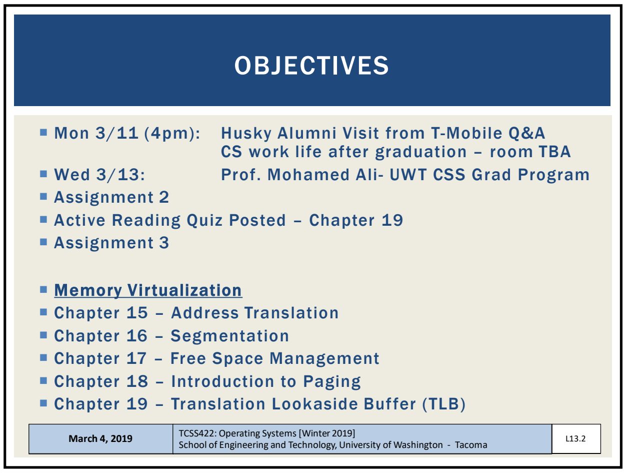
TCSS 422: OPERATING SYSTEMS

**Address Translation,
Segmentation,
Free Space Management,
Intro to Paging**

Wes J. Lloyd
School of Engineering and Technology
University of Washington - Tacoma

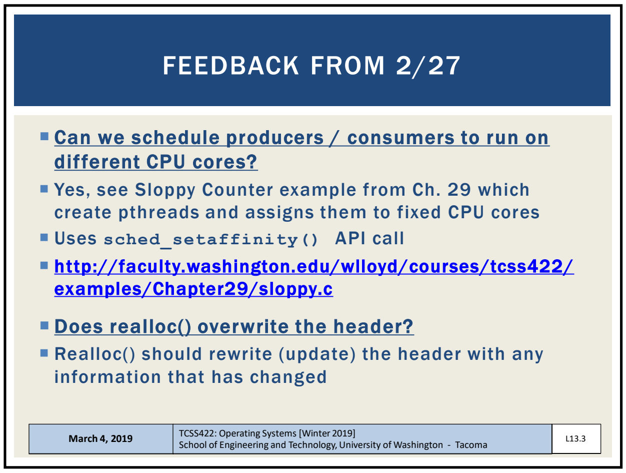
March 4, 2019 TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma



OBJECTIVES

- Mon 3/11 (4pm): Husky Alumni Visit from T-Mobile Q&A CS work life after graduation – room TBA
- Wed 3/13: Prof. Mohamed Ali- UWT CSS Grad Program
- Assignment 2
- Active Reading Quiz Posted – Chapter 19
- Assignment 3
- **Memory Virtualization**
- Chapter 15 – Address Translation
- Chapter 16 – Segmentation
- Chapter 17 – Free Space Management
- Chapter 18 – Introduction to Paging
- Chapter 19 – Translation Lookaside Buffer (TLB)

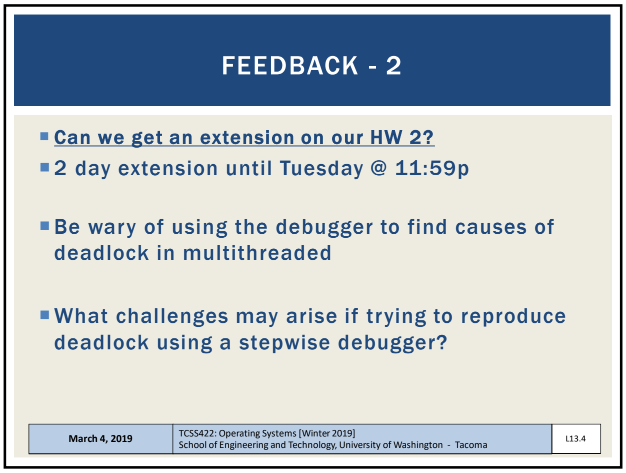
March 4, 2019 TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma L13.2



FEEDBACK FROM 2/27

- **Can we schedule producers / consumers to run on different CPU cores?**
- Yes, see Sloppy Counter example from Ch. 29 which create pthreads and assigns them to fixed CPU cores
- Uses `sched_setaffinity()` API call
- http://faculty.washington.edu/wlloyd/courses/tcss422/examples/Chapter29/sloppy.c
- **Does realloc() overwrite the header?**
- Realloc() should rewrite (update) the header with any information that has changed

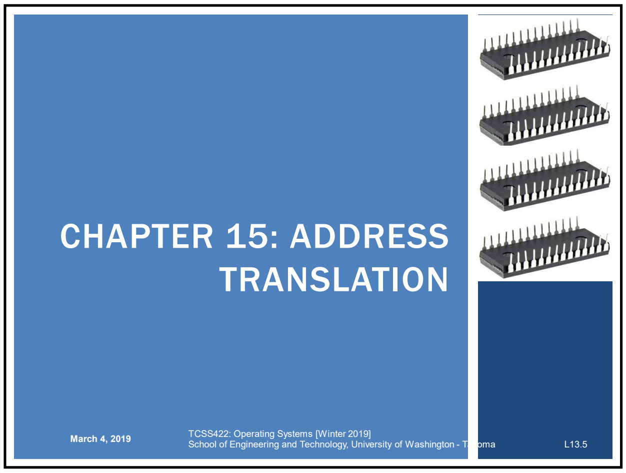
March 4, 2019 TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma L13.3



FEEDBACK - 2

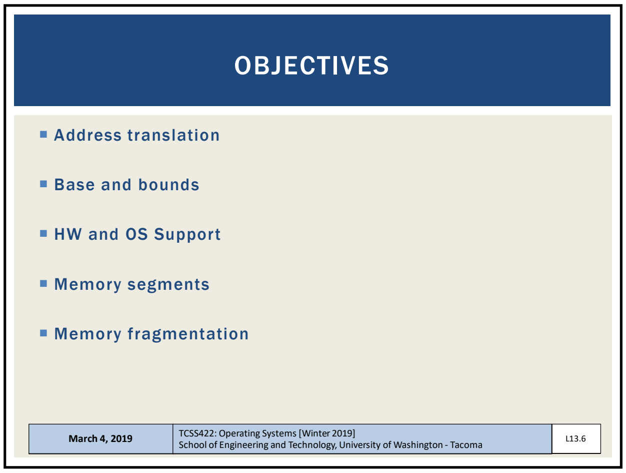
- **Can we get an extension on our HW 2?**
- 2 day extension until Tuesday @ 11:59p
- Be wary of using the debugger to find causes of deadlock in multithreaded
- What challenges may arise if trying to reproduce deadlock using a stepwise debugger?

March 4, 2019 TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma L13.4



CHAPTER 15: ADDRESS TRANSLATION

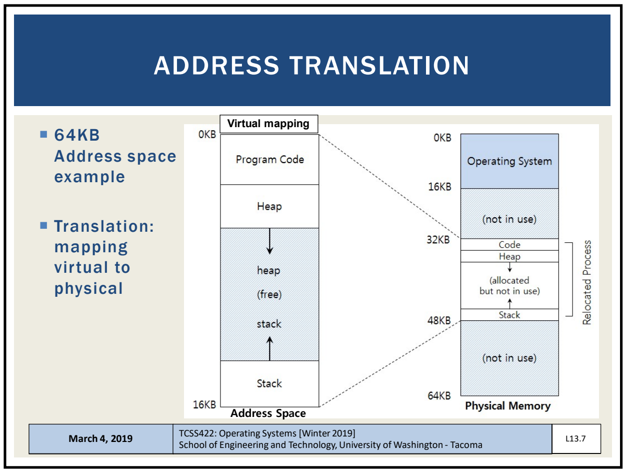
March 4, 2019 TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma L13.5



OBJECTIVES

- Address translation
- Base and bounds
- HW and OS Support
- Memory segments
- Memory fragmentation

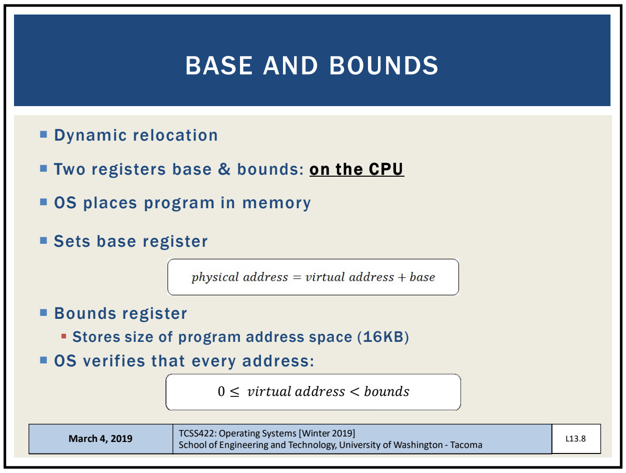
March 4, 2019 TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma L13.6



