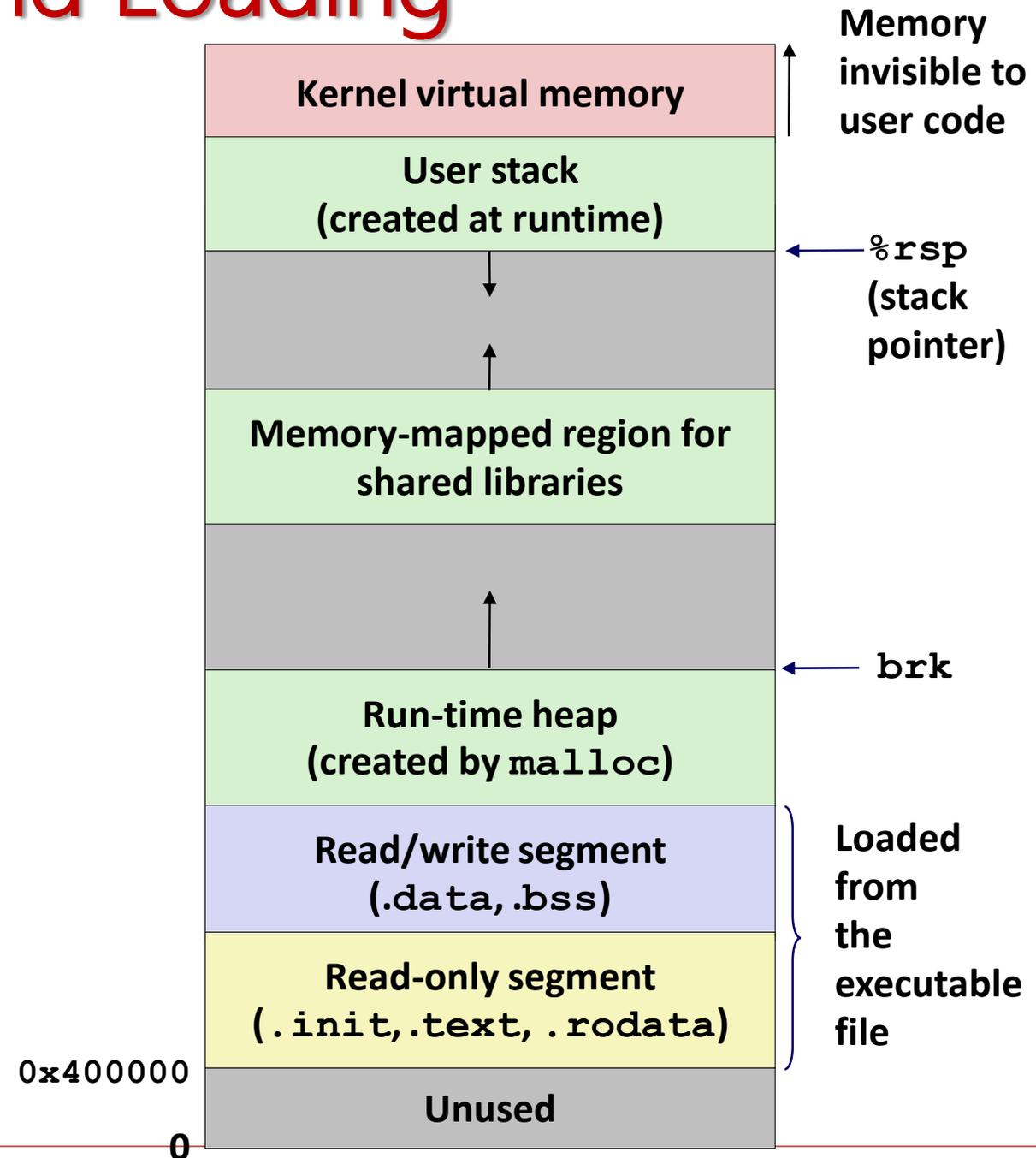
VM as a Tool for Memory Protection

- **Extend PTEs with permission bits**
- **MMU checks these bits on each access**

Physical Address Space

Process i:

	SUP	READ	WRITE	EXEC	Address
VP 0:	No	Yes	No	Yes	PP 6
VP 1:	No	Yes	Yes	Yes	PP 4
VP 2:	Yes	Yes	Yes	No	PP 2

Process j:

	SUP	READ	WRITE	EXEC	Address
VP 0:	No	Yes	No	Yes	PP 9
VP 1:	Yes	Yes	Yes	Yes	PP 6
VP 2:	No	Yes	Yes	Yes	PP 11

PP 2
PP 4
PP 6
PP 8
PP 9
PP 11
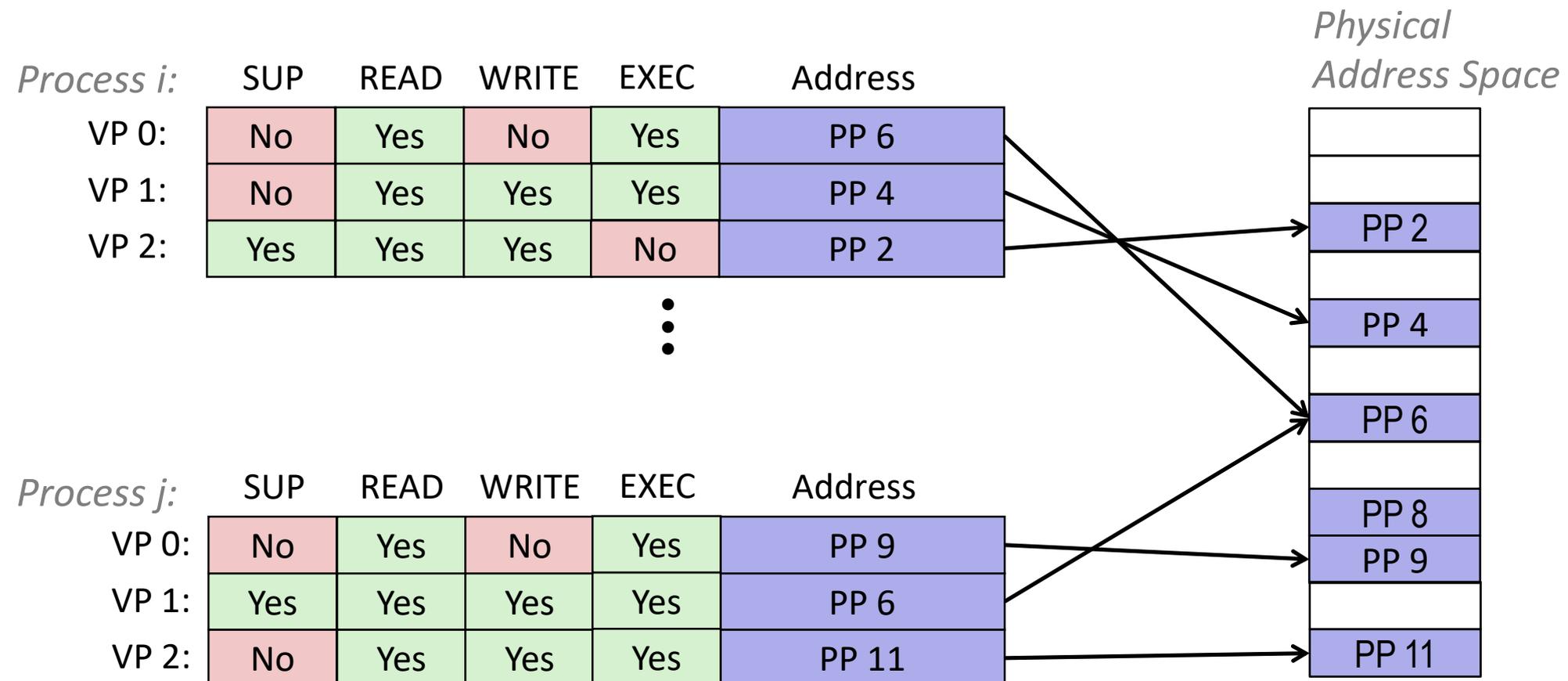
VM Address Translation

- **Virtual Address Space**
  - *V = {0, 1, …, N−1}*

- **Physical Address Space**
  - *P = {0, 1, …, M−1}*

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

Carnegie Mellon University

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

Address Translation With a Page Table

Virtual address

$n-1$	p $p-1$	0
Virtual page number (VPN)	Virtual page offset (VPO)	

Page table
base register
(PTBR)

Physical page table address for the current process

Page table

Valid   Physical page number (PPN)

Valid bit = 0:
Page not in memory
(page fault)

Valid bit = 1

Physical address

$m-1$	p $p-1$	0
Physical page number (PPN)	Physical page offset (PPO)	
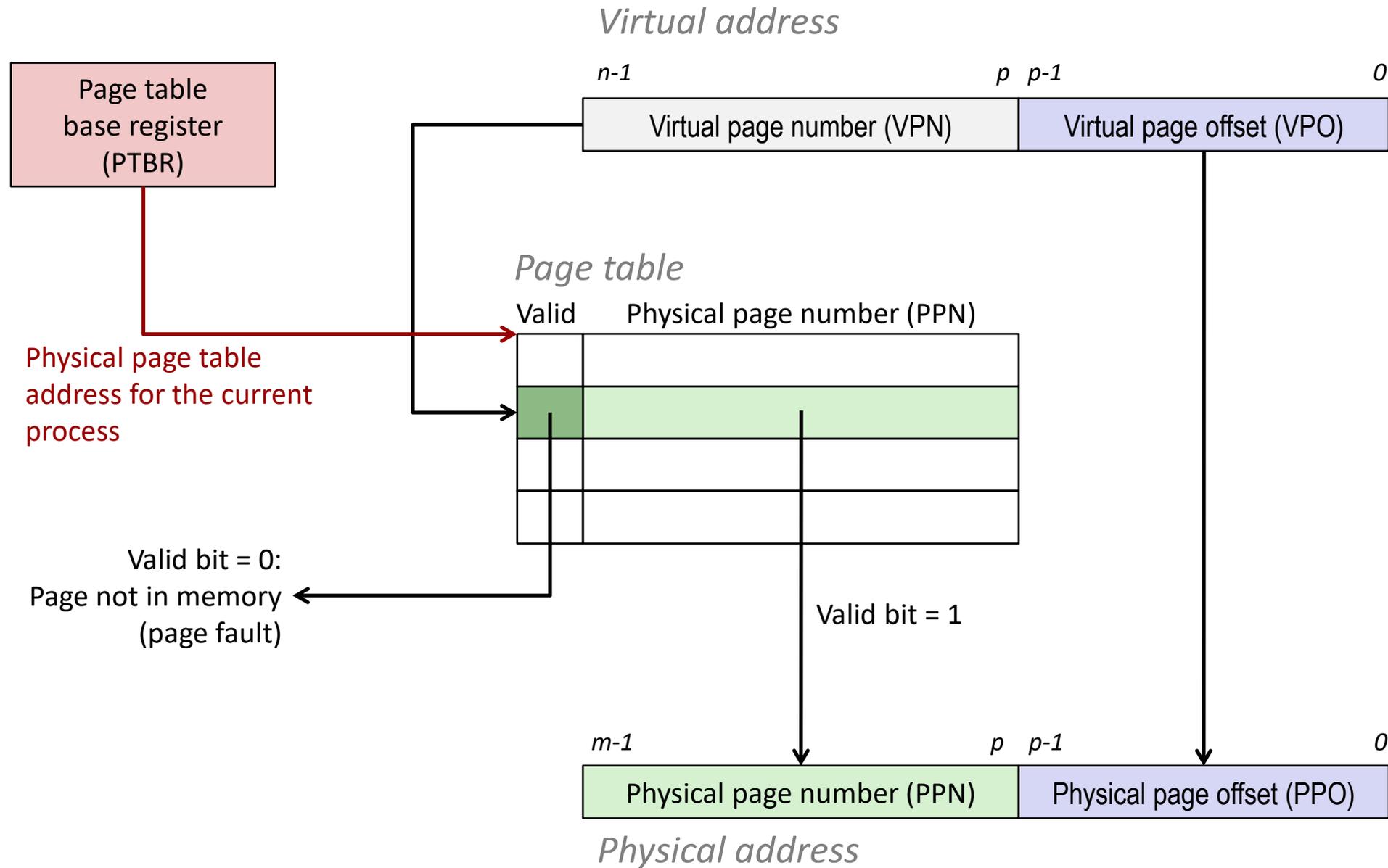
Address Translation: Page Hit
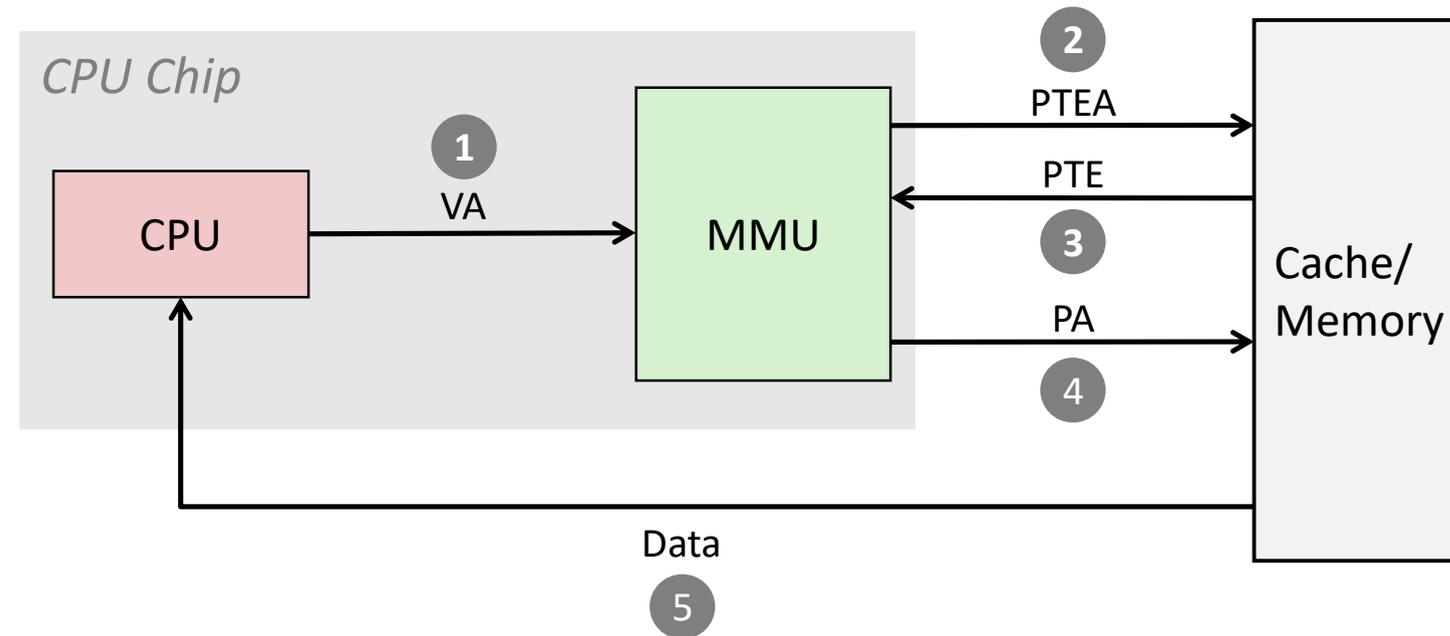
1) Processor sends virtual address to MMU

