CS3523: Operating Systems-II

Bheemarjuna Reddy Tamma

IIT Hyderabad

https://www.iith.ac.in/~tbr

Syllabus

- *What is OS? History of OS? OS Structure*
- *System Calls, System Programming*
- *Process and Thread Management: Processes, Inter Process Communication, Threads, Scheduling, Concurrency, Synchronization primitives, Deadlocks*
- *Remaining part:*

Live classes	Flipped classes
Memory Management: Memory allocation, Paging, Virtual Memory	*I/O, Storage devices File systems and Implementation*

Main Memory

Note: Some slides and/or pictures in the following are adapted from the text books on OS by Silberschatz, Galvin, and Gagne AND Andrew S. Tanenbaum and Albert S.Woodhull. Slides courtesy of Kubiatowicz, AJ Shankar, George Necula, Alex Aiken, Eric Brewer, Ras Bodik, Ion Stoica, Doug Tygar, David Wagner, etc.

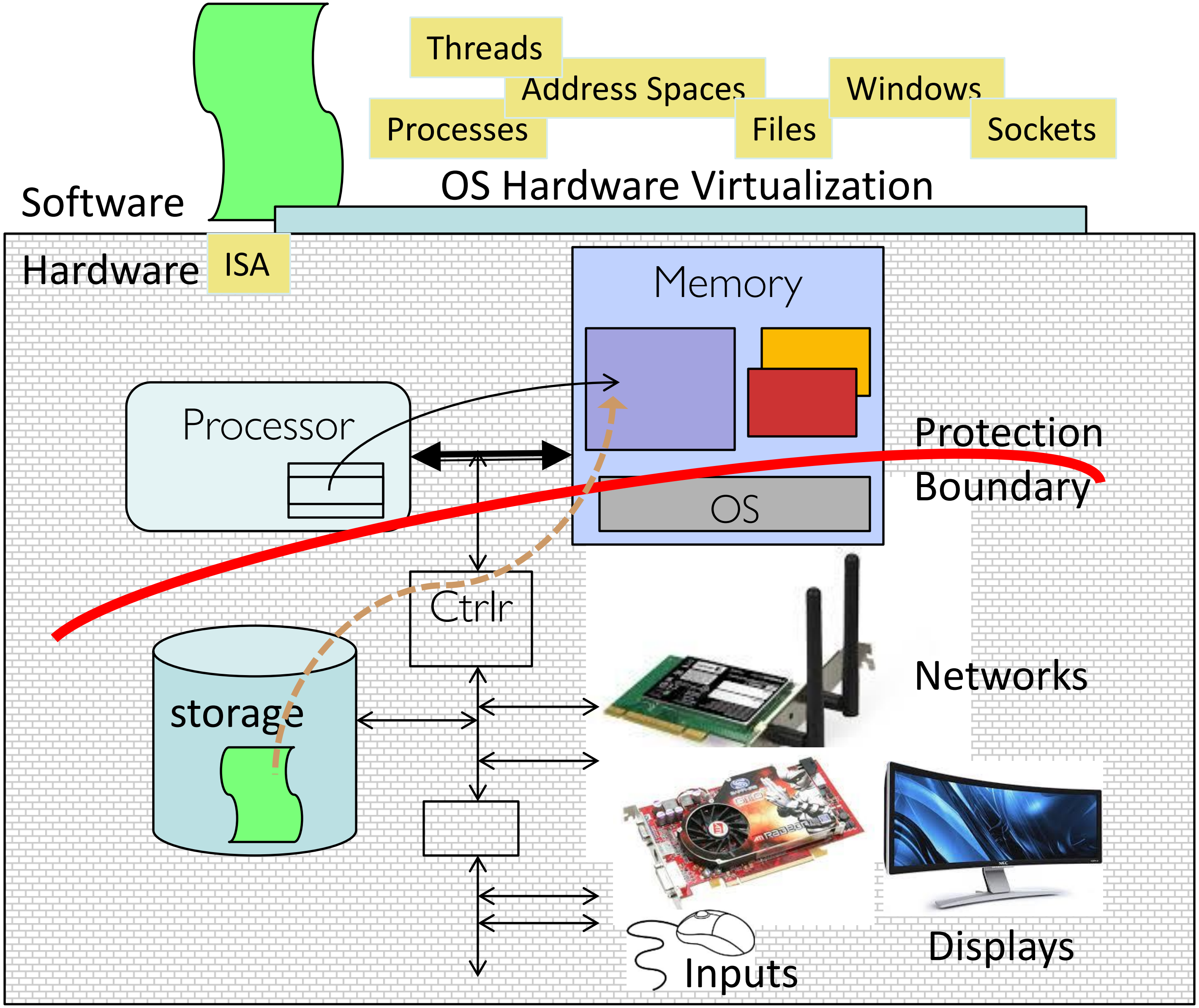
Outline

- Background
- Protection: Address Spaces
  - What is an Address Space?
  - How is it Implemented?
- Address Translation Schemes
  - Segmentation
  - Paging
  - Multi-level translation
  - Paged page tables
  - Inverted page tables
- Comparison among schemes

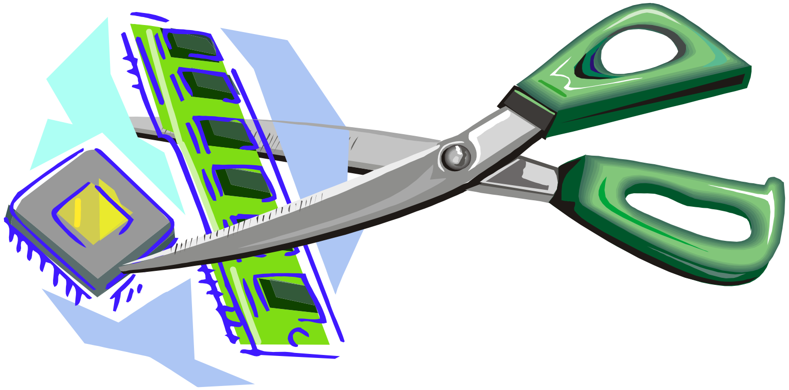
Background

- Program must be brought (from 2nd memory like HDD/SSD) into main memory and placed within a **process** for it to be run
- Main memory, Cache and registers are only storage units that CPU can access directly in almost no time
- Register access is done in one CPU clock (or less)
- Main memory access can take many cycles, causing a **stall** in execution of the running process
- **Cache** sits between main memory and CPU registers
- Memory unit only sees a stream of:
  - addresses + **read** requests, or
  - addresses + data and **write** requests

Loading a program



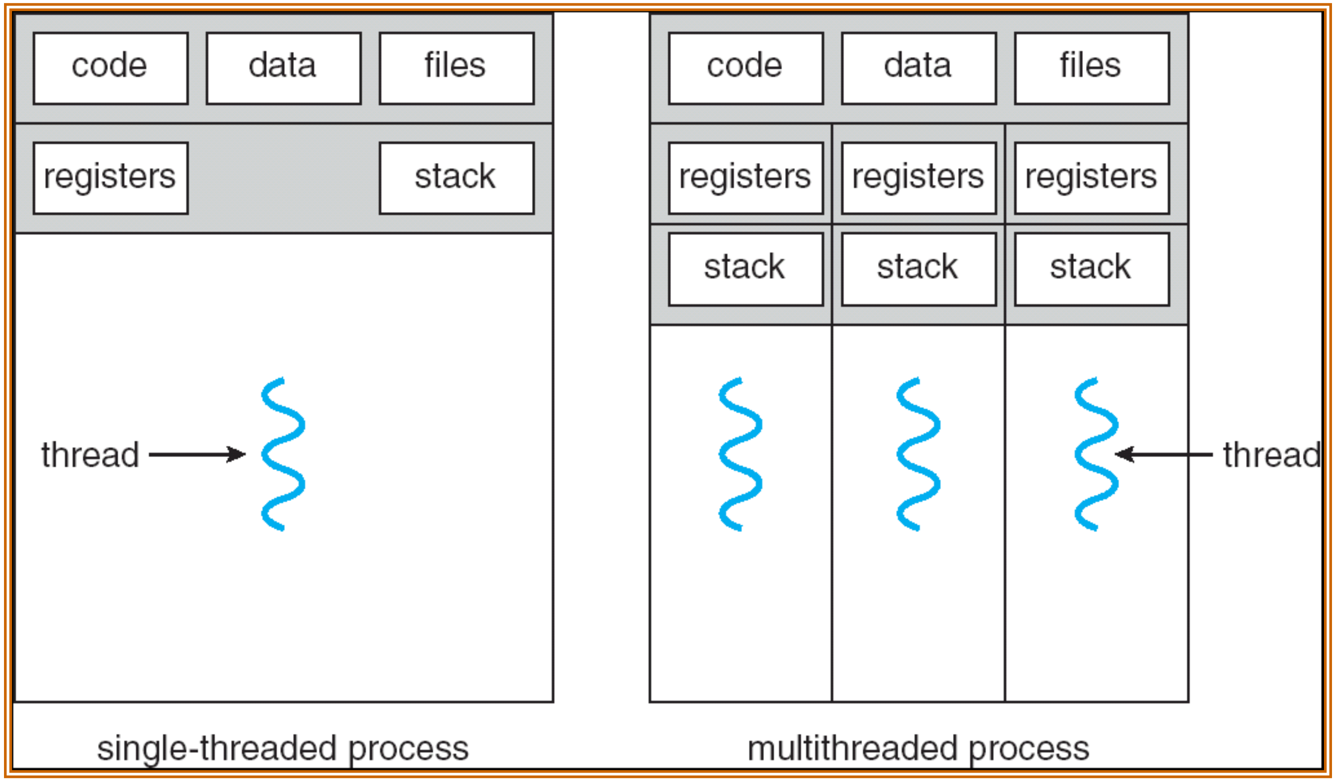
Virtualizing Resources



- Physical Reality:

  Different processes/threads share the same hardware
  - Need to multiplex CPU (already finished: scheduling)
  - Need to multiplex use of main memory (starting today)
  - Need to multiplex disk and other I/O devices (later)
- Why worry about memory sharing/protection?
  - The complete working state of a process and/or kernel is defined by its data in main memory (and registers)
  - Consequently, cannot just let different threads of control use the same memory in general
    - Physics: two different pieces of data can’t occupy the same loc in memory
  - Probably don’t want different threads to even have access to each other’s memory if they are in different processes (protection)

Recall: Single and Multithreaded Processes



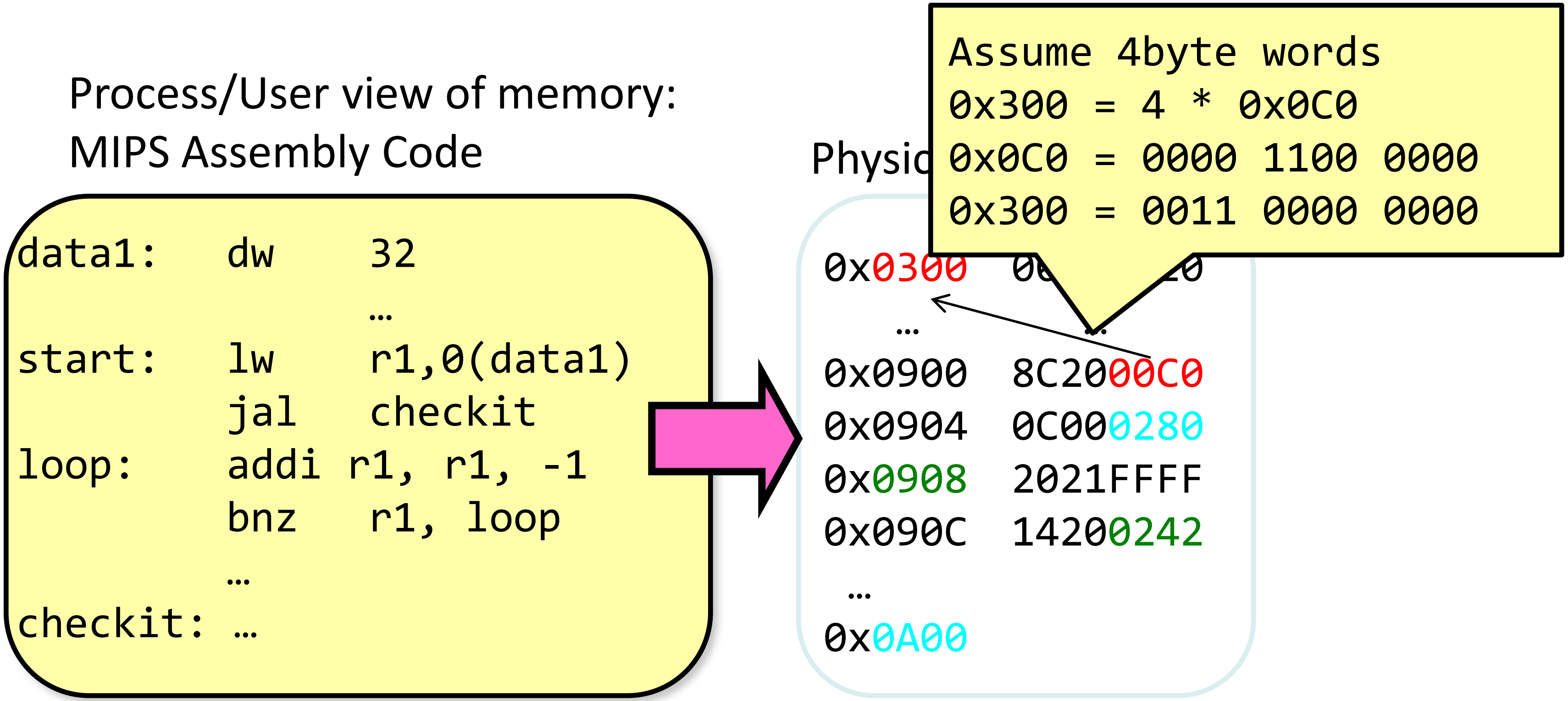
- Threads encapsulate concurrency
  - "Active" component of a process
- Address spaces encapsulate protection
  - "Passive" component of a process
  - Keeps buggy program from trashing the system

Important Aspects of Memory Multiplexing

- Protection:
  - Prevent access to private memory of other processes
    - » Different pages of memory can be given special behavior (Read Only, Invisible to user programs, etc)
    - » Kernel data protected from User programs
    - » Programs protected from themselves!
- Controlled overlap:
  - Separate state of threads should not collide in physical memory. Obviously, unexpected overlap causes chaos!
  - Conversely, would like the ability to overlap when desired (for communication like shared memory IPC)
- Translation:
  - Ability to translate accesses from one address space (logical aka virtual) to a different one (physical)
  - When translation exists, processor uses virtual addresses, physical memory uses physical addresses
  - Side effects:
    - » Can be used to avoid overlap
    - » Can be used to give uniform view of memory to programs

Binding of Instructions and Data to Main Memory

Process/User view of memory:
MIPS Assembly Code



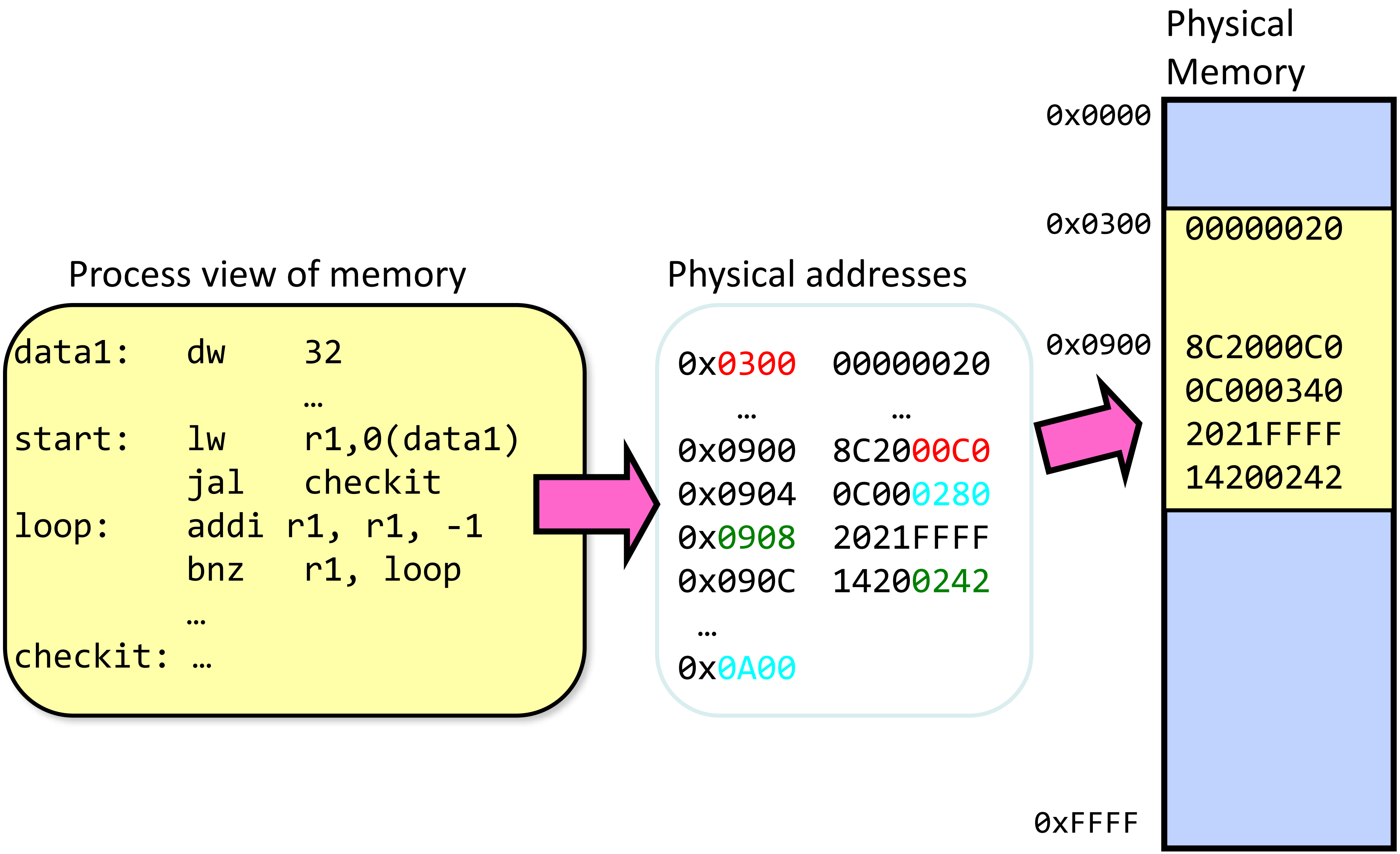
```
data1:   dw    32
               …
start:   lw    r1,0(data1)
         jal   checkit
loop:    addi r1, r1, -1
         bnz   r1, loop
         …
checkit: …
```

Physic

```
0x0300  00      20
   …
0x0900  8C2000C0
0x0904  0C000280
0x0908  2021FFFF
0x090C  14200242
 …
0x0A00
```

```
Assume 4byte words
0x300 = 4 * 0x0C0
0x0C0 = 0000 1100 0000
0x300 = 0011 0000 0000
```

Binding of Instructions and Data to Memory



Process view of memory

```
data1:   dw    32
               …
start:   lw    r1,0(data1)
         jal   checkit
loop:    addi r1, r1, -1
         bnz   r1, loop
         …
checkit: …
```

Physical addresses

```
0x0300   00000020
    …        …
0x0900   8C2000C0
0x0904   0C000280
0x0908   2021FFFF
0x090C   14200242
  …
0x0A00
```

Physical Memory

```
0x0000
0x0300   00000020
0x0900   8C2000C0
         0C000340
         2021FFFF
         14200242
0xFFFF
```

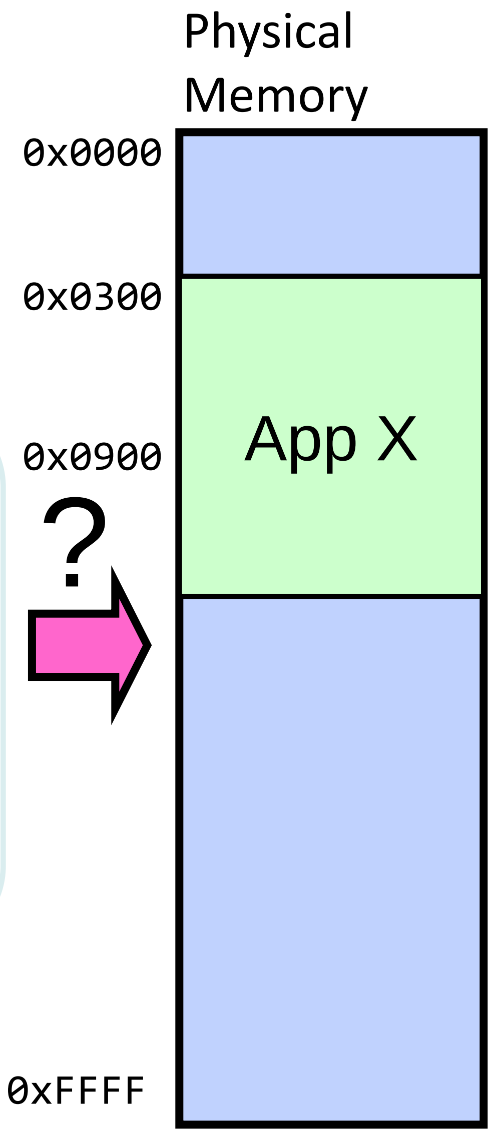
Second copy of program from previous example

Process view of memory

```
data1:   dw     32
                …
start:   lw     r1,0(data1)
         jal    checkit
loop:    addi r1, r1, -1
         bnz    r1, loop
         …
checkit: …
```

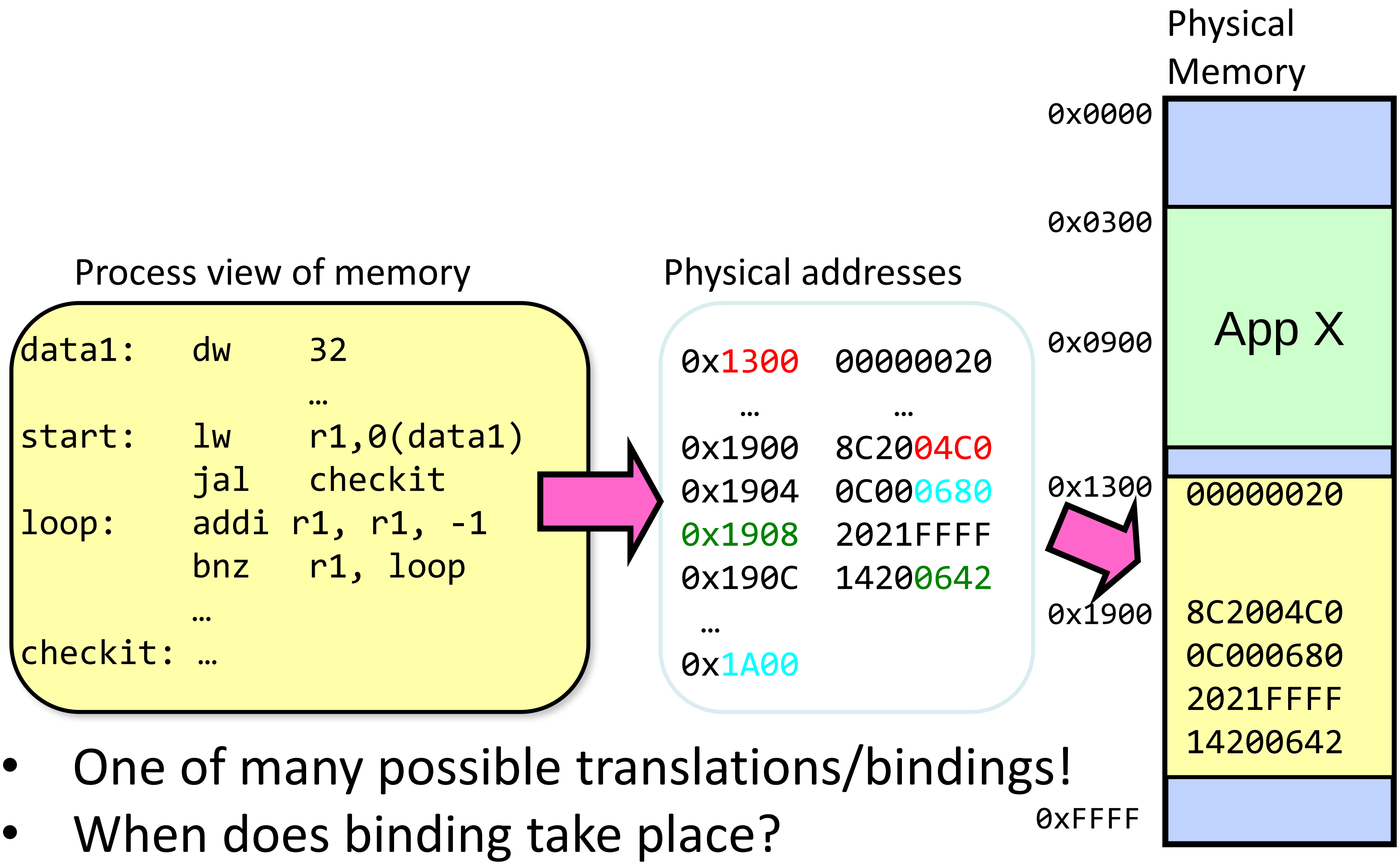
Physical addresses

```
0x0300  00000020
  …        …
0x0900  8C2000C0
0x0904  0C000280
0x0908  2021FFFF
0x090C  14200242
 …
0x0A00
```



Need address translation!

Second copy of program from previous example



Process view of memory

```
data1:   dw    32
               …
start:   lw    r1,0(data1)
         jal   checkit
loop:    addi r1, r1, -1
         bnz   r1, loop
         …
checkit: …
```

Physical addresses

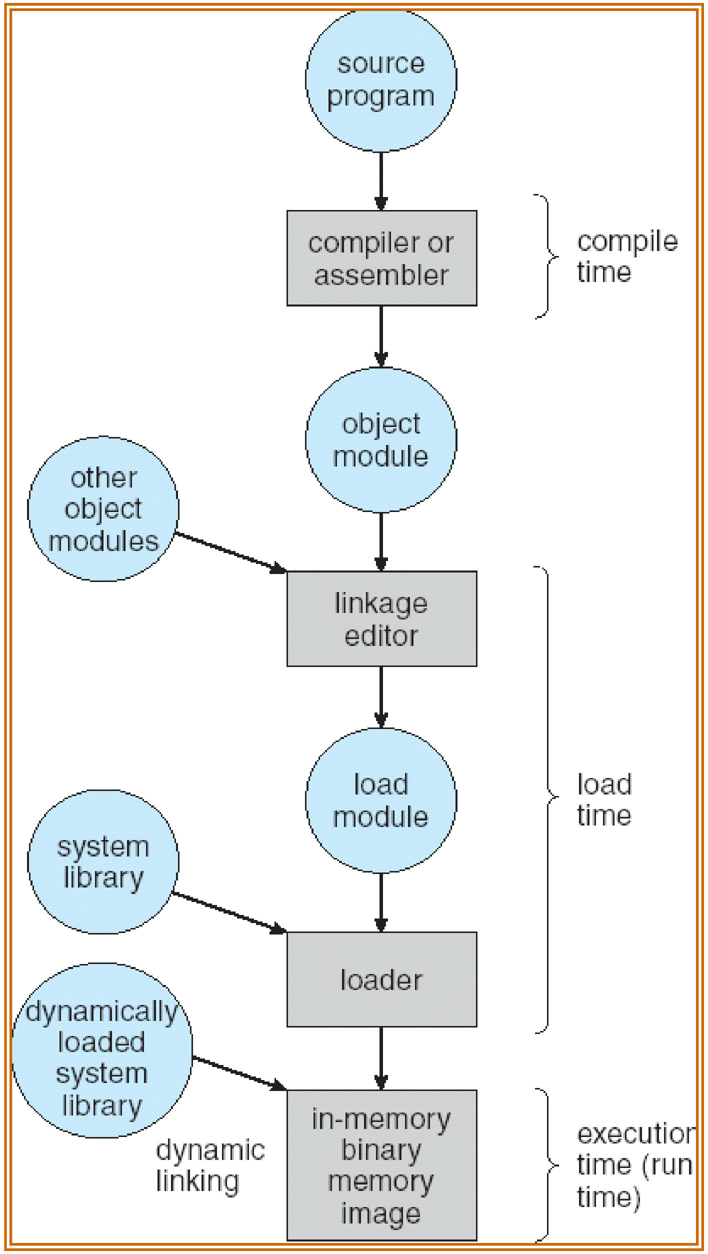
```
0x1300  00000020
   …       …
0x1900  8C2004C0
0x1904  0C000680
0x1908  2021FFFF
0x190C  14200642
 …
0x1A00
```

- One of many possible translations/bindings!
- When does binding take place?

  Compile time, Link/Load time, or Execution time?