ADDRESS TRANSLATION

- 64KB Address space example
- Translation: mapping virtual to physical

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.7



BASE AND BOUNDS

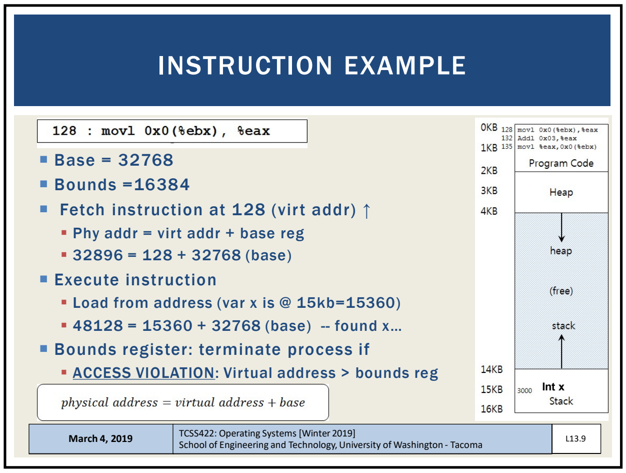
- Dynamic relocation
- Two registers base & bounds: **on the CPU**
- OS places program in memory
- Sets base register

$$physical\ address = virtual\ address + base$$

- Bounds register
  - Stores size of program address space (16KB)
- OS verifies that every address:

$$0 \leq virtual\ address < bounds$$

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.8



INSTRUCTION EXAMPLE

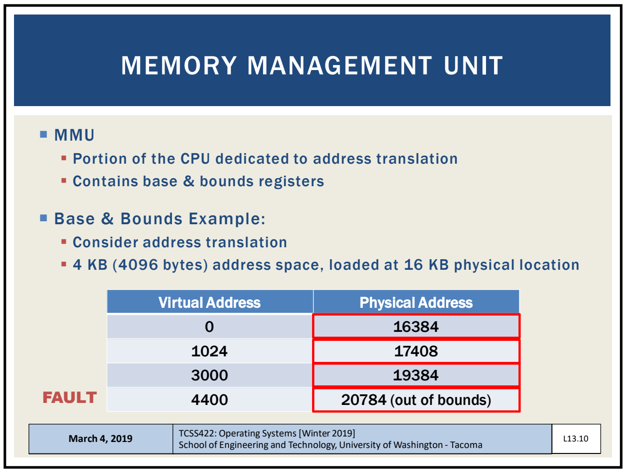
```
128 : movl 0x0(%ebx), %eax
```

- Base = 32768
- Bounds =16384
- Fetch instruction at 128 (virt addr) ↑
  - Phy addr = virt addr + base reg
  - 32896 = 128 + 32768 (base)
- Execute instruction
  - Load from address (var x is @ 15kb=15360)
  - 48128 = 15360 + 32768 (base) -- found x…
- Bounds register: terminate process if
  - ACCESS VIOLATION: Virtual address > bounds reg

$$physical\ address = virtual\ address + base$$

```
128 movl 0x0(%ebx),%eax
132 Addl 0x03,%eax
135 movl %eax,0x0(%ebx)
```

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.9

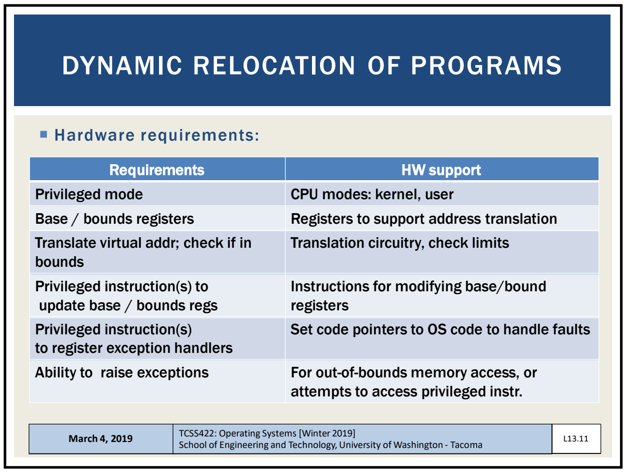


MEMORY MANAGEMENT UNIT

- MMU
  - Portion of the CPU dedicated to address translation
  - Contains base & bounds registers
- Base & Bounds Example:
  - Consider address translation
  - 4 KB (4096 bytes) address space, loaded at 16 KB physical location

	Virtual Address	Physical Address
	0	16384
	1024	17408
	3000	19384
FAULT	4400	20784 (out of bounds)

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.10

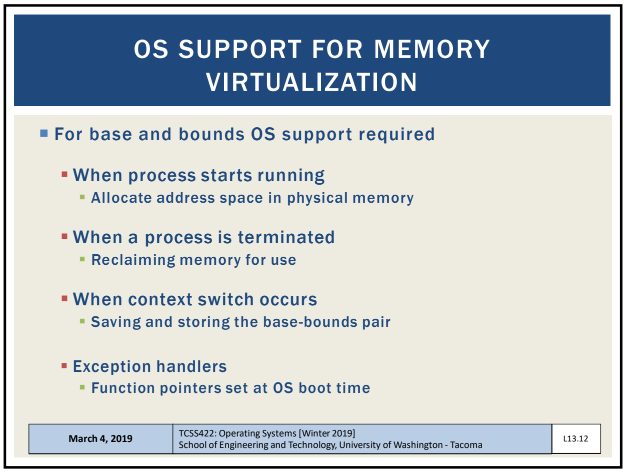


DYNAMIC RELOCATION OF PROGRAMS

- Hardware requirements:

Requirements	HW support
Privileged mode	CPU modes: kernel, user
Base / bounds registers	Registers to support address translation
Translate virtual addr; check if in bounds	Translation circuitry, check limits
Privileged instruction(s) to update base / bounds regs	Instructions for modifying base/bound registers
Privileged instruction(s) to register exception handlers	Set code pointers to OS code to handle faults
Ability to raise exceptions	For out-of-bounds memory access, or attempts to access privileged instr.

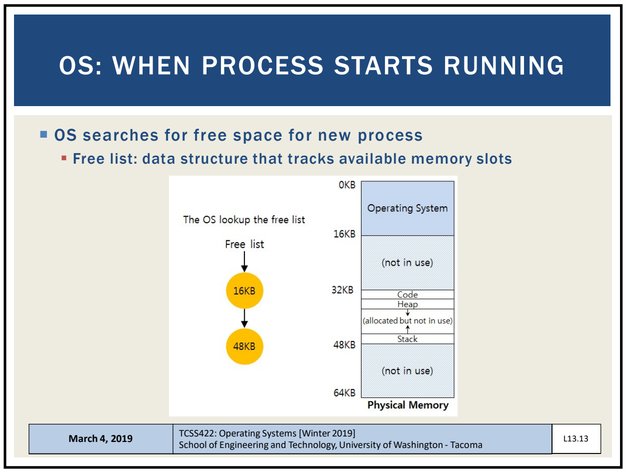
March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.11



OS SUPPORT FOR MEMORY VIRTUALIZATION

- For base and bounds OS support required
  - When process starts running
    - Allocate address space in physical memory
  - When a process is terminated
    - Reclaiming memory for use
  - When context switch occurs
    - Saving and storing the base-bounds pair
  - Exception handlers
    - Function pointers set at OS boot time

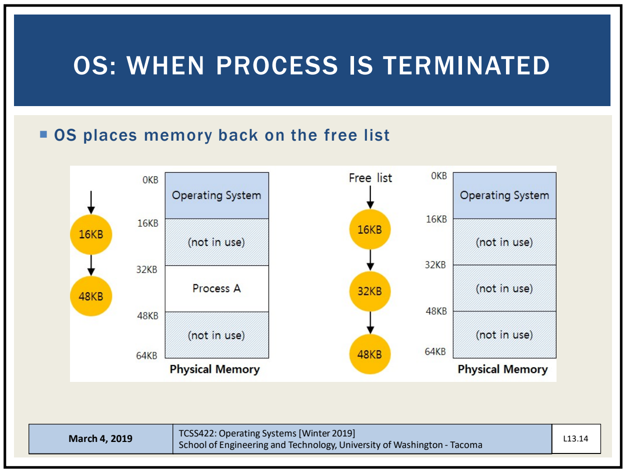
March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.12



OS: WHEN PROCESS STARTS RUNNING

- OS searches for free space for new process
  - Free list: data structure that tracks available memory slots

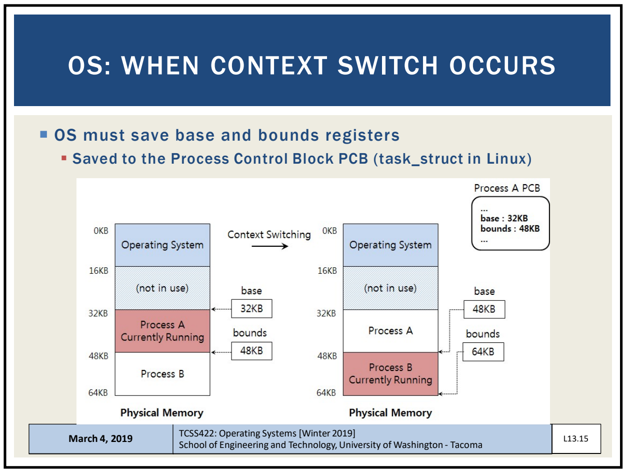
March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.13