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
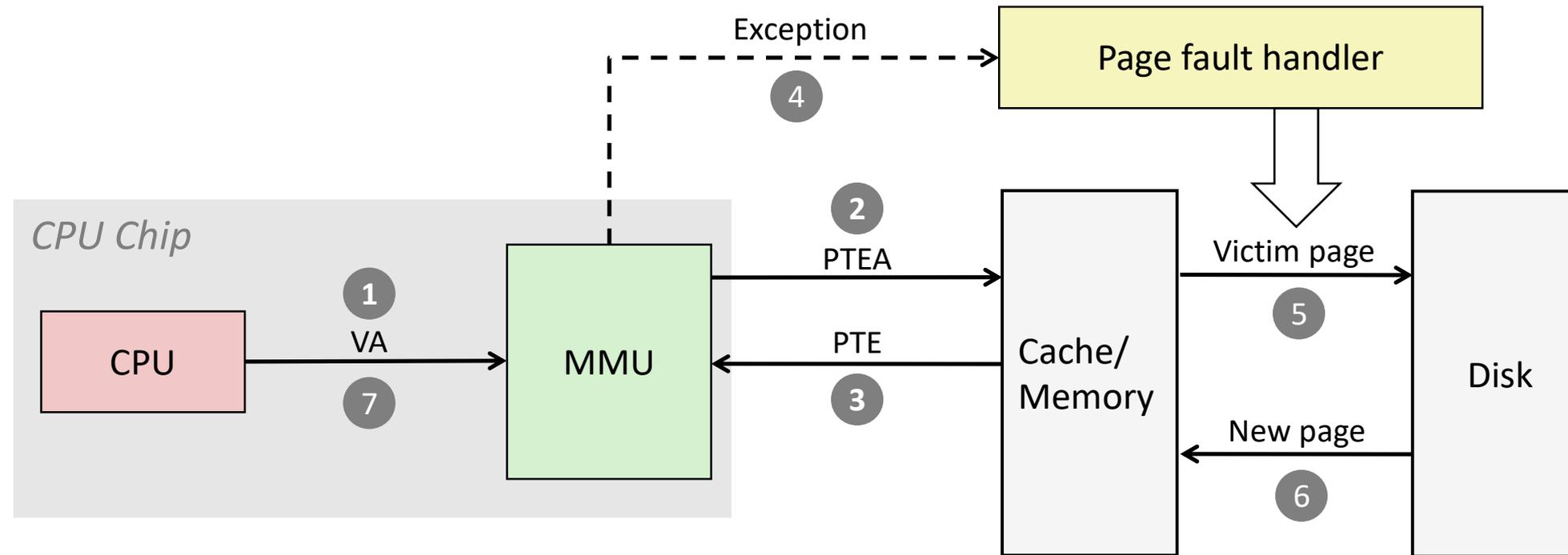
1) Processor sends virtual address to MMU

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

Carnegie Mellon University

Integrating VM and Cache
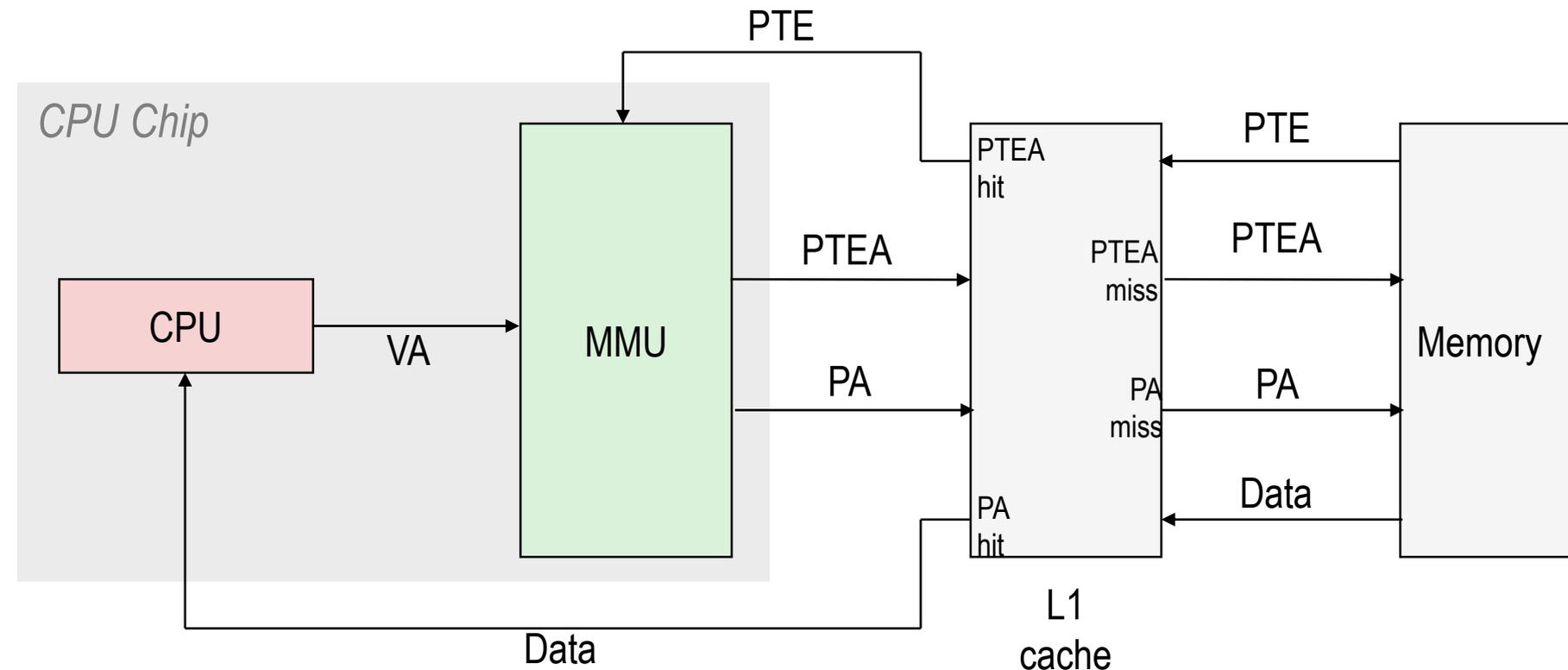
VA: virtual address, PA: physical address, PTE: page table entry, PTEA = PTE address

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 set-associative hardware cache in MMU
  - Maps virtual page numbers to physical page numbers
  - Contains complete page table entries for small number of pages

Carnegie Mellon University

Accessing the TLB

- **MMU uses the VPN portion of the virtual address to access the TLB:**

$T = 2^t$ sets

VPN

TLBT matches tag of line within set

n-1	p+t	p+t-1	p	p-1	0

TLB tag (TLBT)	TLB index (TLBI)	VPO

Set 0:
| v | tag | PTE | | v | tag | PTE |

Set 1:
| v | tag | PTE | | v | tag | PTE |

TLBI selects the set

⋮

Set T-1:
| v | tag | PTE | | v | tag | PTE |

Carnegie Mellon University   30

TLB Hit

CPU Chip

TLB

2 VPN

PTE **3**

1

CPU —VA→ MMU —PA→ Cache/ Memory

4

Data

5
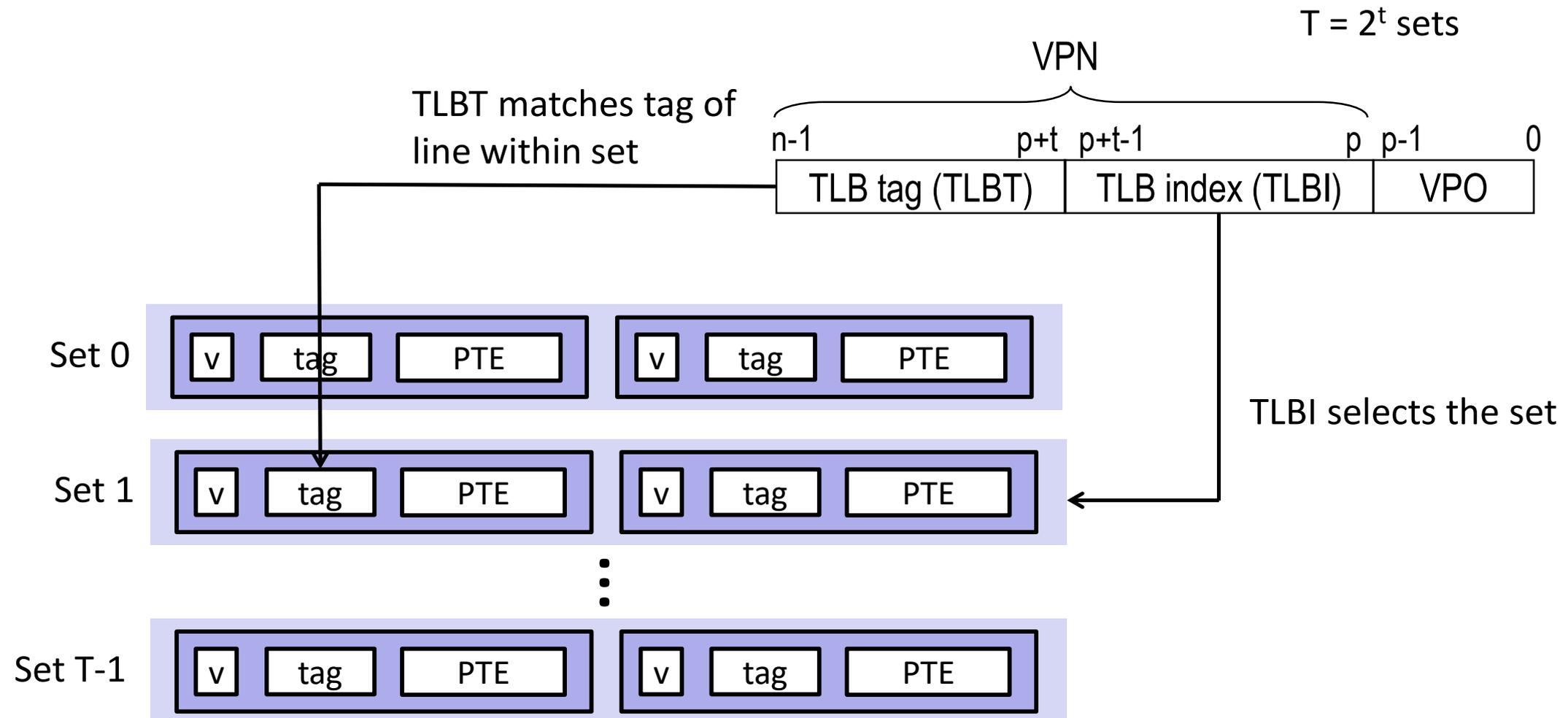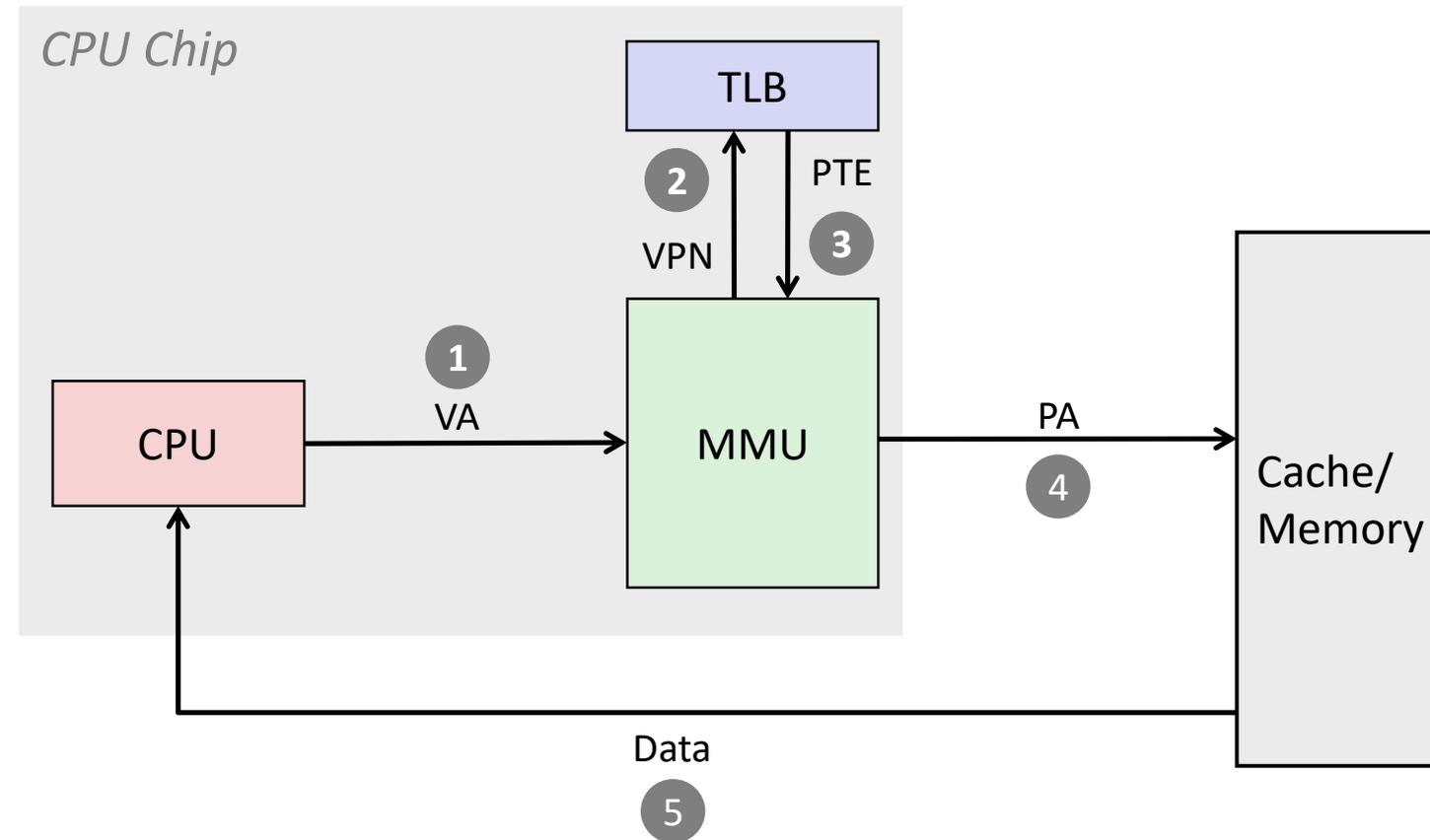
A TLB hit eliminates a memory access