Multi-step Processing of a Program for Execution

- Preparation of a program for execution involves components at:
  - Compile time (i.e., “gcc”)
  - Link/Load time (UNIX “ld” does linking)
  - Execution time (e.g., dynamic libs)
- Dynamic Libraries (DLL)
  - Linking postponed until execution
  - Small piece of code (i.e. the *stub*), locates appropriate memory-resident library routine
  - Stub replaces itself with the address of the routine, and executes routine
  - So, all processes share only one copy of the shared library code
  - It improves both RAM and disk utilization
- Addresses can be bound to final values anywhere in this path
  - Depends on hardware support
  - Also depends on operating system

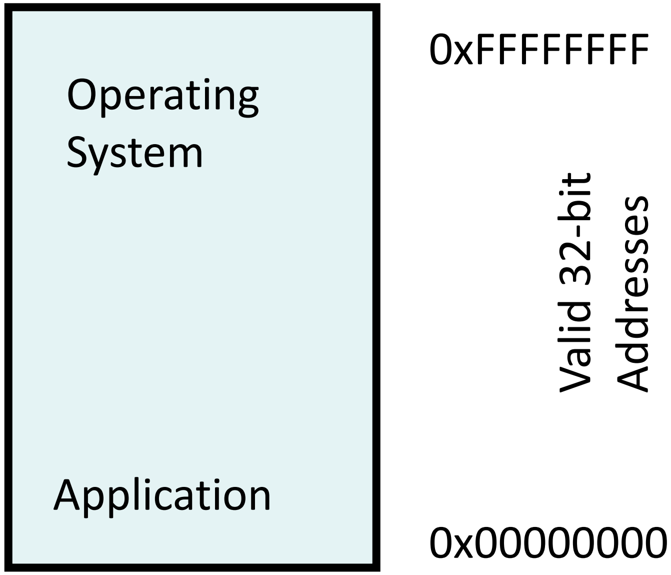


Binding of Instructions and Data to Memory

- Source program generally contains symbolic addresses
  - E.g., *pid, count, i, j*
- Binding (mapping from one address space to other) can happen at 3 different stages and hence addresses may be represented in different ways
  - **Compile time**: If main memory location known a priori, absolute addresses can be generated
    - Must recompile the code if starting location changes
  - **Load time**: Must generate *relocatable* code if main memory location is not known at compile time
    - e.g., “10 bytes from the start of process CODE block”
    - Linker/loader binds relocatable addresses to absolute addresses
      - Symbol table in the compiled file lists address values that need to be modified
    - Must reload the compiled code if starting location changes
  - **Execution time**: Binding delayed until run time if the process can be moved during its execution from one memory segment to another
    - Need hardware support for address mappings (e.g., MMU, *base* and *limit registers, page/segment table*)
    - Most common in general-purpose OSs where compiler generates relocatable addresses and then linker/loader generates logical addresses

Recall: Uniprogramming

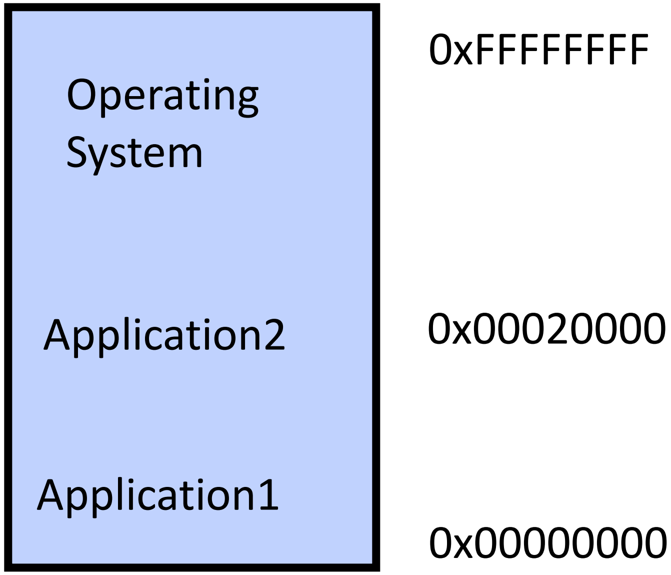
- Uniprogramming (no Translation or Protection)
  - Application always runs at same place in physical memory since only one application at a time
  - Application can access any physical address

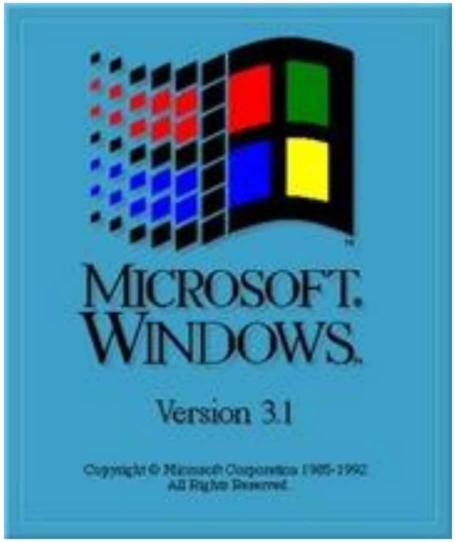


  - Application given illusion of dedicated machine by giving it reality of a dedicated machine

Multiprogramming (primitive stage)

- Multiprogramming without Translation or Protection
  - Must somehow prevent address overlap between threads/apps

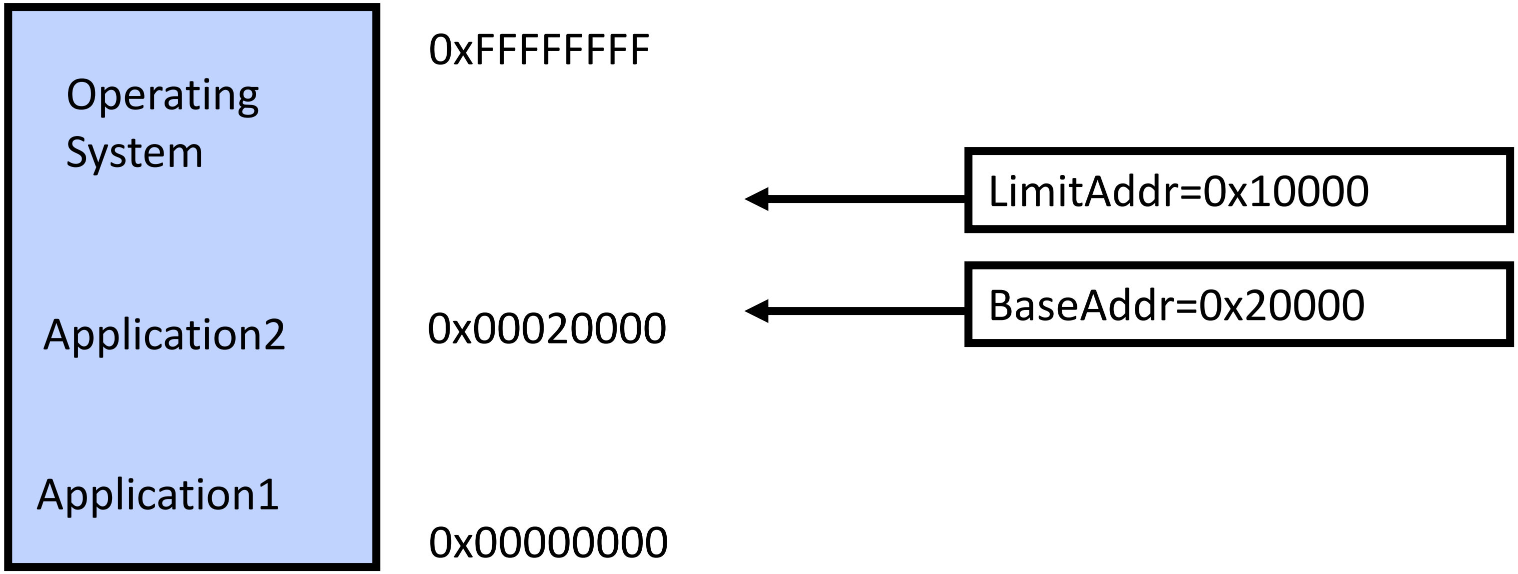




  - Use Loader/Linker: Adjust addresses while program loaded into memory (loads, stores, jumps)
    - Everything adjusted to memory location of program
    - Translation done by a linker-loader (relocation)
    - Common in early days (… till Windows 3.x, 95?)
- With this solution, no protection: bugs in any program can cause other programs to crash or even the OS

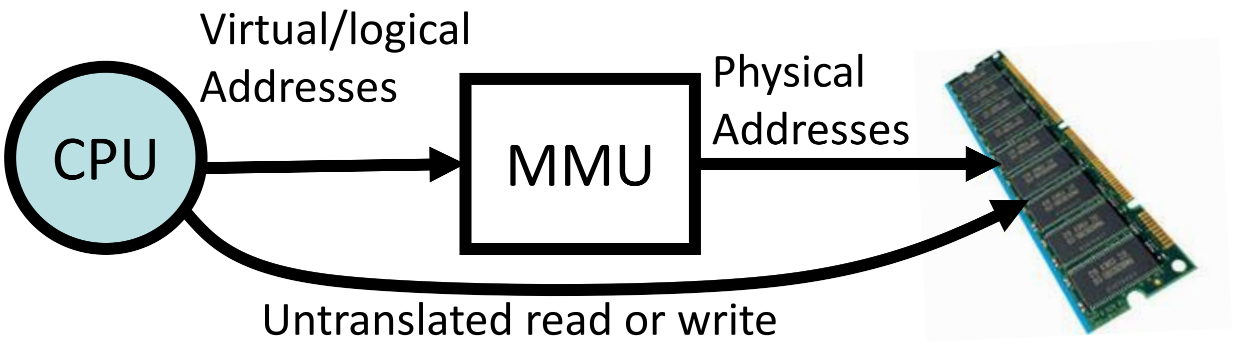
Multiprogramming (with Protection)

- Can we protect programs from each other without address translation?



  - Yes: use two special registers *BaseAddr* and *LimitAddr* to prevent user from straying outside designated area
    - Cause error if user tries to access an illegal address
  - During switch, kernel loads new base/limit from PCB (Process Control Block)
    - User not allowed to change base/limit registers

Recall: General Address translation



- Recall: Address Space:
  - All the addresses and state a process can touch
  - Each process and kernel have different address spaces
- Consequently, two views of memory:
  - View from the CPU (what program sees, logical/virtual memory addresses)
  - View from memory unit (physical memory addresses)
  - Translation box (MMU) converts between the two views

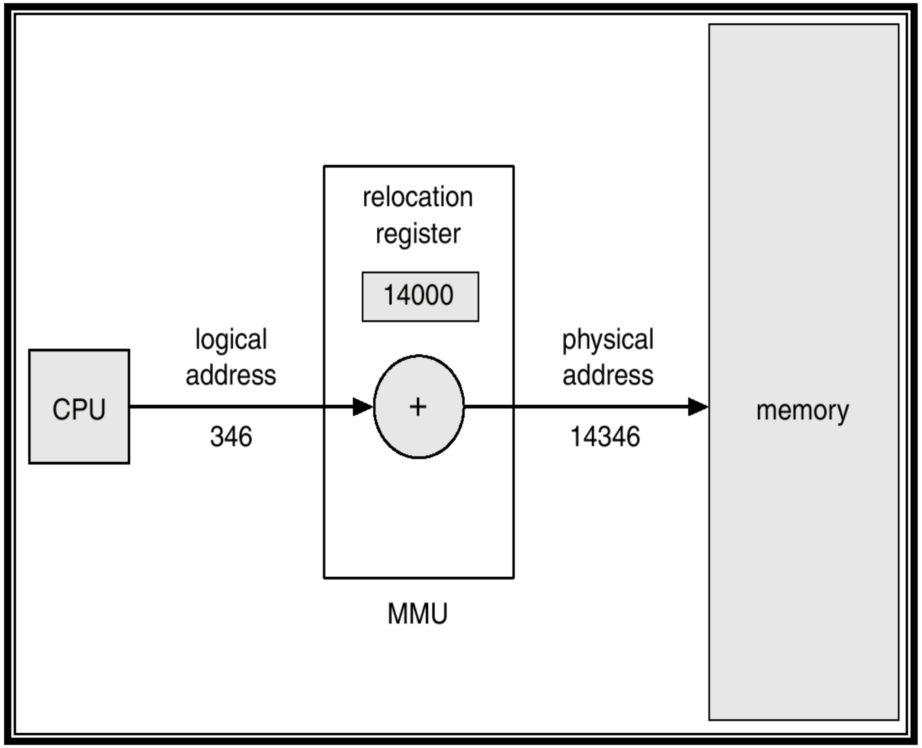
Logical vs Physical Address Space

- The concept of a logical *address space* that is bound to a separate *physical address space* is central to proper memory management.
  - *Logical address* – generated by the CPU; also referred to as *virtual address*.
  - *Physical address* – address seen by the memory unit.

- Logical and physical addresses are the same in compile-time and load-time address-binding schemes
- Logical (virtual) and physical addresses differ in execution-time address-binding scheme

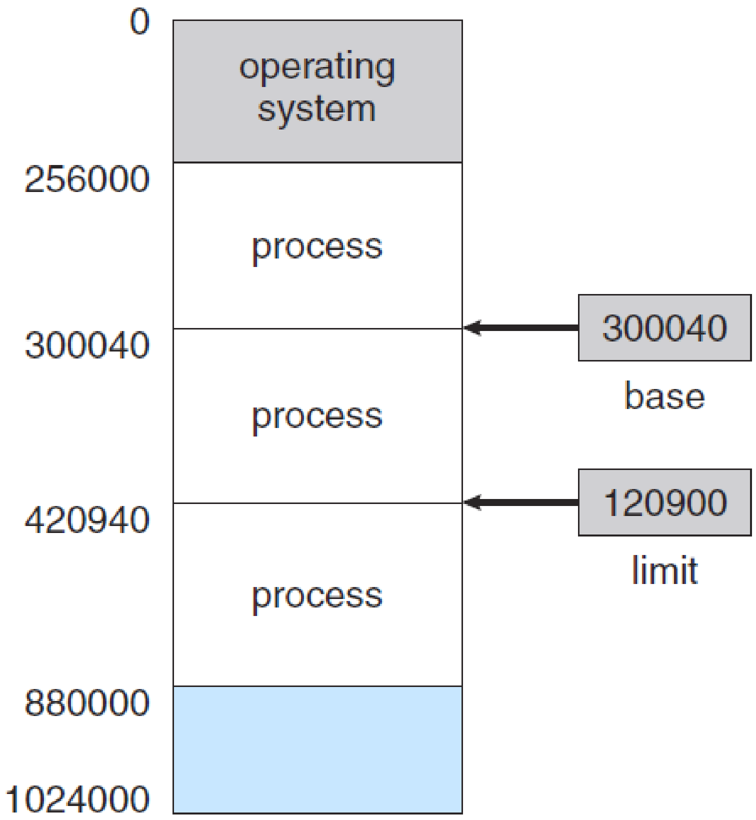
Memory-Management Unit (MMU)

- Hardware device that maps (translates) virtual/logical address to physical address.
- In MMU, the value in the relocation register is added to every address generated by a user process at the time it is sent to memory.
- The user program deals with *logical* addresses; it never sees the *real* physical addresses.

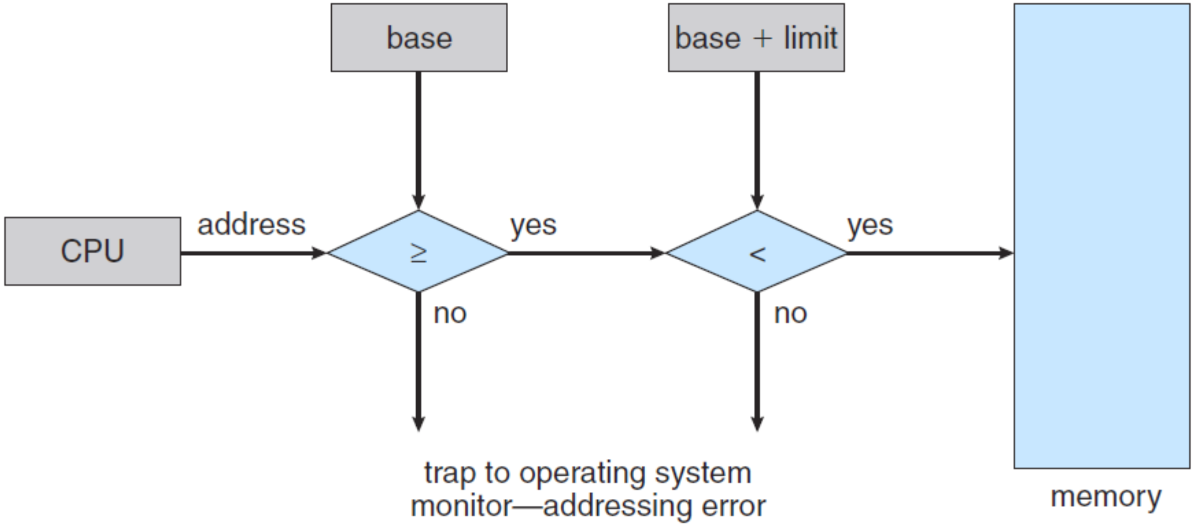


Simple Contiguous Address Space: Base and Limit

- Each program gets a single contiguous memory in main memory
- Two special registers *base* and *limit* define address space of a user program



Simple Contiguous Memory: Base and Limit



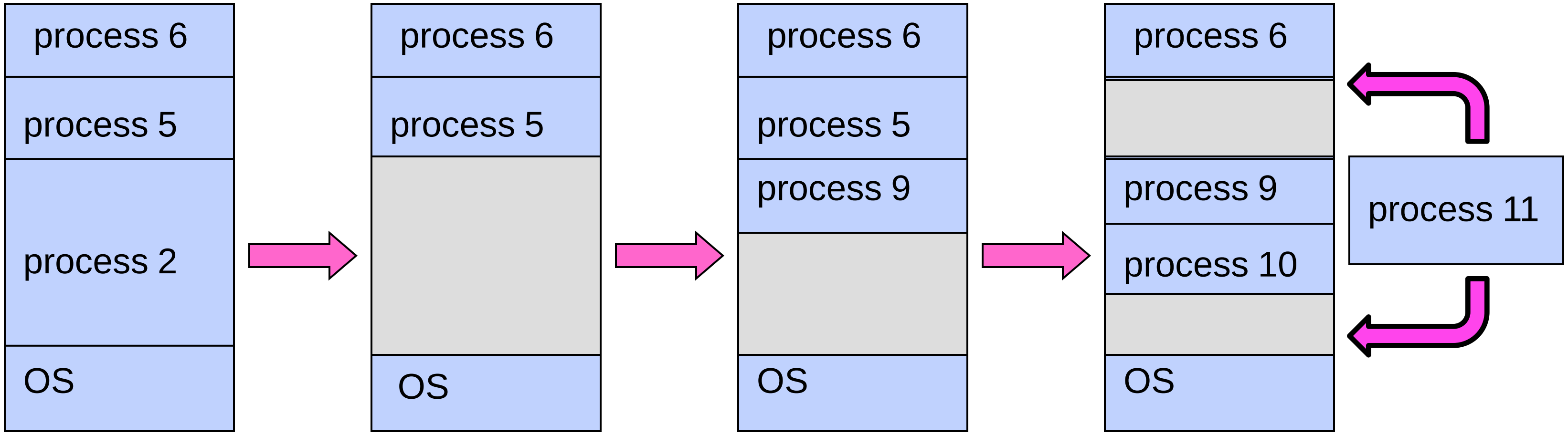
https://en.wikipedia.org/wiki/Cray-1

- We can also use base & bounds/limit registers for dynamic address translation (a simple form of “segmentation”):
  - Alter every address by adding “base”
  - Generate error if address bigger than limit → much easier to implement protection!
- This gives program the illusion that it is running on its own dedicated machine, with memory starting at 0 to some MAX Limit
  - Program gets continuous region of memory
  - Addresses within program do not have to be changed when program placed in different region of DRAM
- During context switch, kernel loads new base/limit from PCB
  - User not allowed to change base/limit registers

Base and Limit contiguous memory discussion

- Base and Limit Pros: Simple, relatively fast
- Provides level of indirection
  - OS can move bits around behind program's back
  - Can be used to correct if program needs to grow beyond its bounds or coalesce fragments of memory
- Only OS gets to change the base and limit!
  - Would defeat protection otherwise
- What gets saved/restored on a context switch?
  - Everything from before + base/limit values
  - Or: How about complete contents of memory (out to disk)?
    - Called "Swapping"
- Hardware cost
  - 2 registers/Adder/Comparator
  - Slows down system because it need to do add/compare on every memory access for fetching instructions and Data

Other Issues with Simple Contiguous Address Method



- Fragmentation problem over time
  - Not every process is same size ⇒ memory becomes fragmented over time with small holes
  - Really bad if want space to grow dynamically (e.g. heap and stack)
- Missing support for sparse address space
  - Would like to have multiple chunks/program (Code, Data, Stack, Heap, etc) → non-contiguous memory allocation
- Hard to do inter-process sharing
  - Want to share code segments when possible
  - Want to share memory between processes

Dynamic Memory Allocation Schemes

A linked list of free holes:

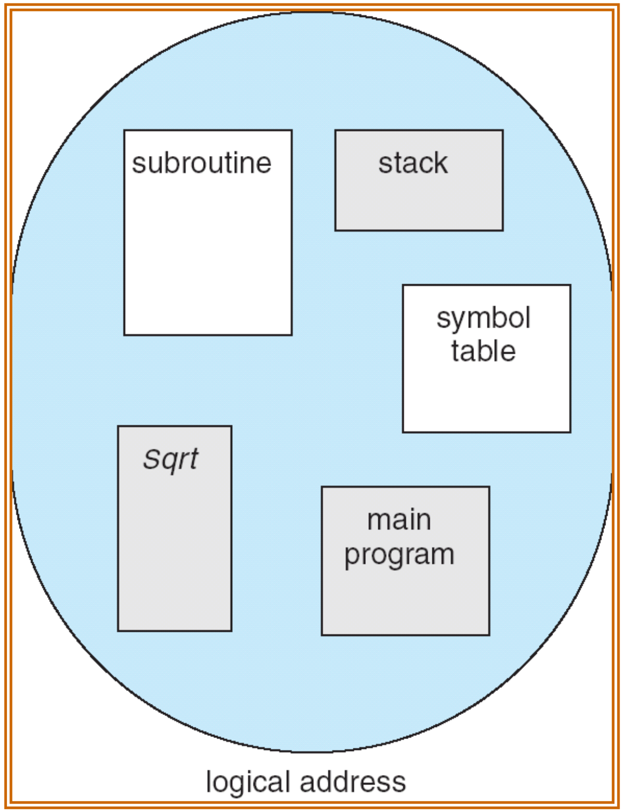
- **First-fit**: Allocate the *first* hole that is big enough
- **Best-fit**: Allocate the *smallest* hole that is big enough; must search entire list, unless ordered by size
  - Produces the smallest leftover hole
- **Worst-fit**: Allocate the *largest* hole; must also search entire list
  - Produces the largest leftover hole

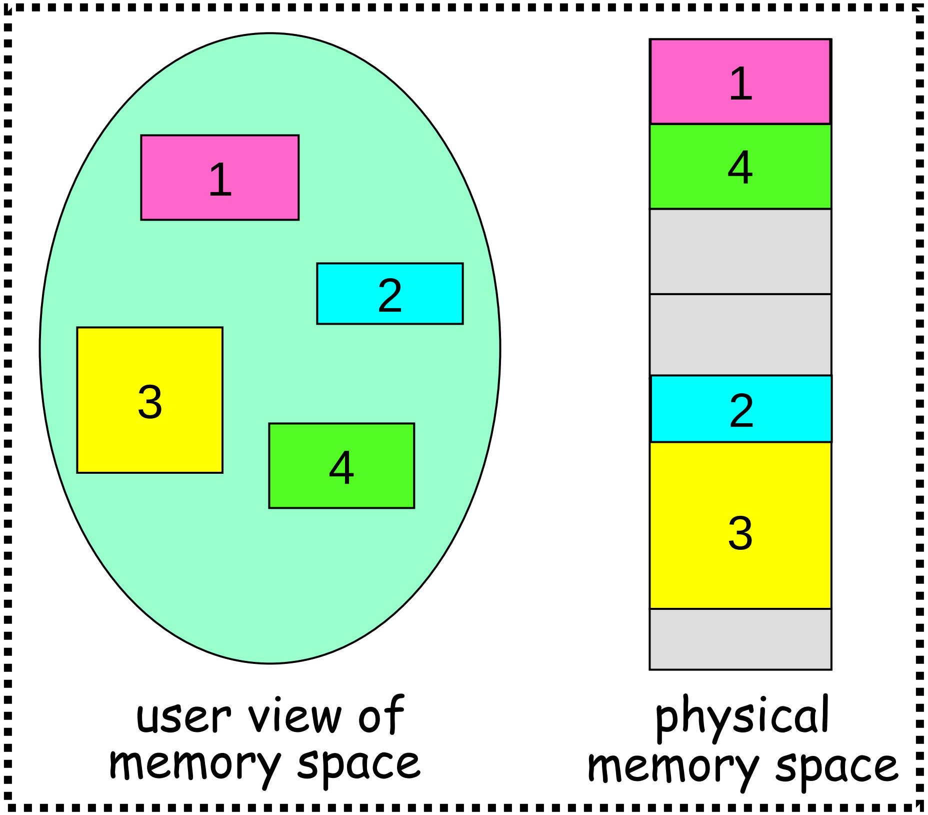
First-fit and best-fit are better than worst-fit in terms of speed and storage utilization. First-fit is faster than best-fit.

Fragmentation

- **External Fragmentation** – total memory space exists to satisfy a request, but it is not contiguous
  - First-fit and best-fit suffer most from this
  - One Solution: Compaction (not always possible) only for execution time binding
  - 2nd Solution: let process to get its DRAM allocated in non-contiguous fashion →Segmentation, Paging
- **Internal Fragmentation** – allocated memory may be slightly larger than requested memory; this size difference is memory internal to a partition, but not being used
  - Hole of 18,464 B, but process' requests 18,462 B
  - Overhead to keep track of hole is substantially larger than the hole itself!

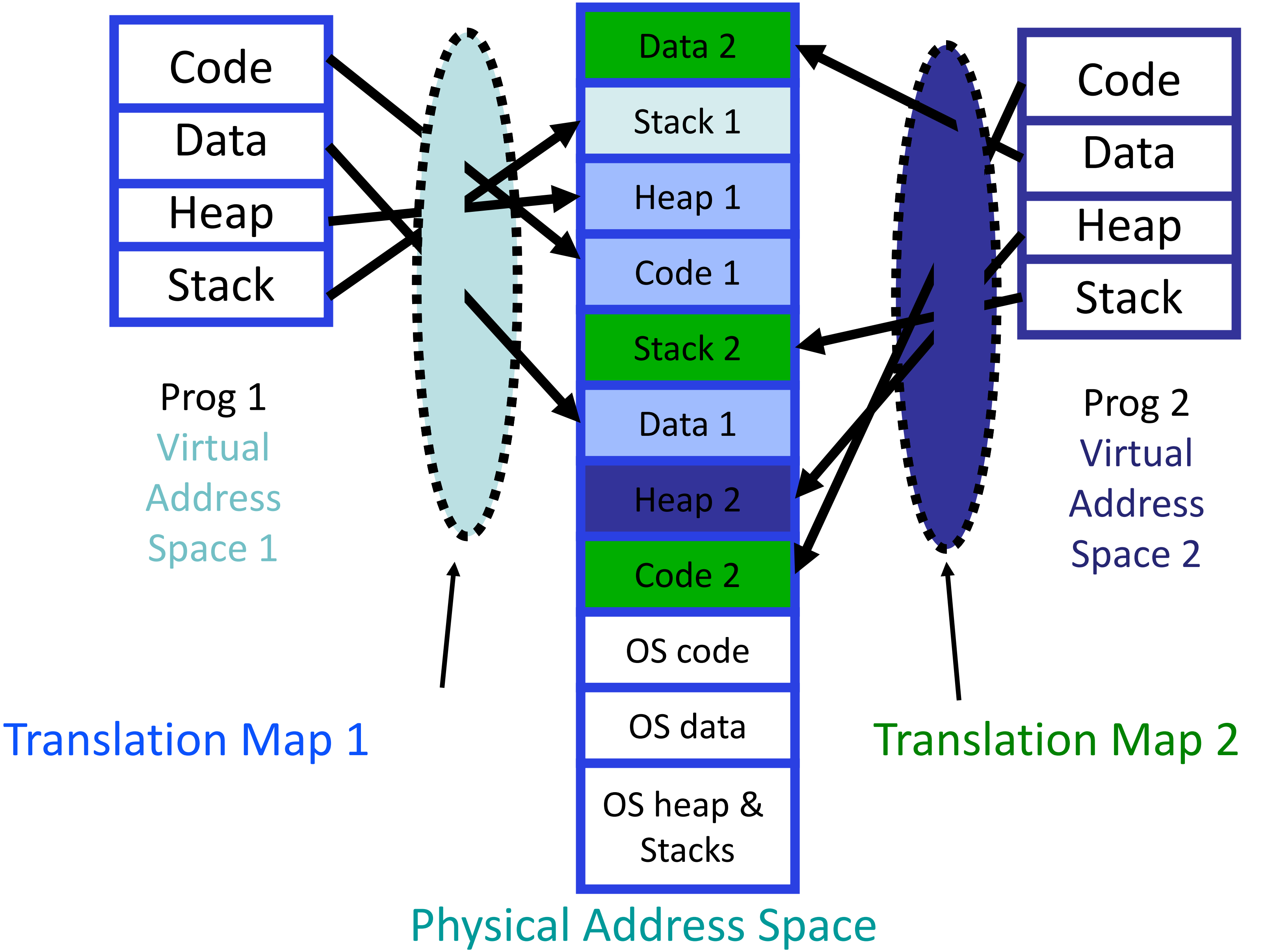
Segmentation



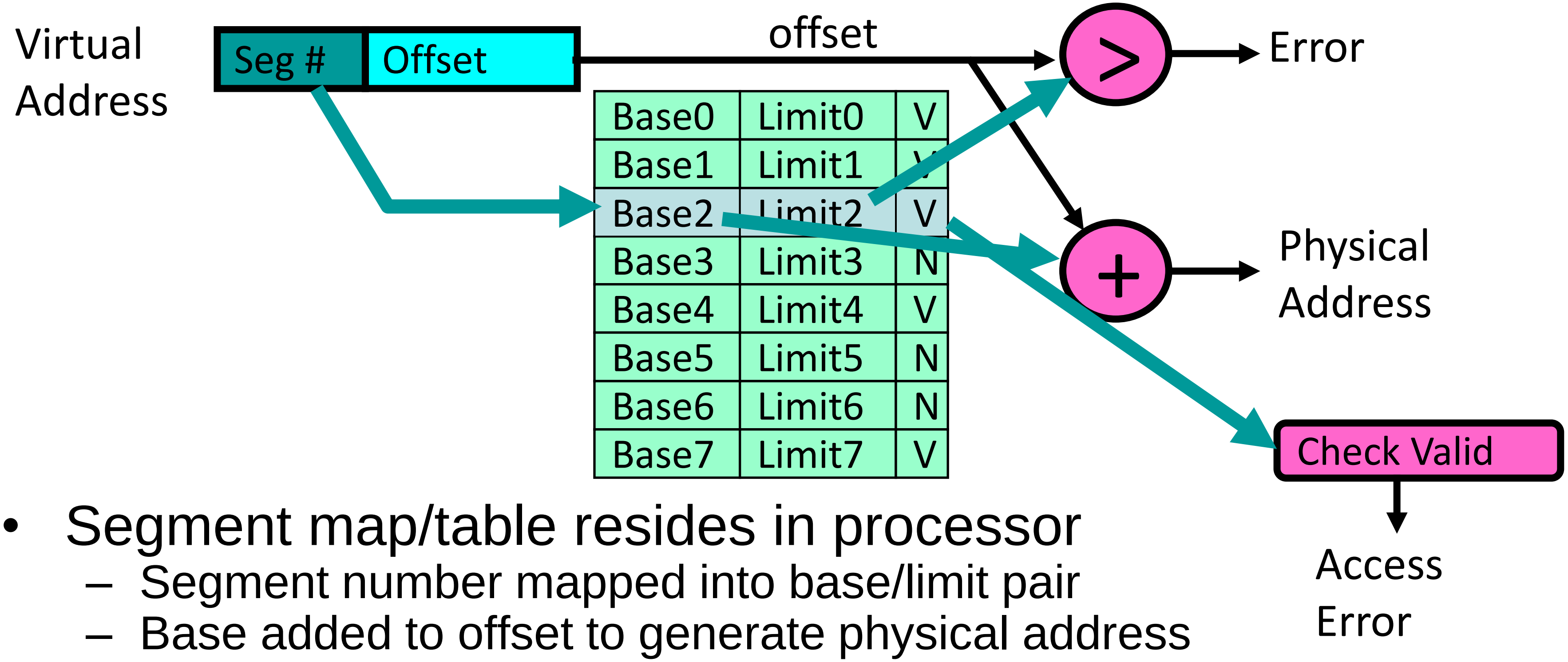


- Logical View: multiple separate segments
  - – Typical: Code, Data, Stack per each thread, Heap
  - – Others: memory sharing, Std C library
- Each segment is given a region of contiguous memory
  - – Has a separate base and limit values
  - – Can reside anywhere in physical memory☺!