OS: WHEN PROCESS IS TERMINATED

- OS places memory back on the free list

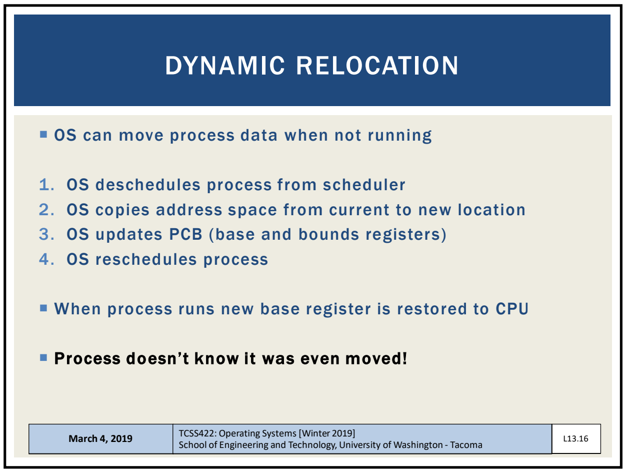
March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.14



OS: WHEN CONTEXT SWITCH OCCURS

- OS must save base and bounds registers
  - Saved to the Process Control Block PCB (task_struct in Linux)

March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.15



DYNAMIC RELOCATION

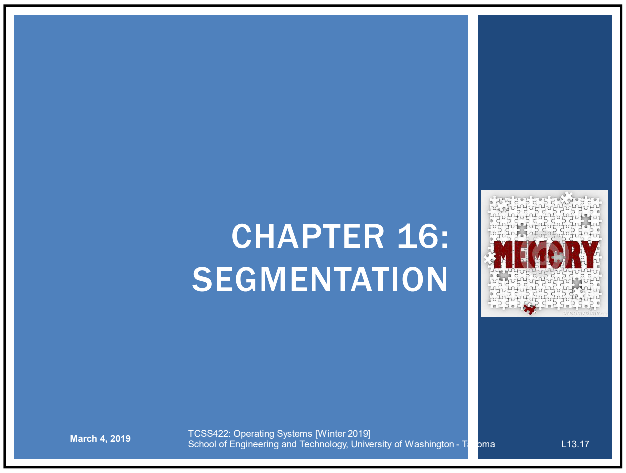
- OS can move process data when not running

1. OS deschedules process from scheduler
2. OS copies address space from current to new location
3. OS updates PCB (base and bounds registers)
4. OS reschedules process

- When process runs new base register is restored to CPU

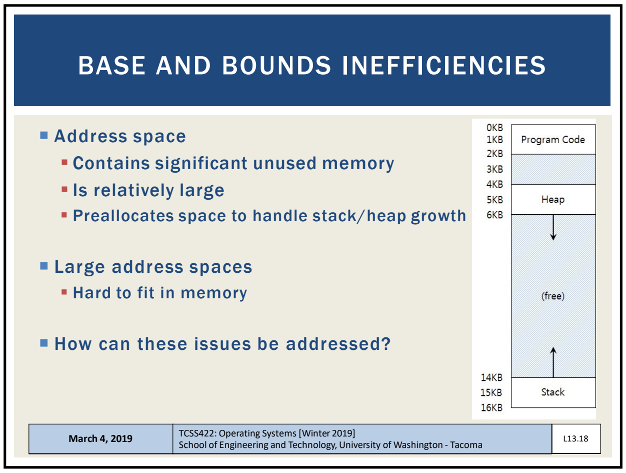
- **Process doesn't know it was even moved!**

March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.16



CHAPTER 16: SEGMENTATION

March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.17



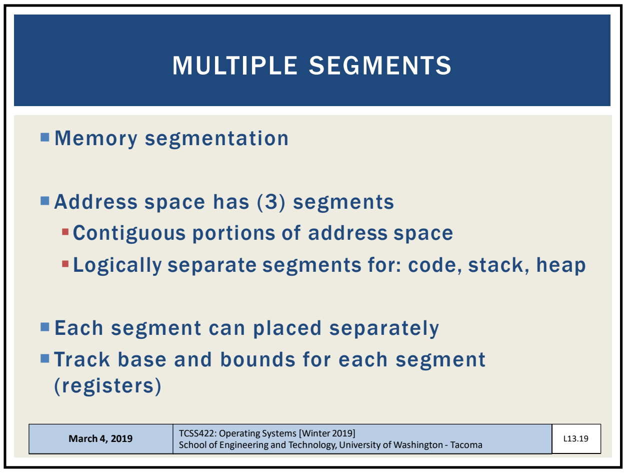
BASE AND BOUNDS INEFFICIENCIES

- Address space
  - Contains significant unused memory
  - Is relatively large
  - Preallocates space to handle stack/heap growth

- Large address spaces
  - Hard to fit in memory

- How can these issues be addressed?

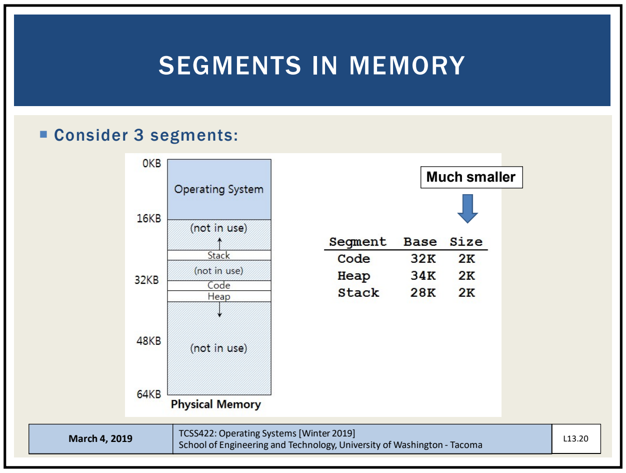
March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.18



MULTIPLE SEGMENTS

- Memory segmentation
- Address space has (3) segments
  - Contiguous portions of address space
  - Logically separate segments for: code, stack, heap
- Each segment can placed separately
- Track base and bounds for each segment (registers)

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.19

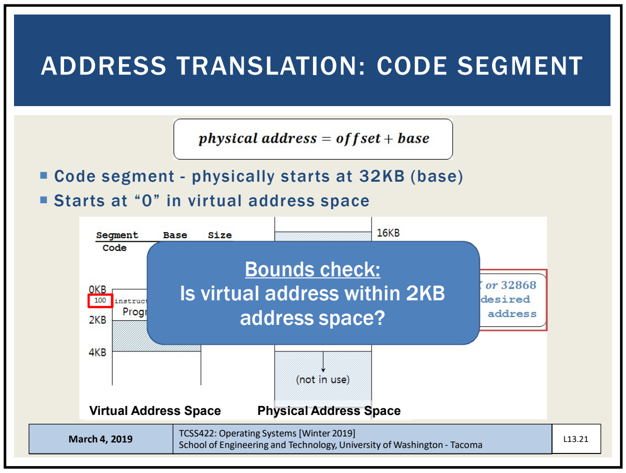


SEGMENTS IN MEMORY

- Consider 3 segments:

Segment	Base	Size
Code	32K	2K
Heap	34K	2K
Stack	28K	2K

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.20



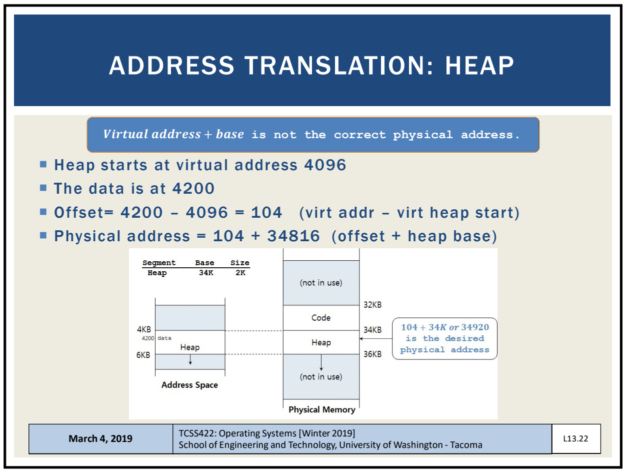
ADDRESS TRANSLATION: CODE SEGMENT

$$physical\ address = offset + base$$

- Code segment - physically starts at 32KB (base)
- Starts at "0" in virtual address space

Bounds check:
Is virtual address within 2KB address space?