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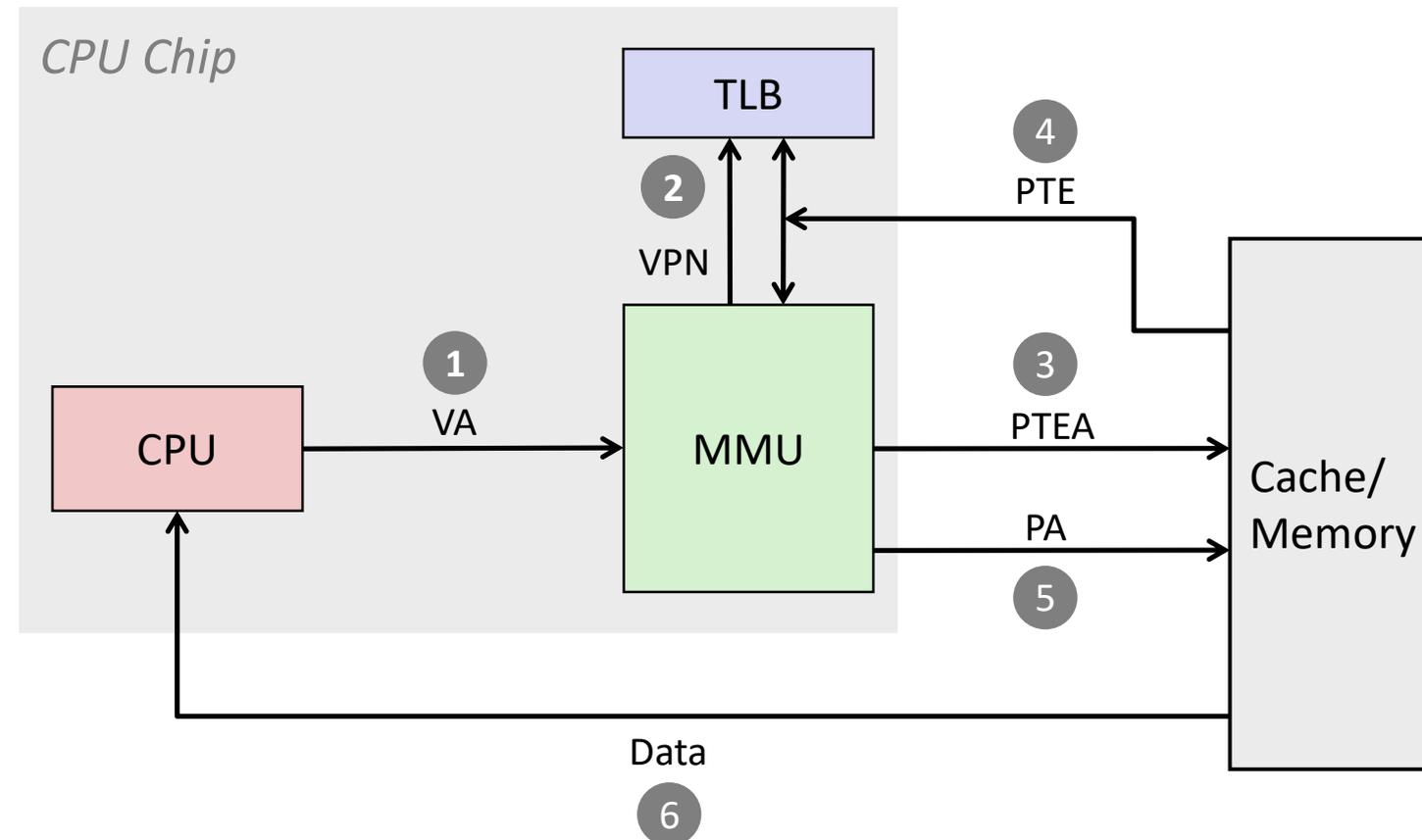
A TLB miss incurs an additional memory access (the PTE)
Fortunately, TLB misses are rare. Why?

Carnegie Mellon University

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!
    - $2^{48} * 2^{-12} * 2^{3} = 2^{39}$ bytes

- **Common solution: Multi-level page table**

- **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)

Level 2 Tables

Level 1 Table

Carnegie Mellon University
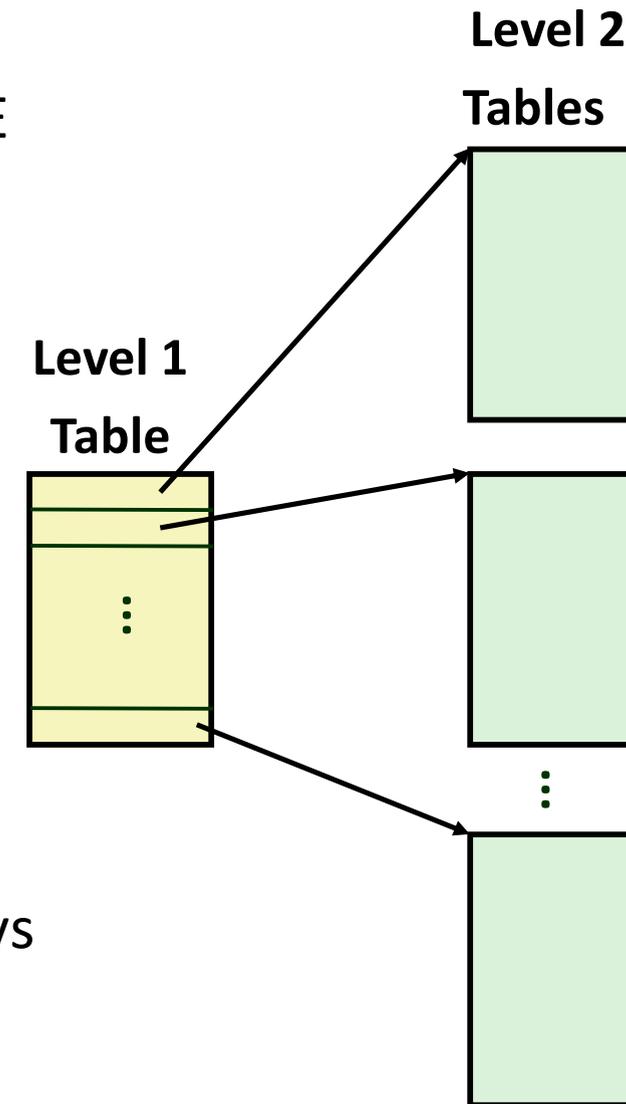
A Two-Level Page Table Hierarchy

Level 1 page table	Level 2 page tables	Virtual memory

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

PTE 0
...
PTE 1023

PTE 0
...
PTE 1023

1023 null PTEs
PTE 1023

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

Gap

1023 unallocated pages

VP 9215

0

2K allocated VM pages for code and data

6K unallocated VM pages

1023 unallocated pages

1 allocated VM page for the stack

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

Translating with a k-level Page Table

Carnegie Mellon University 35
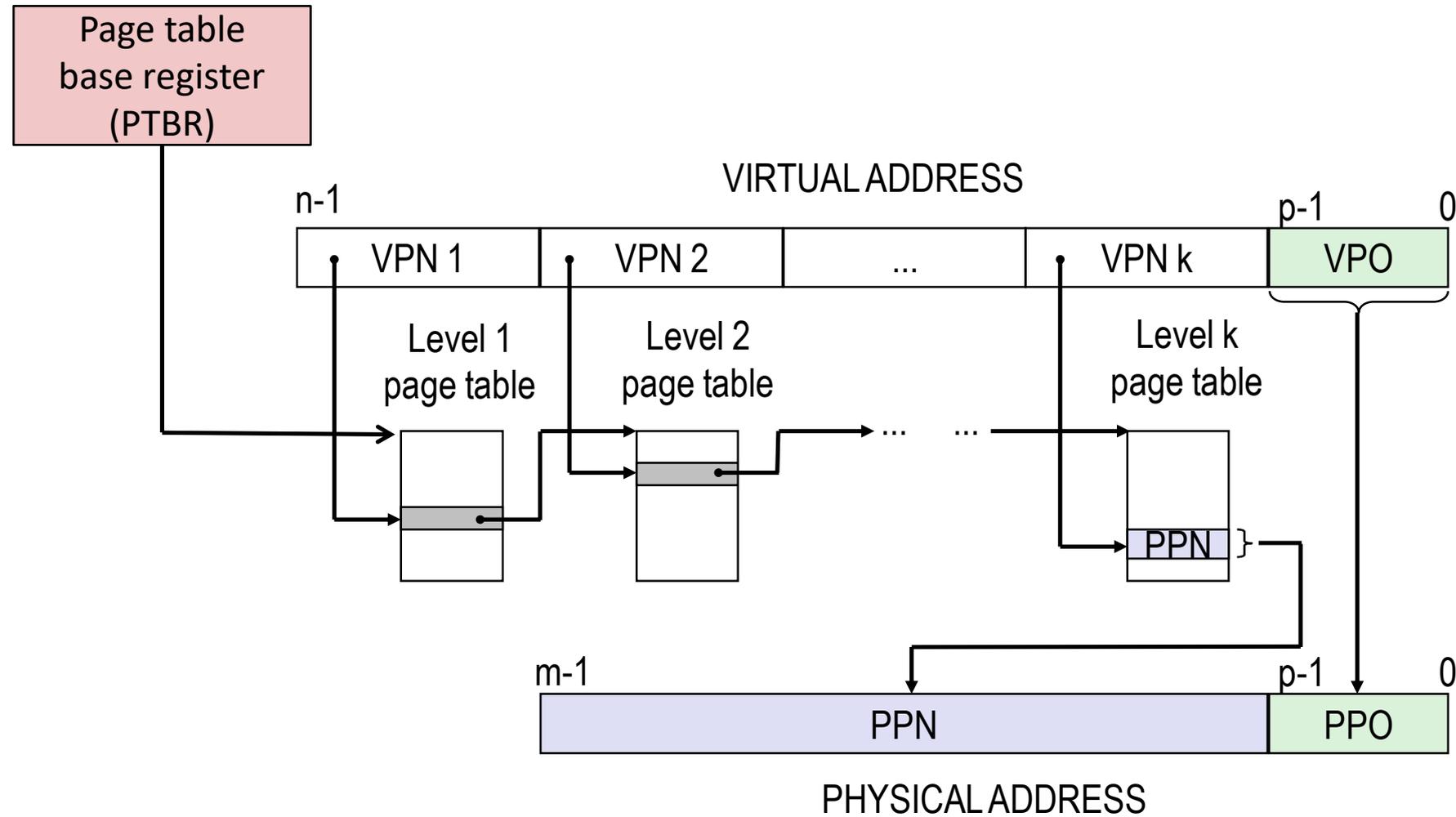
Summary

- **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

Carnegie Mellon University

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

Carnegie Mellon University

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

$\longleftarrow$ **VPN** $\longrightarrow$ $\longleftarrow$ **VPO** $\longrightarrow$

Virtual Page Number   Virtual Page Offset

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

$\longleftarrow$ **PPN** $\longrightarrow$ $\longleftarrow$ **PPO** $\longrightarrow$

Physical Page Number   Physical Page Offset

Carnegie Mellon University
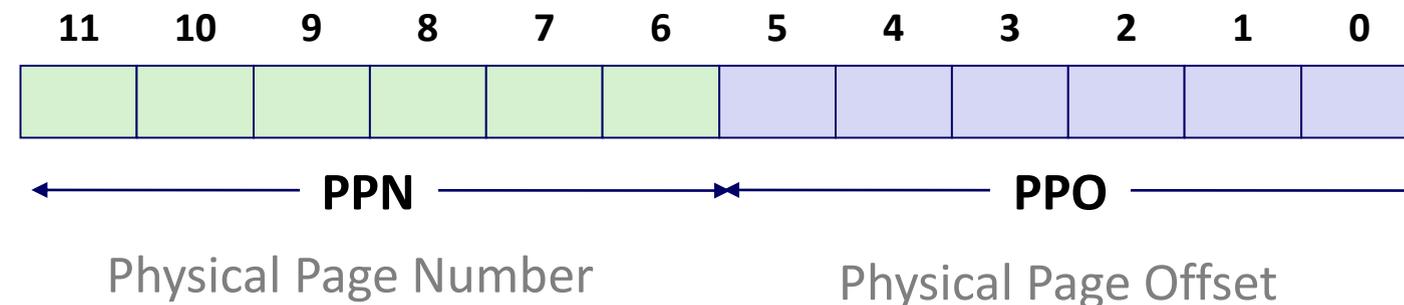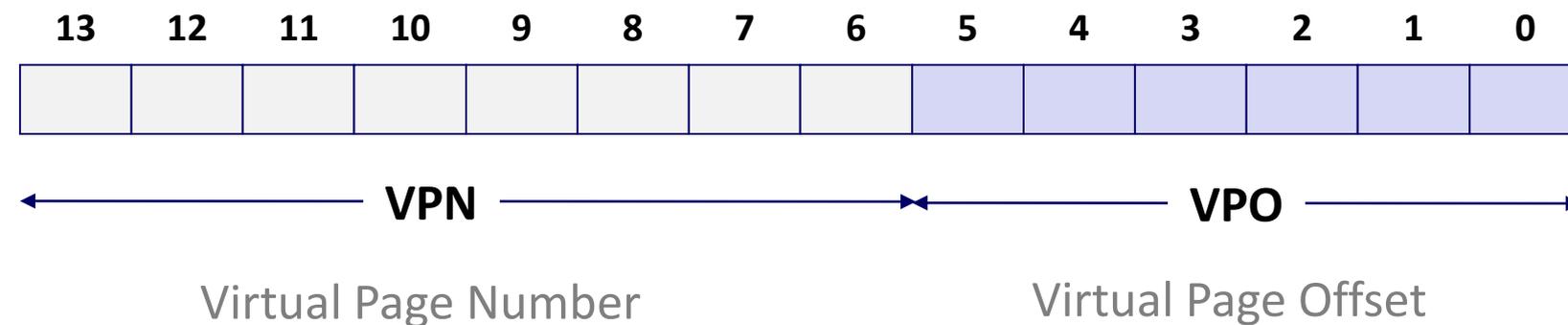
1. Simple Memory System TLB