Recall: General Address Translation



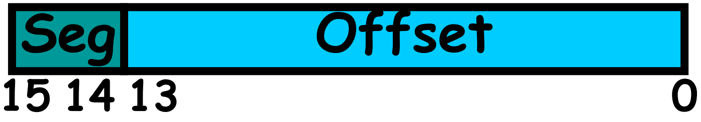
Implementation of Multi-Segment Model



Base0	Limit0	V
Base1	Limit1	V
Base2	Limit2	V
Base3	Limit3	N
Base4	Limit4	V
Base5	Limit5	N
Base6	Limit6	N
Base7	Limit7	V

- Segment map/table resides in processor
  - Segment number mapped into base/limit pair
  - Base added to offset to generate physical address
  - Error check catches offset out-of-range
- As many chunks of physical memory as entries
  - Segment addressed by portion of virtual address
  - However, could be included in instruction instead:
    - x86 example: mov ax, [ds:bx] → ax=m[ds:bx]
- What is “V/N” (valid / not valid)?
  - Can mark segments as invalid; requires check as well

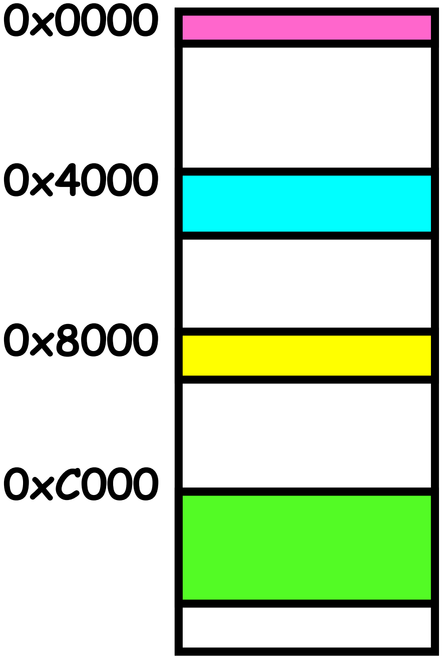
Example: Four Segments (16 bit addr)



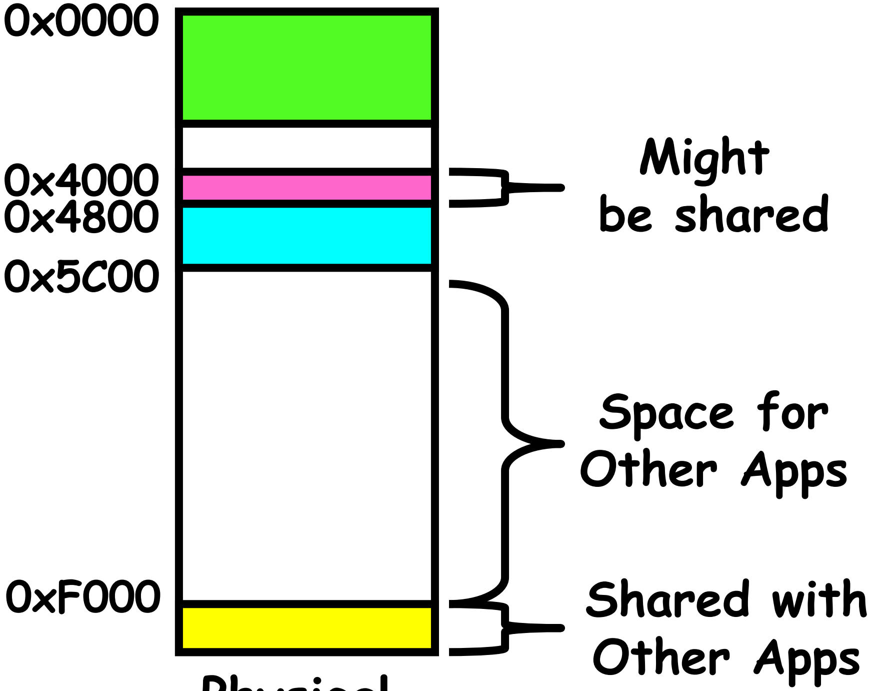
Virtual Address Format

Segment Table

Seg ID #	Base	Limit
0 (code)	0x4000	0x0800
1 (data)	0x4800	0x1400
2 (shared)	0xF000	0x1000
3 (stack)	0x0000	0x3000



Virtual Address Space

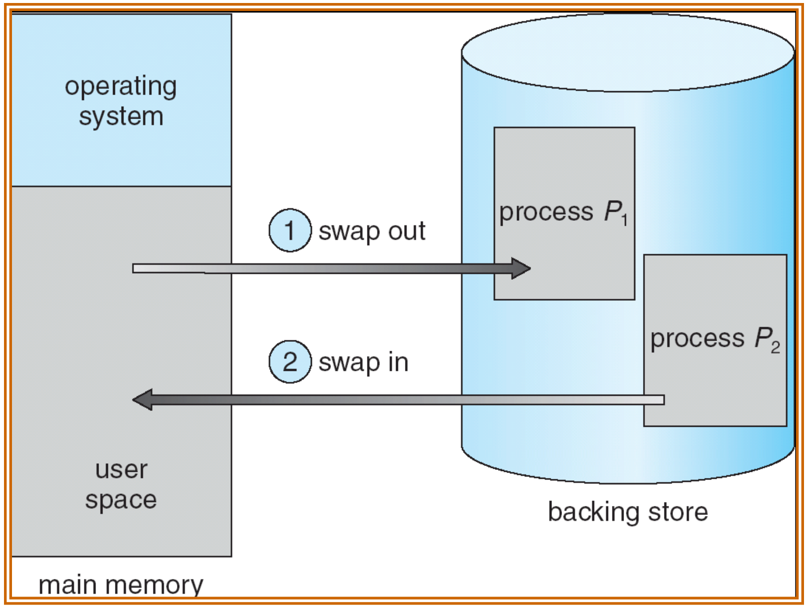


Physical Address Space

Observations about Segmentation

- Sharing of segments among multiple processes to achieve
  - sharing of CODE/libraries and to realize IPC
- Virtual address space has holes
  - Segmentation efficient for sparse address spaces
  - A correct program should never address gaps (except as mentioned in moment)
    - If it does, trap to kernel and dump core → segmentation fault
- When it is OK to address outside valid address range?
  - This is how the stack and heap are allowed to grow
  - For instance, stack takes fault, system automatically increases size of stack
- Need protection mode in segment table
  - For example, code segment would be read-only
  - Data and stack would be read-write (stores allowed)
  - Shared segment could be read-only or read-write
- What must be saved/restored on context switch?
  - Segment table stored in CPU, not in memory (small)
  - Or segment table base register (STBR)
  - Might store all of processes memory onto disk when switched (called “swapping”)

What if not all segments fit into memory?



- Extreme form of Context Switch: Swapping
  - In order to make room for next process, some or all of the previous process is moved to disk
    - Likely need to send out complete segments
  - This greatly increases the cost of context-switching
- Modified versions of swapping are found on many systems (i.e., UNIX, Linux, and Windows)
  - Swapping normally disabled
  - Started if more than threshold amount of DRAM allocated
  - Disabled again once DRAM utilization reduced below threshold

Swapping on Mobile Systems

- Not typically supported
  - Flash memory based
    - Small amount of space
    - Limited number of write cycles
    - Poor throughput between flash memory and CPU on mobile platform
- Instead use other methods to free memory if low
  - iOS ***asks*** apps to voluntarily relinquish allocated memory
    - Read-only data thrown out and reloaded from flash if needed
    - Failure to free can result in termination
  - Android terminates apps if low free memory, but first writes **application state** to flash for fast restart
  - Both OSes support paging as discussed next

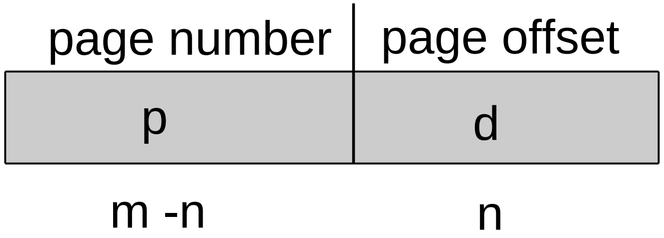
- Desirable alternative to Segmentation?
  - Some way to keep only active portions of a process in memory at any one time
    - Helps in improving the degree of multiprogramming
  - Need finer granularity control over physical memory

Paging: Physical Memory in Fixed Size Chunks

- Problems with segmentation?
  - Must fit variable-sized chunks into physical memory
  - May move processes multiple times to fit everything
  - Limited options for swapping to disk
- Fragmentation: wasted space
  - External: free gaps between allocated chunks
  - Internal: don't need all memory within allocated chunks
- Solution to fragmentation from segments?
  - Allocate physical memory in fixed size chunks ("**page frames**")
    - Frame size is a power of 2, between 512 B and 1 GB!
  - Every chunk of physical memory is equivalent
    - Can use simple vector of bits to keep track of free chunks and handle allocation:
      00110001110001101 … 110010
    - Each bit represents page of physical memory
      1⇒allocated, 0⇒free
- Divide logical memory of process into blocks of same size called **pages**
- Should pages be as big as our previous segments?
  - No: Can lead to lots of internal fragmentation
    - Typically have small pages (4K-16K)
  - Consequently: need multiple pages per each segment of the process
- To run a program of size ***N*** pages, need to find ***N*** free frames in the main memory to load it from the secondary memory (HDD/SSD)

Address Translation in Paging

- Logical Address generated by CPU is divided into:
  - **Page number** (***p***) – used as an index into a **page table** which contains base address of each page (aka page frame ID) in physical memory
  - **Page offset** (***d***) – combined with base address to define the physical memory address that is sent to the memory unit



page number

page offset

p

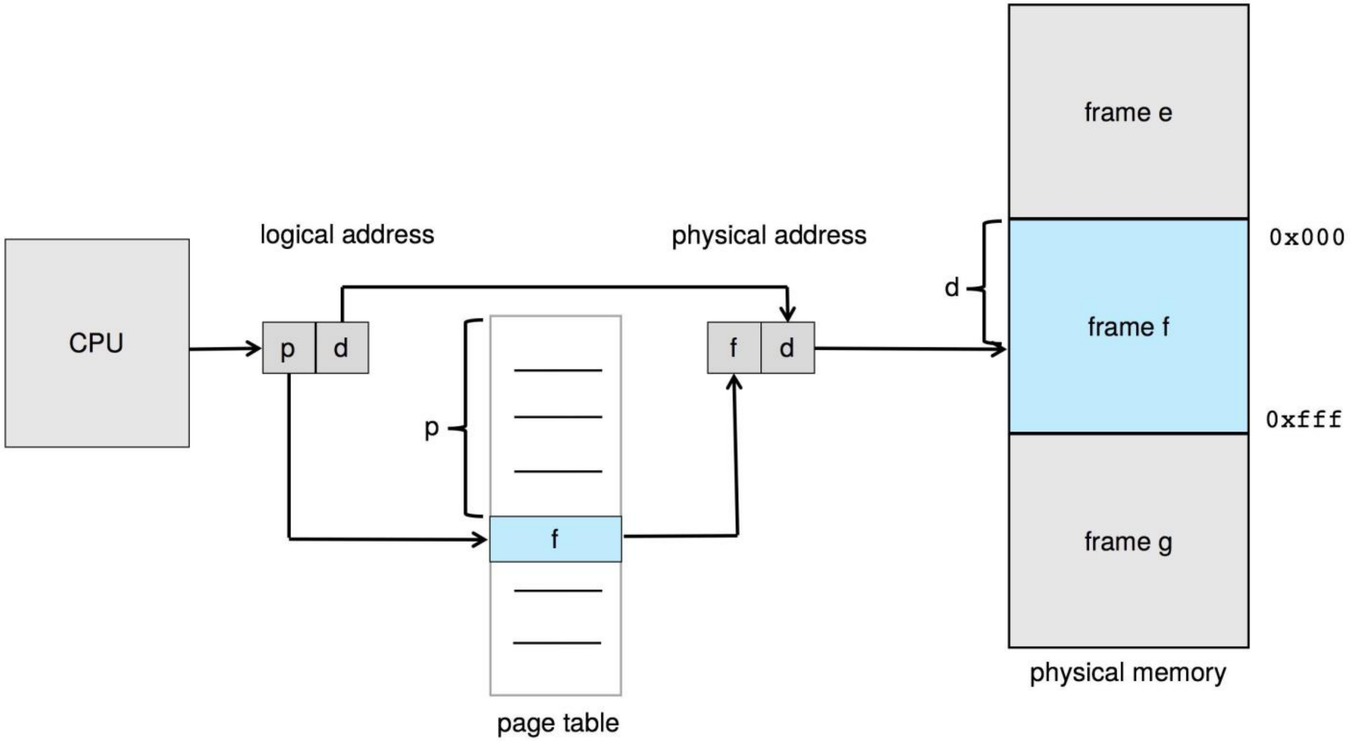
d

m -n

n

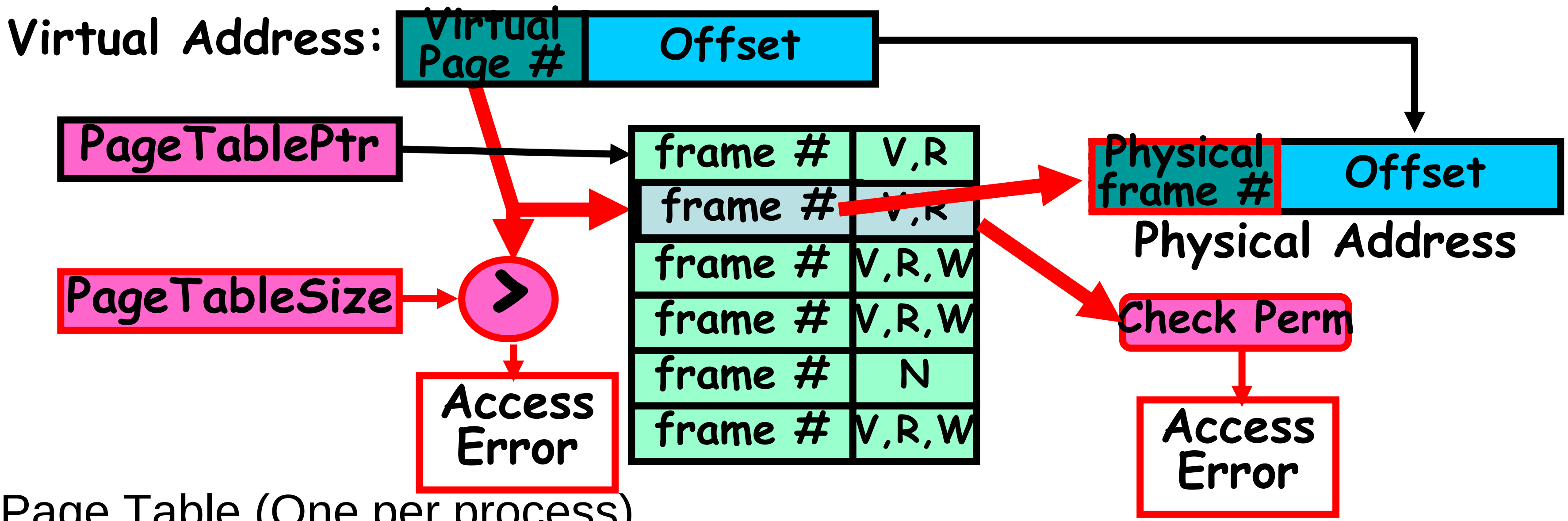
For a given logical address space 2^m with page size of 2^n

Paging Hardware



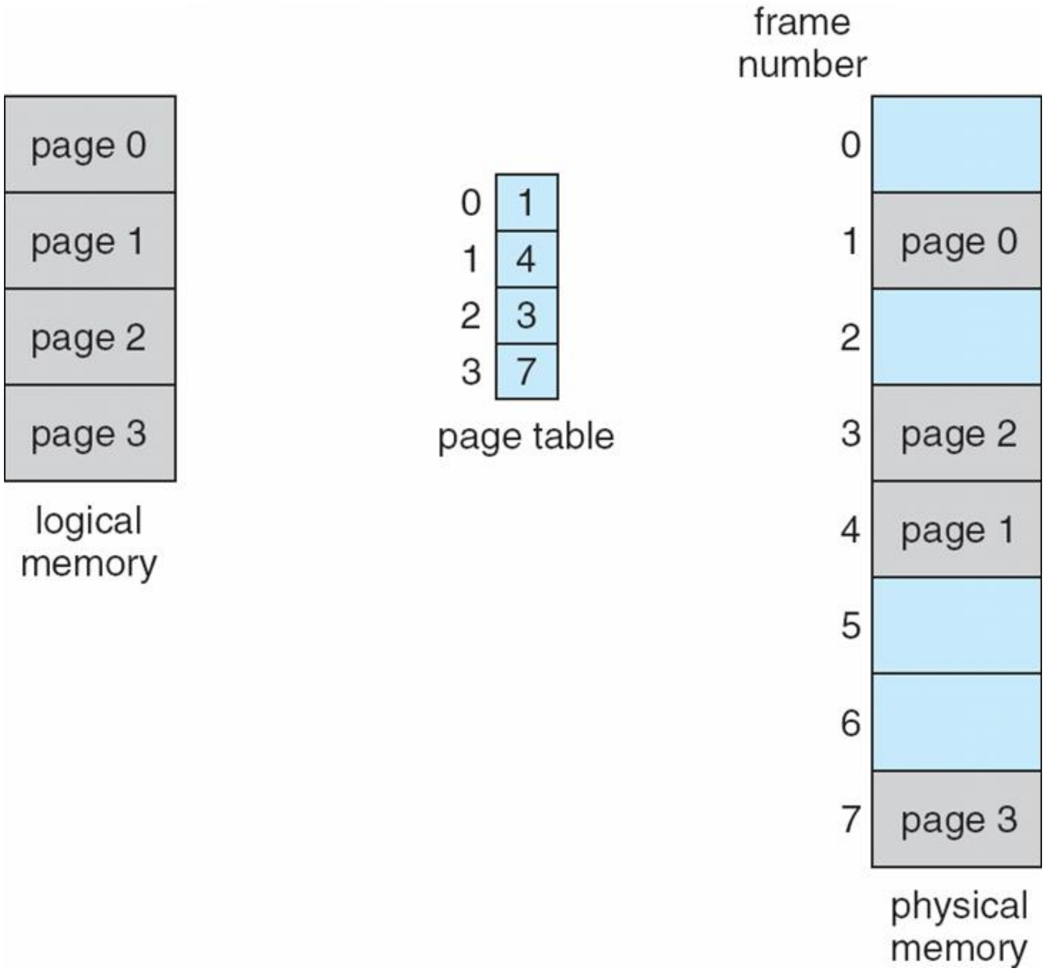
- **Page table** to translate logical to physical addresses
- Backing store (HDD) likewise split into pages of same size or a cluster of pages

How to Implement Paging?

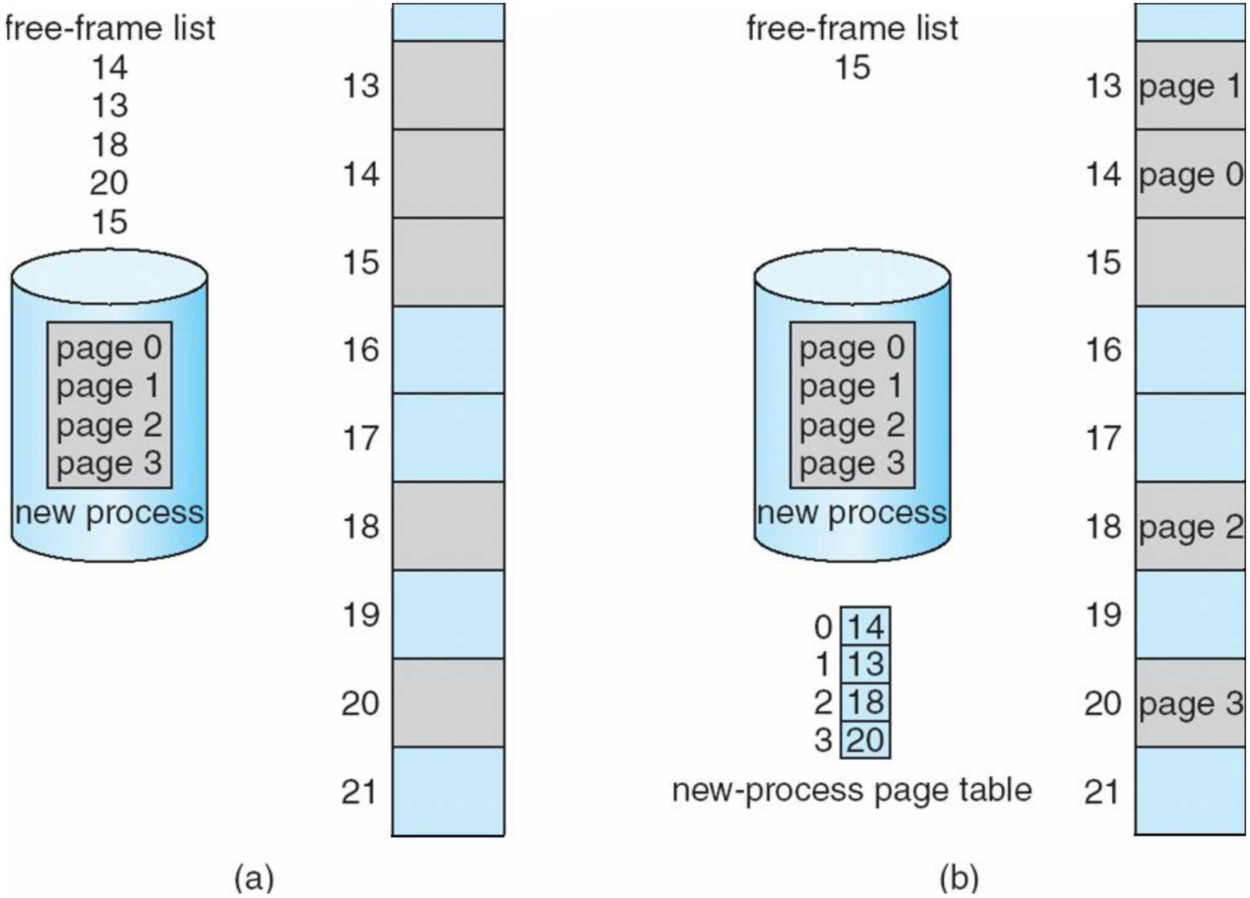


- Page Table (One per process)
  - Resides in physical memory!
  - Contains physical page (page frame) ID (#) and permissions for each virtual page
    - Permissions include: Valid bits, Read, Write, etc
- Virtual address mapping
  - Offset from Virtual address copied to Physical Address
    - Example: 10 bit offset ⇒ 1024-byte pages
  - Virtual page # (or simply page no) is all remaining bits
    - Example for 32-bits: 32-10 = 22 bits, i.e. 4 million entries
    - Physical page (frame) # ID copied from table to get physical address
  - Check Page Table bounds and permissions
  - Comparison with Segmentation hardware?

Paging Example



Free Frames

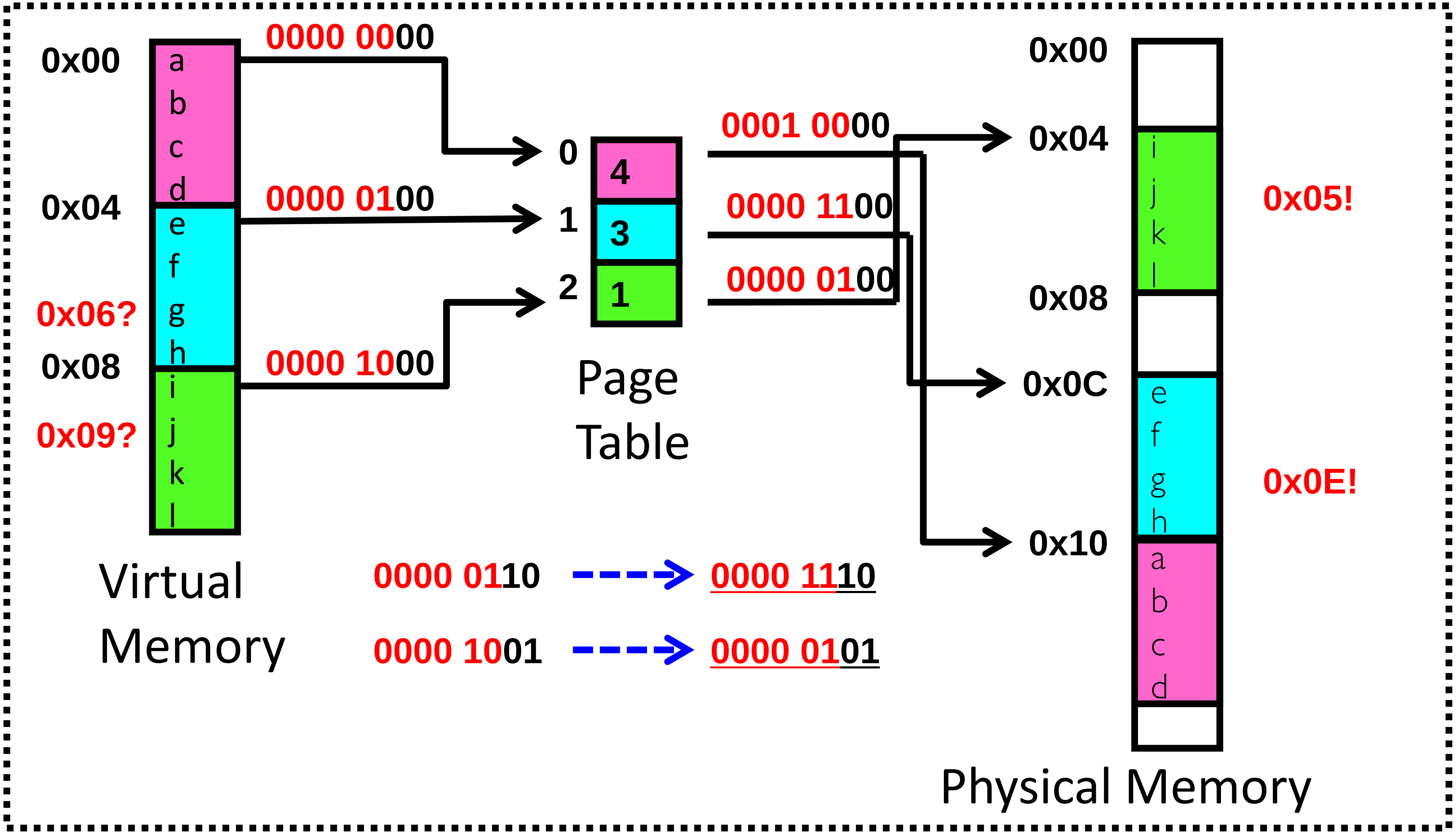


Before allocation

After allocation

Simple Page Table Example

Example (4 byte pages)