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.21



ADDRESS TRANSLATION: HEAP

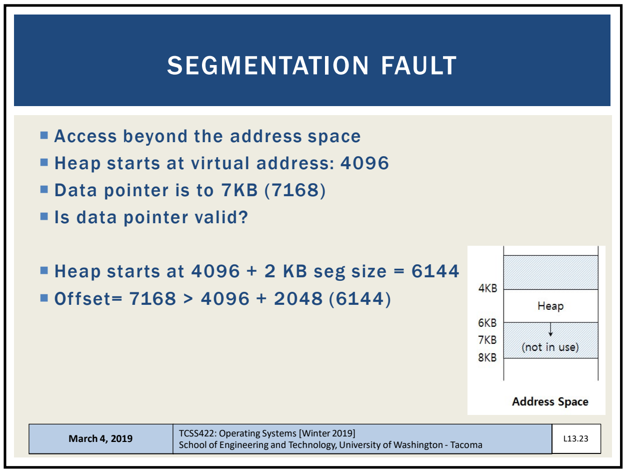
$Virtual\ address + base$ is not the correct physical address.

- Heap starts at virtual address 4096
- The data is at 4200
- Offset= 4200 – 4096 = 104 (virt addr – virt heap start)
- Physical address = 104 + 34816 (offset + heap base)

Segment	Base	Size
Heap	34K	2K

$104 + 34K\ or\ 34920$ is the desired physical address

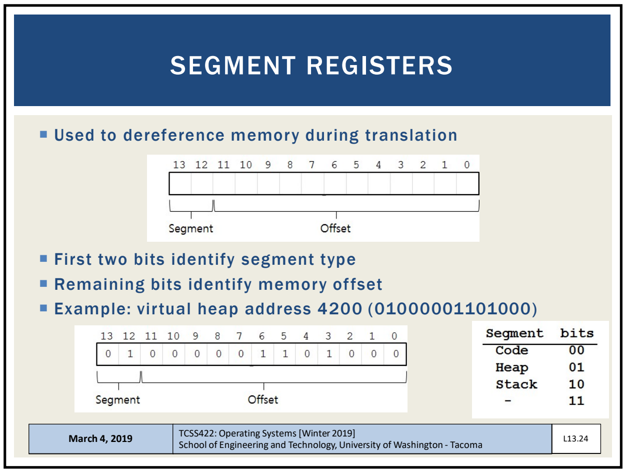
March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.22



SEGMENTATION FAULT

- Access beyond the address space
- Heap starts at virtual address: 4096
- Data pointer is to 7KB (7168)
- Is data pointer valid?
- Heap starts at 4096 + 2 KB seg size = 6144
- Offset= 7168 > 4096 + 2048 (6144)

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.23



SEGMENT REGISTERS

- Used to dereference memory during translation

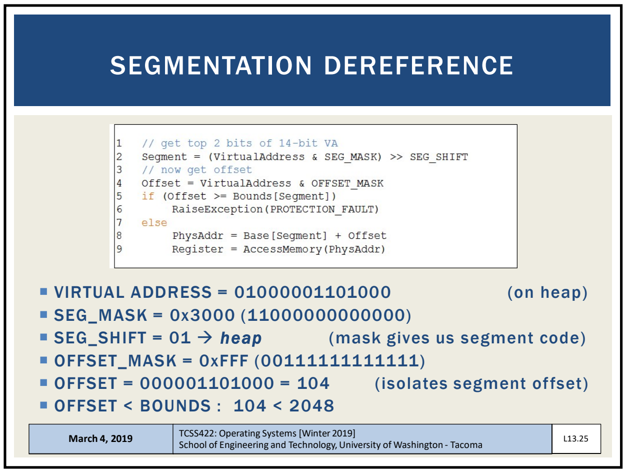
13	12	11	10	9	8	7	6	5	4	3	2	1	0

- First two bits identify segment type
- Remaining bits identify memory offset
- Example: virtual heap address 4200 (01000001101000)

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

Segment	bits
Code	00
Heap	01
Stack	10
-	11

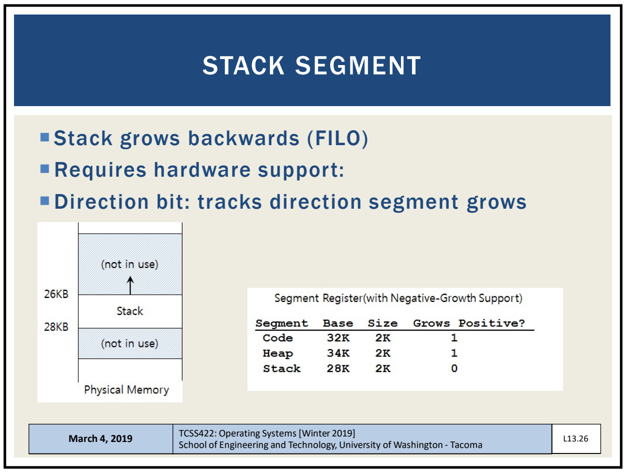
March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.24



SEGMENTATION DEREFERENCE

```
// get top 2 bits of 14-bit VA
Segment = (VirtualAddress & SEG_MASK) >> SEG_SHIFT
// now get offset
Offset = VirtualAddress & OFFSET_MASK
if (Offset >= Bounds[Segment])
     RaiseException(PROTECTION_FAULT)
else
     PhysAddr = Base[Segment] + Offset
     Register = AccessMemory(PhysAddr)
```

- VIRTUAL ADDRESS = 01000001101000 (on heap)
- SEG_MASK = 0x3000 (11000000000000)
- SEG_SHIFT = 01 → *heap* (mask gives us segment code)
- OFFSET_MASK = 0xFFF (00111111111111)
- OFFSET = 000001101000 = 104 (isolates segment offset)
- OFFSET < BOUNDS : 104 < 2048



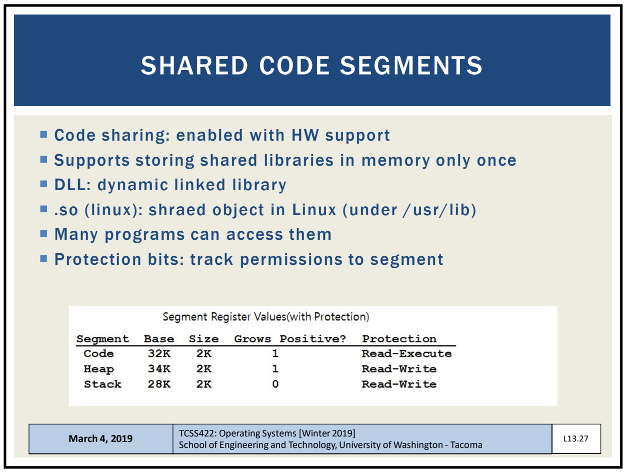
STACK SEGMENT

- Stack grows backwards (FILO)
- Requires hardware support:
- Direction bit: tracks direction segment grows

(not in use)
26KB
Stack
28KB
(not in use)
Physical Memory

Segment Register(with Negative-Growth Support)

Segment	Base	Size	Grows Positive?
Code	32K	2K	1
Heap	34K	2K	1
Stack	28K	2K	0

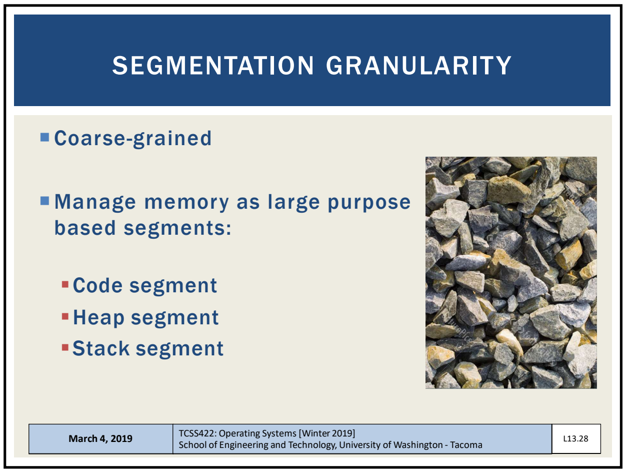


SHARED CODE SEGMENTS

- Code sharing: enabled with HW support
- Supports storing shared libraries in memory only once
- DLL: dynamic linked library
- .so (linux): shraed object in Linux (under /usr/lib)
- Many programs can access them
- Protection bits: track permissions to segment

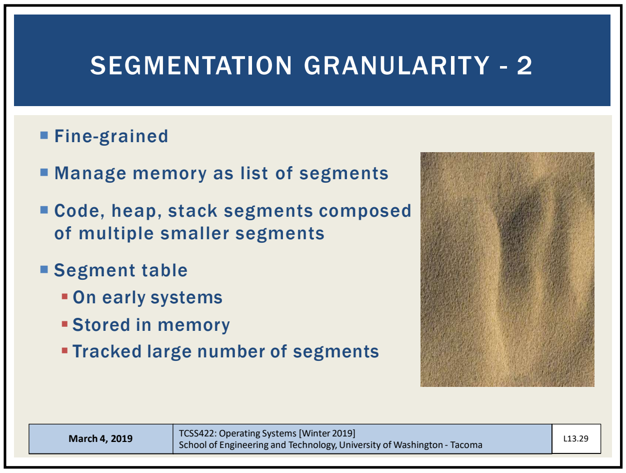
Segment Register Values(with Protection)