- **16 entries**
- **4-way associative**
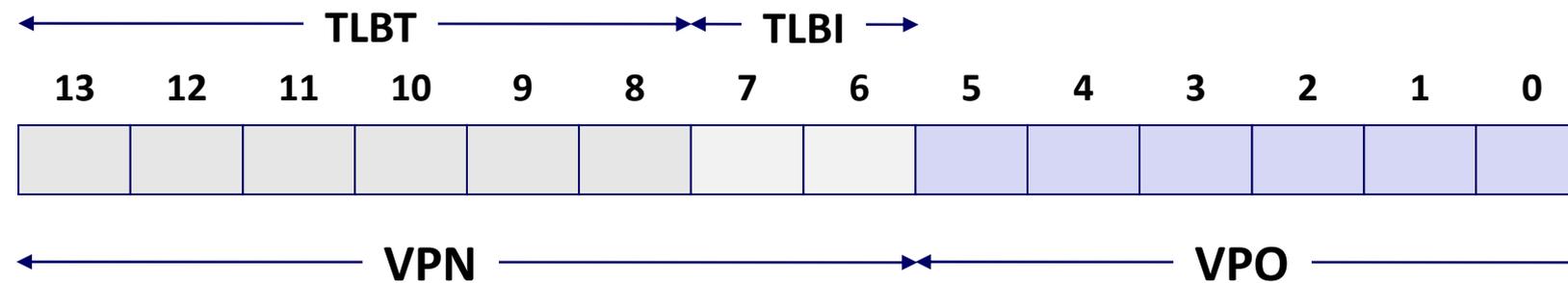
Set	Tag	PPN	Valid	Tag	PPN	Valid	Tag	PPN	Valid	Tag	PPN	Valid
0	03	–	0	09	0D	1	00	–	0	07	02	1
1	03	2D	1	02	–	0	04	–	0	0A	–	0
2	02	–	0	08	–	0	06	–	0	03	–	0
3	07	–	0	03	0D	1	0A	34	1	02	–	0

Carnegie Mellon University

2. Simple Memory System Page Table

Only show first 16 entries (out of 256)

VPN	PPN	Valid
00	28	1
01	–	0
02	33	1
03	02	1
04	–	0
05	16	1
06	–	0
07	–	0

VPN	PPN	Valid
08	13	1
09	17	1
0A	09	1
0B	–	0
0C	–	0
0D	2D	1
0E	11	1
0F	0D	1

Carnegie Mellon University

3. Simple Memory System Cache

- **16 lines, 4-byte block size**
- **Physically addressed**
- **Direct mapped**
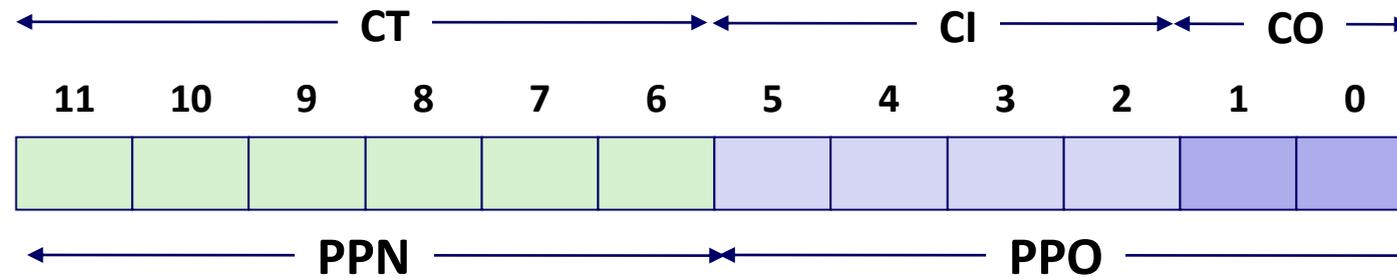
Idx	Tag	Valid	B0	B1	B2	B3
0	19	1	99	11	23	11
1	15	0	–	–	–	–
2	1B	1	00	02	04	08
3	36	0	–	–	–	–
4	32	1	43	6D	8F	09
5	0D	1	36	72	F0	1D
6	31	0	–	–	–	–
7	16	1	11	C2	DF	03

Idx	Tag	Valid	B0	B1	B2	B3
8	24	1	3A	00	51	89
9	2D	0	–	–	–	–
A	2D	1	93	15	DA	3B
B	0B	0	–	–	–	–
C	12	0	–	–	–	–
D	16	1	04	96	34	15
E	13	1	83	77	1B	D3
F	14	0	–	–	–	–

Carnegie Mellon University

Address Translation Example #1

Virtual Address: 0x03D4

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

VPN ← → | VPO ← →
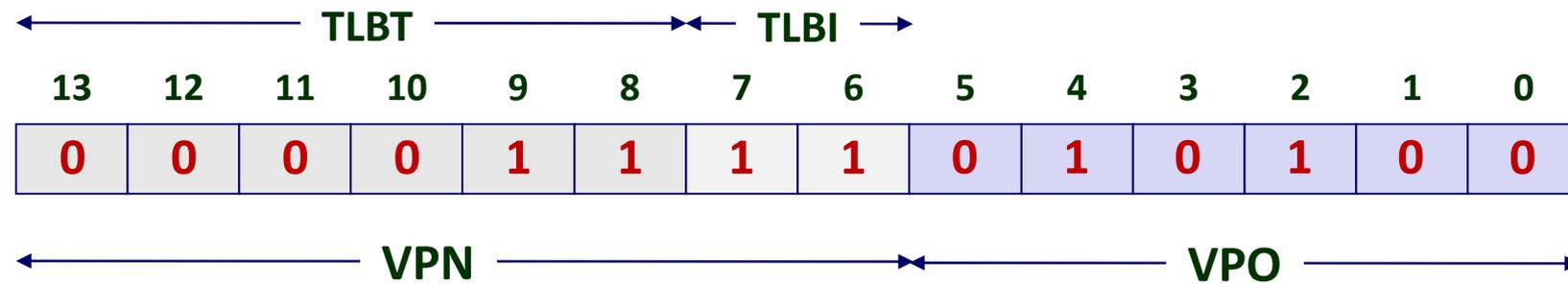
VPN **0x0F**   TLBI **0x3**   TLBT **0x03**   TLB Hit? **Y**   Page Fault? **N**   PPN: **0x0D**

Physical Address

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

PPN ← → | PPO ← →
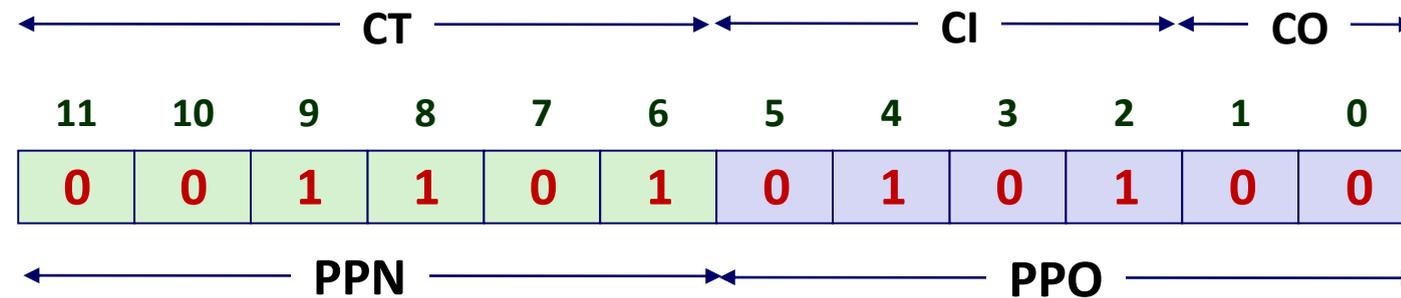
CO **0**   CI **0x5**   CT **0x0D**   Hit? **Y**   Byte: **0x36**

Address Translation Example #2

Virtual Address: 0x0020

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

TLBT ← → TLBI

VPN ← → VPO

VPN **0x00**   TLBI 0   TLBT **0x00**   TLB Hit? N   Page Fault? N   PPN: **0x28**

Physical Address

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

CT ← → CI ← → CO

PPN ← → PPO

CO **0**   CI **0x8**   CT **0x28**   Hit? N   Byte: **Mem**
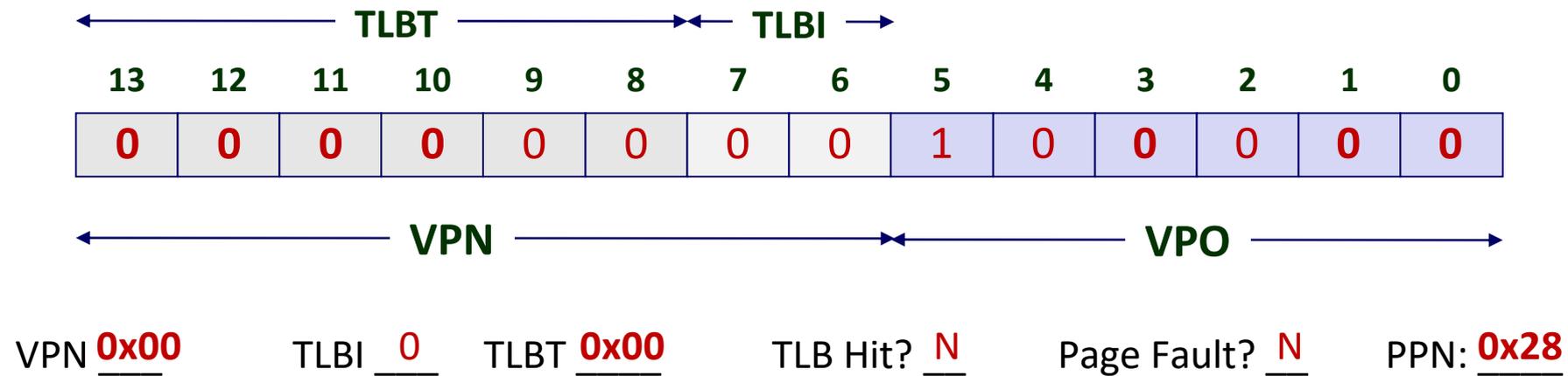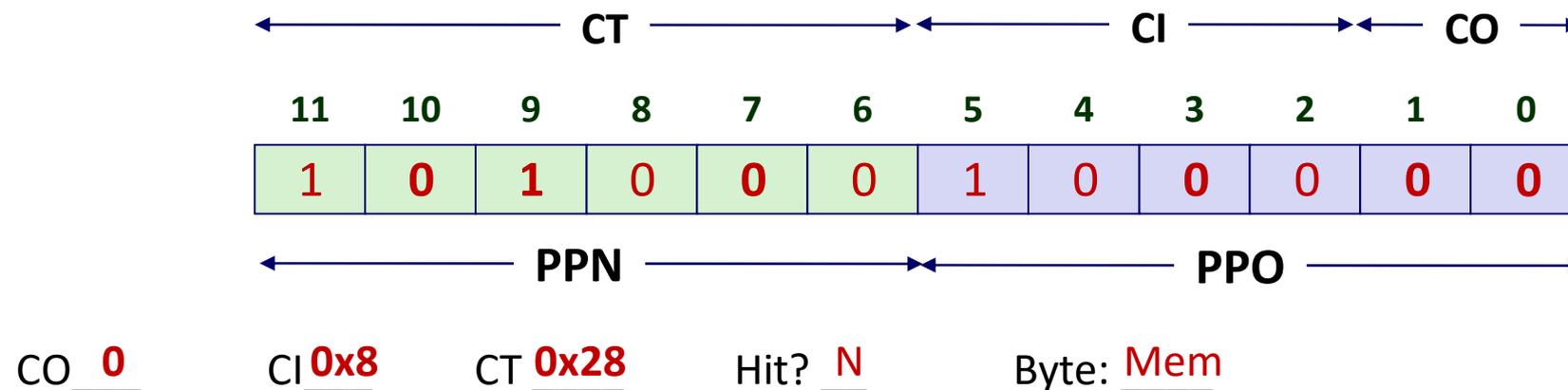
Address Translation Example #3

Virtual Address: 0x0020

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

VPN ——————— VPO

VPN **0x00**   TLBI **0**   TLBT **0x00**   TLB Hit? **N**   Page Fault? **N**   PPN: **0x28**

Physical Address

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

PPN ——————— PPO

CO **0**   CI **0x8**   CT **0x28**   Hit? **N**   Byte: **Mem**
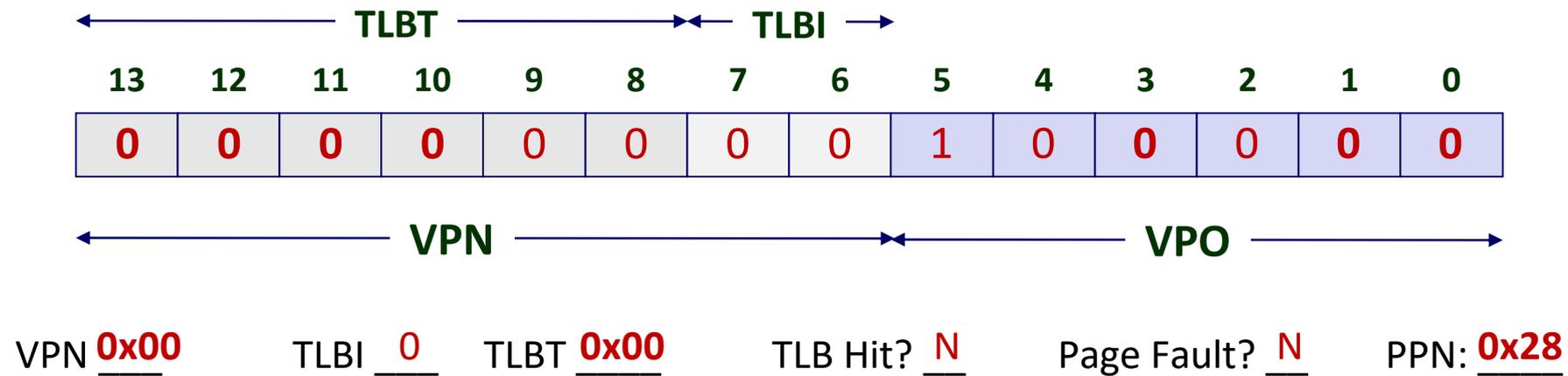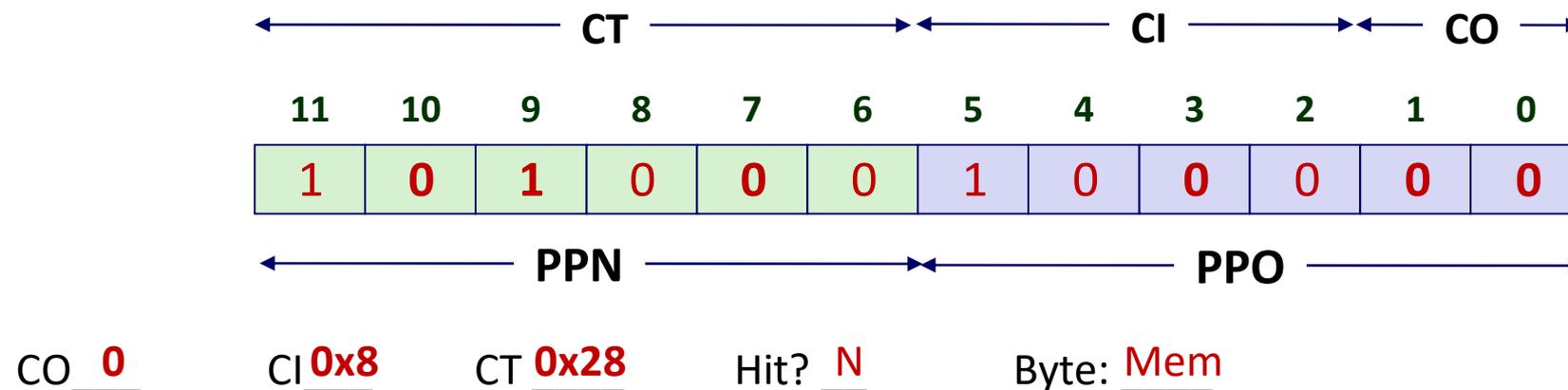
Intel Core i7 Memory System