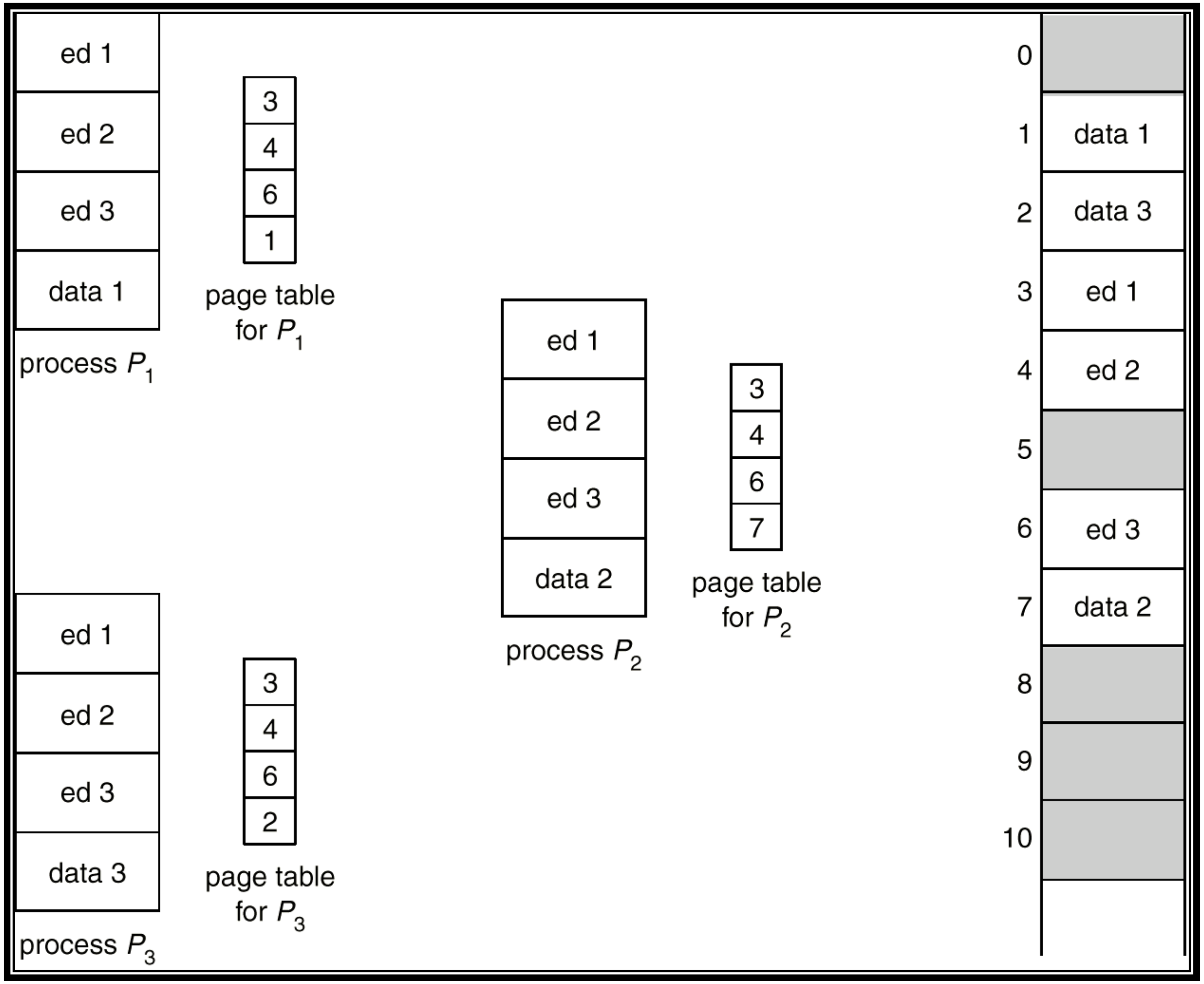


Shared Pages

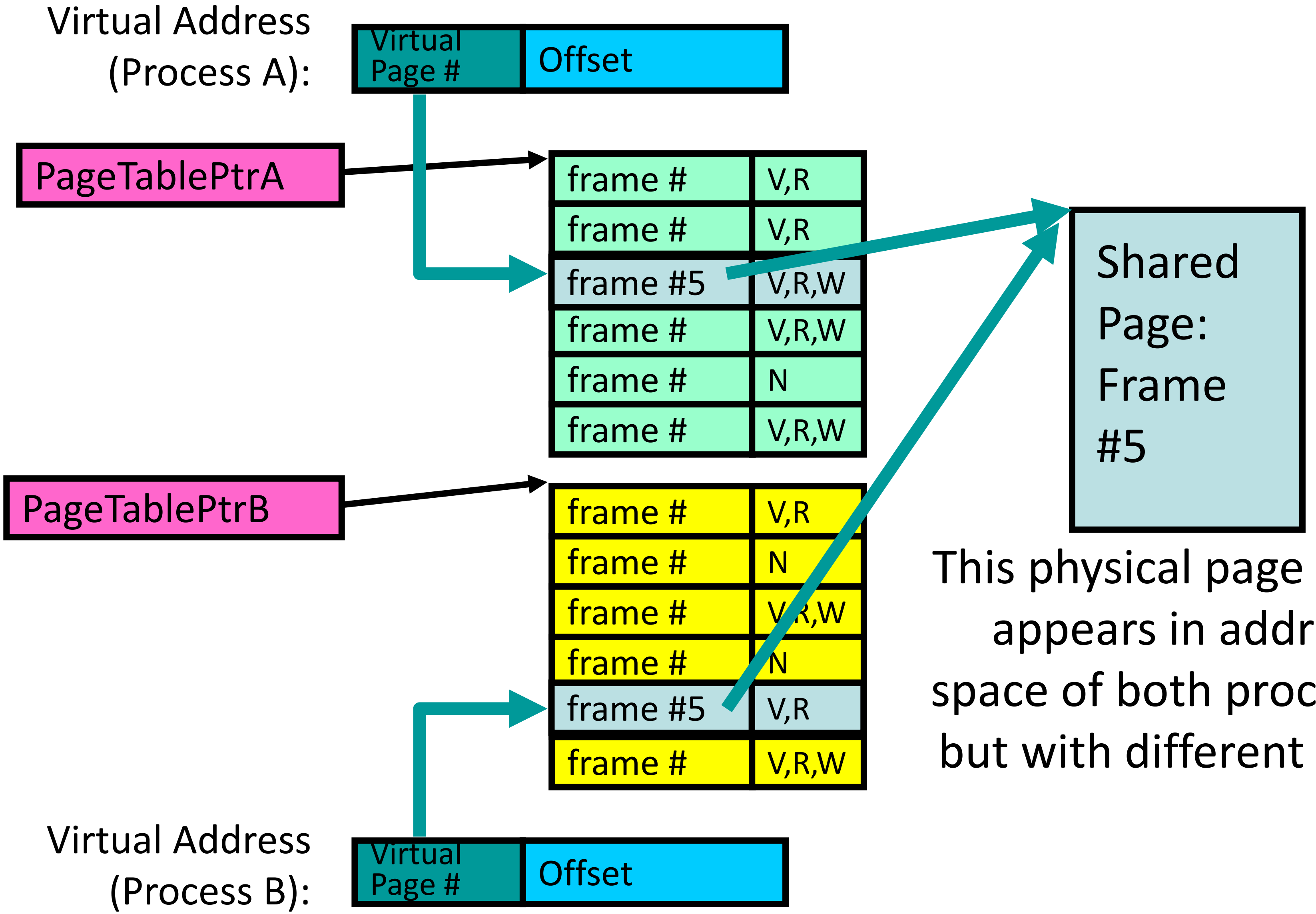
- Shared code
  - One copy of read-only (reentrant) code shared among processes
  - E.g., text editors, compilers, window systems

- Private code and data
  - Each process keeps a separate copy of the code and data

Shared Pages Example



What about Sharing?



Virtual Address (Process A):

Virtual Page #	Offset

PageTablePtrA

frame #	V,R
frame #	V,R
frame #5	V,R,W
frame #	V,R,W
frame #	N
frame #	V,R,W

PageTablePtrB

frame #	V,R
frame #	N
frame #	V,R,W
frame #	N
frame #5	V,R
frame #	V,R,W

Virtual Address (Process B):

Virtual Page #	Offset

Shared Page: Frame #5

This physical page frame appears in address space of both processes, but with different rights

Paging -- Internal fragmentation

- Page size = 2,048 bytes
- Process size = 72,766 bytes
- 35 pages + 1,086 bytes
- Internal fragmentation of 2,048 - 1,086 = 962 bytes
- Worst case fragmentation = 1 frame with 1 byte
- On average fragmentation = 1 / 2 frame size
- So small frame sizes desirable?
- But each page table entry takes memory to track
- Page sizes growing over time
  - Solaris supports two page sizes – 8 KB and 4 MB

What is in a PTE?

- What is in a Page Table Entry (or PTE)?
  - Pointer to next-level page table or to actual page in the main memory
  - Permission bits: valid, read-only, read-write, execute-only, etc
- Example: Intel x86 architecture PTE:
  - Addressing format (10, 10, 12-bit offset)
  - Top-level page tables called “Directories”

Page Frame Number (Physical Page Number)	Free (OS)	0	L	D	A	PCD	PWT	U	W	P
31-12	11-9	8	7	6	5	4	3	2	1	0

P: Present (same as “valid” bit in other architectures)

W: Writeable

U: User accessible

PWT: Page write transparent: external cache write-through

PCD: Page cache disabled (page cannot be cached)

A: Accessed: page has been accessed recently

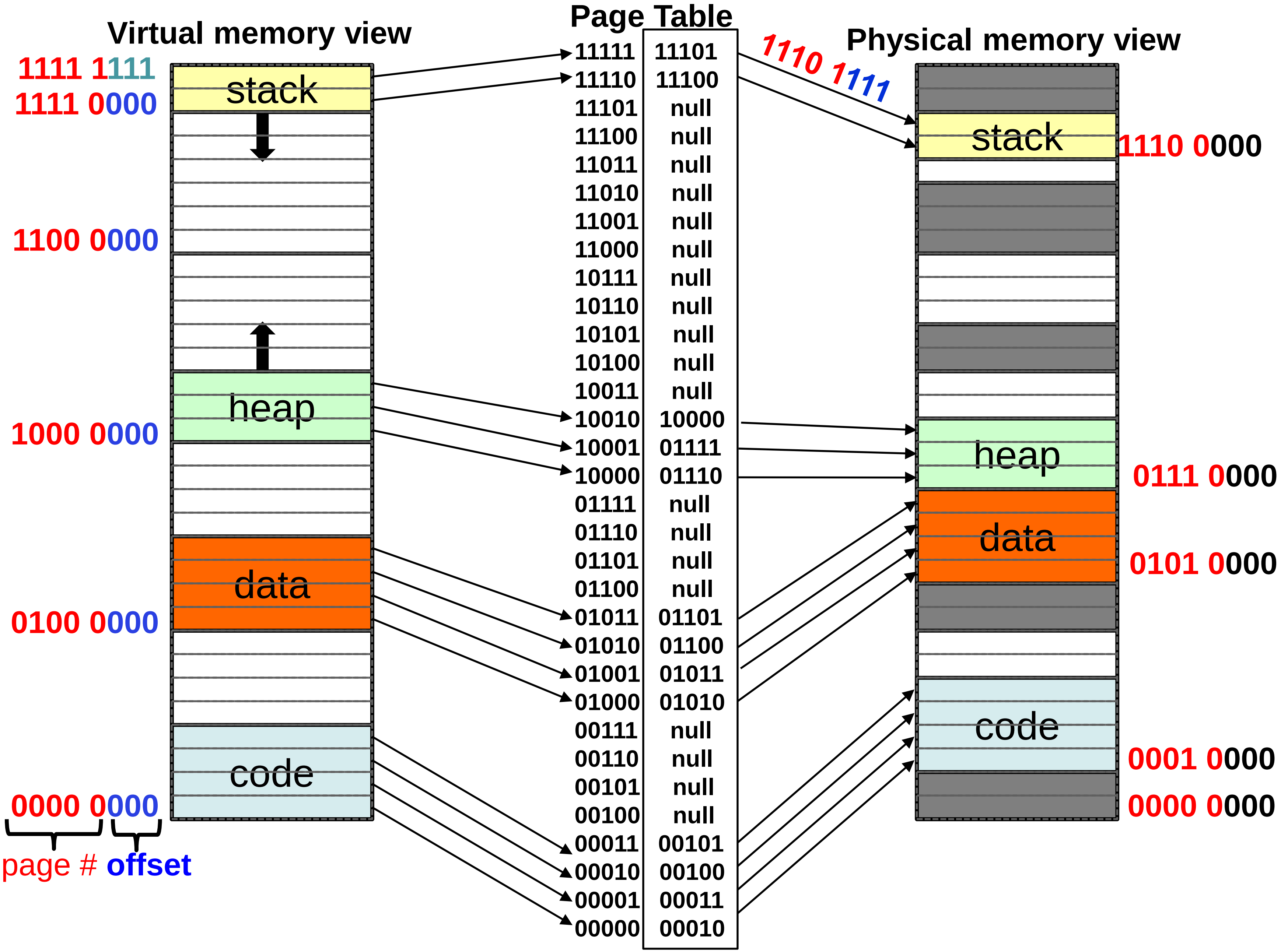
D: Dirty (PTE only): page has been modified recently

L: L=1⇒4MB page (directory only).
Bottom 22 bits of virtual address serve as offset

Examples of how to use a PTE

- How do we use the PTE?
  - Invalid PTE can imply different things:
    - Region of address space is actually invalid in virtual address space or
    - Page/directory is just somewhere else than in main memory
  - Validity checked first
    - OS can use other (say) 31 bits for location info
- Usage Example: Demand Paging
  - Keep only active pages in memory
  - Place others on disk and mark their PTEs invalid
- Usage Example: Copy on Write
  - UNIX fork gives *copy* of parent address space to child
    - Address spaces disconnected after child created
  - How to do this cheaply?
    - Make copy of parent's page tables (point at same memory)
    - Mark entries in both sets of page tables as read-only
    - Page fault on write access to a page creates two copies of that page
- Usage Example: Zero Fill On Demand
  - New data pages must carry no information (say be zeroed)
  - Mark PTEs as invalid; page fault on use gets zeroed page
  - Often, OS creates zeroed pages in background

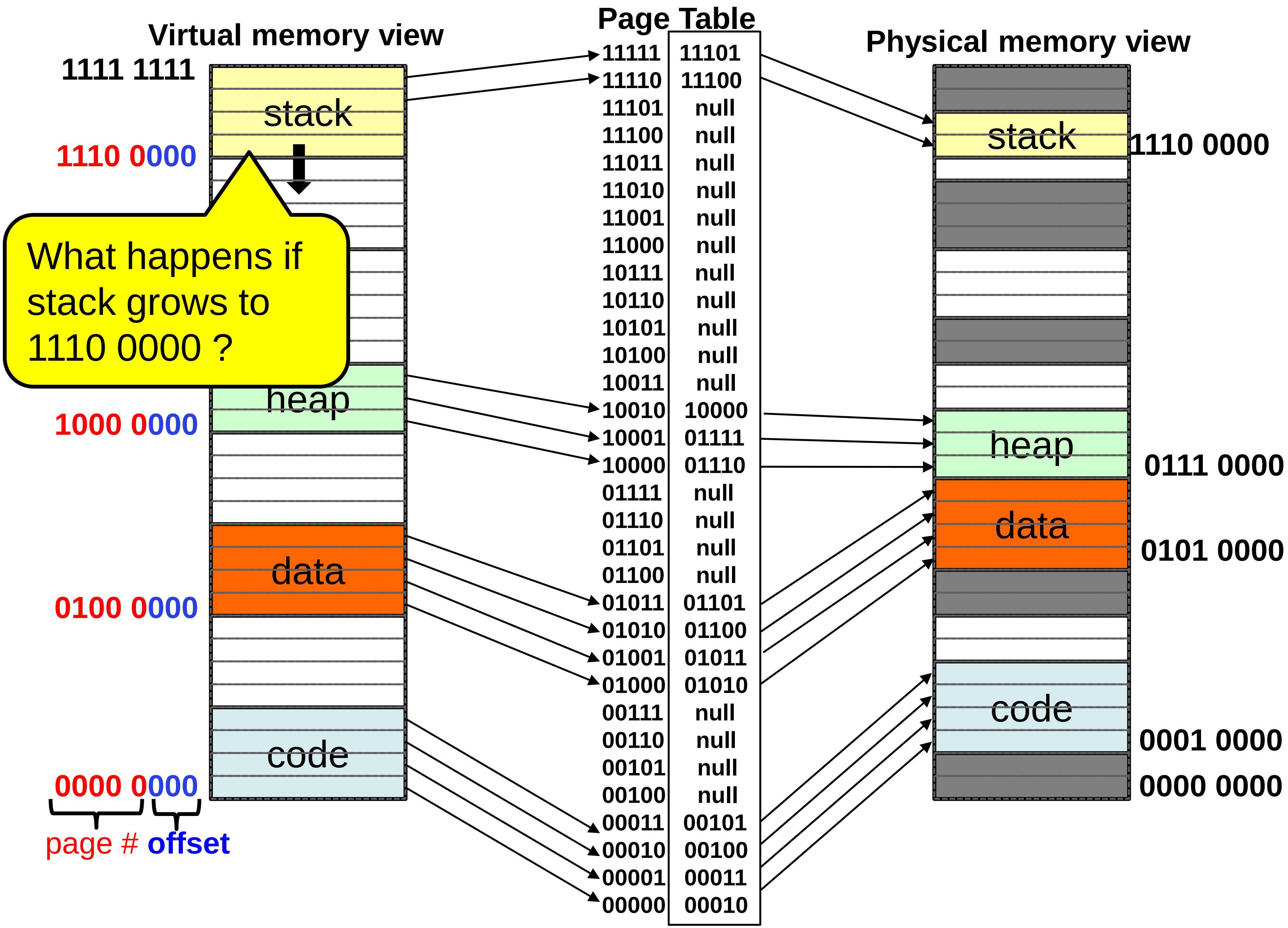
Summary: Paging



Page Table

11111	11101
11110	11100
11101	null
11100	null
11011	null
11010	null
11001	null
11000	null
10111	null
10110	null
10101	null
10100	null
10011	null
10010	10000
10001	01111
10000	01110
01111	null
01110	null
01101	null
01100	null
01011	01101
01010	01100
01001	01011
01000	01010
00111	null
00110	null
00101	null
00100	null
00011	00101
00010	00100
00001	00011
00000	00010

Summary: Paging

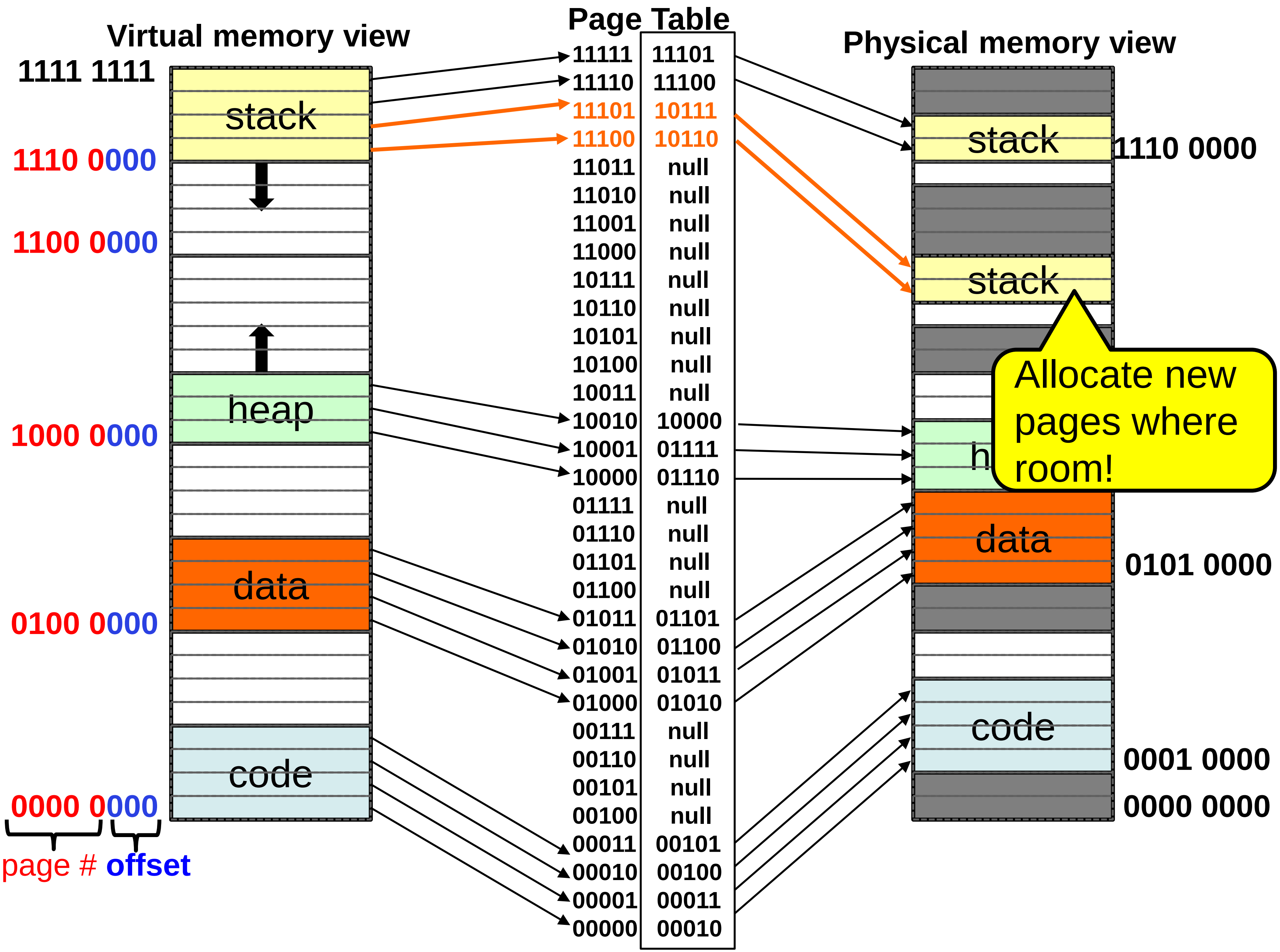


What happens if stack grows to 1110 0000 ?

Page Table

11111	11101
11110	11100
11101	null
11100	null
11011	null
11010	null
11001	null
11000	null
10111	null
10110	null
10101	null
10100	null
10011	null
10010	10000
10001	01111
10000	01110
01111	null
01110	null
01101	null
01100	null
01011	01101
01010	01100
01001	01011
01000	01010
00111	null
00110	null
00101	null
00100	null
00011	00101
00010	00100
00001	00011
00000	00010

Summary: Paging



Page Table

11111	11101
11110	11100
11101	10111
11100	10110
11011	null
11010	null
11001	null
11000	null
10111	null
10110	null
10101	null
10100	null
10011	null
10010	10000
10001	01111
10000	01110
01111	null
01110	null
01101	null
01100	null
01011	01101
01010	01100
01001	01011
01000	01010
00111	null
00110	null
00101	null
00100	null
00011	00101
00010	00100
00001	00011
00000	00010

Allocate new pages where room!

Paging Discussion

- What needs to be switched on a context switch?
  - Page table pointer and limit/length pointer
- Analysis
  - Pros
    - Page Table: An array of structures
    - Simplifies free-space allocation using bitmap
    - Easy to share by setting flags of those frames in PTs
    - Copy-on-Write, Zero Fill on Demand
  - Con: What if address space is sparse?
    - Stack and heap grow dynamically, so cause sparsity
    - E.g., on UNIX, code starts at 0, stack starts at (2^{31}-1)
    - With 1K pages, need 4 million page table entries!
      - With PTBR, PT needs to be stored contiguously in DRAM!
    - Multi-threading: more stacks, each needs to grow!
  - Con: What if table is really big?
    - 64-bit virtual address space → PT is almost empty!
    - Not all pages used all the time ⇒ would be nice to have only the working set of page table in memory
- Better data structure than arrays for lookups in sparse address space?
  - Trees and hash tables
    - Multi-level translation: Multi-level paging or combining paging and segmentation

Page Table Structures

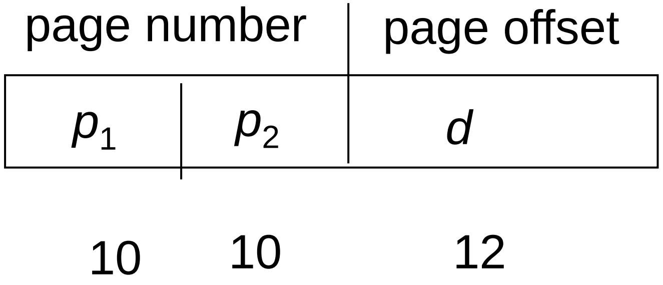
- Hierarchical Paging
- Paged Segmentation
- Hashed Page Tables
- Inverted Page Tables

Hierarchical Page Tables

- Break up the logical address space into multiple page tables.
- A simple technique is a two-level page table.

Two-Level Paging Example

- A logical address (on 32-bit machine with 4KB page size and 4B page table entry) is divided into:
  - a page number consisting of 20 bits.
  - a page offset consisting of 12 bits.
- Page table is too big (2^{22} B) to fit in one page frame (4 KB)
  - No of frames needed to store page table is 2^{10}
  - Page table of Page table contains 2^{10} entries and fits in 1 page frame
- Since the page table is paged, the page number is further divided into:
  - a 10-bit page number.
  - a 10-bit page offset.
- Thus, a logical address is as follows:



page number | page offset

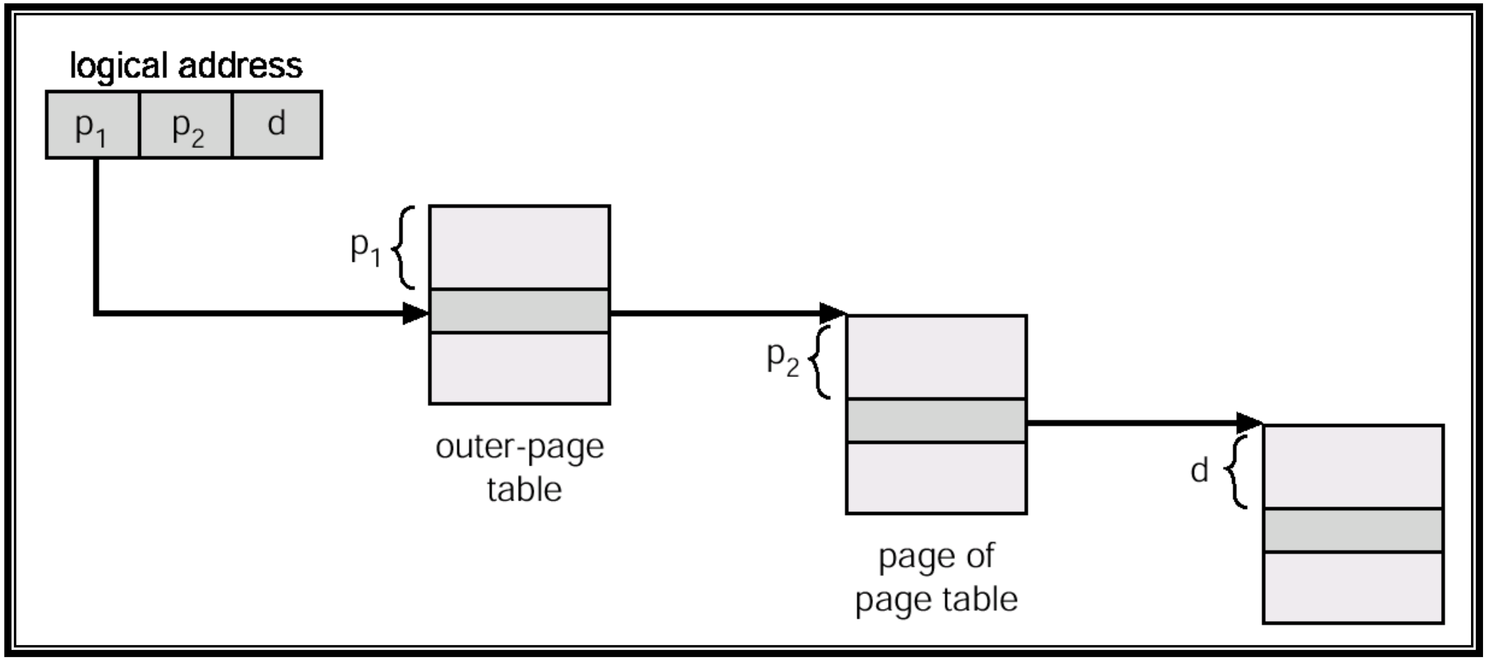
p_1 | p_2 | d

10 | 10 | 12

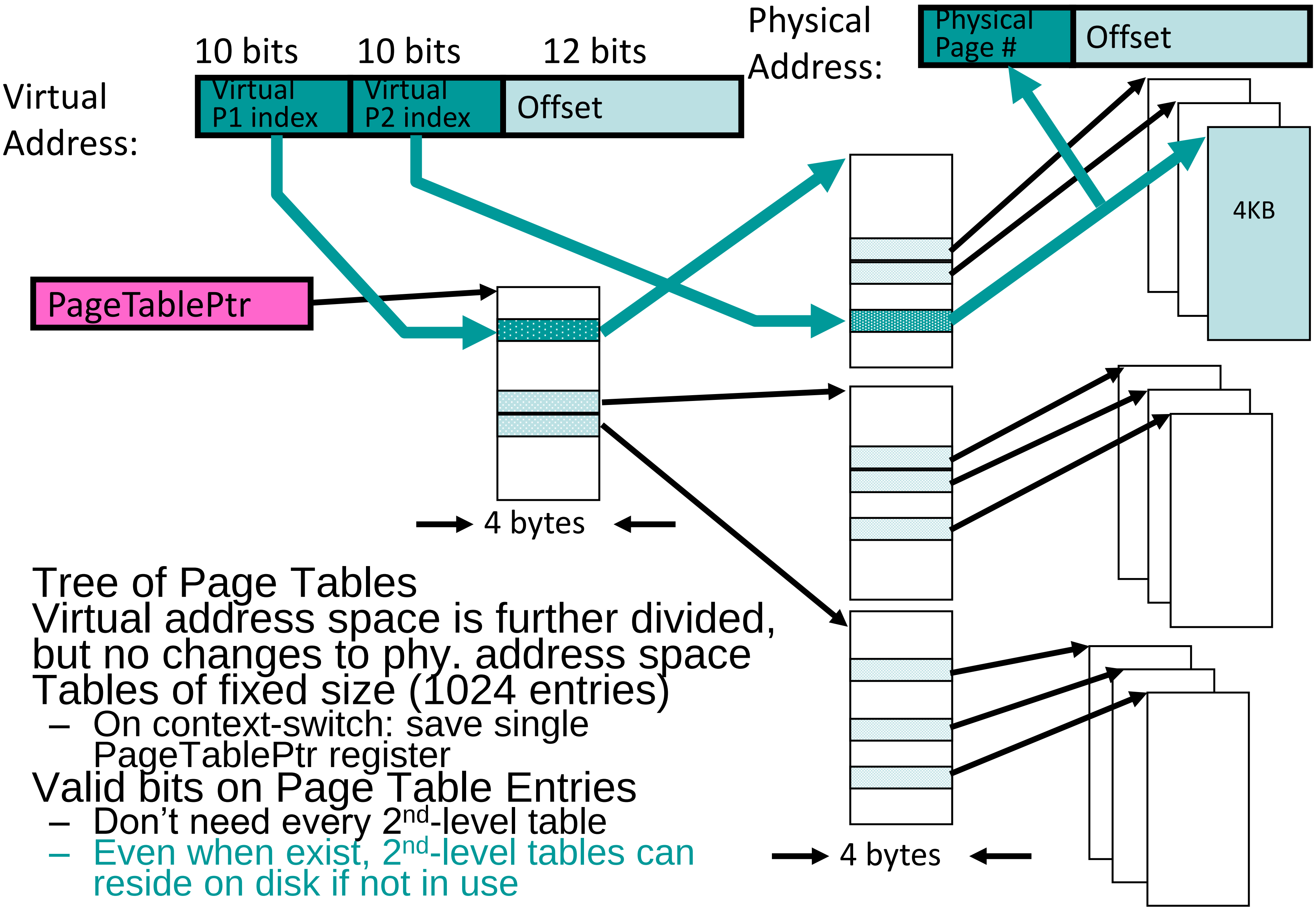
where p_1 is an index into the outer page table, and p_2 is the displacement within the page of the outer page table.