Segment	Base	Size	Grows Positive?	Protection
Code	32K	2K	1	Read-Execute
Heap	34K	2K	1	Read-Write
Stack	28K	2K	0	Read-Write



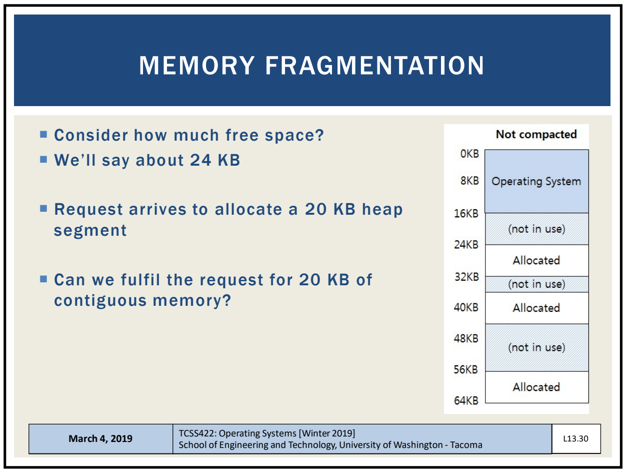
SEGMENTATION GRANULARITY

- Coarse-grained
- Manage memory as large purpose based segments:
  - Code segment
  - Heap segment
  - Stack segment



SEGMENTATION GRANULARITY - 2

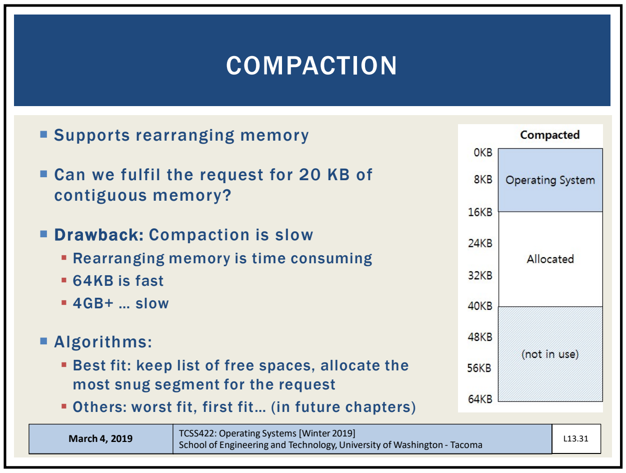
- Fine-grained
- Manage memory as list of segments
- Code, heap, stack segments composed of multiple smaller segments
- Segment table
  - On early systems
  - Stored in memory
  - Tracked large number of segments



MEMORY FRAGMENTATION

- Consider how much free space?
- We'll say about 24 KB
- Request arrives to allocate a 20 KB heap segment
- Can we fulfil the request for 20 KB of contiguous memory?

Not compacted
0KB
8KB
Operating System
16KB
(not in use)
24KB
Allocated
32KB
(not in use)
40KB
Allocated
48KB
(not in use)
56KB
Allocated
64KB

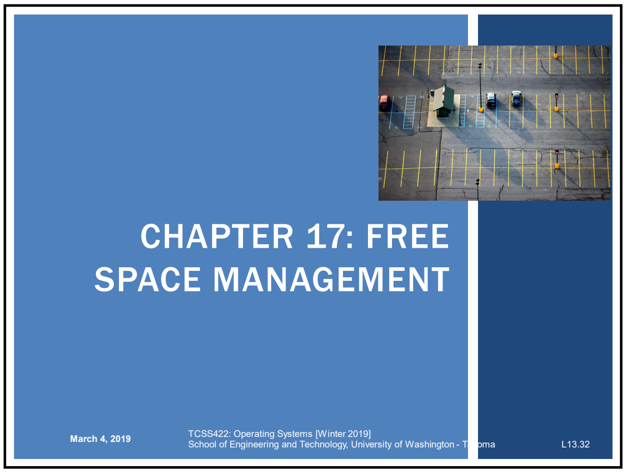


COMPACTION

- Supports rearranging memory
- Can we fulfil the request for 20 KB of contiguous memory?
- **Drawback:** Compaction is slow
  - Rearranging memory is time consuming
  - 64KB is fast
  - 4GB+ … slow
- Algorithms:
  - Best fit: keep list of free spaces, allocate the most snug segment for the request
  - Others: worst fit, first fit… (in future chapters)

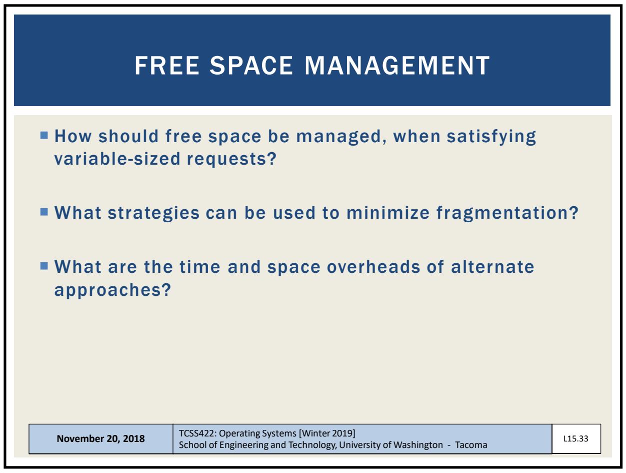
Compacted
0KB
8KB
Operating System
16KB
24KB
Allocated
32KB
40KB
48KB
(not in use)
56KB
64KB

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.31



CHAPTER 17: FREE SPACE MANAGEMENT

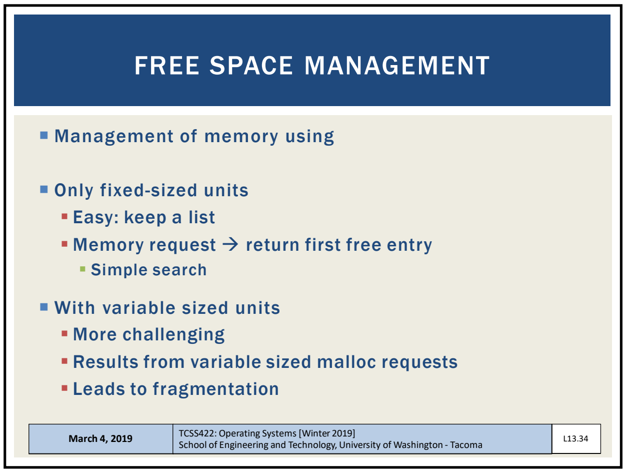
March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.32



FREE SPACE MANAGEMENT

- How should free space be managed, when satisfying variable-sized requests?
- What strategies can be used to minimize fragmentation?
- What are the time and space overheads of alternate approaches?

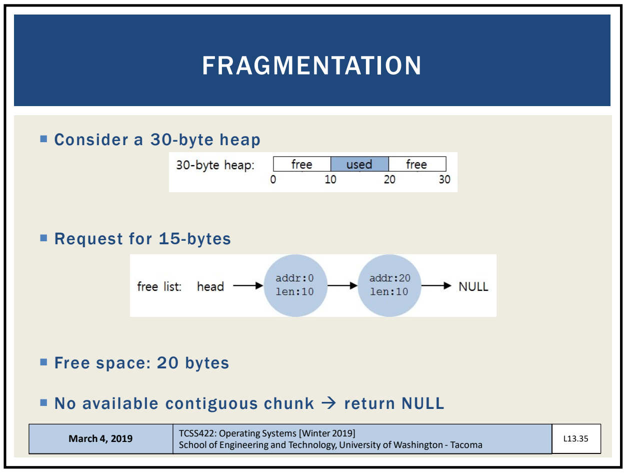
November 20, 2018 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L15.33



FREE SPACE MANAGEMENT

- Management of memory using
- Only fixed-sized units
  - Easy: keep a list
  - Memory request → return first free entry
    - Simple search
- With variable sized units
  - More challenging
  - Results from variable sized malloc requests
  - Leads to fragmentation