Processor package
Core x4
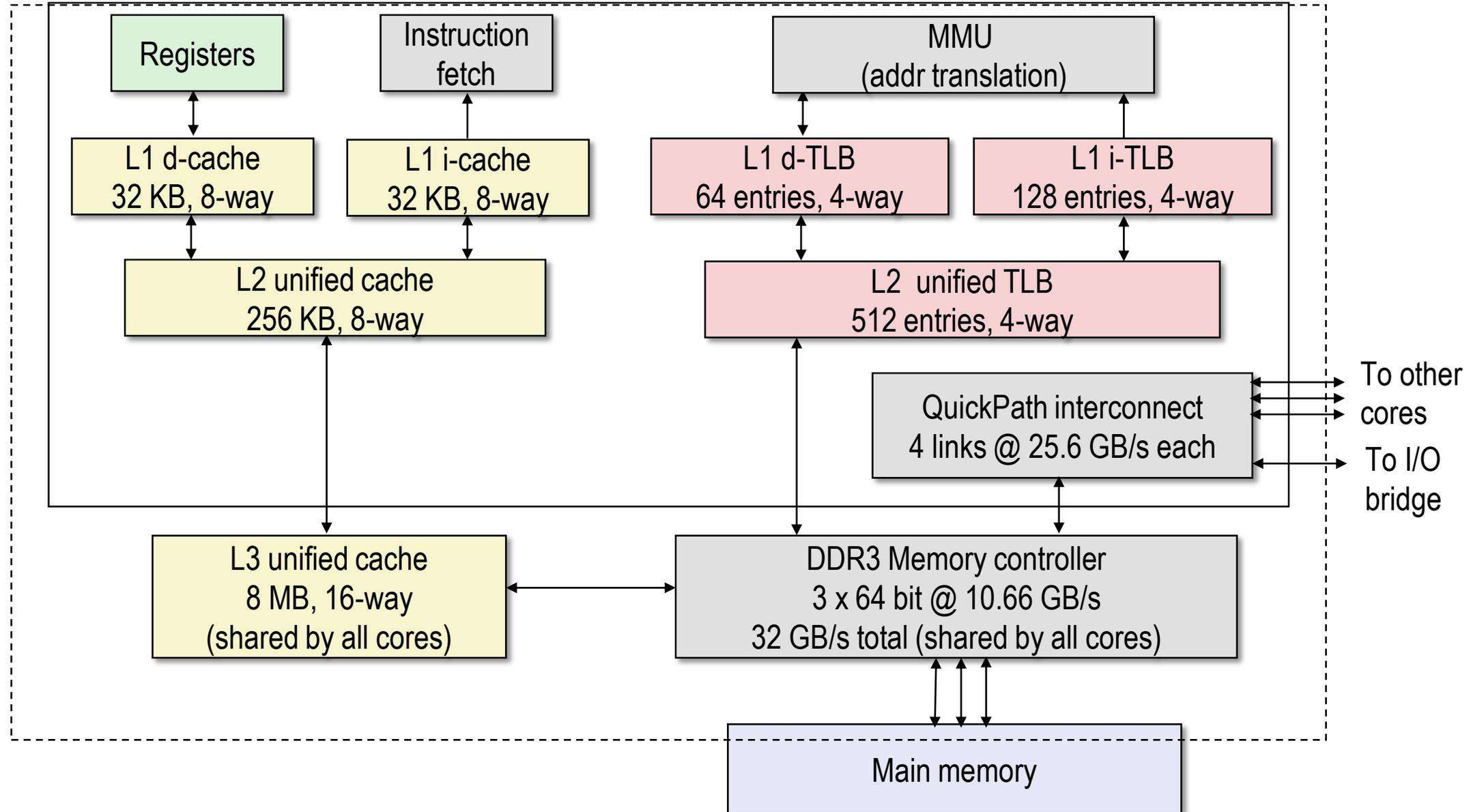
Carnegie Mellon University

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
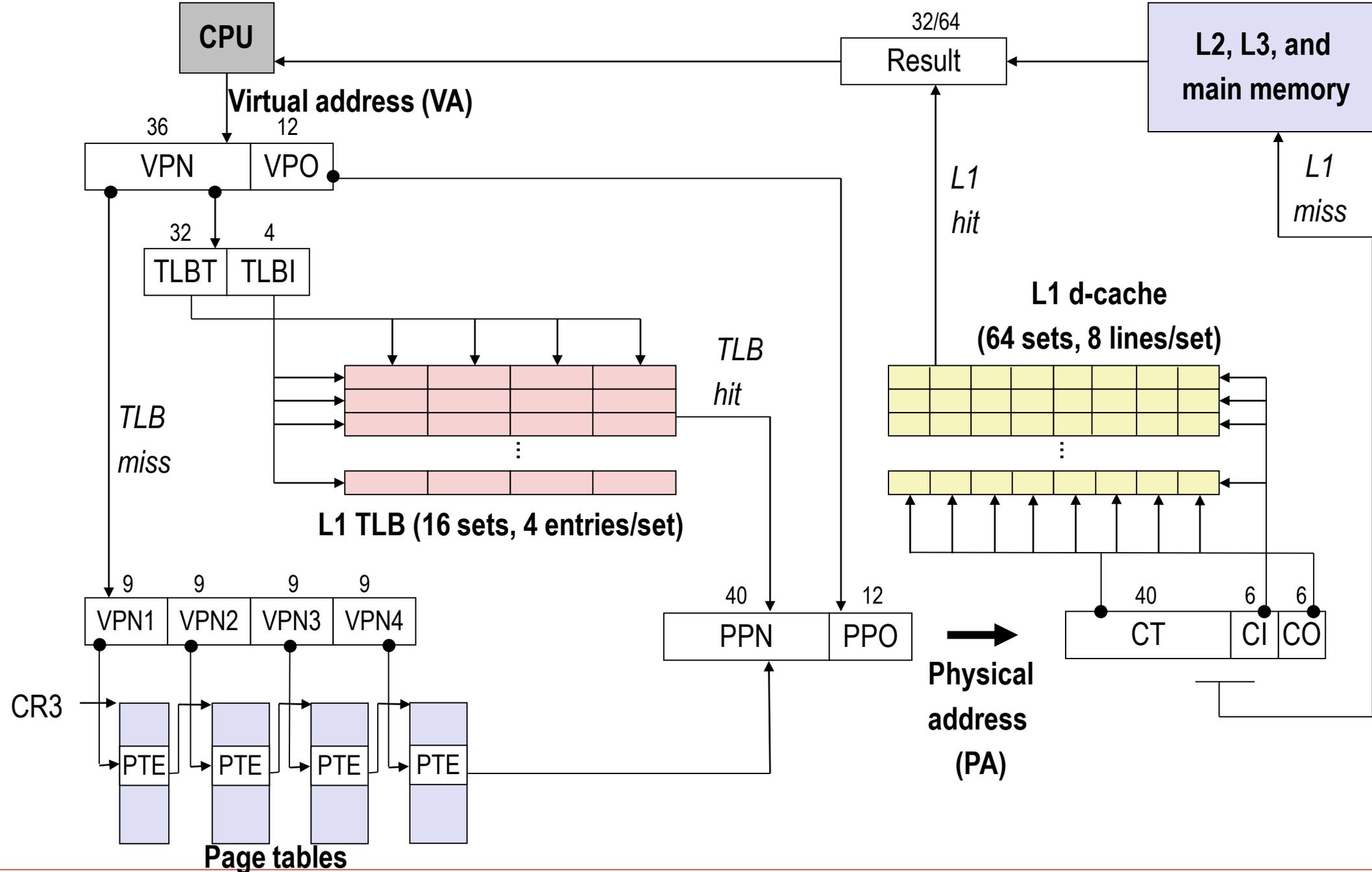 - **P = 2^p** : Page size (bytes)

- **Components of the virtual address (VA)**
  - **TLBI**: TLB index
  - **TLBT**: TLB tag
  - **VPO**: Virtual page offset
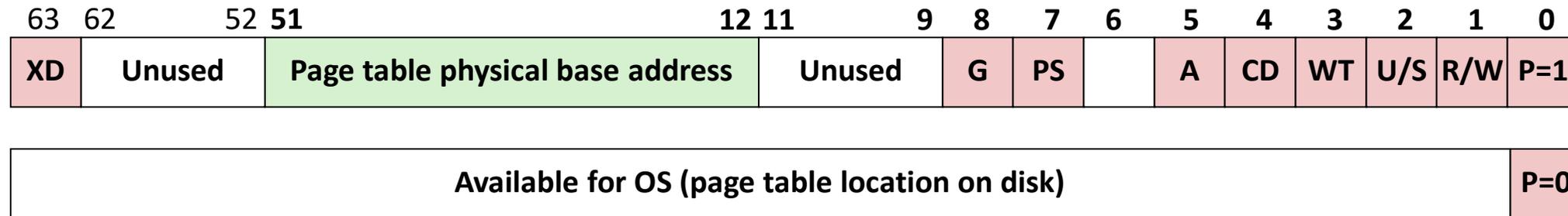 - **VPN**: Virtual page number

- **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

Core i7 Level 1-3 Page Table Entries

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

					P=0
Available for OS (page table location on disk)					P=0

Each entry references a 4K child page table. Significant fields:

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.

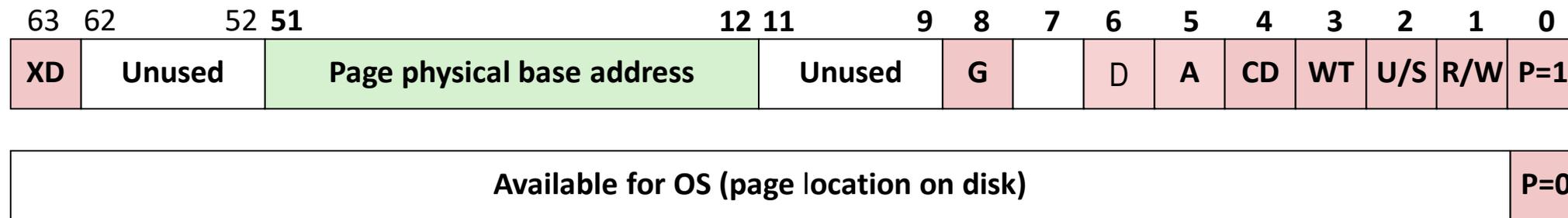WT: Write-through or write-back cache policy for the child page table.

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).

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

XD: Disable or enable instruction fetches from all pages reachable from this PTE.

Core i7 Level 4 Page Table Entries

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

	0
Available for OS (page location on disk)	P=0

Each entry references a 4K child page. Significant fields:

P: Child page is present in memory (1) or not (0)

R/W: Read-only or read-write access permission for child page

U/S: User or supervisor mode access

WT: Write-through or write-back cache policy for this page

A: Reference bit (set by MMU on reads and writes, cleared by software)

D: Dirty bit (set by MMU on writes, cleared by software)
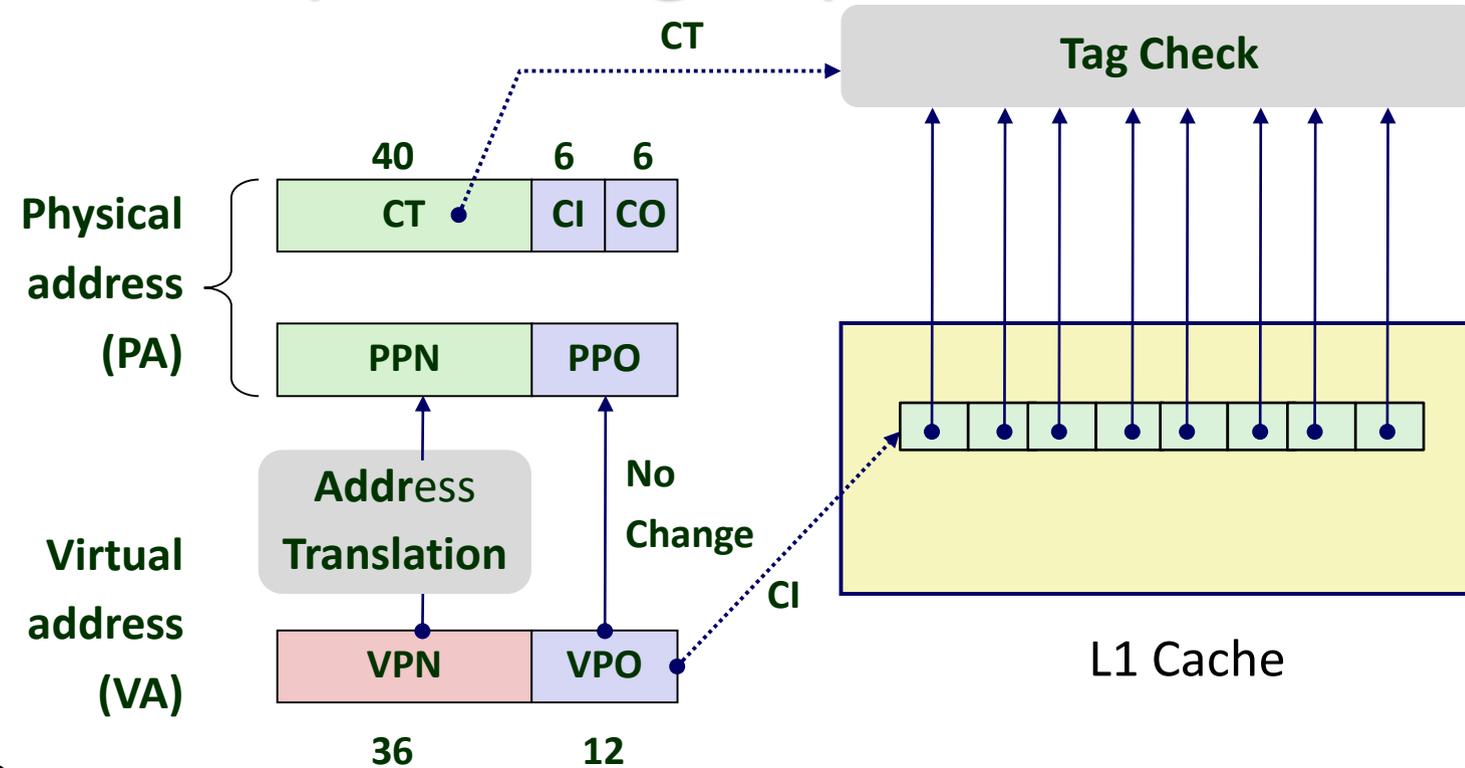
Page physical base address: 40 most significant bits of physical page address (forces pages to be 4KB aligned)

XD: Disable or enable instruction fetches from this page.