Address-Translation Scheme

- Address-translation scheme for a two-level 32-bit paging architecture

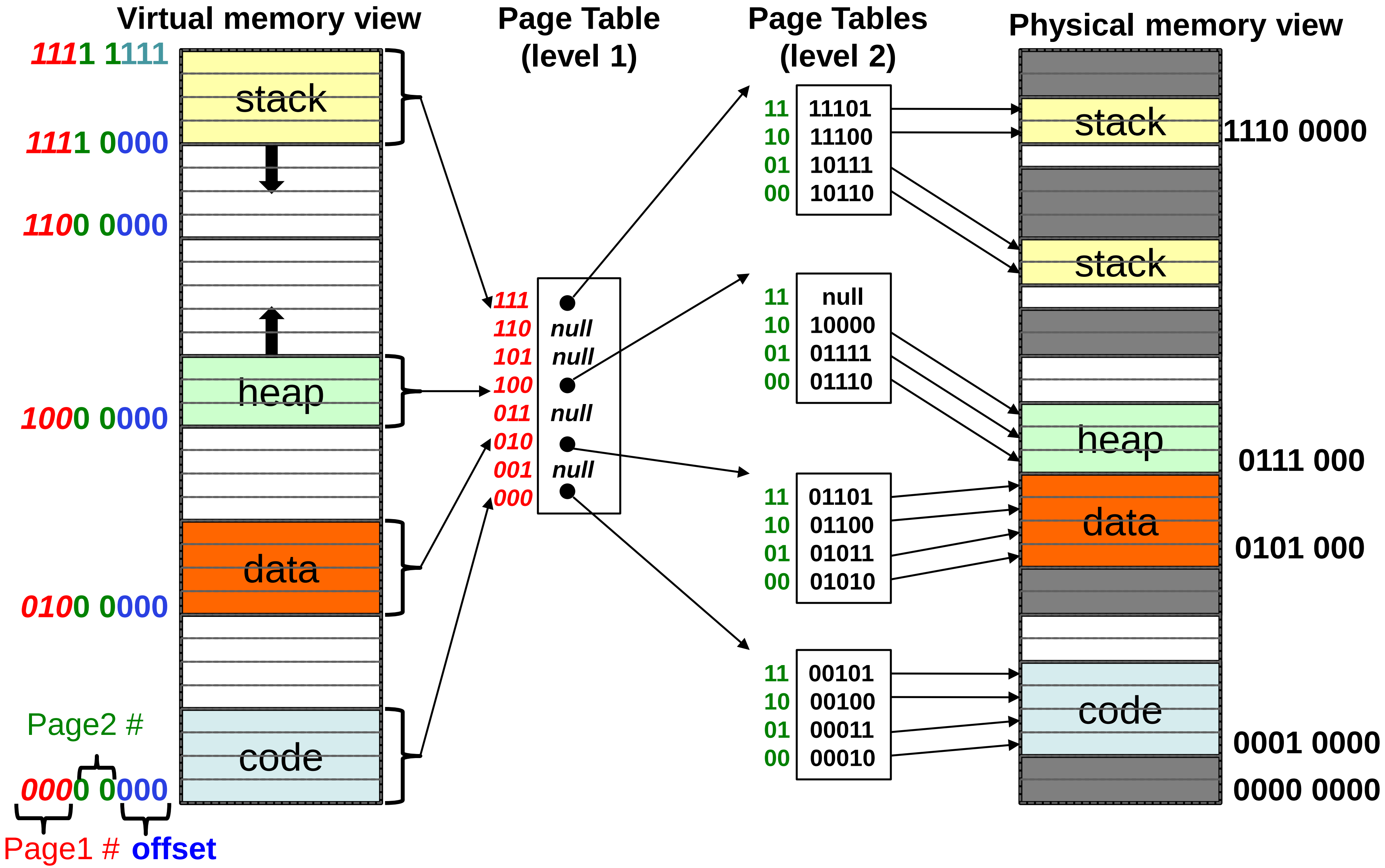


Fix for sparse address space: two-level page table



- Tree of Page Tables
- Virtual address space is further divided, but no changes to phy. address space
- Tables of fixed size (1024 entries)
  - On context-switch: save single PageTablePtr register
- Valid bits on Page Table Entries
  - Don’t need every 2nd-level table
  - Even when exist, 2nd-level tables can reside on disk if not in use

Summary: Two-Level Paging



Virtual memory view

1111 1111

1111 0000

1100 0000

1000 0000

0100 0000

0000 0000

stack

heap

data

code

Page2 #

Page1 #

offset

Page Table (level 1)

111	●
110	null
101	null
100	●
011	null
010	●
001	null
000	●

Page Tables (level 2)

11	11101
10	11100
01	10111
00	10110

11	null
10	10000
01	01111
00	01110

11	01101
10	01100
01	01011
00	01010

11	00101
10	00100
01	00011
00	00010

Physical memory view

stack

1110 0000

stack

heap

0111 000

data

0101 000

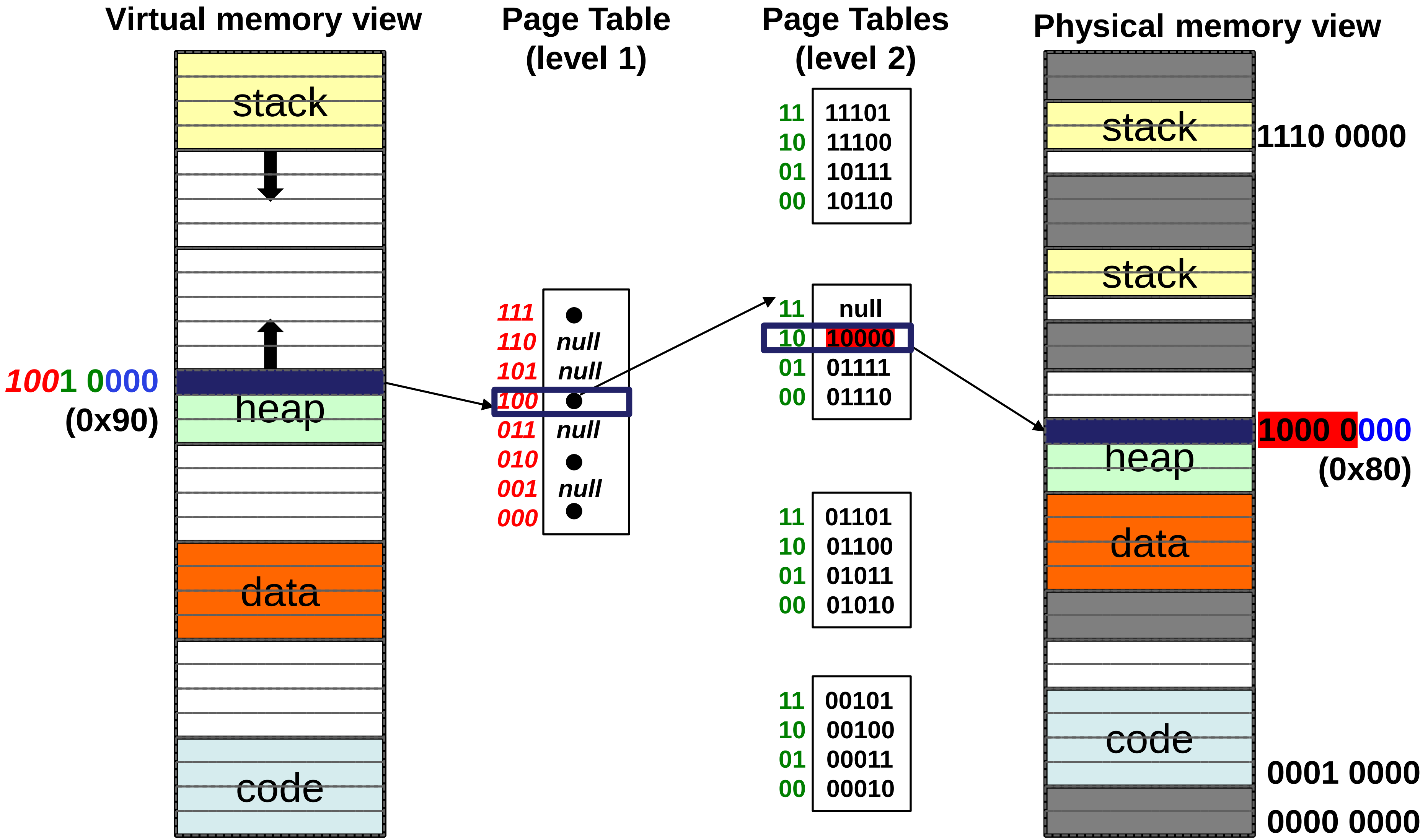
code

0001 0000

0000 0000

Compare with pure paging

Summary: Two-Level Paging



IA64: 64bit addresses: Six-level page table?!?

	7 bits	9 bits	9 bits	9 bits	9 bits	9 bits	12 bits
64bit Virtual Address:	Virtual P1 index	Virtual P2 index	Virtual P3 index	Virtual P4 index	Virtual P5 index	Virtual P6 index	Offset

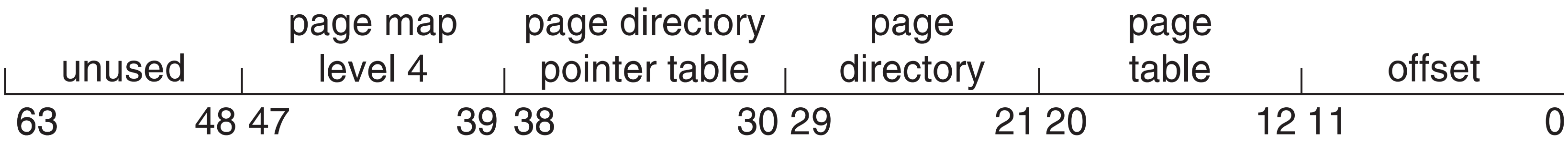
No!

Too slow

Too many almost-empty tables

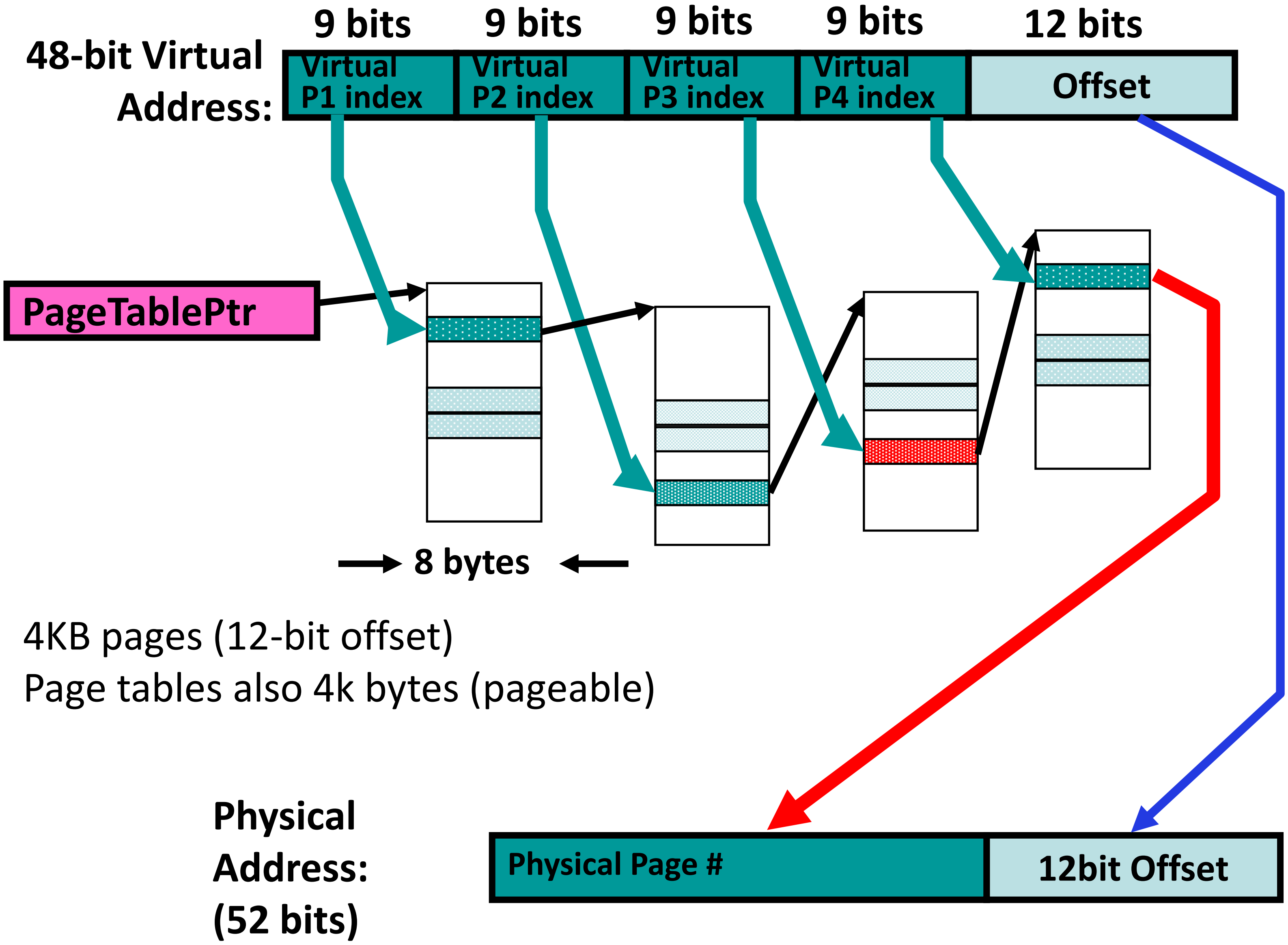
Intel/AMD: x86-64

- 64-bit addressing is ginormous (16 Exa Bytes)
- In practice only implement 48-bit addressing
  - Four levels of paging hierarchy
  - Page sizes of 4 KB, 2 MB, 1 GB
  - 2 MB/1GB are Superpages
  - Superpage: A set of contiguous pages in one page table so we can save on page table storage



unused	page map level 4	page directory pointer table	page directory	page table	offset
63 48	47 39	38 30	29 21	20 12	11 0

X86_64: Four-level page table!



48-bit Virtual Address:

9 bits | 9 bits | 9 bits | 9 bits | 12 bits

Virtual P1 index | Virtual P2 index | Virtual P3 index | Virtual P4 index | Offset

PageTablePtr

8 bytes

4KB pages (12-bit offset)

Page tables also 4k bytes (pageable)

Physical Address: (52 bits)

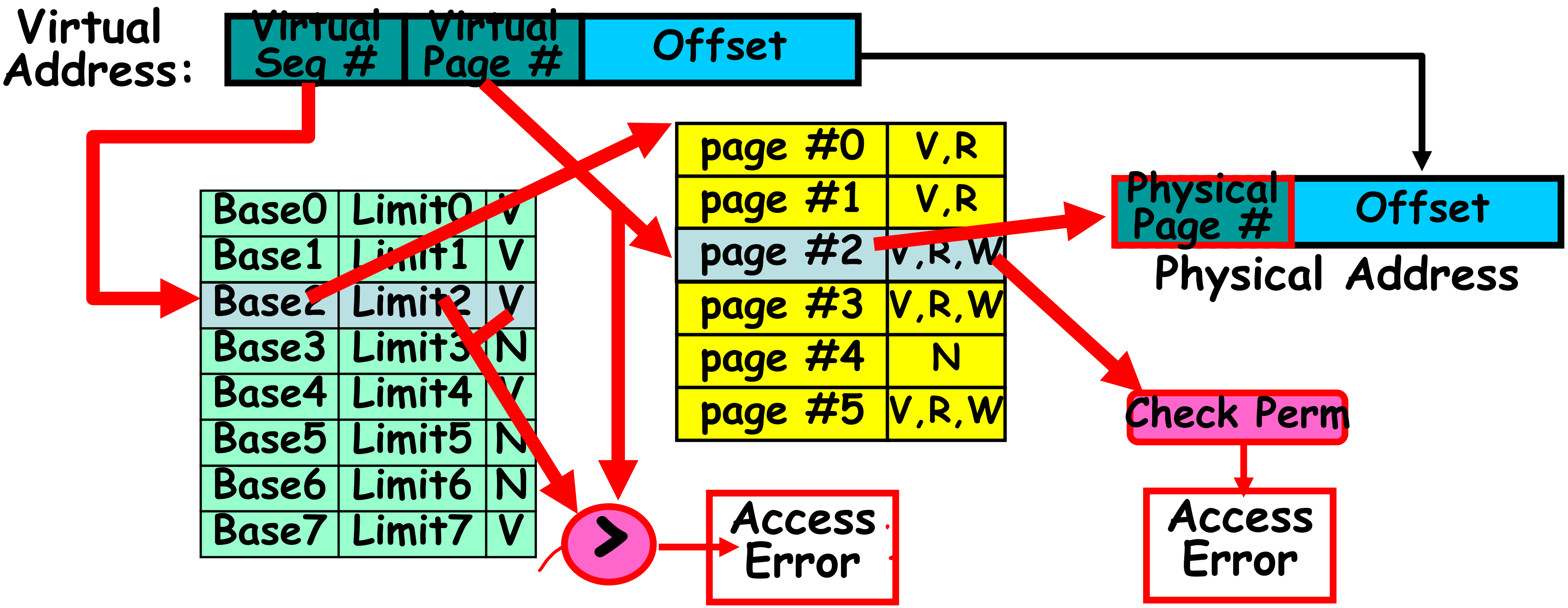
Physical Page # | 12bit Offset

Observations on Multi-level Paging

- Two-level paging
  - TLB cache hit → 1 memory access (discussed later)
  - But, TLB cache miss → 3 memory accesses
- Beyond 32-bit addressing, two-level paging is not sufficient
  - E.g., 64-bit addressing need 6-level paging!!
- N-Level paging → N+1 memory accesses on TLB miss!
- So, multi-level paging is very inefficient (slow)
- Alternatives are
  - Paged Segmentation
  - Hashed page tables
  - Inverted page tables

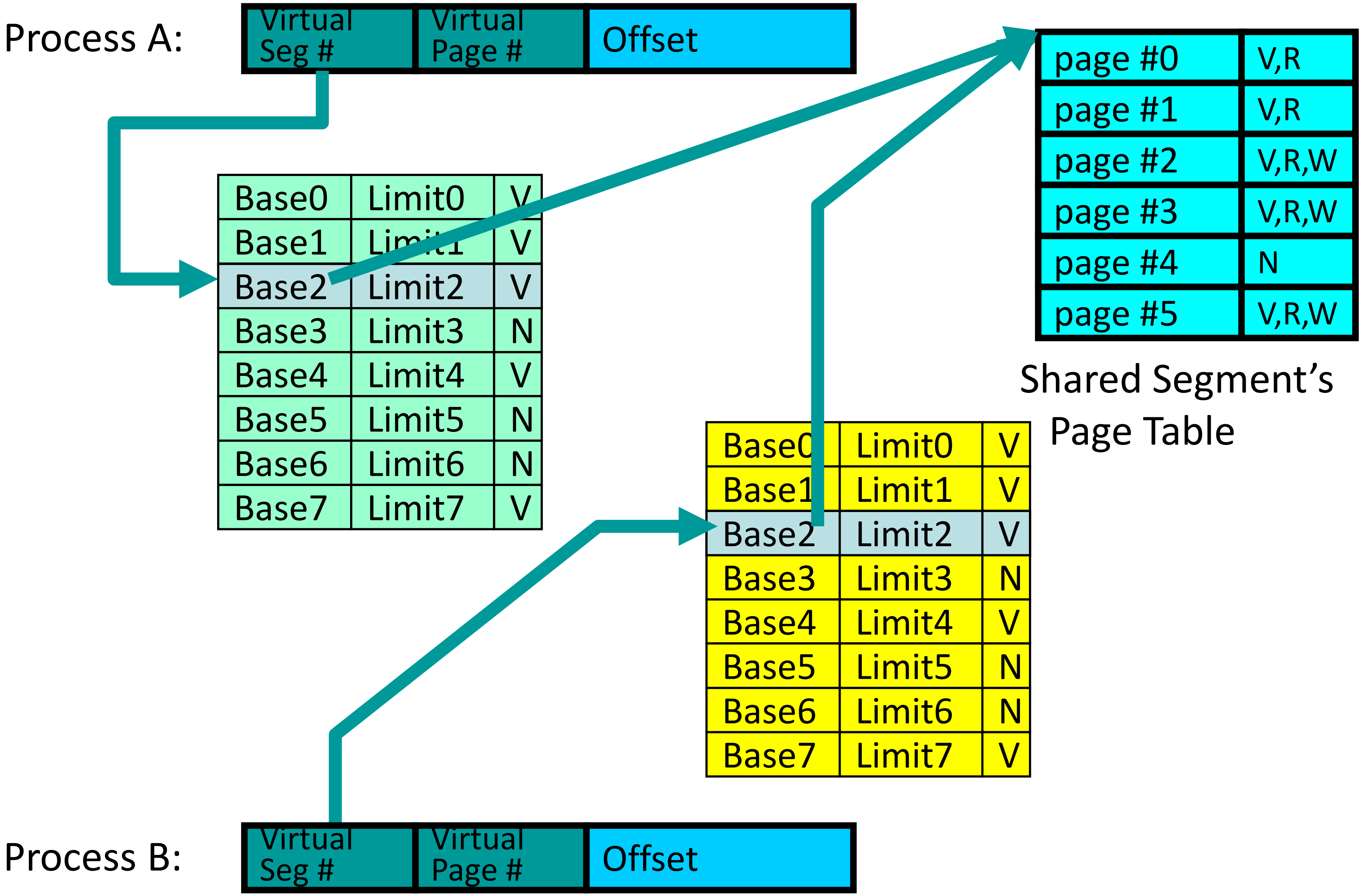
Multi-level Translation: Segments + Pages

- What about a tree of tables?
  - Lowest level page table⇒memory still allocated with bitmap
  - Higher levels often segmented → Paged Segmentation
- Could have any number of levels. Example (top segment):



- What must be saved/restored on context switch?
  - Contents of top-level segment registers (for this example)
  - Pointer to top-level table (in multi-level paging)

What about Sharing (Complete Segment)?



Process A:

Virtual Seg #	Virtual Page #	Offset

Base0	Limit0	V
Base1	Limit1	V
Base2	Limit2	V
Base3	Limit3	N
Base4	Limit4	V
Base5	Limit5	N
Base6	Limit6	N
Base7	Limit7	V

page #0	V,R
page #1	V,R
page #2	V,R,W
page #3	V,R,W
page #4	N
page #5	V,R,W

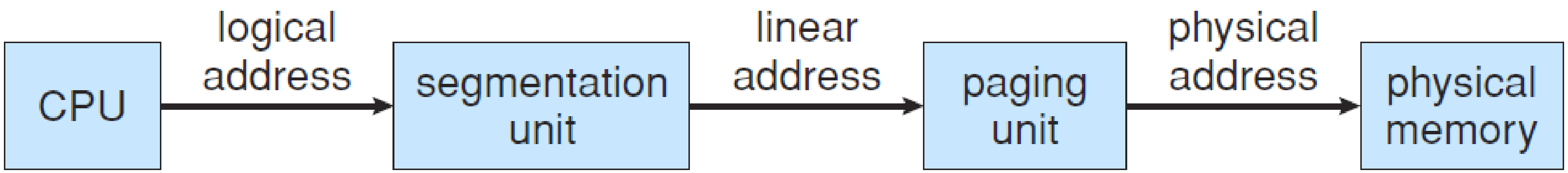
Shared Segment's Page Table

Base0	Limit0	V
Base1	Limit1	V
Base2	Limit2	V
Base3	Limit3	N
Base4	Limit4	V
Base5	Limit5	N
Base6	Limit6	N
Base7	Limit7	V

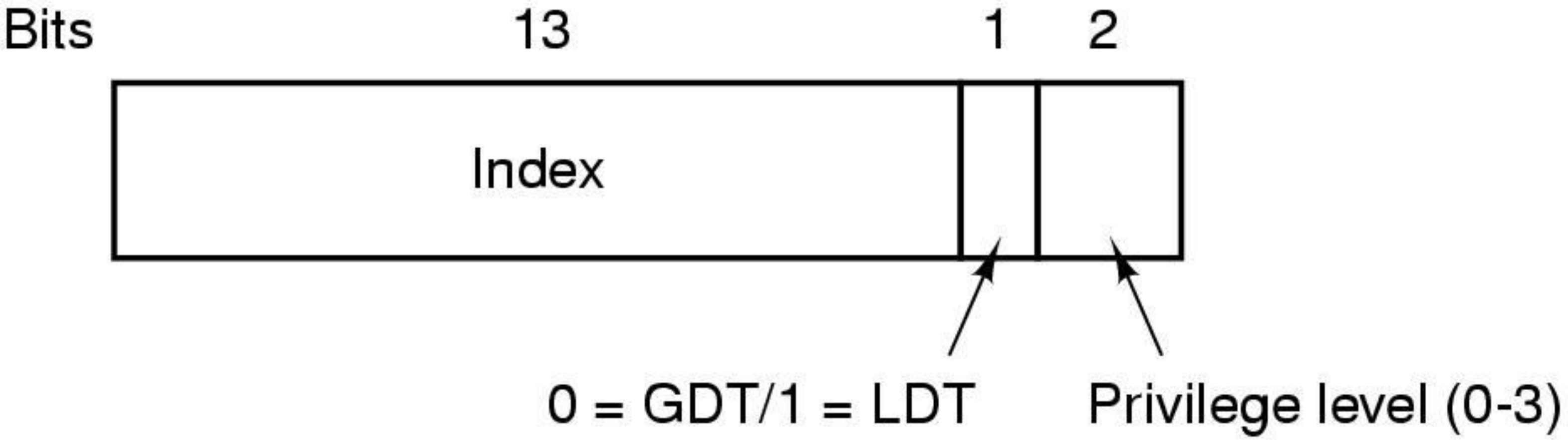
Process B:

Virtual Seg #	Virtual Page #	Offset

Segmentation with Paging: x86 32-bit



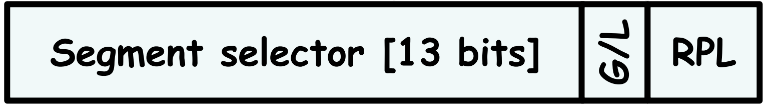
- Segment size: 4GB (32-bit)
- 16 K segments: 8K segments private (13-bit) and rest shared (13-bit)



A Pentium selector

X86 Segment Descriptors (32-bit Protected Mode)

- Segments are either implicit in the instruction (say for code segments) or actually part of the instruction
  - There are 6 registers: SS, CS, DS, ES, FS, GS
- What is in a segment register?
  - A *pointer* to the actual segment description:



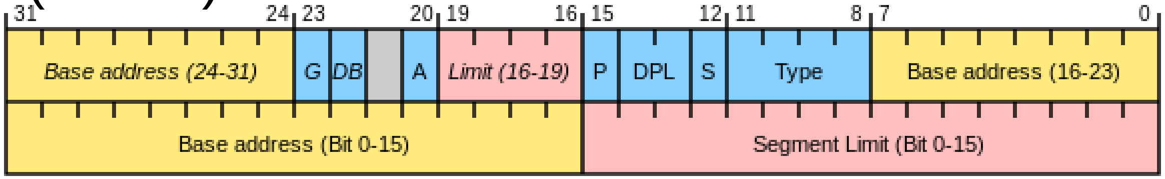
Segment selector [13 bits]

G/L

RPL

G/L selects between GDT and LDT tables (global vs local descriptor tables)

- Two registers: GDTR and LDTR hold pointers to the global and local descriptor tables in memory
  - Includes length of table (for < 2^{13}) entries
- Descriptor format (64 bits):