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.34



FRAGMENTATION

- Consider a 30-byte heap

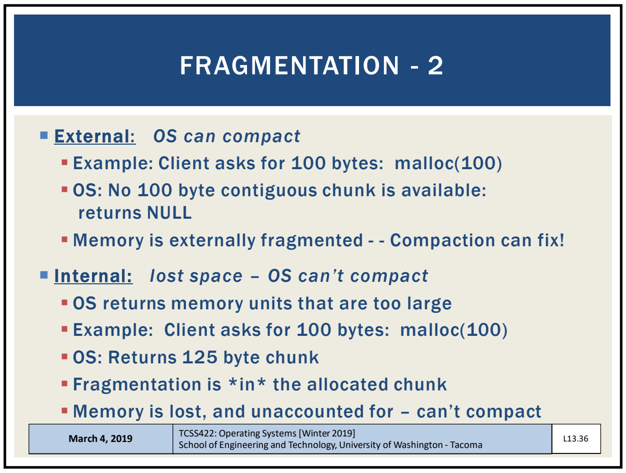
30-byte heap:
free
used
free
0
10
20
30

- Request for 15-bytes

free list:
head
addr:0
len:10
addr:20
len:10
NULL

- Free space: 20 bytes
- No available contiguous chunk → return NULL

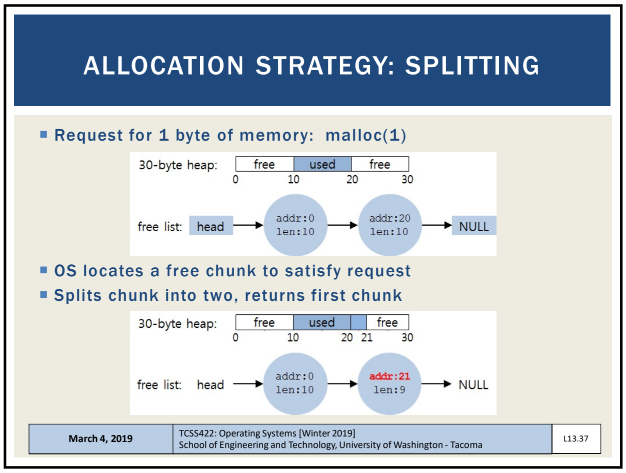
March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.35



FRAGMENTATION - 2

- **External:** *OS can compact*
  - Example: Client asks for 100 bytes: malloc(100)
  - OS: No 100 byte contiguous chunk is available: returns NULL
  - Memory is externally fragmented - - Compaction can fix!
- **Internal:** *lost space – OS can't compact*
  - OS returns memory units that are too large
  - Example: Client asks for 100 bytes: malloc(100)
  - OS: Returns 125 byte chunk
  - Fragmentation is *in* the allocated chunk
  - Memory is lost, and unaccounted for – can't compact

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.36



ALLOCATION STRATEGY: SPLITTING

- Request for 1 byte of memory: malloc(1)

30-byte heap: free | used | free
0 10 20 30

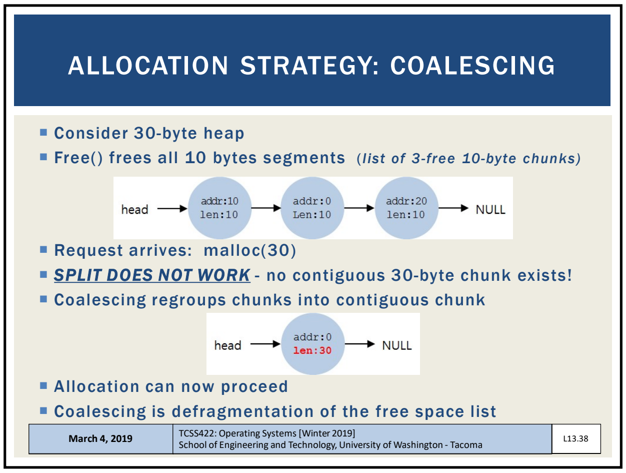
free list: head → addr:0 len:10 → addr:20 len:10 → NULL

- OS locates a free chunk to satisfy request
- Splits chunk into two, returns first chunk

30-byte heap: free | used | | free
0 10 20 21 30

free list: head → addr:0 len:10 → addr:21 len:9 → NULL

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.37



ALLOCATION STRATEGY: COALESCING

- Consider 30-byte heap
- Free() frees all 10 bytes segments *(list of 3-free 10-byte chunks)*

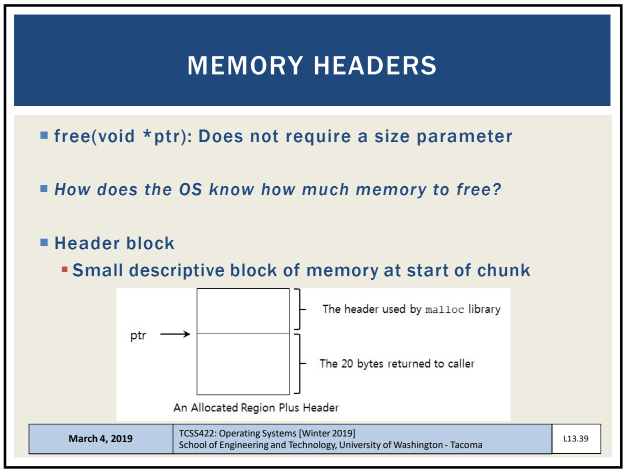
head → addr:10 len:10 → addr:0 Len:10 → addr:20 len:10 → NULL

- Request arrives: malloc(30)
- ***SPLIT DOES NOT WORK*** - no contiguous 30-byte chunk exists!
- Coalescing regroups chunks into contiguous chunk

head → addr:0 len:30 → NULL

- Allocation can now proceed
- Coalescing is defragmentation of the free space list

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.38



MEMORY HEADERS

- free(void *ptr): Does not require a size parameter
- *How does the OS know how much memory to free?*
- Header block
  - Small descriptive block of memory at start of chunk

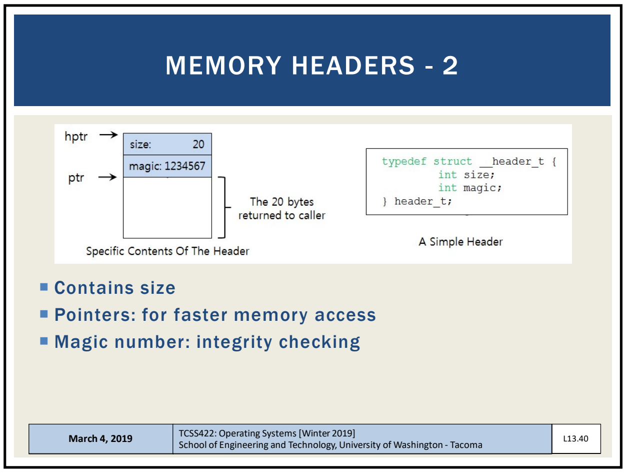
ptr →

The header used by malloc library

The 20 bytes returned to caller

An Allocated Region Plus Header

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.39



MEMORY HEADERS - 2

hptr → size: 20

magic: 1234567

ptr →

The 20 bytes returned to caller

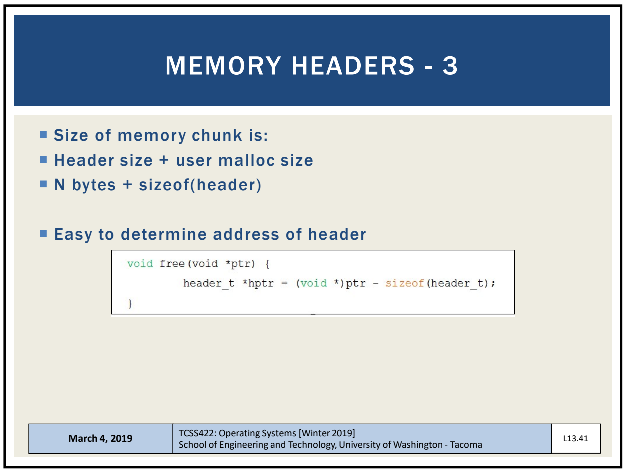
Specific Contents Of The Header

```
typedef struct __header_t {
        int size;
        int magic;
} header_t;
```

A Simple Header

- Contains size
- Pointers: for faster memory access
- Magic number: integrity checking

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.40

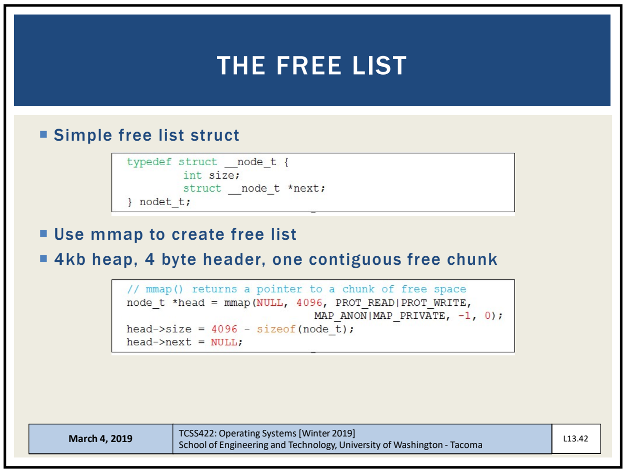


MEMORY HEADERS - 3

- Size of memory chunk is:
- Header size + user malloc size
- N bytes + sizeof(header)
- Easy to determine address of header

```
void free(void *ptr) {
        header_t *hptr = (void *)ptr - sizeof(header_t);
}
```