Core i7 Page Table Translation

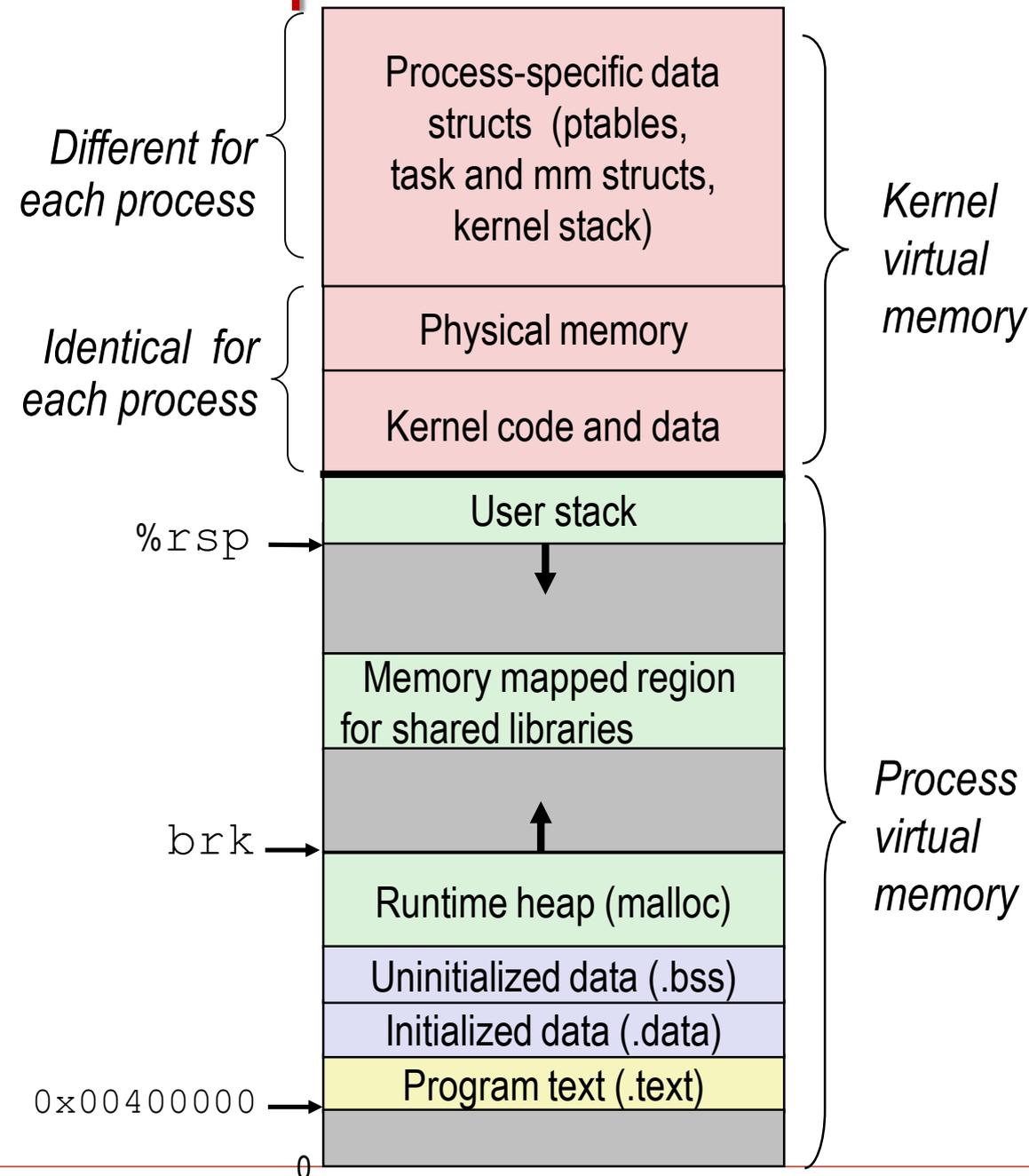**Carnegie Mellon University**
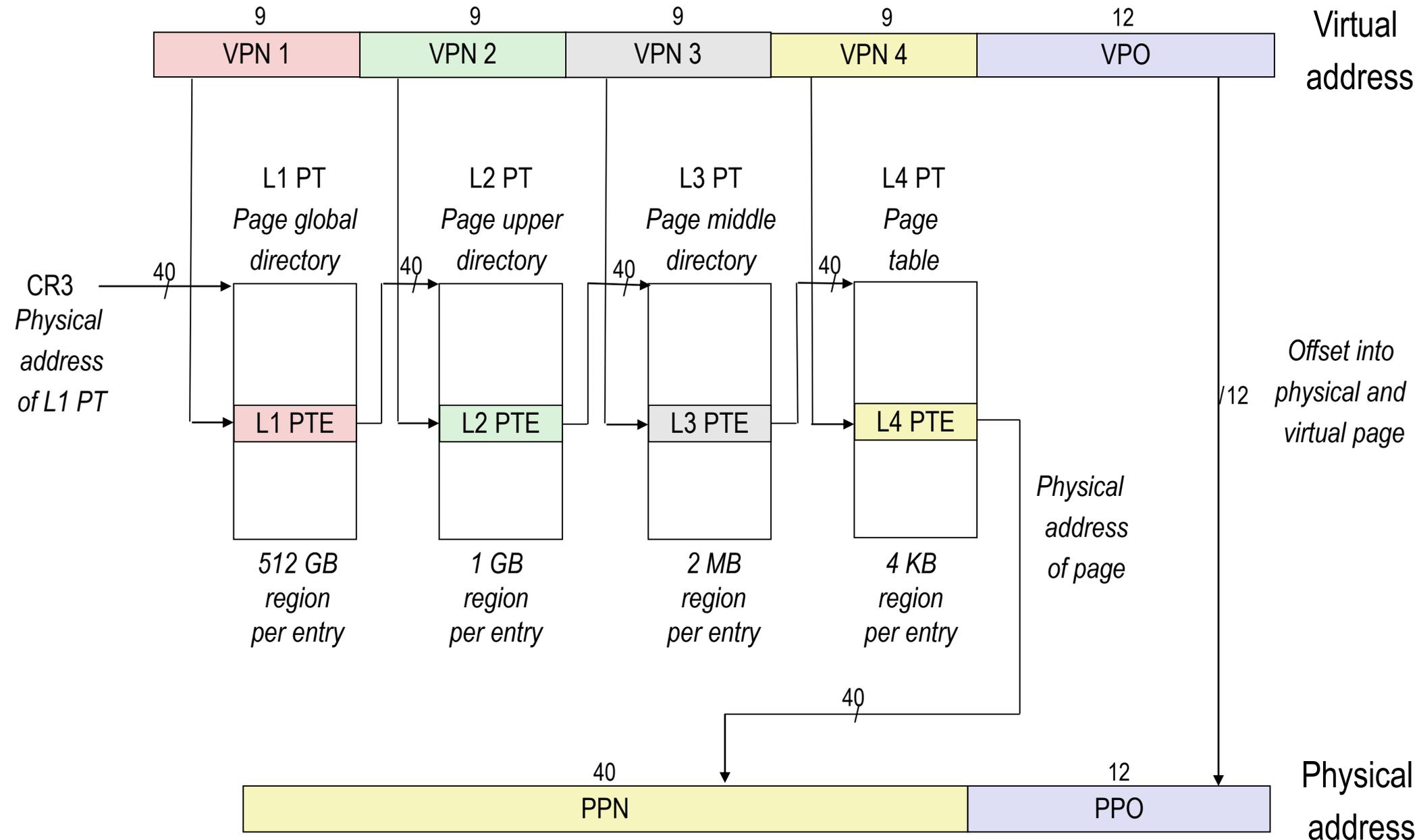
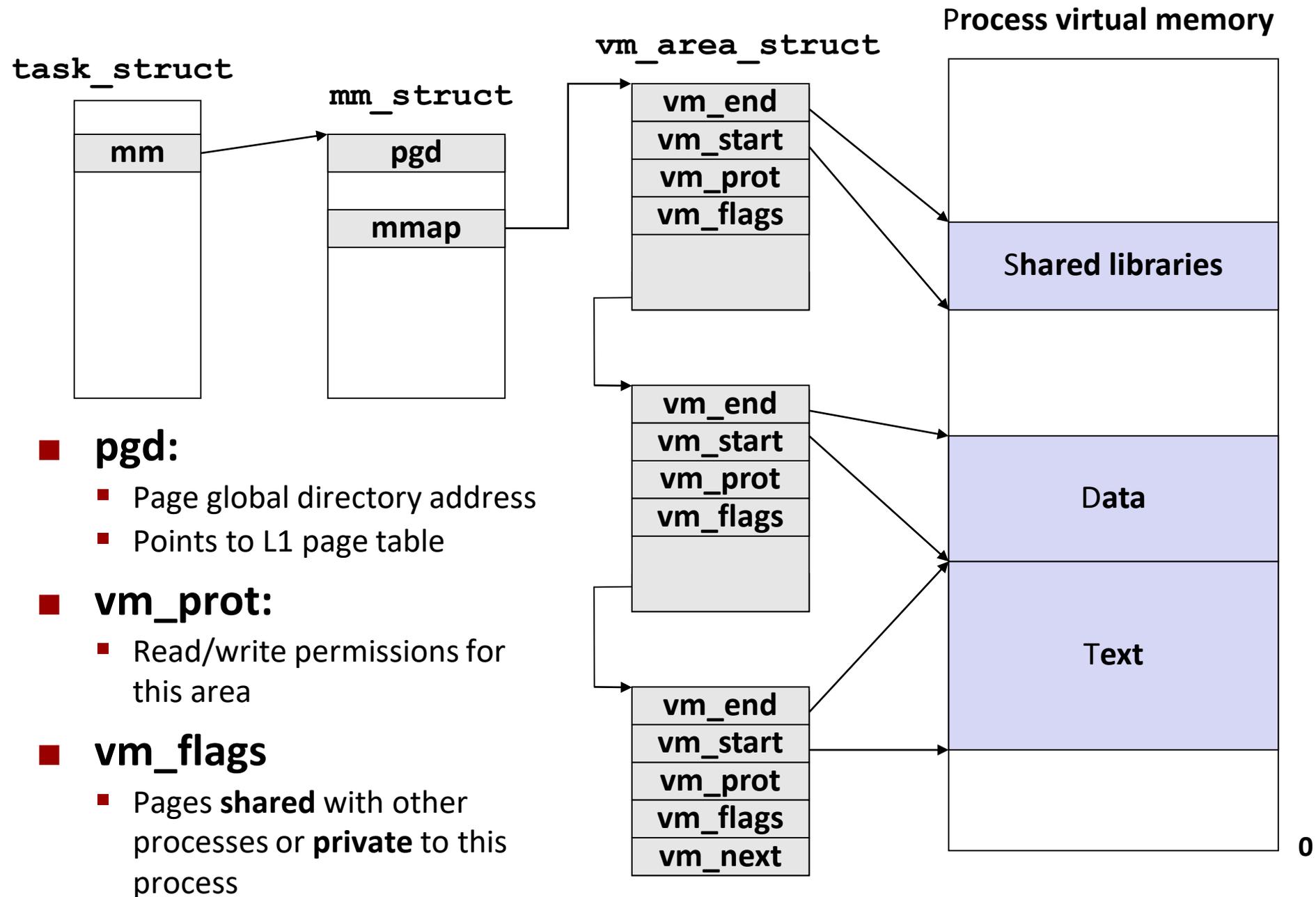# Cute Trick for Speeding Up L1 Access

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

Carnegie Mellon University 51

Virtual Address Space of a Linux Process

Different for each process

Process-specific data structs (ptables, task and mm structs, kernel stack)

Identical for each process

Physical memory

Kernel code and data

Kernel virtual memory

User stack

%rsp

Memory mapped region for shared libraries

brk

Runtime heap (malloc)

Uninitialized data (.bss)

Initialized data (.data)

0x00400000

Program text (.text)

0

Process virtual memory

Linux Organizes VM as Collection of "Areas"

task_struct

mm_struct

vm_area_struct

Process virtual memory

mm

pgd

mmap

vm_end
vm_start
vm_prot
vm_flags

vm_end
vm_start
vm_prot
vm_flags

vm_end
vm_start
vm_prot
vm_flags
vm_next

Shared libraries

Data

Text

0

- **pgd:**
  - Page global directory address
  - Points to L1 page table

- **vm_prot:**
  - Read/write permissions for this area

- **vm_flags**
  - Pages **shared** with other processes or **private** to this process