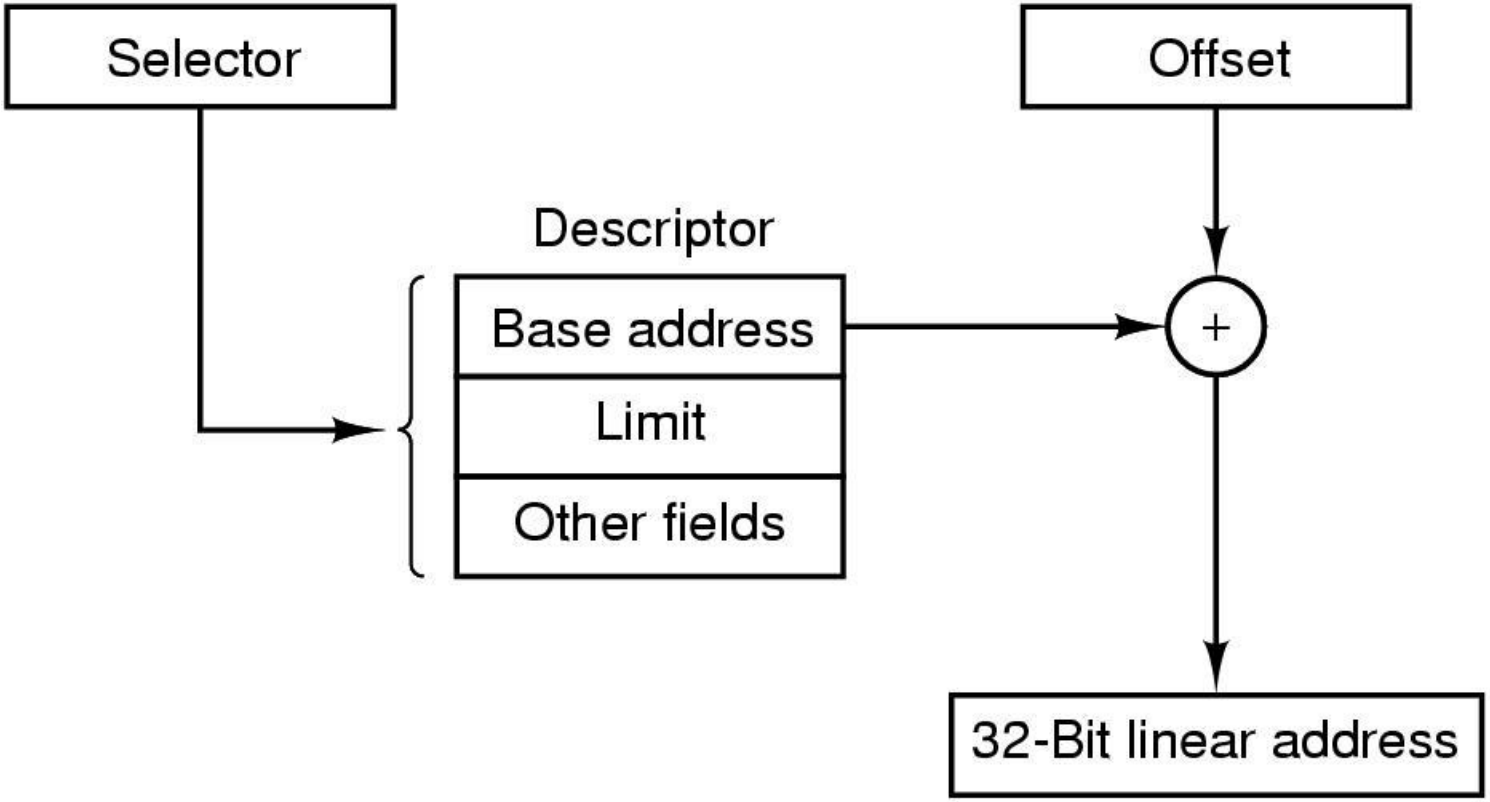
31 24 23 20 19 16 15 12 11 8 7 0

Base address (24-31) G DB A *Limit (16-19)* P DPL S Type Base address (16-23)

Base address (Bit 0-15) Segment Limit (Bit 0-15)

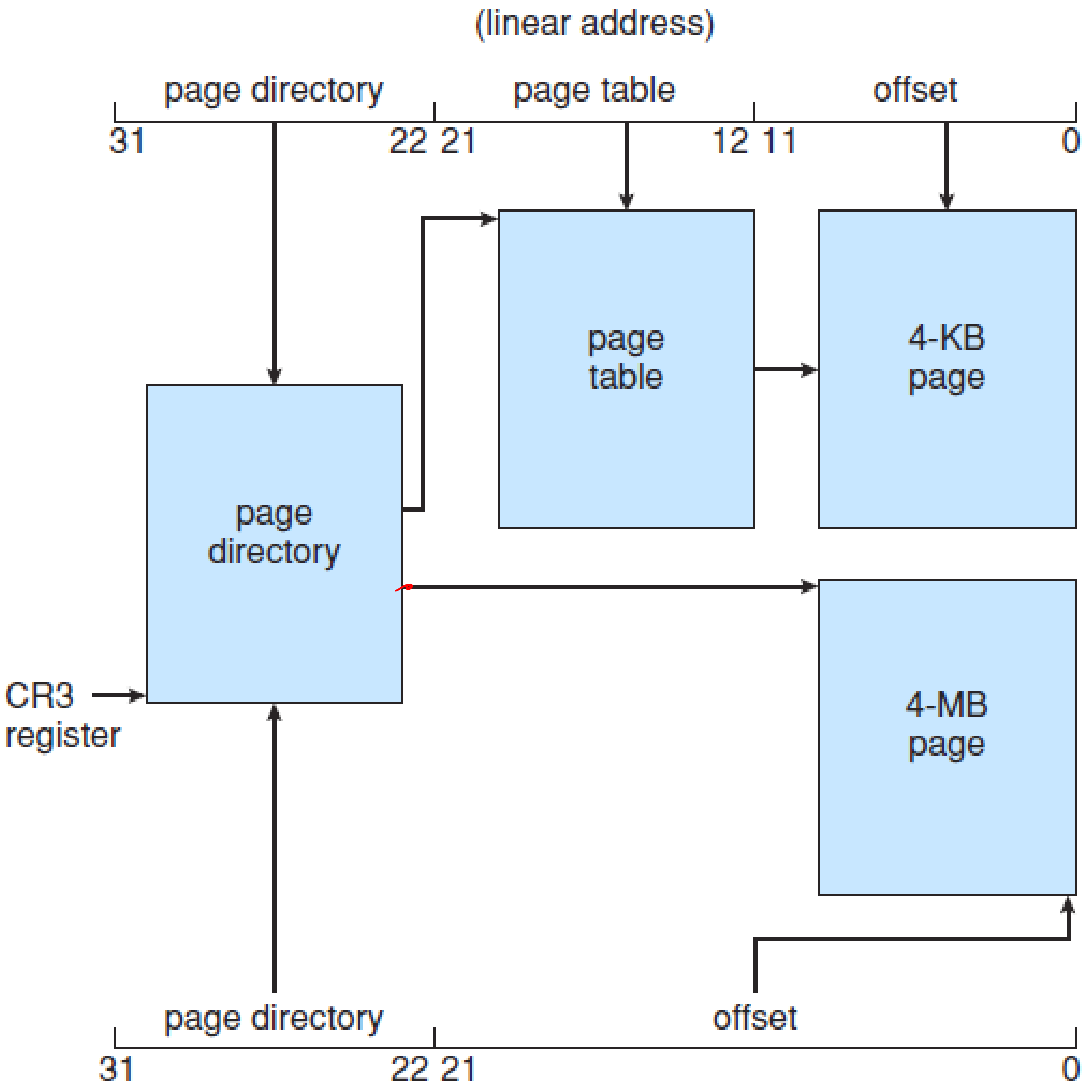
G: Granularity of segment [Limit Size] (0: 16bit, 1: 4KB unit)
DB: Default operand size (0: 16bit, 1: 32bit)
A: Freely available for use by software
P: Segment present
DPL: Descriptor Privilege Level
S: System Segment (0: System, 1: code or data)
Type: Code, Data, Segment

Segmentation with Paging: Pentium



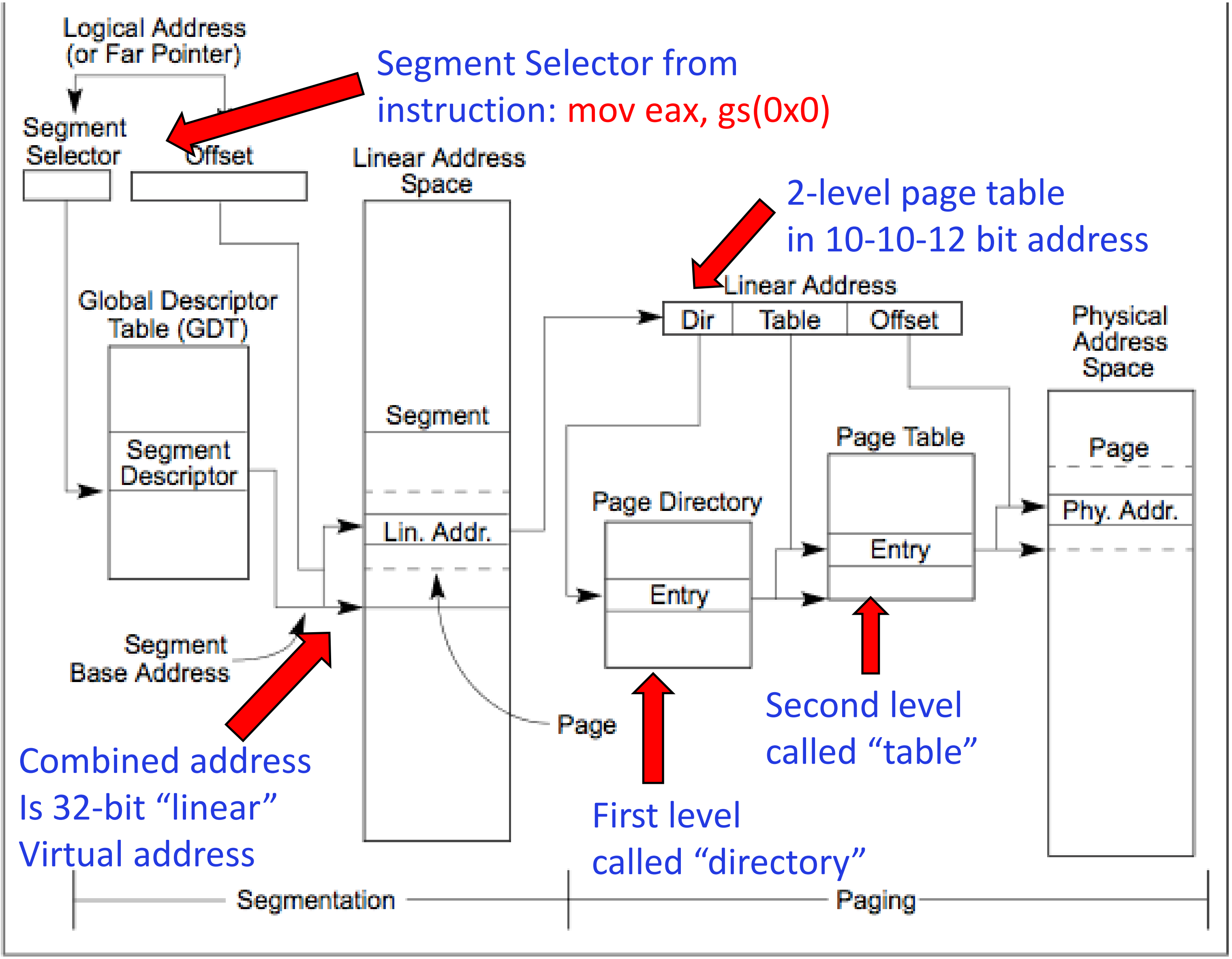
Conversion of a logical address=(selector, offset) pair to a linear address

Paging in x86 32-bit Arch



Superpage

Making it real:
X86 Memory model with segmentation (16/32-bit)



Segment Selector from instruction: mov eax, gs(0x0)

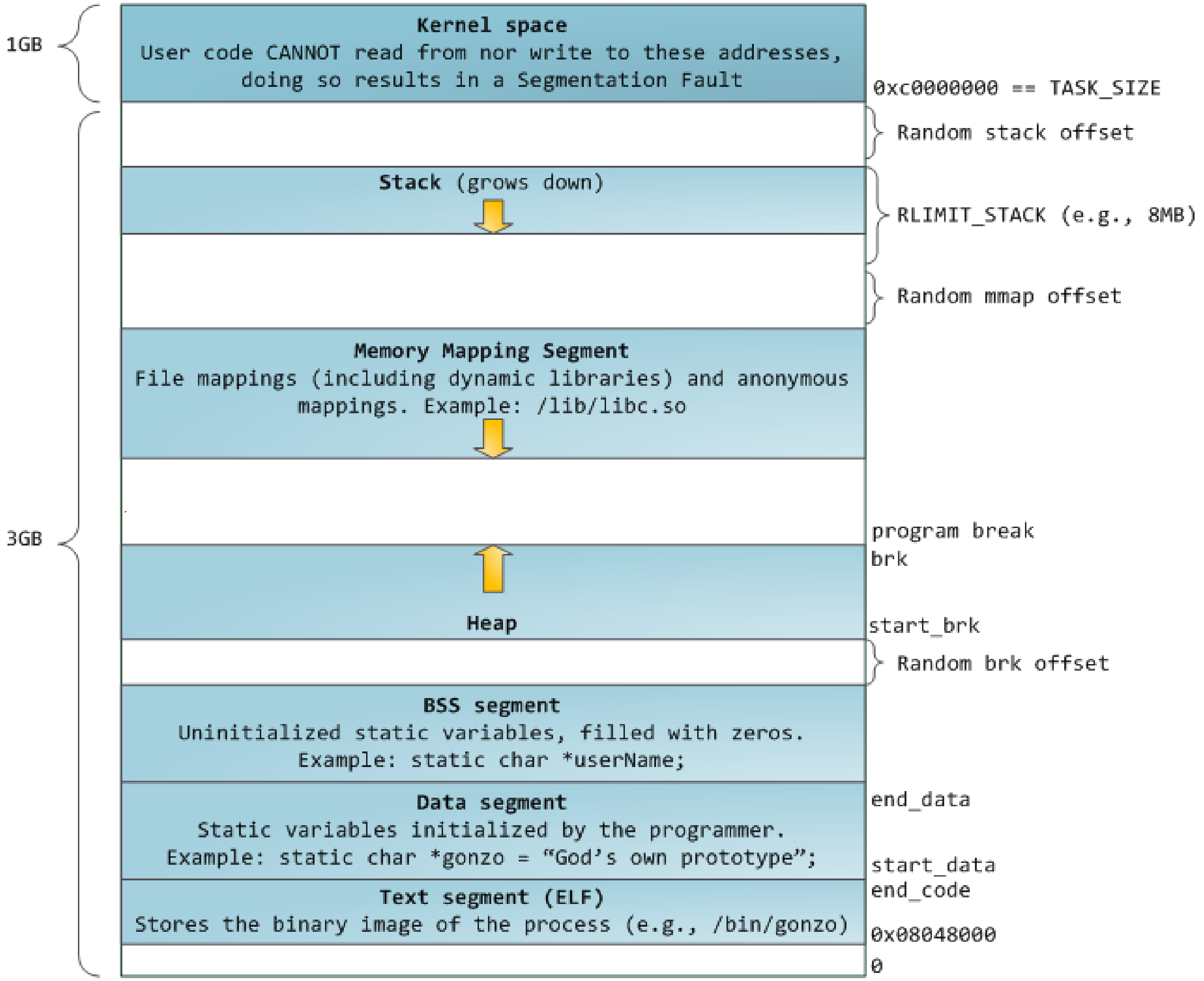
2-level page table in 10-10-12 bit address

Combined address Is 32-bit “linear” Virtual address

First level called “directory”

Second level called “table”

Example: Memory Layout for Linux (32-bit)



1GB

Kernel space
User code CANNOT read from nor write to these addresses, doing so results in a Segmentation Fault

0xc0000000 == TASK_SIZE

Random stack offset

Stack (grows down)

RLIMIT_STACK (e.g., 8MB)

Random mmap offset

Memory Mapping Segment
File mappings (including dynamic libraries) and anonymous mappings. Example: /lib/libc.so

3GB

program break
brk

Heap

start_brk

Random brk offset

BSS segment
Uninitialized static variables, filled with zeros.
Example: static char *userName;

end_data

Data segment
Static variables initialized by the programmer.
Example: static char *gonzo = "God's own prototype";

start_data
end_code

Text segment (ELF)
Stores the binary image of the process (e.g., /bin/gonzo)

0x08048000
0

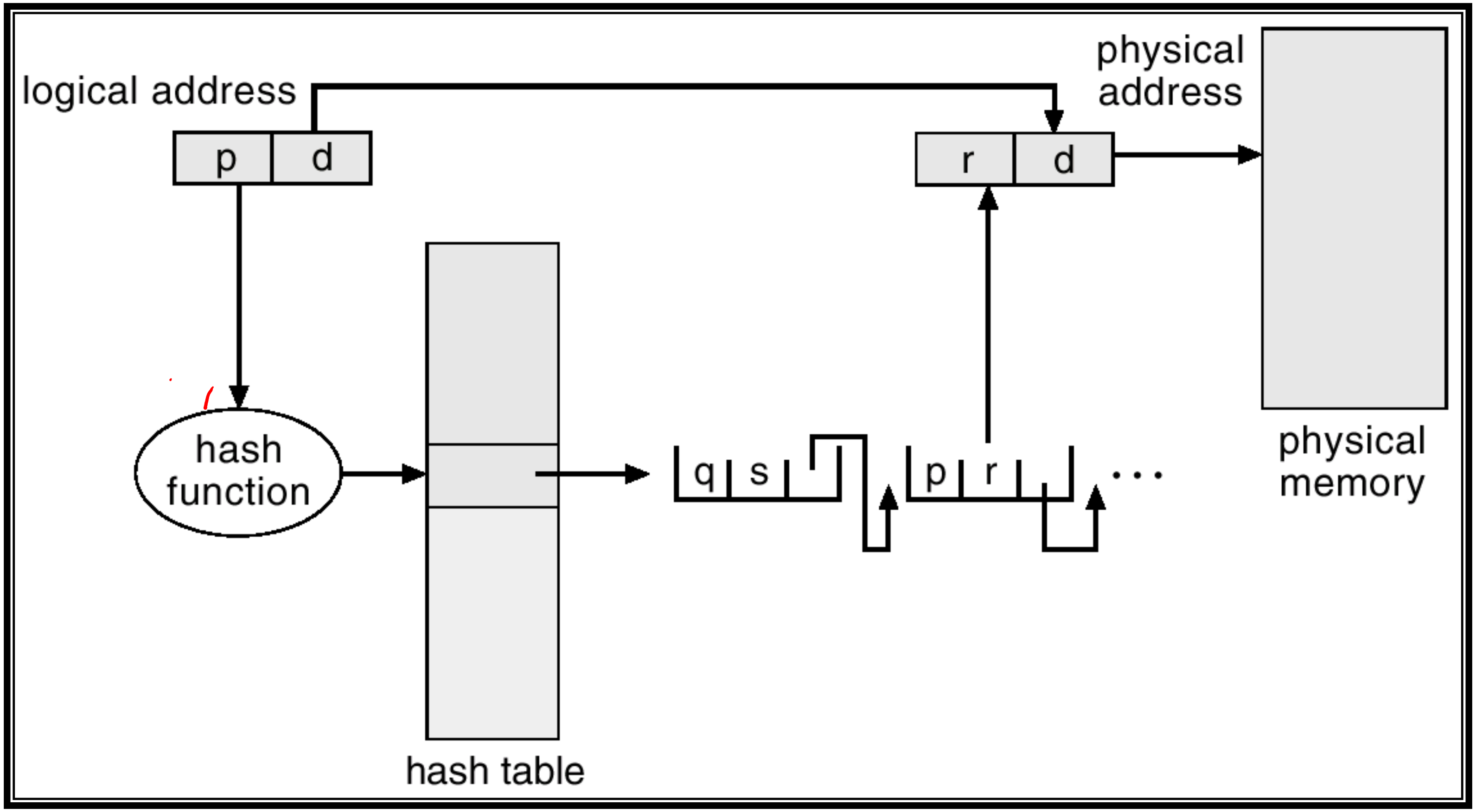
Address Space Layout Randomization (ASLR)

http://static.duartes.org/img/blogPosts/linuxFlexibleAddressSpaceLayout.png

Hashed Page Tables

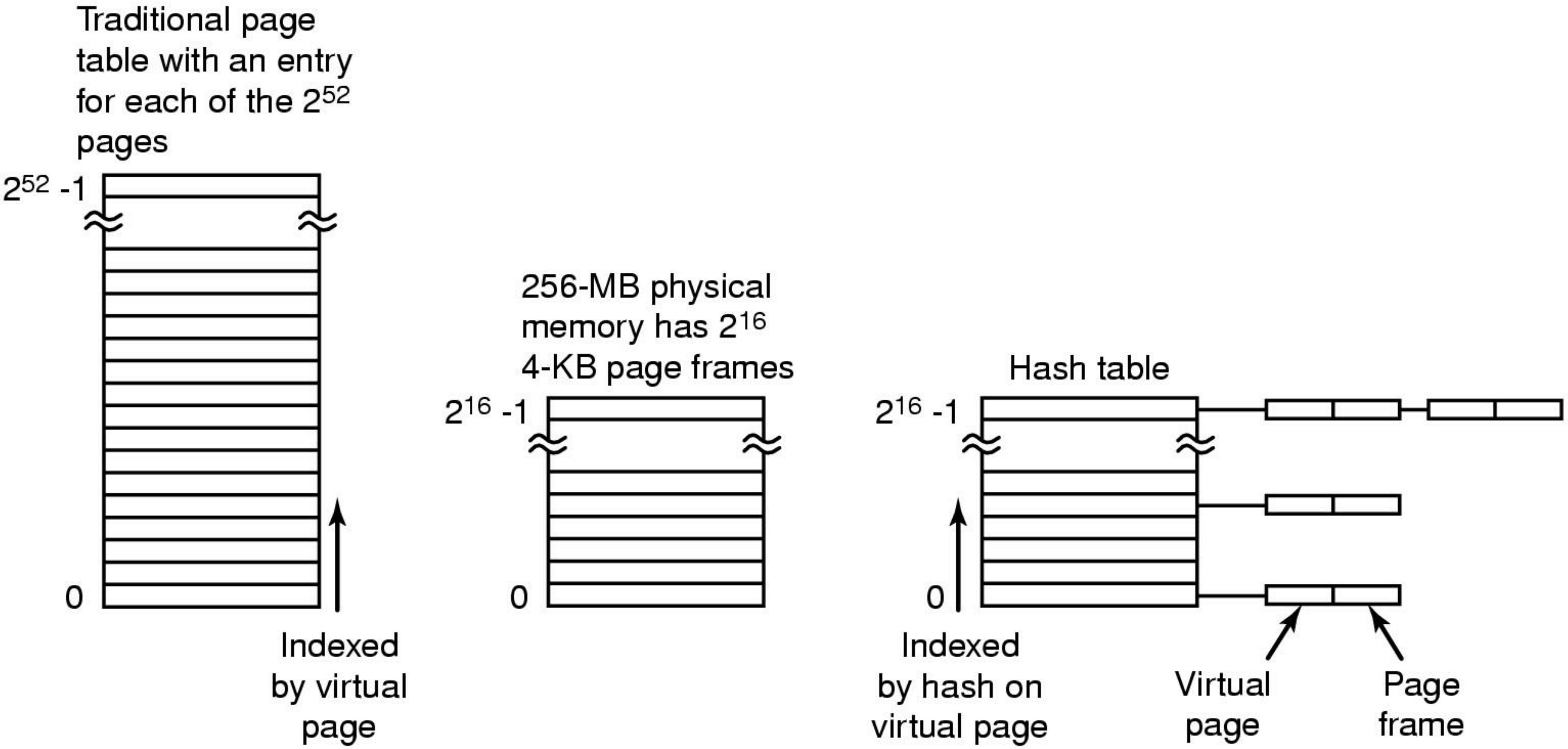
- Common in address spaces > 32 bits
- Efficient for sparse logical address spaces
- The virtual page number is hashed into a page table.
- Page table is a function of physical page frames and at maximum occupies one page frame in the main memory
- This page table contains a chain of elements hashing to the same location.
- Virtual page numbers are compared in this chain searching for a match.
  - If a match is found, the corresponding page frame is extracted.
- Reduces no. of memory accesses, but increases time needed to search inside the table
  - TLB and sparse address spaces help

Hashed Page Table



How many memory access?

Hash Tables

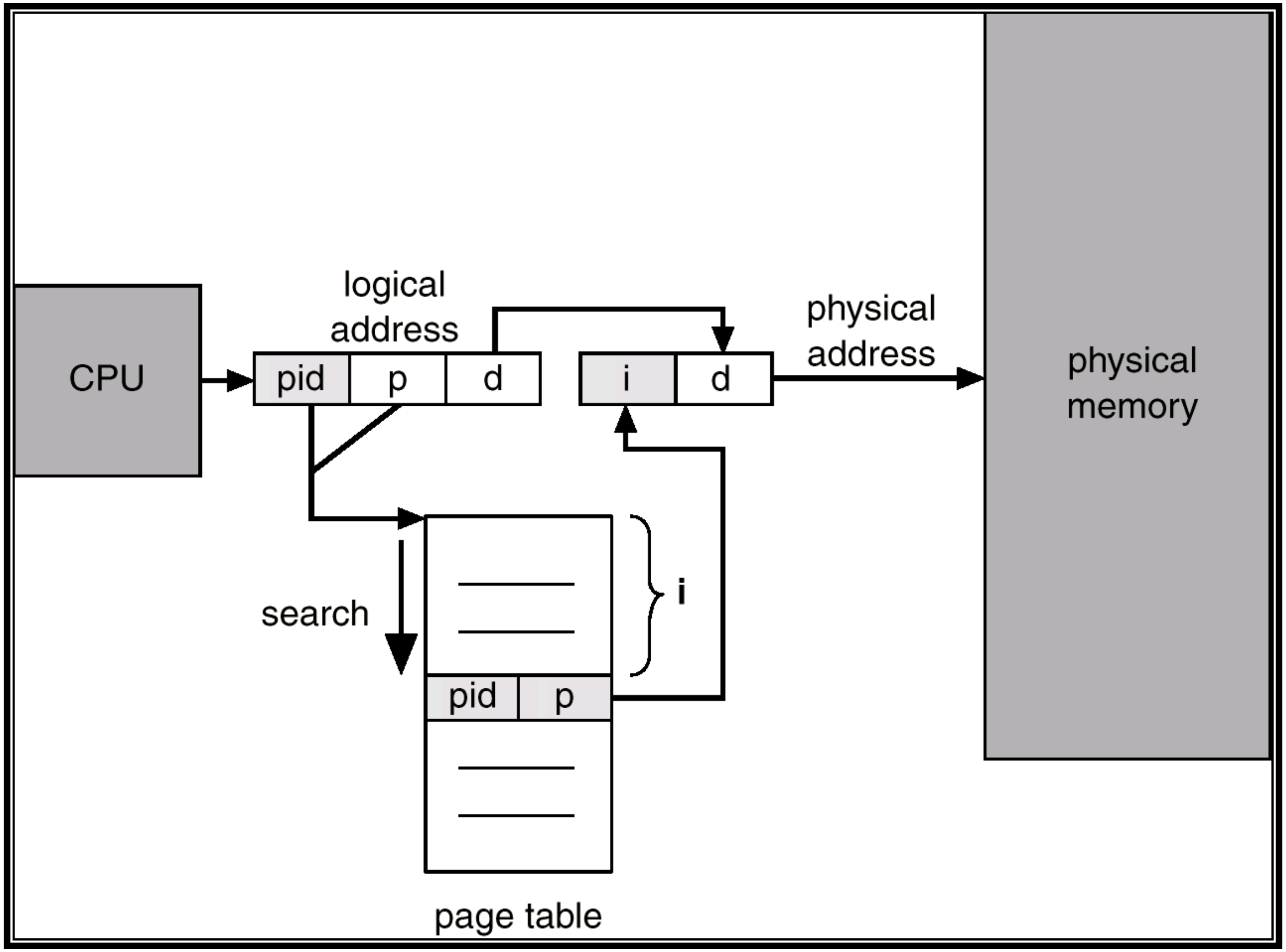

Traditional page table with an entry for each of the 2^{52} pages
2^{52} -1
0
Indexed by virtual page
256-MB physical memory has 2^{16} 4-KB page frames
2^{16} -1
0
Hash table
2^{16} -1
0
Indexed by hash on virtual page
Virtual page
Page frame

Comparison of a traditional page table with Hash table

Inverted Page Table

- One entry for each real page (frame) of main memory.
- Page table entry (PTE) consists of the virtual address of the page stored in that real memory location, with information about the process that owns that page (ASID like pid).
- Decreases memory needed to keep page table per process: now one inverted page table for whole system!
- But increases time needed to search the table when a page reference occurs → no more indexing possible like in Page tables, Segment tables
  - Use hash table to limit the search to one — or at most a few — page table entries.

Inverted Page Table Architecture



Multi-level Translation Analysis

- Pros:
  - Only need to allocate as many page table entries as we need for application
    - In other words, sparse address spaces are easy to manage
  - Easy memory allocation using bitmap
  - Easy sharing
    - Share at segment or page level (need additional reference counting)
- Cons:
  - One pointer per page (typically 4K – 16K pages today)
  - Page tables (or segment tables) need to be contiguous
    - However, previous example keeps tables to exactly one page in size
  - Two (or more, if >2 levels) lookups per reference
    - Seems very expensive!