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.41



THE FREE LIST

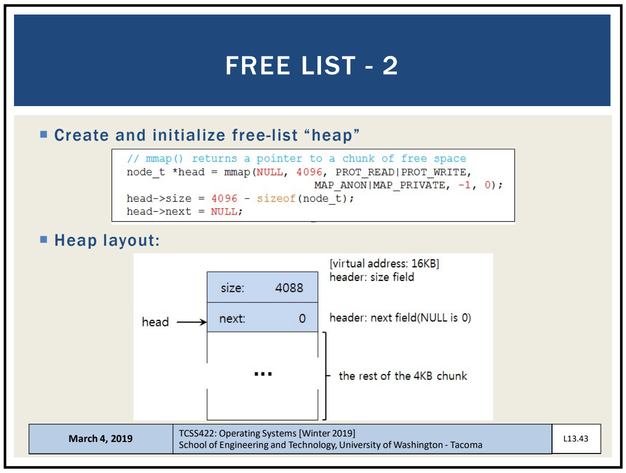
- Simple free list struct

```
typedef struct __node_t {
        int size;
        struct __node_t *next;
} nodet_t;
```

- Use mmap to create free list
- 4kb heap, 4 byte header, one contiguous free chunk

```
// mmap() returns a pointer to a chunk of free space
node_t *head = mmap(NULL, 4096, PROT_READ|PROT_WRITE,
                              MAP_ANON|MAP_PRIVATE, -1, 0);
head->size = 4096 - sizeof(node_t);
head->next = NULL;
```

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.42



FREE LIST - 2

- Create and initialize free-list “heap”

```
// mmap() returns a pointer to a chunk of free space
node_t *head = mmap(NULL, 4096, PROT_READ|PROT_WRITE,
                              MAP_ANON|MAP_PRIVATE, -1, 0);
head->size = 4096 - sizeof(node_t);
head->next = NULL;
```

- Heap layout:

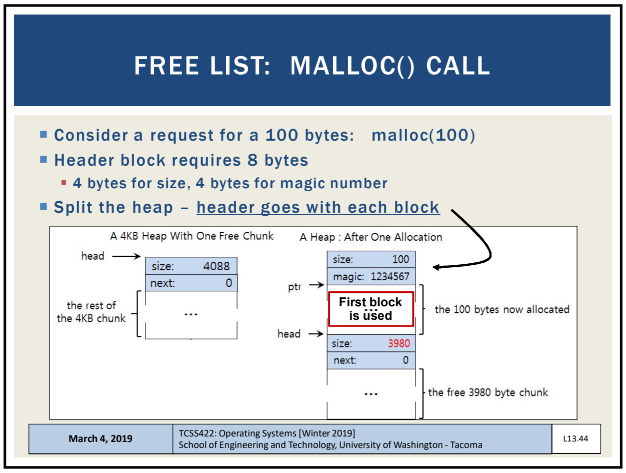
head → size: 4088 | next: 0 | …

[virtual address: 16KB]
header: size field

header: next field(NULL is 0)

the rest of the 4KB chunk

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.43



FREE LIST: MALLOC() CALL

- Consider a request for a 100 bytes: malloc(100)
- Header block requires 8 bytes
  - 4 bytes for size, 4 bytes for magic number
- Split the heap – header goes with each block

A 4KB Heap With One Free Chunk

head → size: 4088 | next: 0 | …

the rest of the 4KB chunk

A Heap : After One Allocation

size: 100 | magic: 1234567

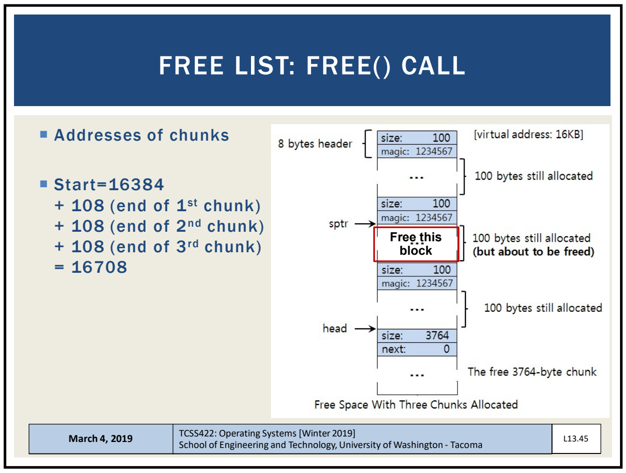
ptr → First block is used

the 100 bytes now allocated

head → size: 3980 | next: 0 | …

the free 3980 byte chunk

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.44



FREE LIST: FREE() CALL

- Addresses of chunks

- Start=16384
  + 108 (end of 1st chunk)
  + 108 (end of 2nd chunk)
  + 108 (end of 3rd chunk)
  = 16708

8 bytes header: size: 100 | magic: 1234567

[virtual address: 16KB]

… 100 bytes still allocated

size: 100 | magic: 1234567

sptr → Free this block

100 bytes still allocated **(but about to be freed)**

size: 100 | magic: 1234567

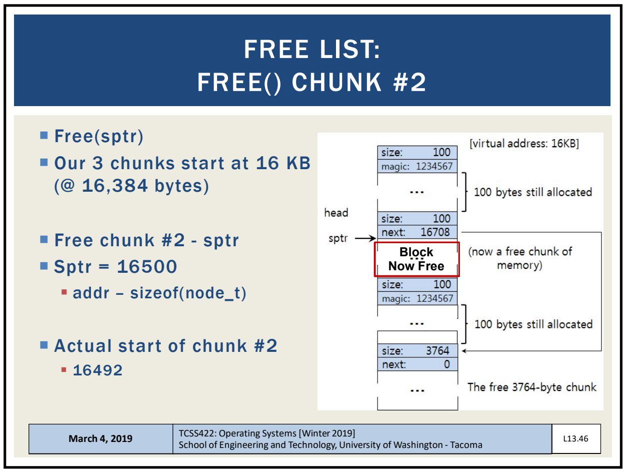
… 100 bytes still allocated

head → size: 3764 | next: 0

… The free 3764-byte chunk

Free Space With Three Chunks Allocated

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.45



FREE LIST: FREE() CHUNK #2

- Free(sptr)
- Our 3 chunks start at 16 KB (@ 16,384 bytes)

- Free chunk #2 - sptr
- Sptr = 16500
  - addr – sizeof(node_t)

- Actual start of chunk #2
  - 16492

[virtual address: 16KB]

size: 100 | magic: 1234567

… 100 bytes still allocated

head

size: 100 | next: 16708

sptr → Block Now Free

(now a free chunk of memory)

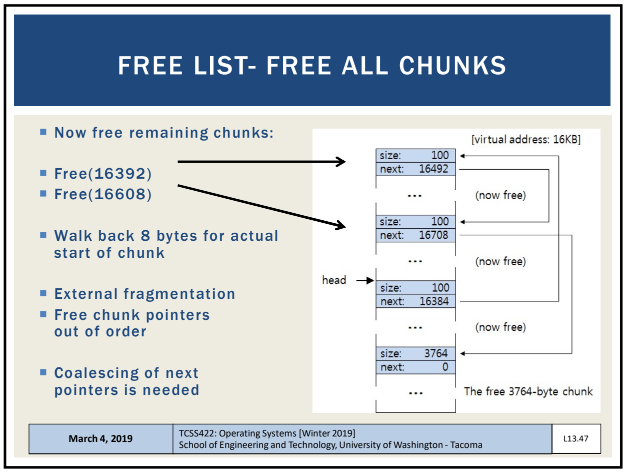
size: 100 | magic: 1234567

… 100 bytes still allocated

size: 3764 | next: 0

… The free 3764-byte chunk

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.46



FREE LIST- FREE ALL CHUNKS

- Now free remaining chunks:

- Free(16392)
- Free(16608)

- Walk back 8 bytes for actual start of chunk

- External fragmentation
- Free chunk pointers out of order

- Coalescing of next pointers is needed

[virtual address: 16KB]

size: 100 | next: 16492

… (now free)

size: 100 | next: 16708

… (now free)