Carnegie Mellon University

Linux Page Fault Handling

vm_area_struct

Process virtual memory

| vm_end |
| vm_start |
| vm_prot |
| vm_flags |

shared libraries

1
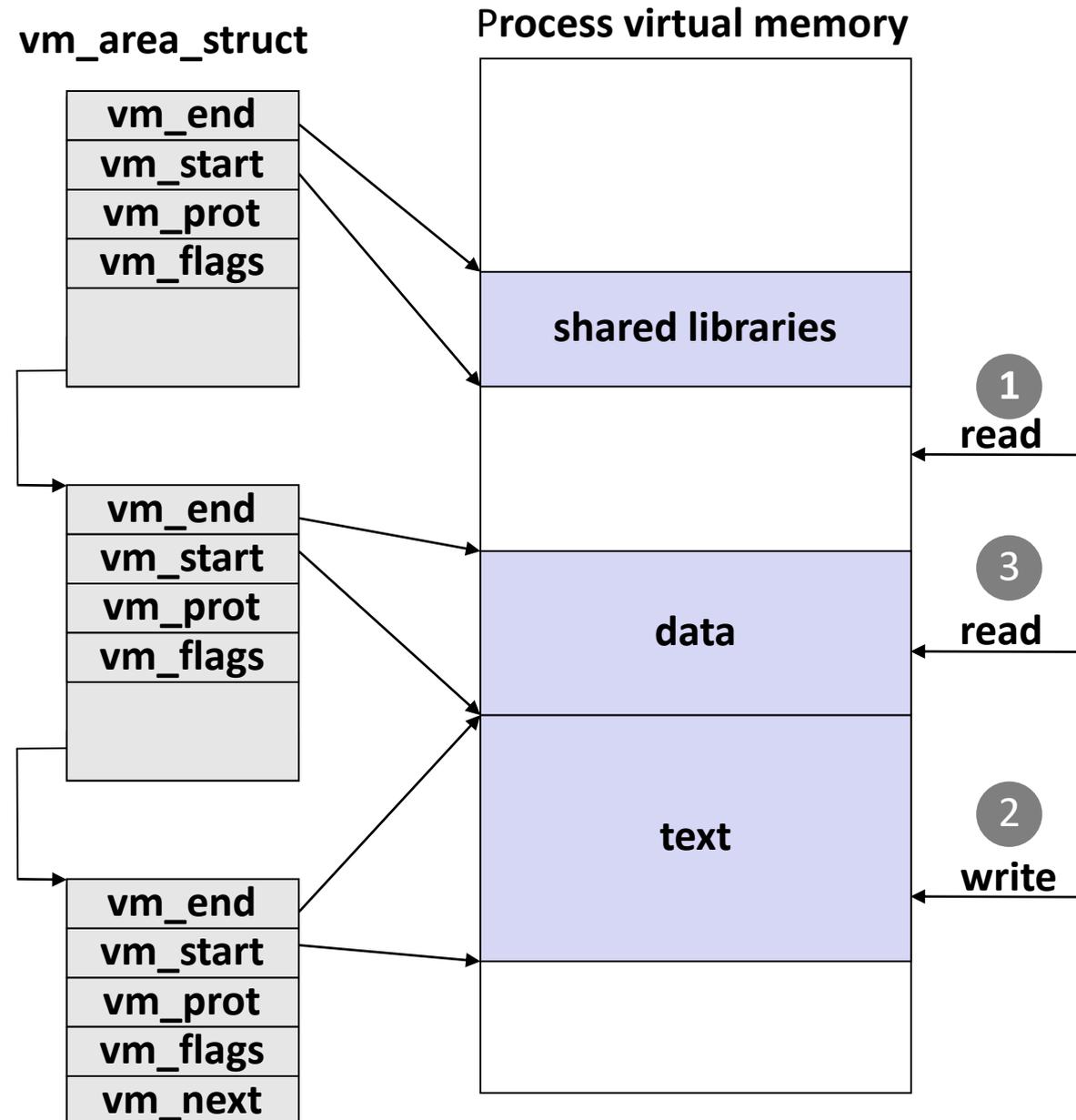
read → Segmentation fault: accessing a non-existing page

| vm_end |
| vm_start |
| vm_prot |
| vm_flags |

data

3

read → Normal page fault

text

| vm_end |
| vm_start |
| vm_prot |
| vm_flags |
| vm_next |

2

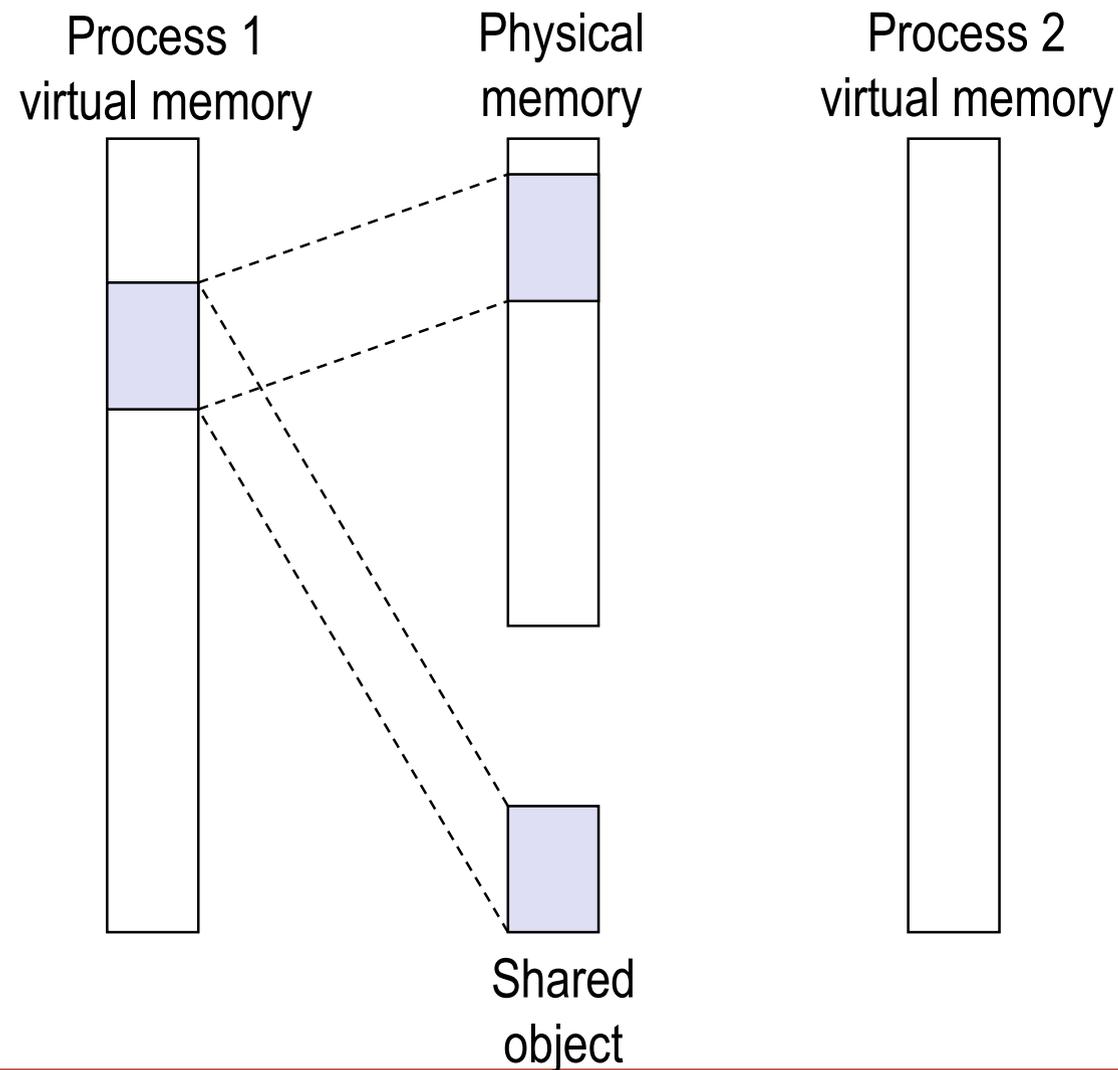**write** → Protection exception: e.g., violating permission by writing to a read-only page (Linux reports as Segmentation fault)

Carnegie Mellon University
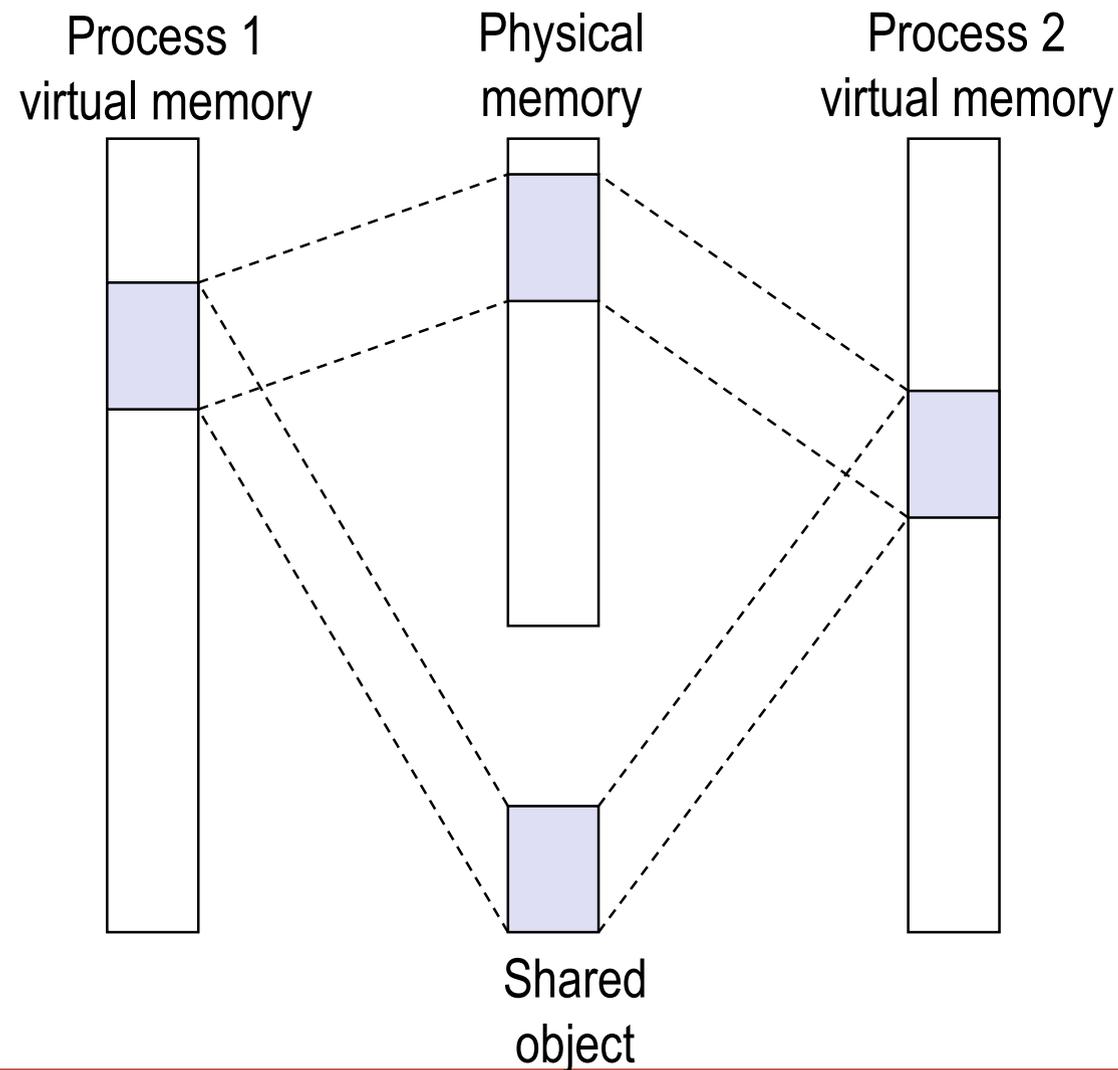
Memory Mapping

- **VM areas initialized by associating them with disk objects.**
  - Process is known as *memory mapping.*

- **Area can be *backed by* (i.e., get its initial values from) :**
  - *Regular file* on disk (e.g., an executable object file)
    - Initial page bytes come from a section of a file
  - *Anonymous file* (e.g., nothing)
    - First fault will allocate a physical page full of 0's (*demand-zero page*)
    - Once the page is written to (*dirtied*), it is like any other page

- **Dirty pages are copied back and forth between memory and a special *swap file*.**

Carnegie Mellon University

Sharing Revisited: Shared Objects

Process 1
virtual memory

Physical
memory

Process 2
virtual memory

Shared
object

- **Process 1 maps the shared object.**

Sharing Revisited: Shared Objects

Process 1
virtual memory

Physical
memory

Process 2
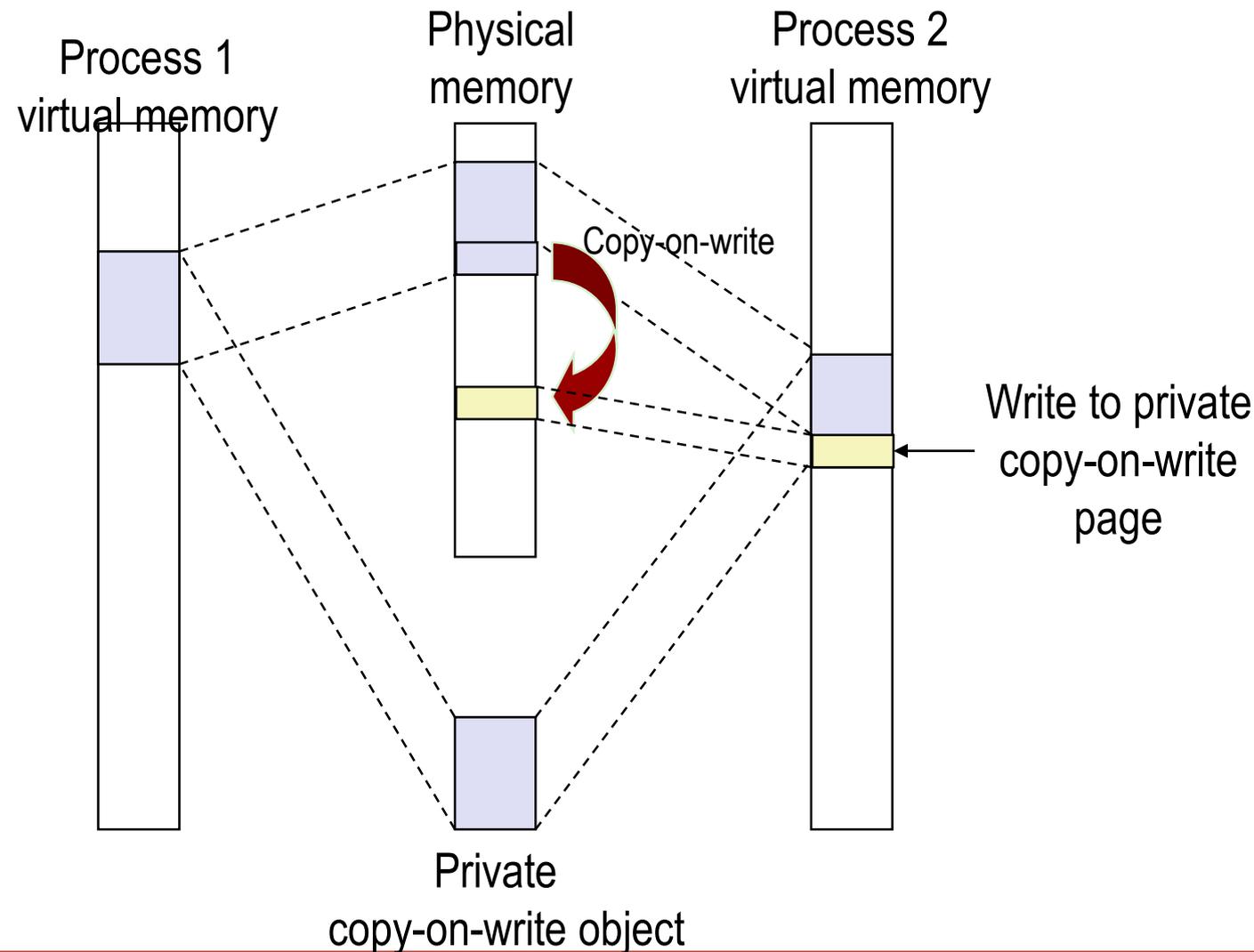virtual memory

Shared
object

- **Process 2 maps the shared object.**

- Notice how the virtual addresses can be different.