Comparison of Memory Mgmt Schemes

Diff schemes: contiguous allocation, paging, segmentation and multi-level translation

- Hardware support: Registers, MMU, tables
- Performance: Mapping delay can be reduced by TLB
- Fragmentation: Avoid external fragmentation to improve degree of multiprogramming
- Relocation: To avoid external fragmentation, use relocatable (virtual) addresses
- Swapping: To let more programs to run on limited RAM, but not on Flash-based systems like tablets/Smart phones
- Sharing: To reduce memory footprint of processes
- Protection: Helps sharing and avoids accidental errors by keeping protection bits in Page/Segment tables

Address Translation Comparison

	Advantages	**Disadvantages**
Pure Segmentation	Fast context switching: Segment mapping maintained by CPU	External fragmentation
Pure Paging (single-level page)	No external fragmentation, fast easy allocation	Large table size ~ virtual memory Internal fragmentation
Paged segmentation	Table size ~ # of pages in virtual memory, fast easy allocation, inner page tables can be kept in disk	Multiple memory references per page access
Two-level Paging		
Inverted Table	Table size ~ # of pages in physical memory	Hash function more complex No cache locality of page table

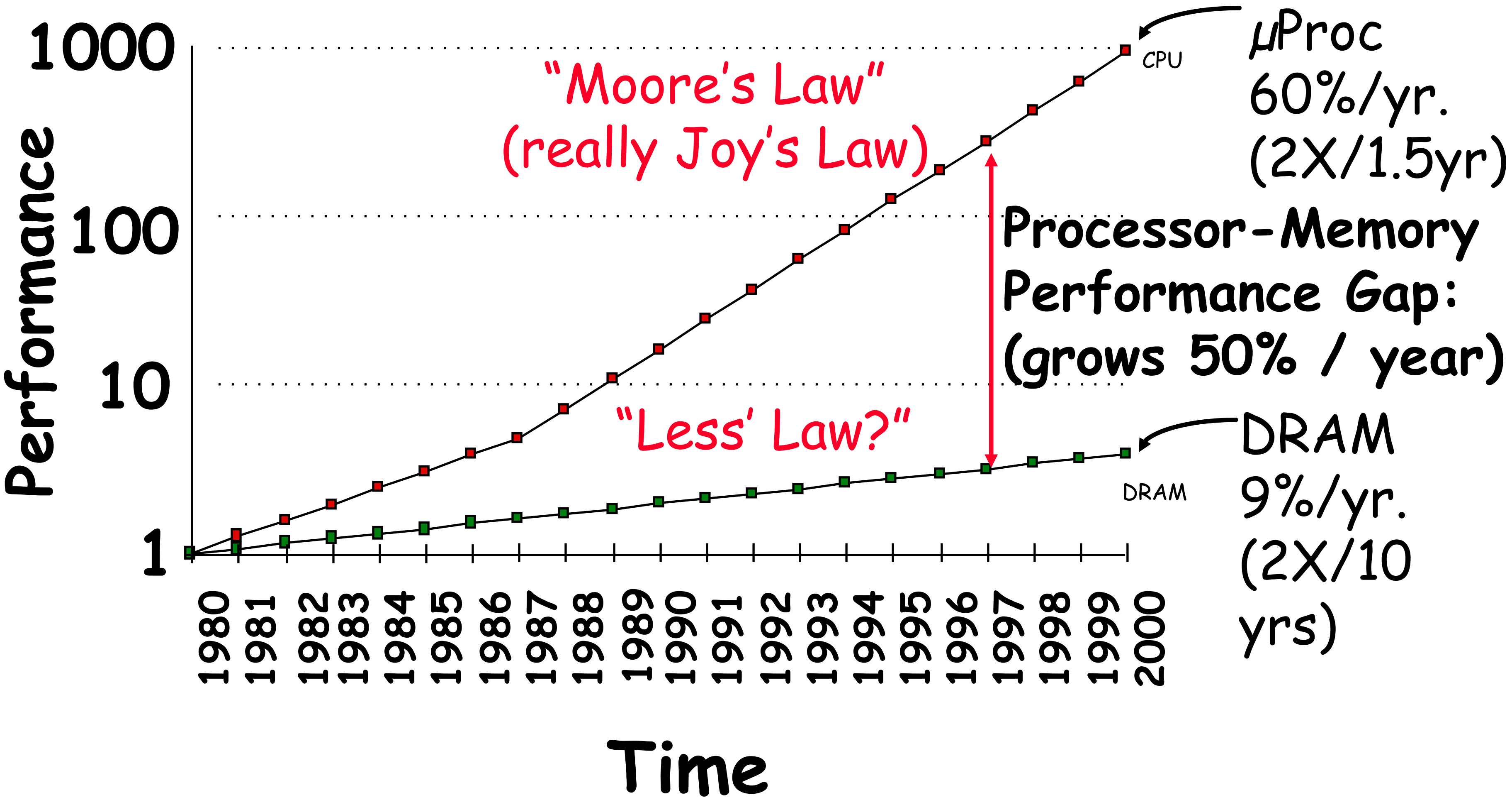
Caching Concept

- Cache: a repository for copies that can be accessed more quickly than the original
  - Make frequent case fast and infrequent case less dominant
- Caching underlies many of the techniques that are used today to make computers fast
  - Can cache: memory locations, address translations, pages, file blocks, file names, network routes, etc…
- Only good if:
  - Frequent case fast enough and
  - Infrequent case not too expensive
- Important measure: Average Access time =

  (Hit Rate x Hit Time) + (Miss Rate x Miss Time)

Why Bother with Caching?

Processor-DRAM Memory Gap (latency)

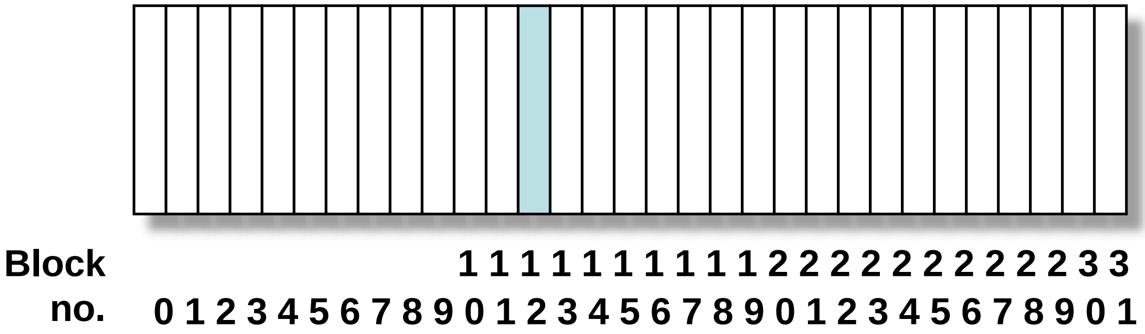


Processor-Memory Performance Gap: (grows 50% / year)

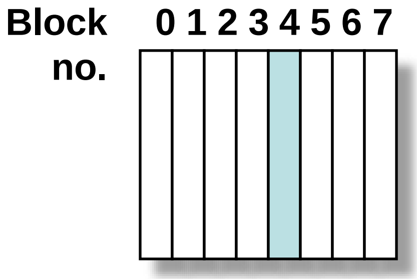
Review: Where does a Block Get Placed in a Cache?

- Example: Block 12 placed in Cache of size 8 blocks

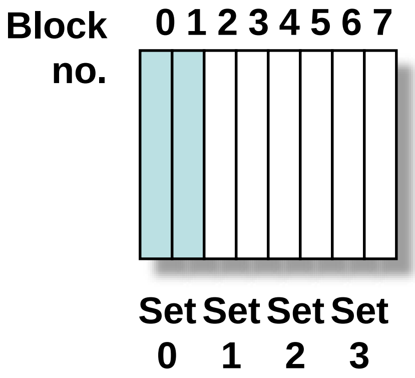
32-Block Physical Address Space:



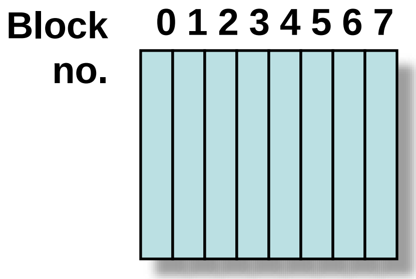
Direct mapped:
block 12 can go only into block 4 (12 mod 8)



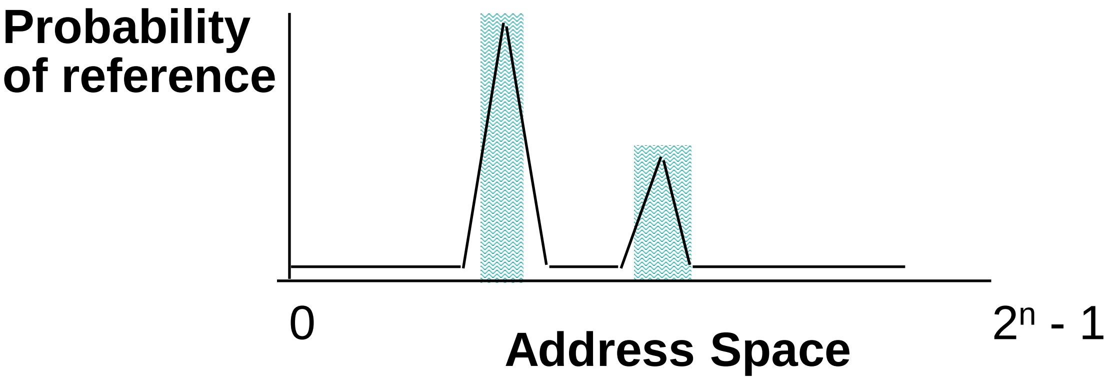
Set associative:
block 12 can go anywhere in set 0 (12 mod 4)



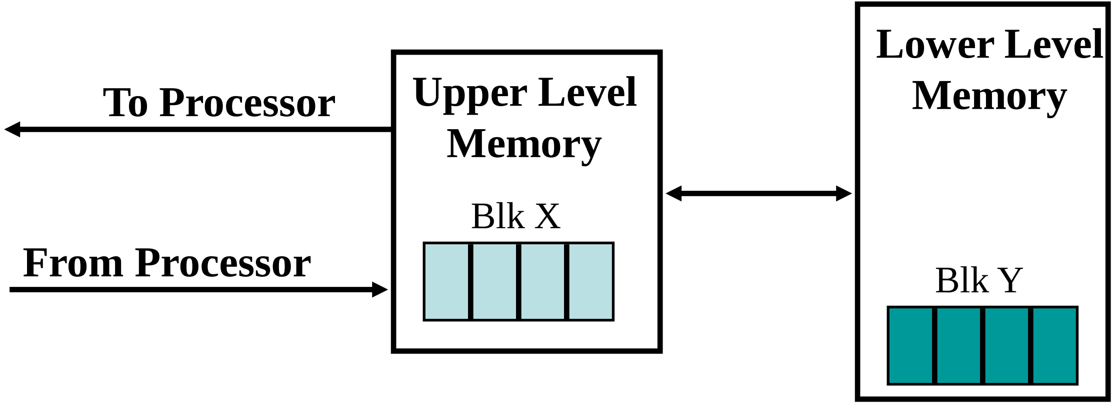
Fully associative:
block 12 can go anywhere



Why Does Caching Help? Locality!

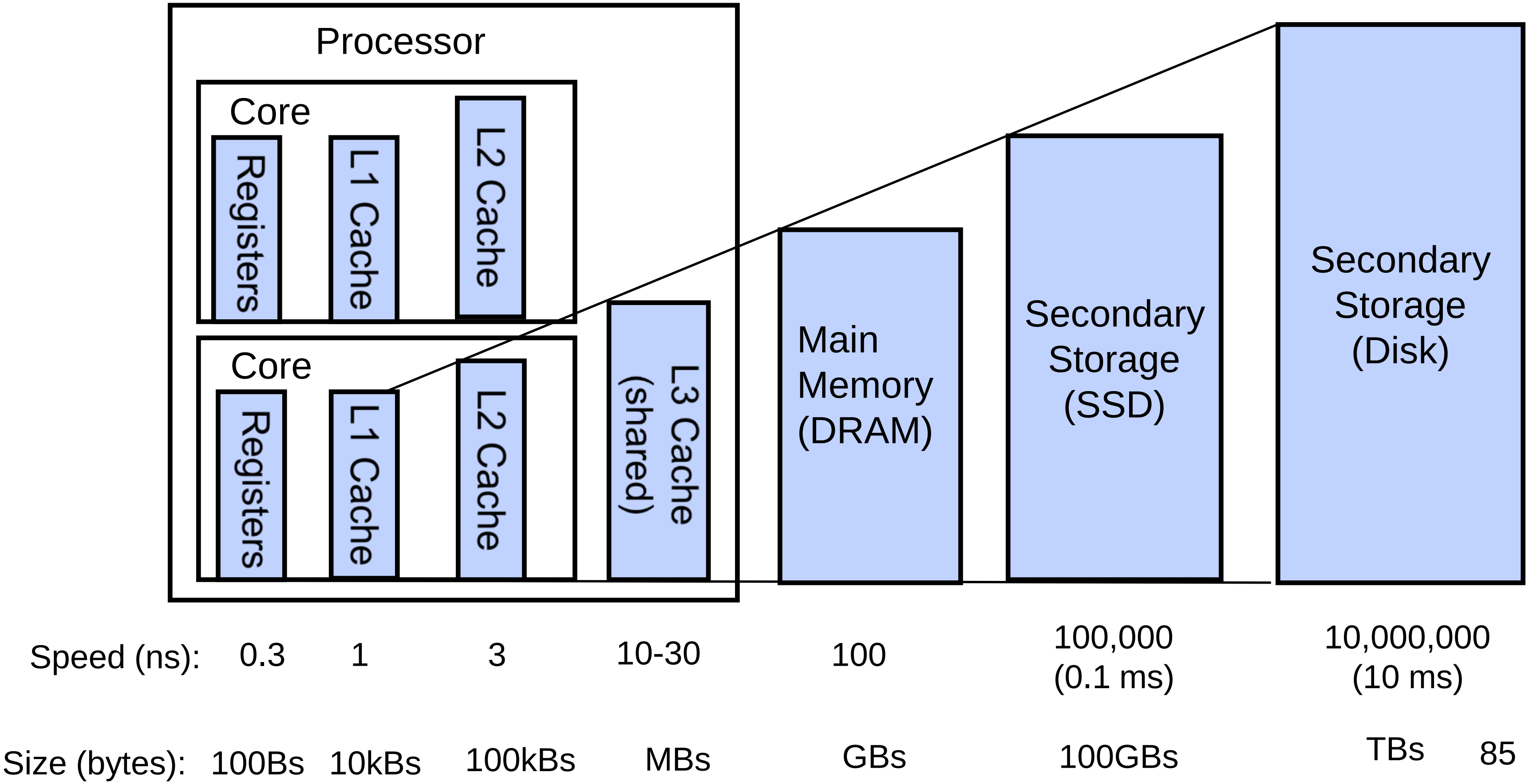


- Temporal Locality (Locality in Time):
  - – Keep recently accessed data items closer to processor
- Spatial Locality (Locality in Space):
  - – Move contiguous blocks to the upper levels

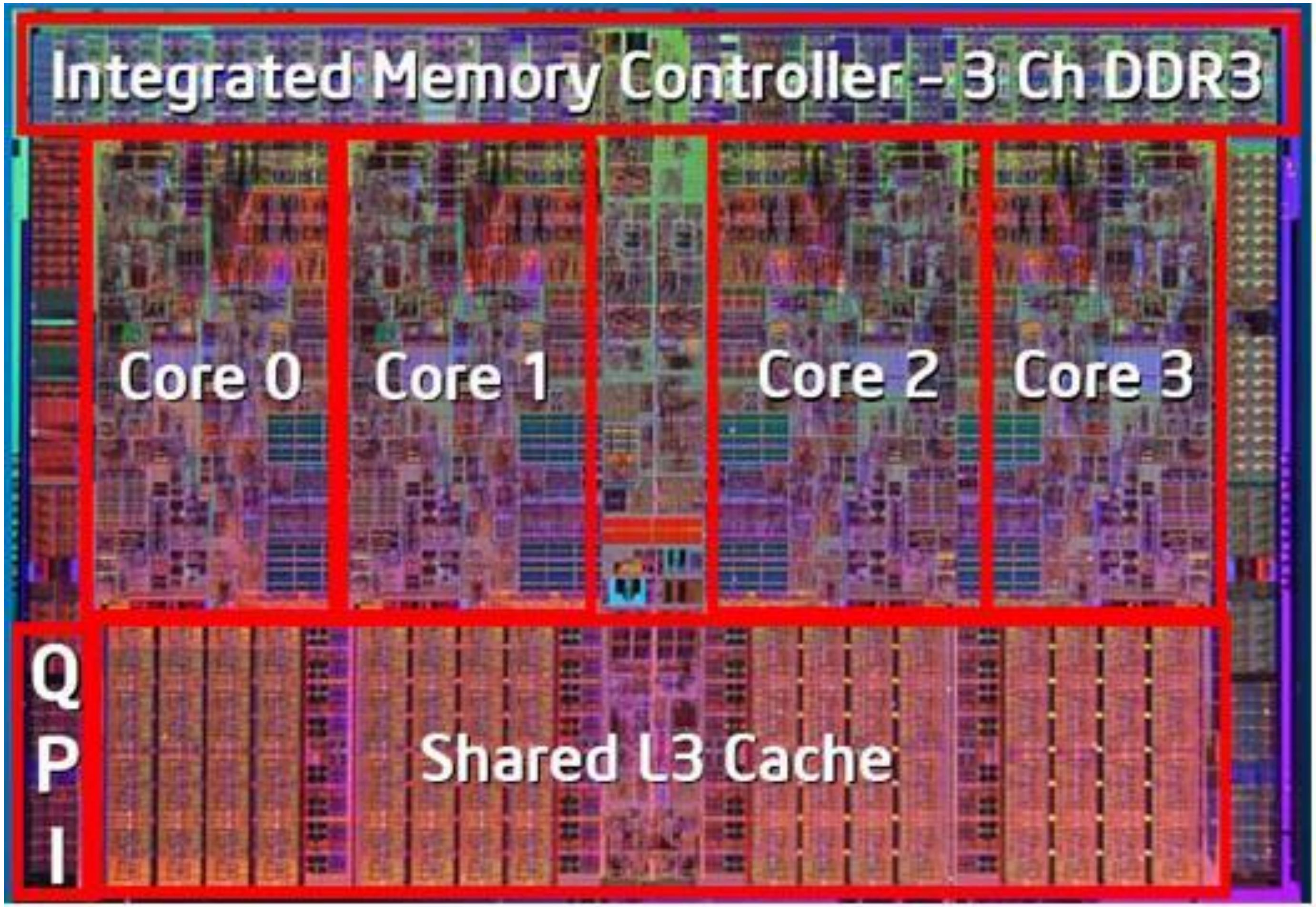


Review: Memory Hierarchy of a Modern Computer System

- Take advantage of the principle of locality to:
  - Present as much memory as in the cheapest technology
  - Provide access at speed offered by the fastest technology



Intel i7


Integrated Memory Controller – 3 Ch DDR3
Core 0
Core 1
Core 2
Core 3
QPI
Shared L3 Cache

A Summary on Sources of Cache Misses

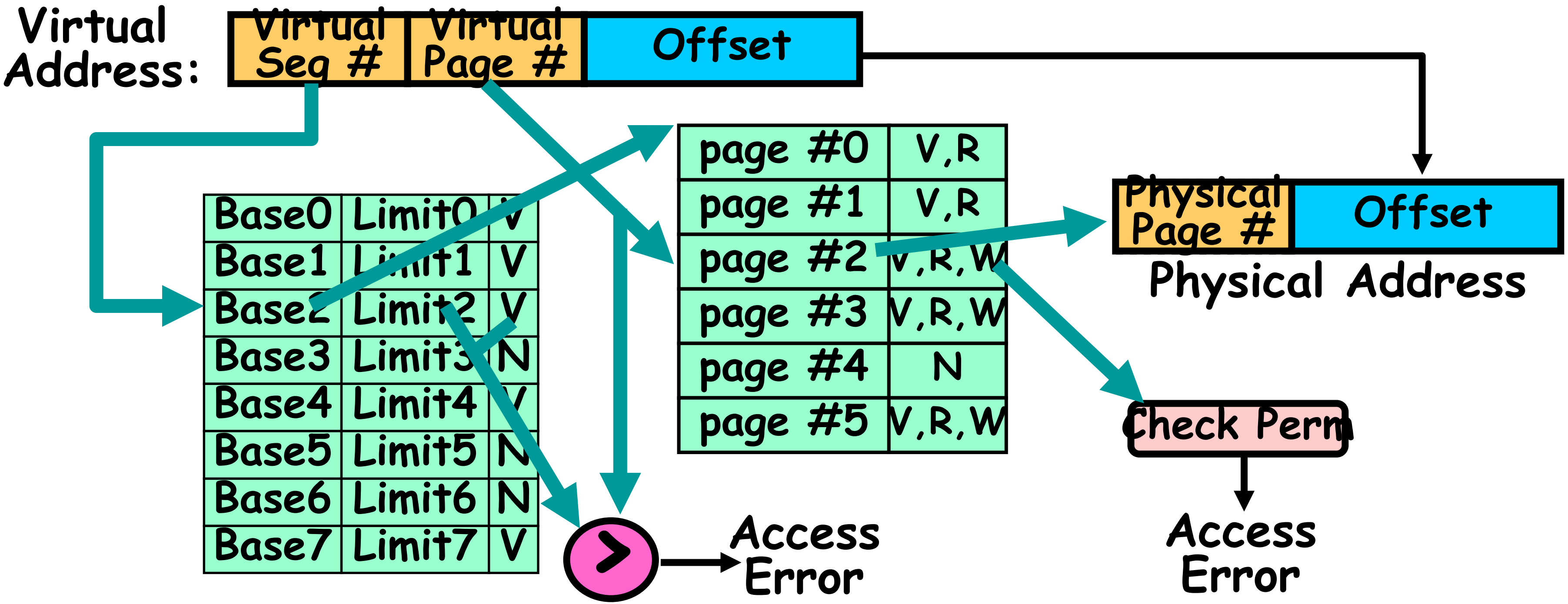
- Compulsory (cold start): first reference to a block
  - "Cold" fact of life: not a whole lot you can do about it
  - Note: When running "billions" of instruction, Compulsory Misses are insignificant
- Capacity:
  - Cache cannot contain all blocks access by the program
  - Solution: increase cache size
- Conflict (collision):
  - Multiple memory locations mapped to same cache location
  - Solutions: increase cache size, or increase associativity
- Two others:
  - Coherence (Invalidation): other process (e.g., I/O) updates memory
  - Policy: Due to non-optimal replacement policy

Other Caching Questions

- What entry gets replaced on cache miss?
  - Easy for Direct Mapped: Only one possibility
  - Set Associative or Fully Associative:
    - Random
    - LRU (Least Recently Used)
- What happens on a write?
  - Write through: The information is written to both the cache and to the block in the lower-level memory
  - Write back: The information is written only to the block in the cache
    - Modified cache block is written to main memory only when it is replaced
    - Question is block clean or dirty?

Major Reason to Deal with Caching

- Page table is kept in main memory. *Page-table base register* (PTBR) points to the page table.
- *Page-table length register* (PRLR) indicates size of the page table.



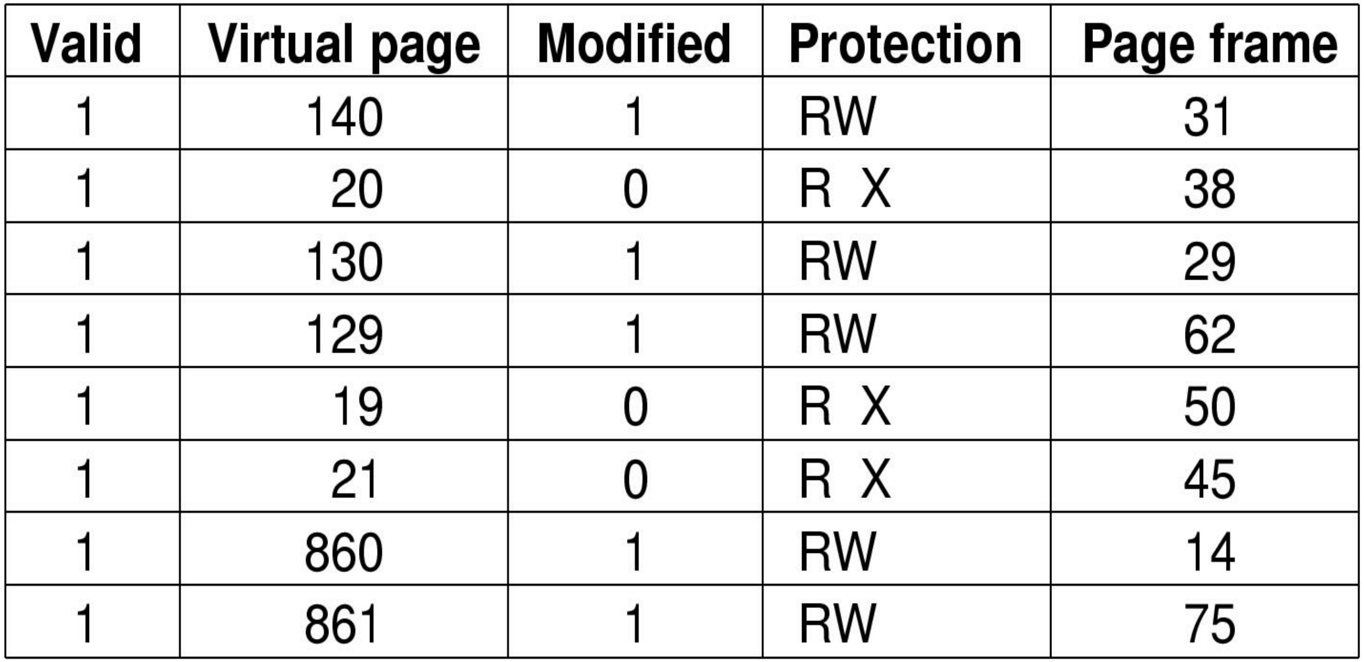
Base0	Limit0	V
Base1	Limit1	V
Base2	Limit2	V
Base3	Limit3	N
Base4	Limit4	V
Base5	Limit5	N
Base6	Limit6	N
Base7	Limit7	V

page #0	V,R
page #1	V,R
page #2	V,R,W
page #3	V,R,W
page #4	N
page #5	V,R,W

- Too expensive to translate on every access
  - At least two extra DRAM accesses per actual DRAM access
  - Or: perhaps I/O if page table partially on disk!
- Solution? Cache translations!
  - Translation Cache: TLB (“Translation Lookaside Buffer”)

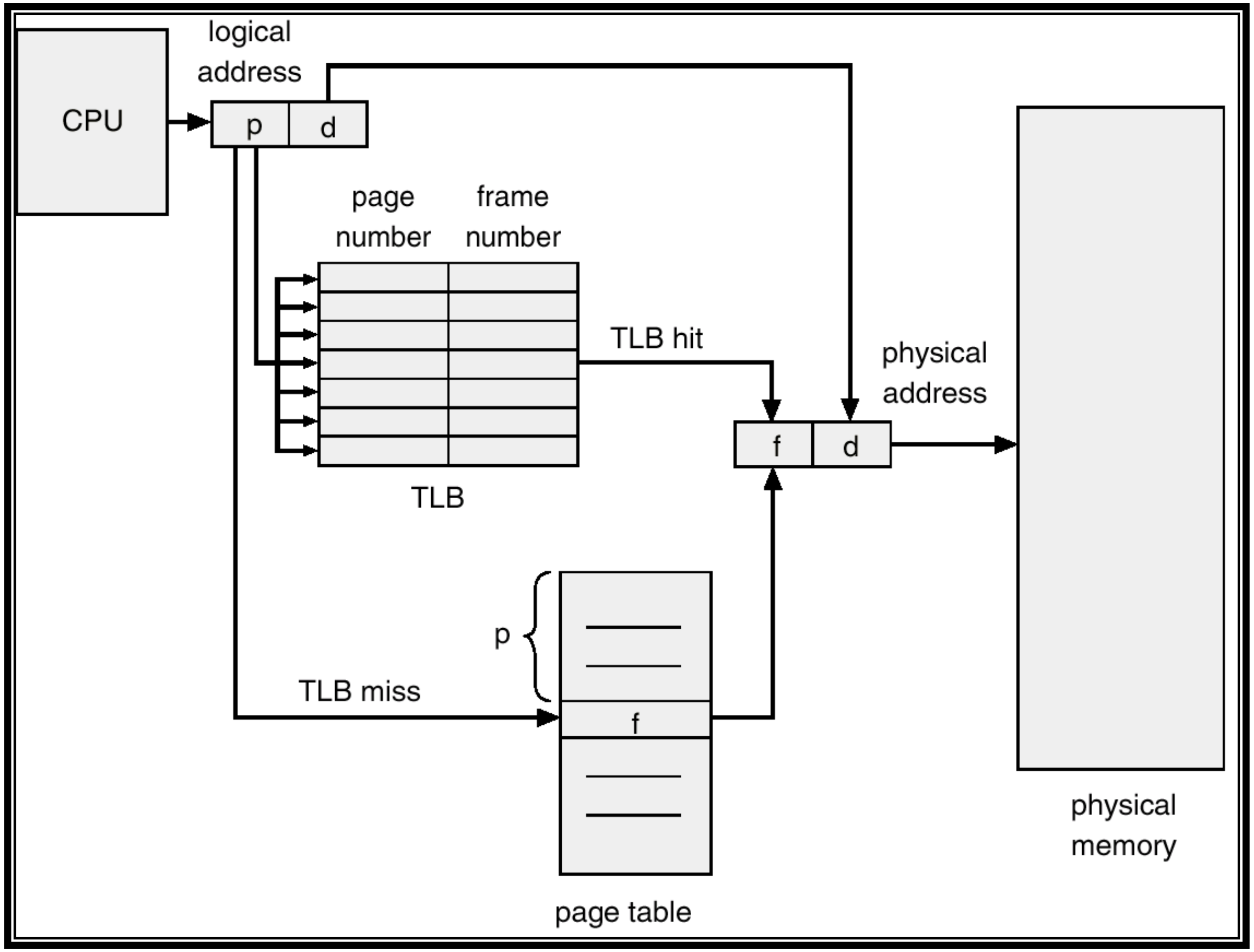
TLB: Cache for Address Translation

- Relatively small number of entries (< 512)
- Fully associative memory – parallel search (since misses are expensive)
- TLB entries contain virtual page ID, PTE and optional pid
- TLB is logically in front of cache
  - needs to be overlapped with cache access to be really fast



Valid	Virtual page	Modified	Protection	Page frame
1	140	1	RW	31
1	20	0	R X	38
1	130	1	RW	29
1	129	1	RW	62
1	19	0	R X	50
1	21	0	R X	45
1	860	1	RW	14
1	861	1	RW	75

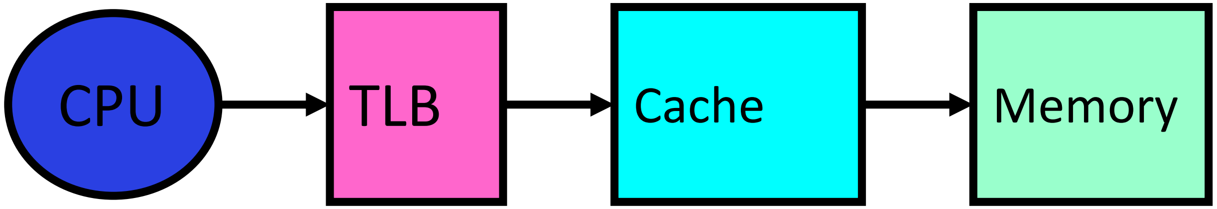
Paging Hardware With TLB



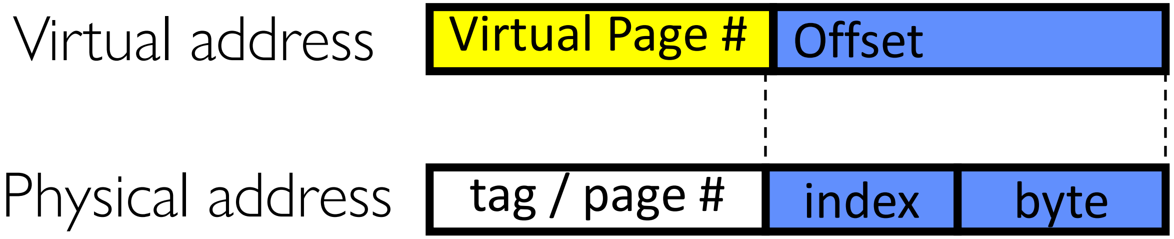
Cache, to be precise, multiple levels of caches are not shown in above diagram!

Intel i7 has L1 Data TLB, L1 Text (Code) TLB and L2 TLB!

What TLB Organization Makes Sense?

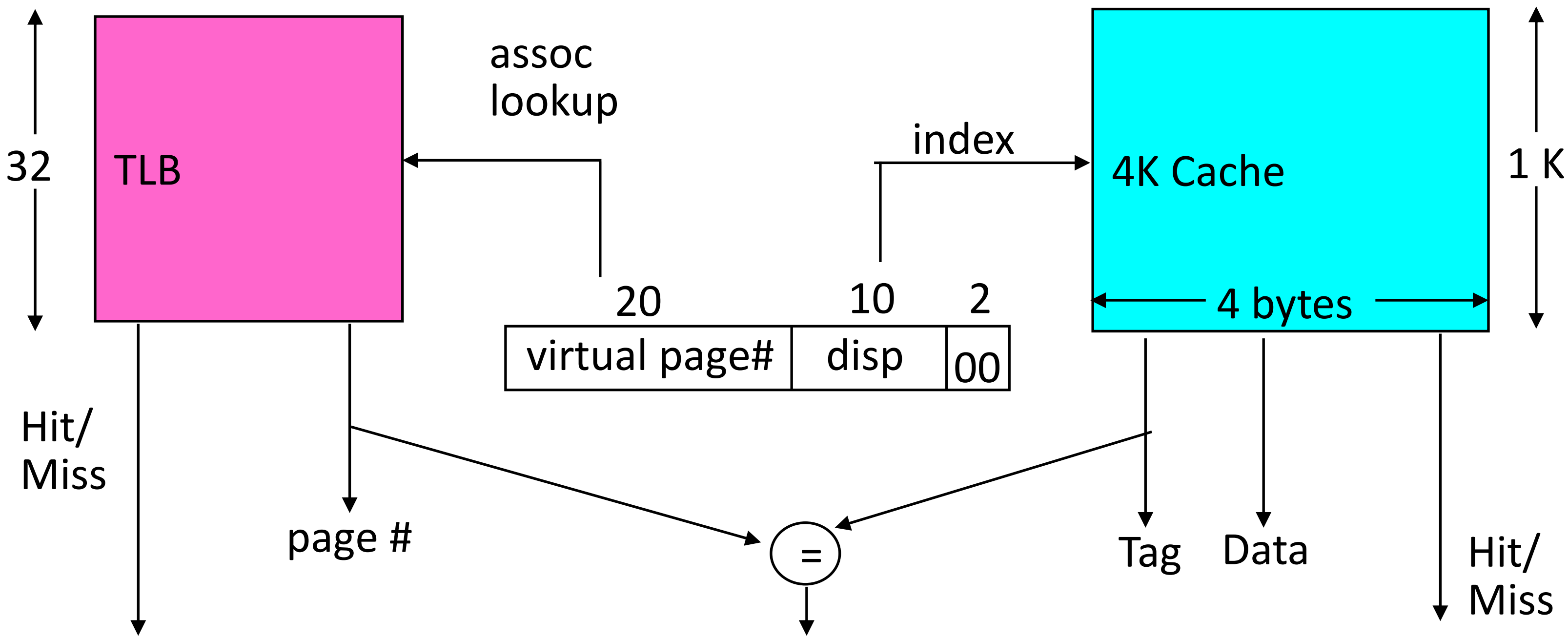


- Needs to be really fast
  - Critical path of memory access
    - » In simplest view: before the cache
    - » Thus, this adds to access time (reducing cache speed)
- Reducing translation time further…
  - Overlap TLB lookup with cache access!
    - Works because offset available early from virtual address itself
    - Offset in virtual address exactly covers "cache index" and "byte select"
    - Thus can select the cached byte(s) in parallel to perform address translation



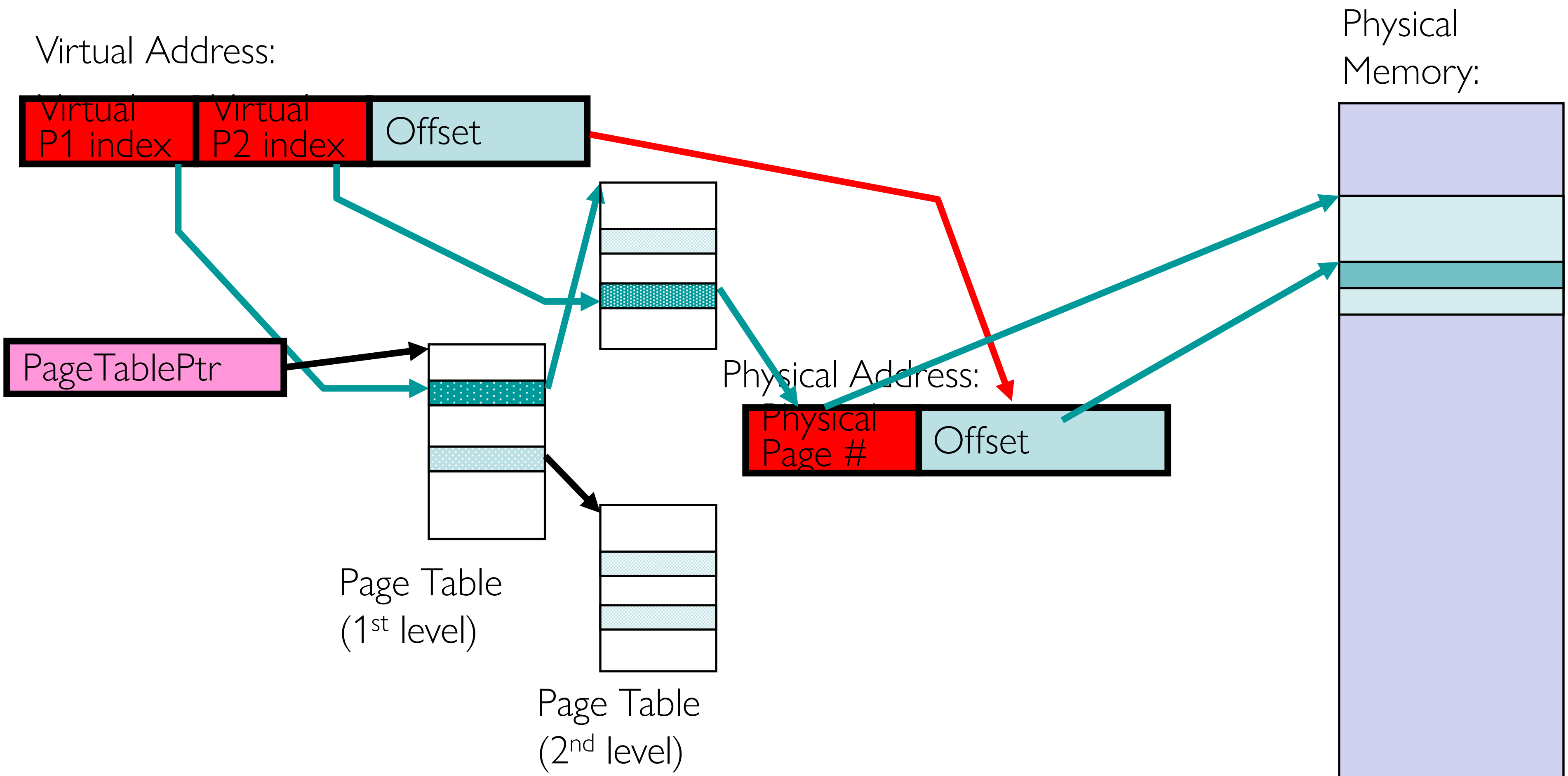
Overlapping TLB & Cache Access

- Here is how this might work with a 4K cache:

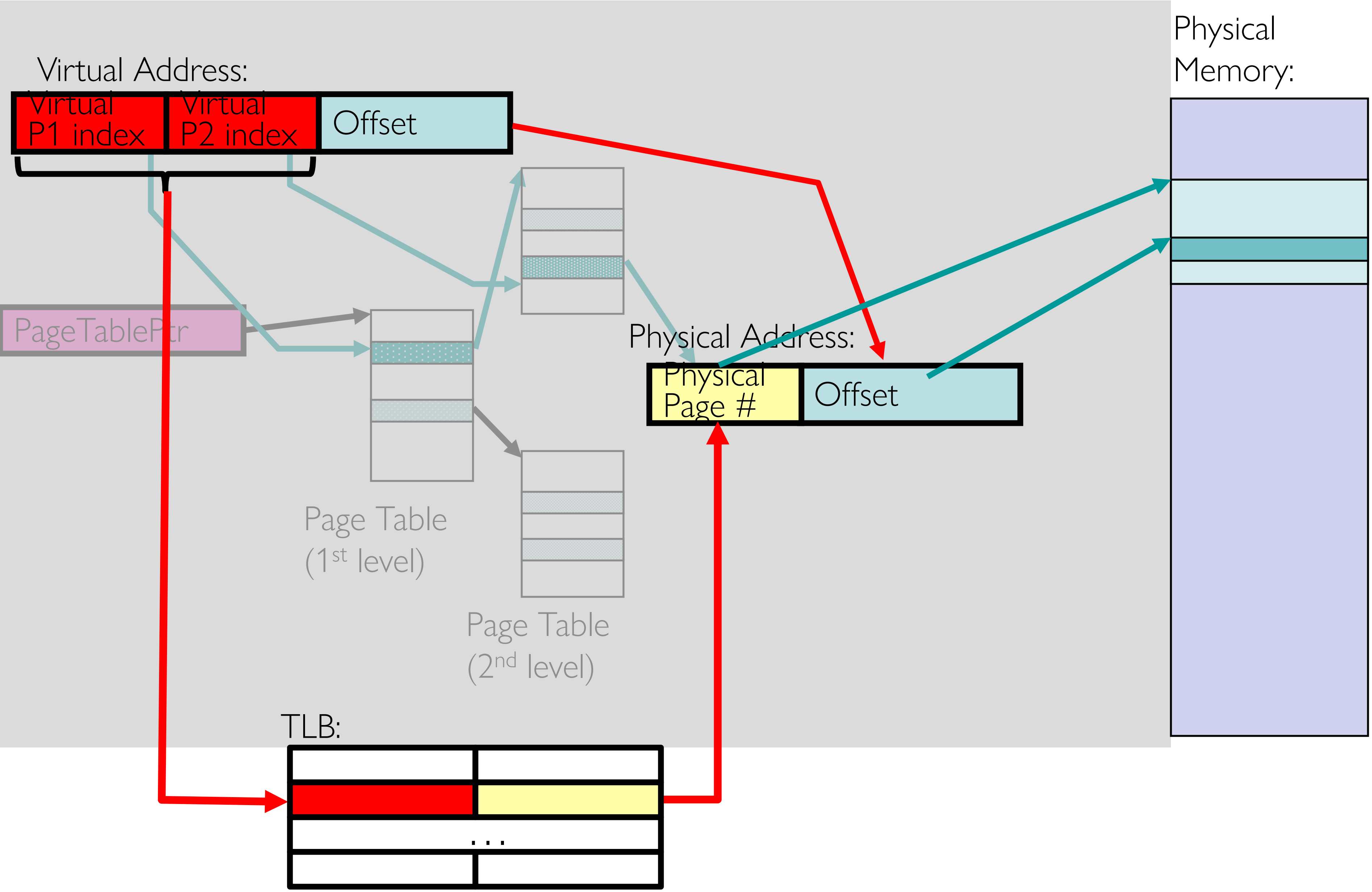


- Another option: Virtual Caches
  - Tags in cache are virtual addresses
  - Translation only happens on cache misses

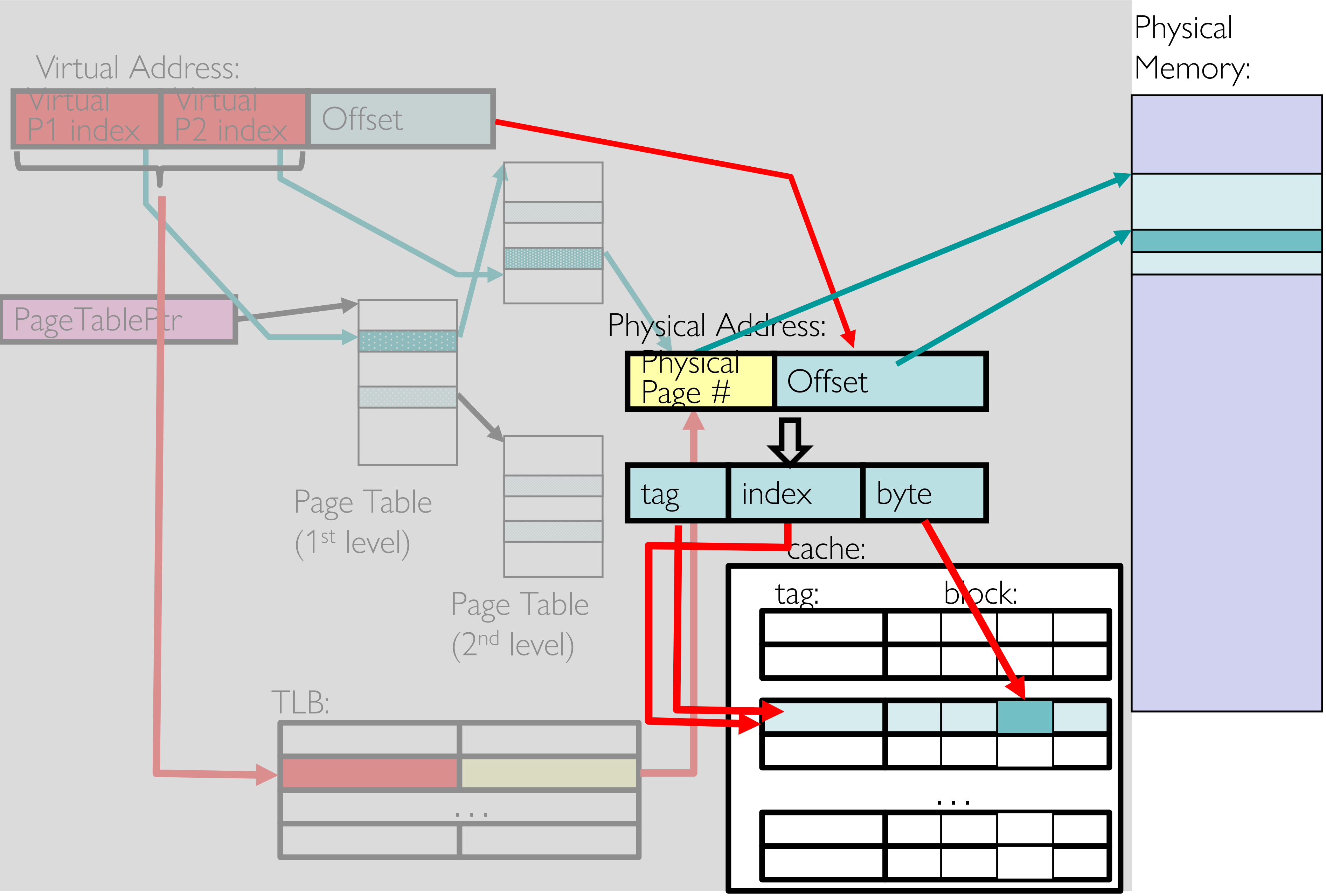
Putting Everything Together: Address Translation



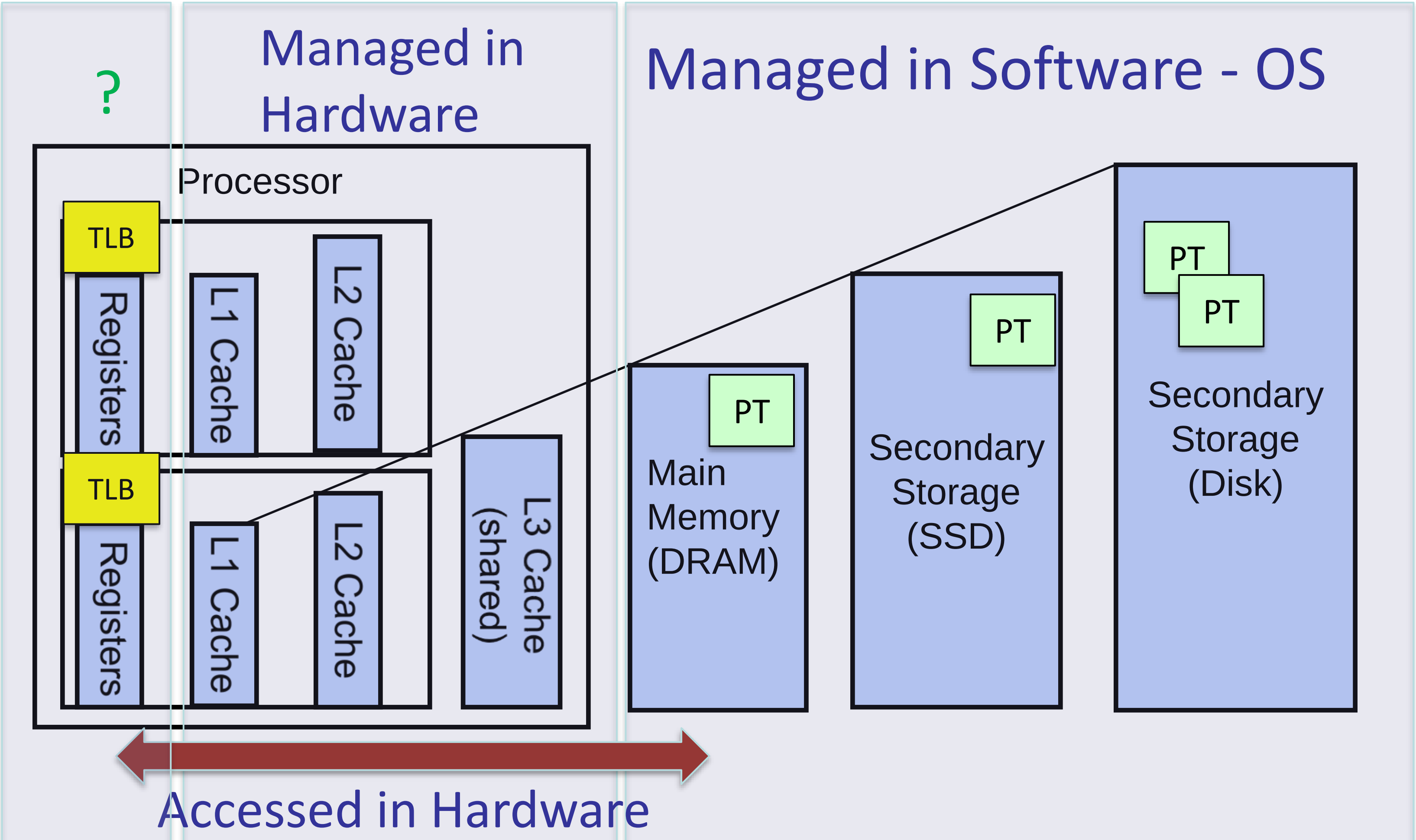
Putting Everything Together: TLB



Putting Everything Together: Cache



Management & Access to the Memory Hierarchy



?

Managed in Hardware

Managed in Software - OS

Processor

TLB

Registers

L1 Cache

L2 Cache

TLB

Registers

L1 Cache

L2 Cache

L3 Cache (shared)

PT

Main Memory (DRAM)

PT

Secondary Storage (SSD)

PT

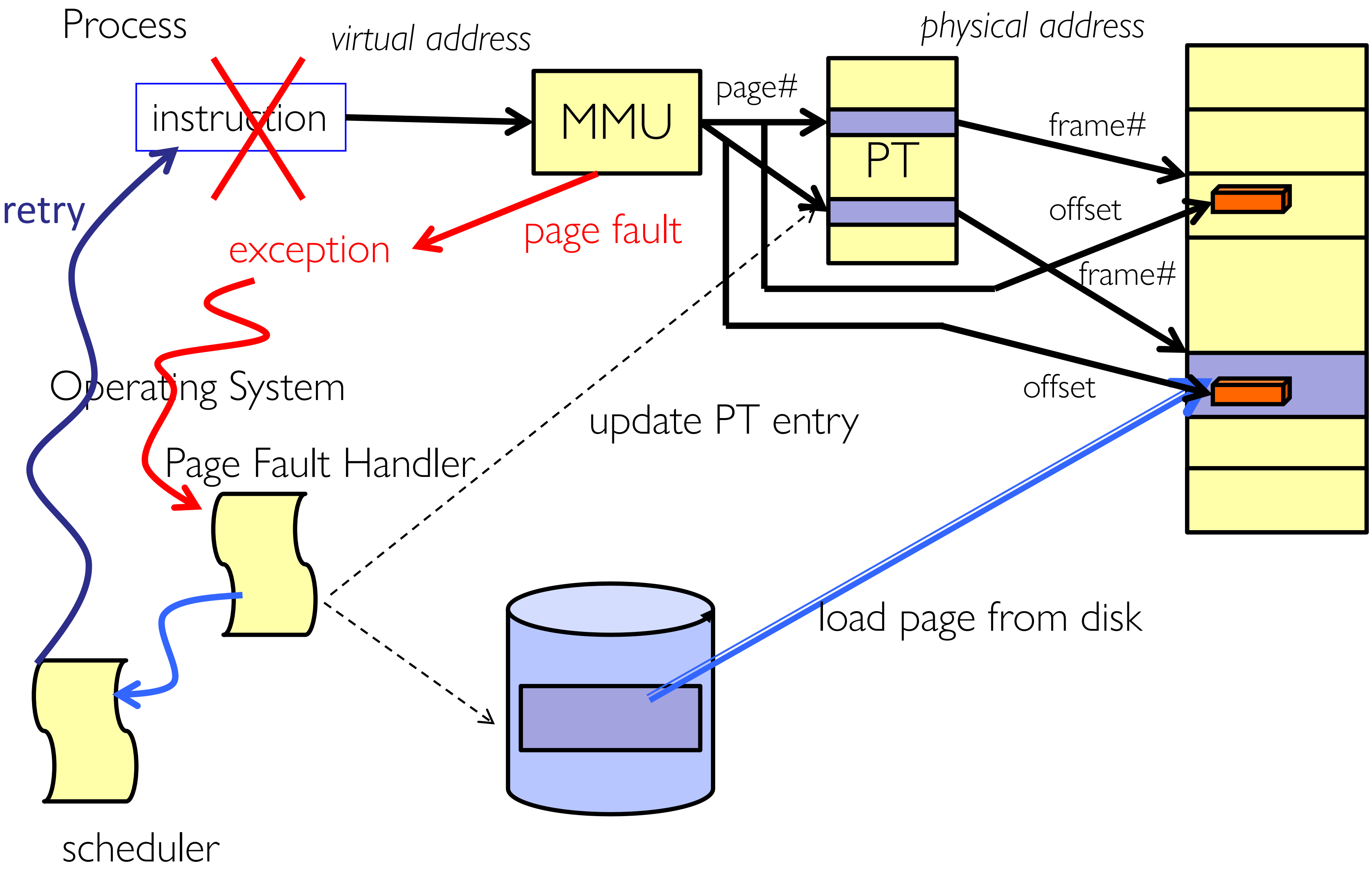
PT

Secondary Storage (Disk)

Accessed in Hardware

Speed (ns):	0.3	1	3	10-30	100	100,000 (0.1 ms)	10,000,000 (10 ms)
Size (bytes):	100Bs	10kBs	100kBs	MBs	GBs	100GBs	TBs

Next: What happens when …



What Actually Happens on a TLB Miss?

- Hardware traversed page tables:
  - On TLB miss, hardware in MMU looks at current page table to fill TLB (may walk multiple levels)
    - If PTE valid, hardware fills TLB and processor never knows
    - If PTE marked as invalid, causes Page Fault, after which kernel decides what to do afterwards
- Software traversed Page tables (like MIPS)
  - On TLB miss, processor receives TLB fault
  - Kernel traverses page table to find PTE
    - If PTE valid, fills TLB and returns from fault
    - If PTE marked as invalid, internally calls Page Fault handler
- Most chip sets provide hardware traversal
  - Modern operating systems tend to have more TLB faults since they use translation for many things
  - Examples:
    - shared segments
    - user-level portions of an operating system

Effective Access Time

- TLB Lookup = 20 ms
- Assume memory access time is 100 ms
- Hit ratio – percentage of times that a page number is found in the associative registers in TLB; ratio related to number of associative registers.
- Hit ratio = α
- Effective Access Time (EAT)

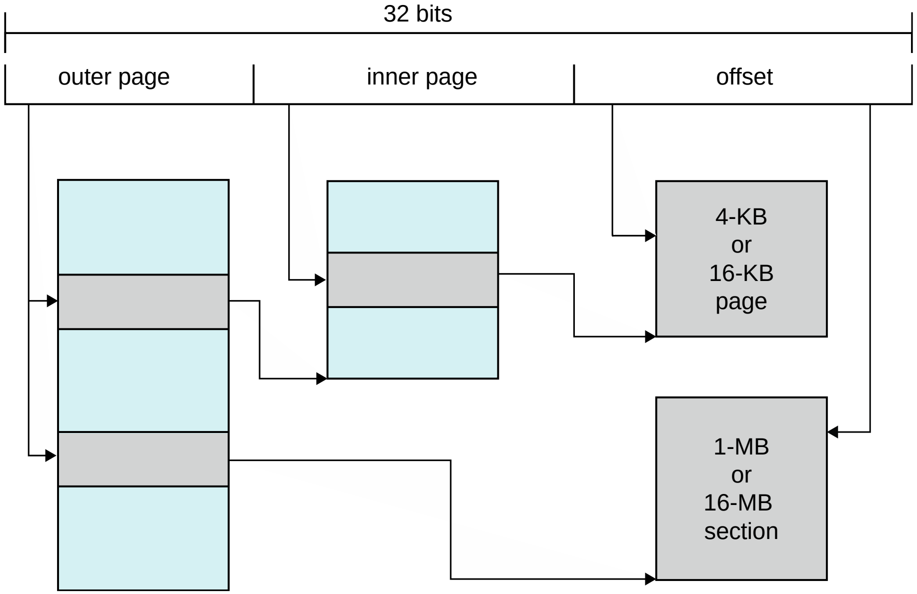
$$EAT = (100+ 20)\ \alpha + (200 + 20)(1 - \alpha)$$

What happens on a Context Switch?

- Need to do something, since TLBs map virtual addresses to physical addresses
  - Address Space just changed, so TLB entries no longer valid!
- Options?
  - Invalidate TLB aka shootdown: simple but might be expensive
    - » What if switching frequently between processes?
  - Include unique process ID in TLB (ASID field along with virtual page #)
    - » This is an architectural solution: needs hardware support
- What if translation (page) tables change?
  - For example, to move page from memory to disk or vice versa…
  - Must invalidate TLB entry!
    - » Otherwise, might think that page is still in memory!
  - Called TLB Consistency

ARM Architecture

- Dominant mobile platform chip (Apple iOS and Google Android)
- Modern, energy efficient, 32-bit CPU
- 4 KB and 16 KB pages
- 1 MB and 16 MB pages (termed **sections**)
- One-level paging for sections, two-level for smaller pages
- Two levels of TLBs
  - Outer level has two micro TLBs (one data, one instruction)
  - Inner is single main TLB
  - First micro TLB is checked, on miss inner, main TLB is checked, and on miss page table walk performed by CPU



Summary

- Memory is a resource that must be shared
  - Controlled Overlap: only shared when appropriate
  - Translation: Change Virtual Addresses into Physical Addresses
  - Protection: Prevent unauthorized Sharing of resources
- Segment Mapping
  - Segment registers within processor
  - Segment ID associated with each access
    - Often comes from portion of virtual address
    - Can come from bits in instruction instead (x86)
  - Each segment contains base and limit information
    - Offset (rest of address) adjusted by adding base
- Page Tables
  - Memory divided into fixed-sized chunks of memory
  - Virtual page number from virtual address mapped through page table to physical page number
  - Offset of virtual address same as physical address
  - Large page tables can be placed into virtual memory
- Multi-Level Tables
  - Virtual address mapped to series of tables
  - Permit sparse population of address space
- Inverted page table
  - Size of page table related to physical memory size
- Effective access time