Sharing Revisited: Private Copy-on-write (COW) Objects

Process 1
virtual memory

Physical
memory

Process 2
virtual memory

Private
copy-on-write
area

Private
copy-on-write object
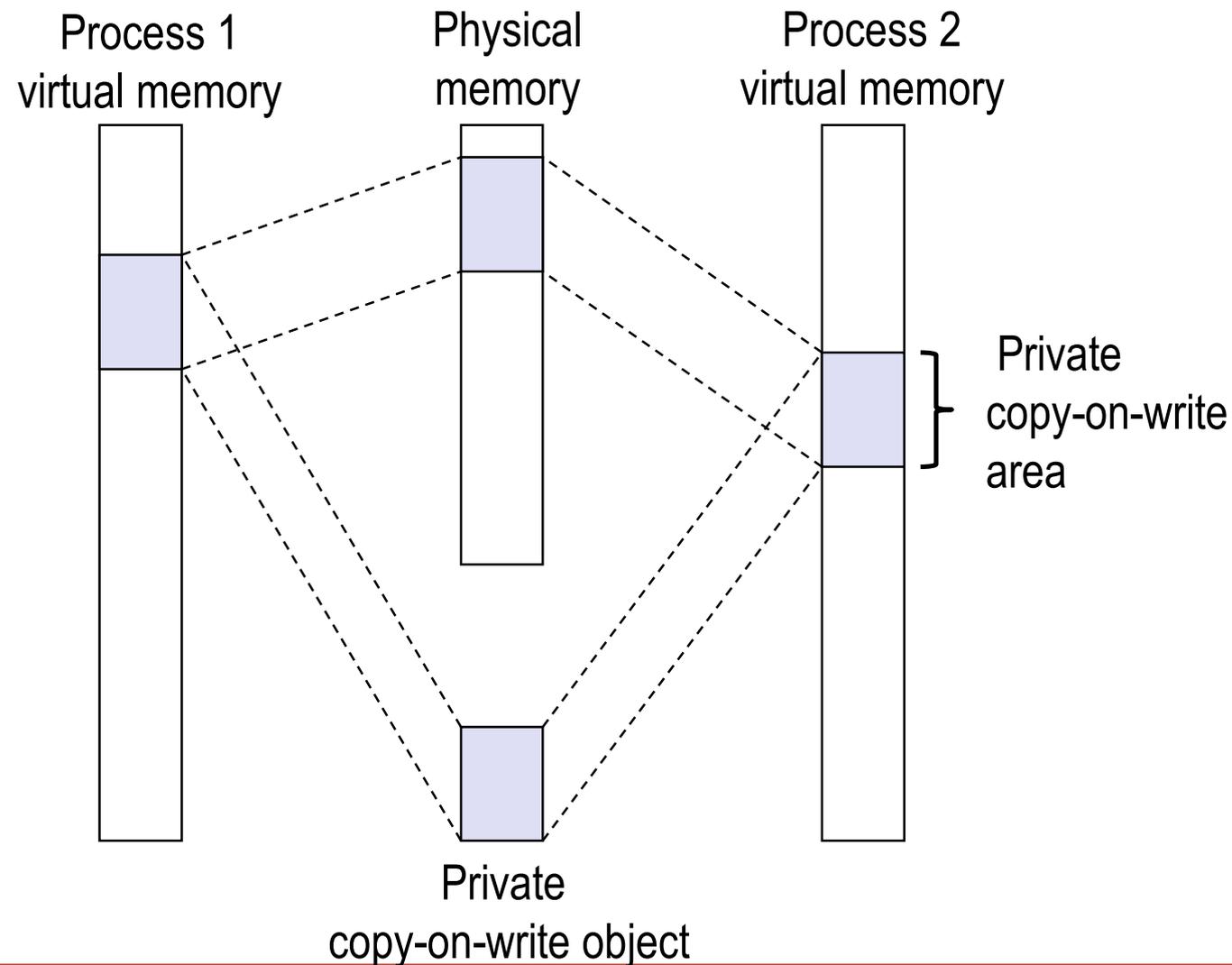
- **Two processes mapping a *private copy-on-write (COW)* object.**

- **Area flagged as private copy-on-write**

- **PTEs in private areas are flagged as read-only**

Carnegie Mellon University

Sharing Revisited: Private Copy-on-write (COW) Objects

Process 1
virtual memory

Physical
memory

Process 2
virtual memory

Copy-on-write

Write to private
copy-on-write
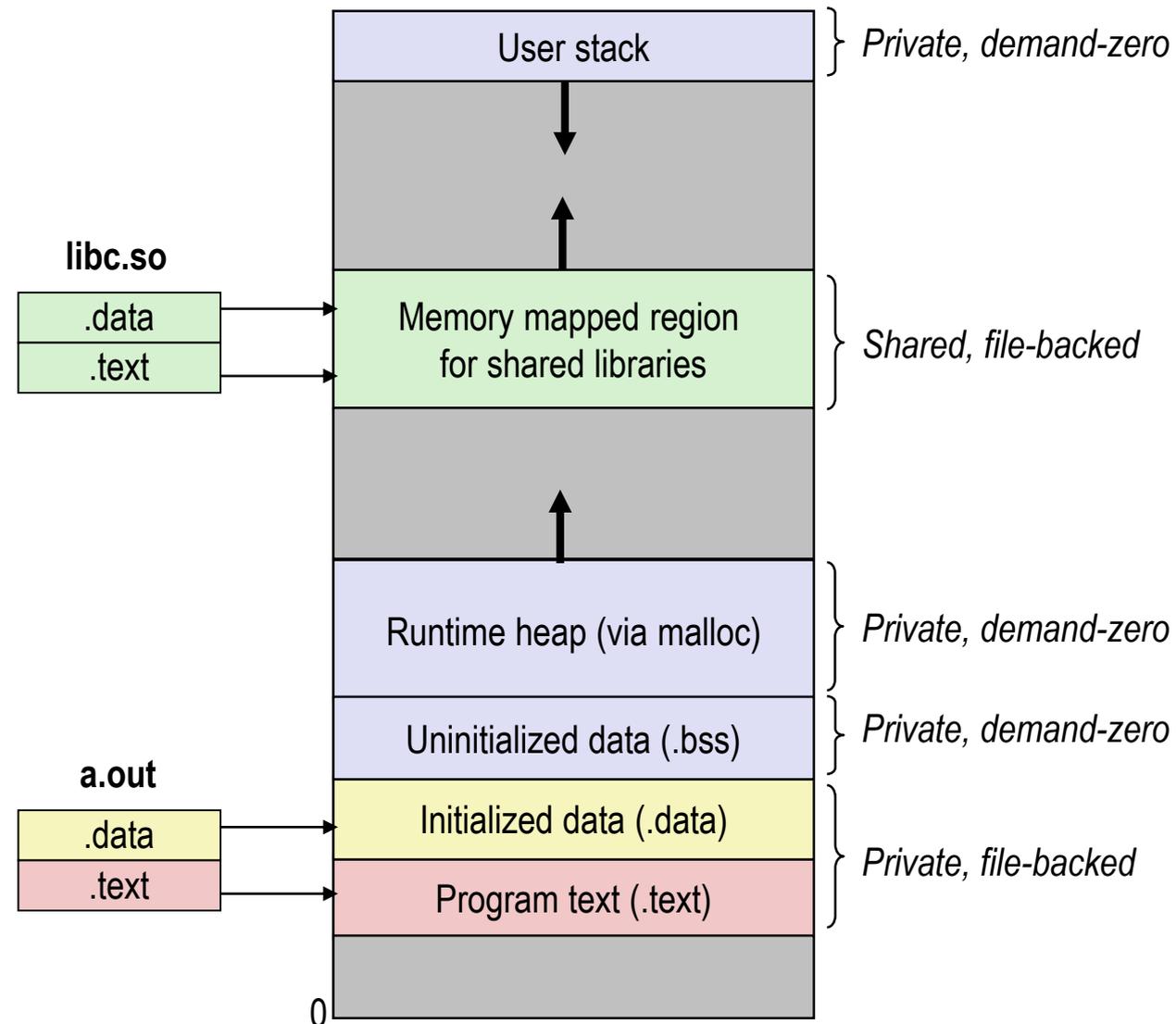page

Private
copy-on-write object

- **Instruction writing to private page triggers protection fault.**

- **Handler creates new R/W page.**

- **Instruction restarts upon handler return.**

- **Copying deferred as long as possible!**

Carnegie Mellon University

The **fork** Function Revisited

- **VM and memory mapping explain how `fork` provides private address space for each process.**

- **To create virtual address for new new process**
  - Create exact copies of current `mm_struct`, `vm_area_struct`, and page tables.
  - Flag each page in both processes as read-only
  - Flag each `vm_area_struct` in both processes as private COW

- **On return, each process has exact copy of virtual memory**

- **Subsequent writes create new pages using COW mechanism.**

The `execve` Function Revisited

User stack	} *Private, demand-zero*
↓	
↑	
Memory mapped region for shared libraries	} *Shared, file-backed*
↑	
Runtime heap (via malloc)	} *Private, demand-zero*
Uninitialized data (.bss)	} *Private, demand-zero*
Initialized data (.data)	} *Private, file-backed*
Program text (.text)	

libc.so
- .data
- .text

a.out
- .data
- .text

0

- **To load and run a new program `a.out` in the current process using `execve`:**

- **Free `vm_area_struct`'s and page tables for old areas**

- **Create `vm_area_struct`'s and page tables for new areas**
  - Programs and initialized data backed by object files.
  - `.bss` and stack backed by anonymous files .

- **Set PC to entry point in `.text`**
  - Linux will fault in code and data pages as needed.
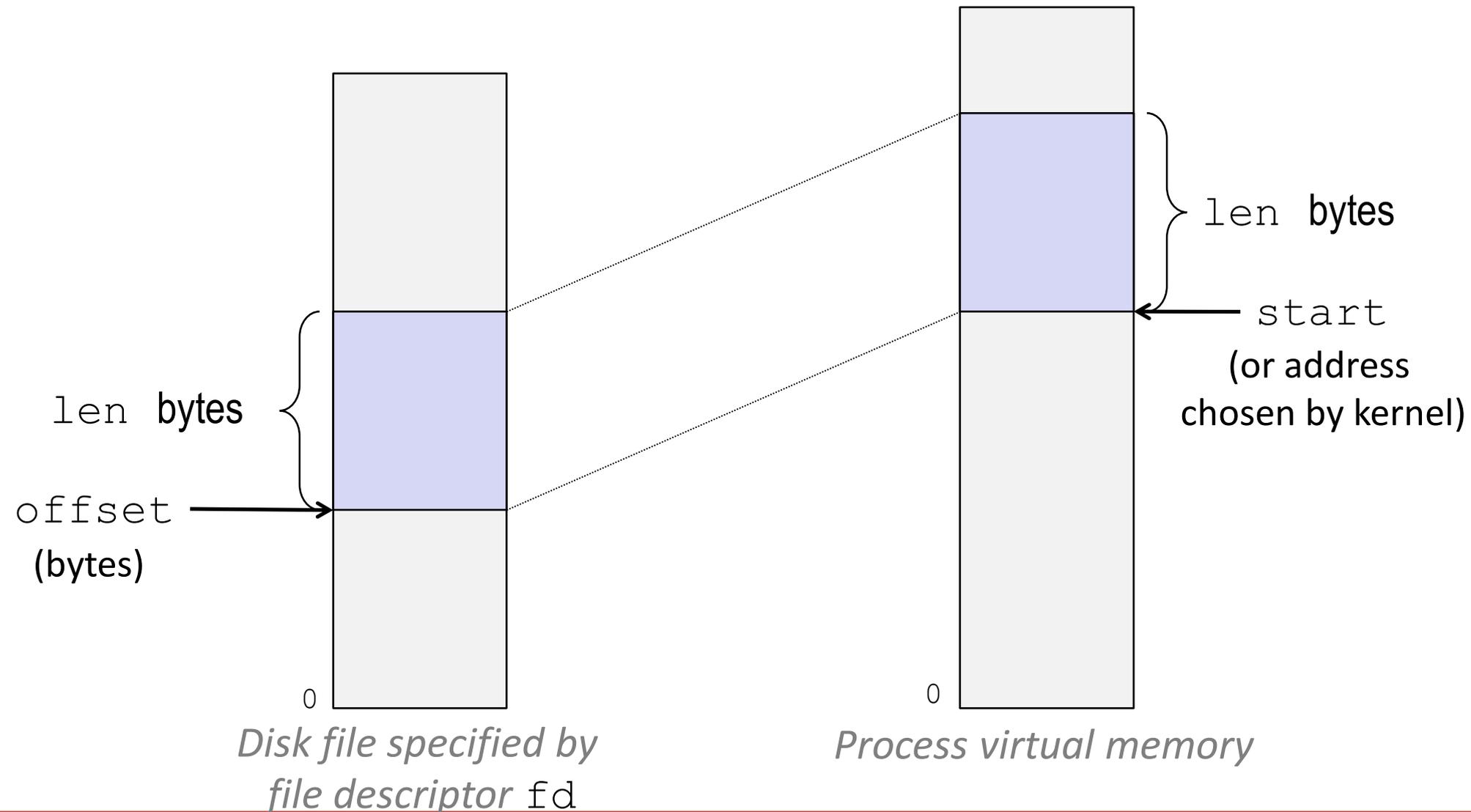
User-Level Memory Mapping

```
void *mmap(void *start, int len,
           int prot, int flags, int fd, int offset)
```

- **Map `len` bytes starting at offset `offset` of the file specified by file description `fd`, preferably at address `start`**
  - **`start`:** may be 0 for "pick an address"
  - **`prot`**: PROT_READ, PROT_WRITE, ...
  - **`flags`**: MAP_ANON, MAP_PRIVATE, MAP_SHARED, ...

- **Return a pointer to start of mapped area (may not be `start`)**

User-Level Memory Mapping

```
void *mmap(void *start, int len,
           int prot, int flags, int fd, int offset)
```


len bytes

start
(or address
chosen by kernel)

len bytes

offset
(bytes)

*Disk file specified by
file descriptor* `fd`

Process virtual memory

Carnegie Mellon University

Example: Using mmap to Copy Files

- Copying a file to `stdout` without transferring data to user space .

```c
#include "csapp.h"

void mmapcopy(int fd, int size)
{

    /* Ptr to memory mapped area */
    char *bufp;

    bufp = Mmap(NULL, size,
            PROT_READ,
            MAP_PRIVATE,
            fd, 0);
    Write(1, bufp, size);
    return;
}
```
mmapcopy.c

```c
/* mmapcopy driver */
int main(int argc, char **argv)
{
    struct stat stat;
    int fd;

    /* Check for required cmd line arg */
    if (argc != 2) {
        printf("usage: %s <filename>\n",
            argv[0]);
        exit(0);
    }

    /* Copy input file to stdout */
    fd = Open(argv[1], O_RDONLY, 0);
    Fstat(fd, &stat);
    mmapcopy(fd, stat.st_size);
    exit(0);
}
```
mmapcopy.c

18-600 Foundations of Computer Systems

Lecture 16:
"Dynamic Memory Allocation"

October 23, 2017

Next Time ...

> Required Reading Assignment:
> • **Chapter 9 of CS:APP (3rd edition) by Randy Bryant & Dave O'Hallaron.**

Electrical & Computer ENGINEERING

Carnegie Mellon University