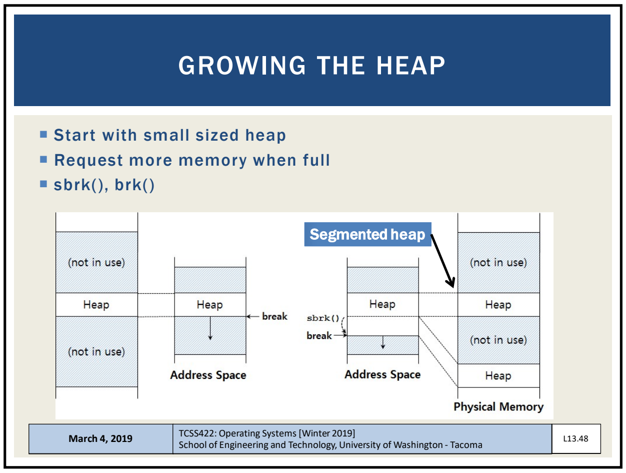
head → size: 100 | next: 16384

… (now free)

size: 3764 | next: 0

… The free 3764-byte chunk

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.47



GROWING THE HEAP

- Start with small sized heap
- Request more memory when full
- sbrk(), brk()

March 4, 2019 | TCSS422: Operating Systems [Winter 2019] School of Engineering and Technology, University of Washington - Tacoma | L13.48

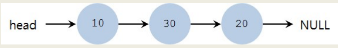
MEMORY ALLOCATION STRATEGIES

- **Best fit**
  - Traverse free list
  - Identify all candidate free chunks
  - Note which is smallest (has best fit)
  - When splitting, “leftover” pieces are small (and potentially less useful -- fragmented)

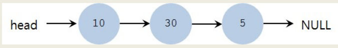
- **Worst fit**
  - Traverse free list
  - Identify largest free chunk
  - Split largest free chunk, leaving a <u>still large free chunk</u>

EXAMPLES

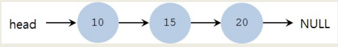
- Allocation request for 15 bytes



- Result of Best Fit



- Result of Worst Fit



MEMORY ALLOCATION STRATEGIES - 2

- **First fit**
  - Start search at beginning of free list
  - Find first chunk large enough for request
  - Split chunk, returning a “fit” chunk, saving the remainder
  - Avoids full free list traversal of best and worst fit

- **Next fit**
  - Similar to first fit, but start search at last search location
  - Maintain a pointer that “cycles” through the list
  - Helps balance chunk distribution vs. first fit
  - Find first chunk, that is large enough for the request, and split
  - Avoids full free list traversal

SEGREGATED LISTS

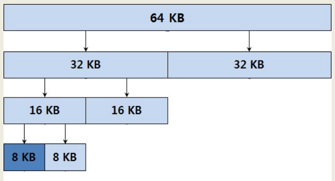
- For popular sized requests e.g. for kernel objects such as locks, inodes, etc.
- Manage as segregated free lists
- Provide object caches: stores pre-initialized objects

- How much memory should be dedicated for specialized requests (object caches)?

- If a given cache is low in memory, can request “*slabs*” of memory from the general allocator for caches.
- General allocator will reclaim slabs when not used

BUDDY ALLOCATION

- Binary buddy allocation
  - Divides free space by two to find a block that is big enough to accommodate the request; the next split is too small…
- Consider a 7KB request



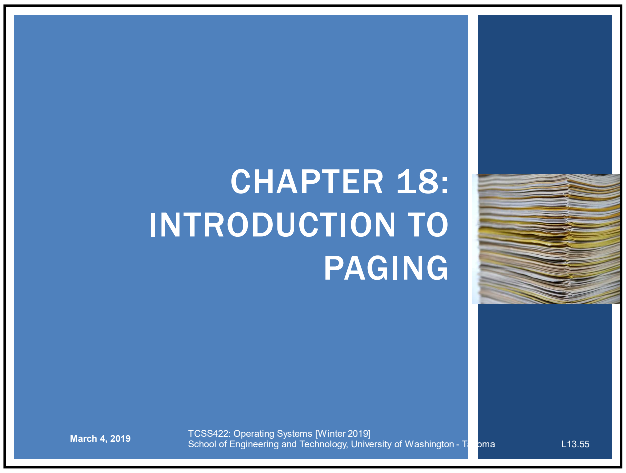
64KB free space for 7KB request

BUDDY ALLOCATION - 2

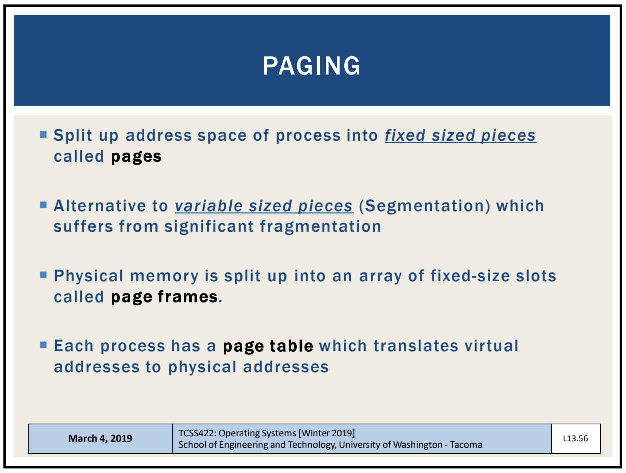
- Buddy allocation: suffers from internal fragmentation

- Allocated fragments, typically too large

- Coalescing is simple
  - Two adjacent blocks are promoted up

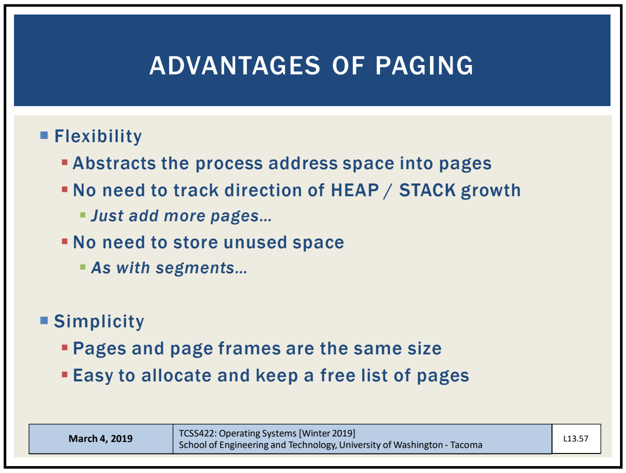


CHAPTER 18: INTRODUCTION TO PAGING



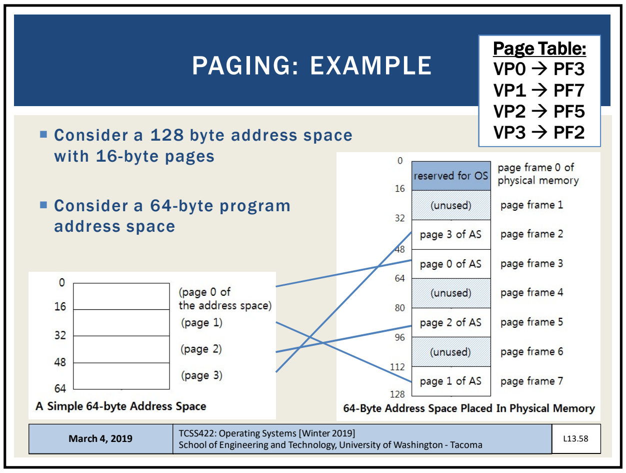
PAGING

- Split up address space of process into ***fixed sized pieces*** called **pages**
- Alternative to ***variable sized pieces*** (Segmentation) which suffers from significant fragmentation
- Physical memory is split up into an array of fixed-size slots called **page frames**.
- Each process has a **page table** which translates virtual addresses to physical addresses



ADVANTAGES OF PAGING

- Flexibility
  - Abstracts the process address space into pages
  - No need to track direction of HEAP / STACK growth
    - *Just add more pages…*
  - No need to store unused space
    - *As with segments…*
- Simplicity
  - Pages and page frames are the same size
  - Easy to allocate and keep a free list of pages



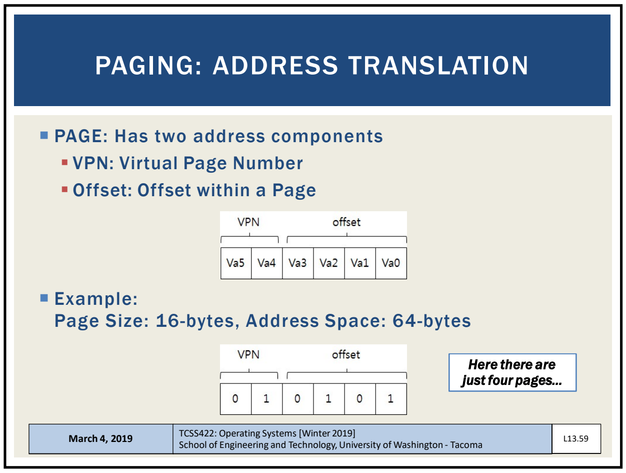
PAGING: EXAMPLE

Page Table:
VP0 → PF3
VP1 → PF7
VP2 → PF5
VP3 → PF2

- Consider a 128 byte address space with 16-byte pages
- Consider a 64-byte program address space

A Simple 64-byte Address Space

64-Byte Address Space Placed In Physical Memory

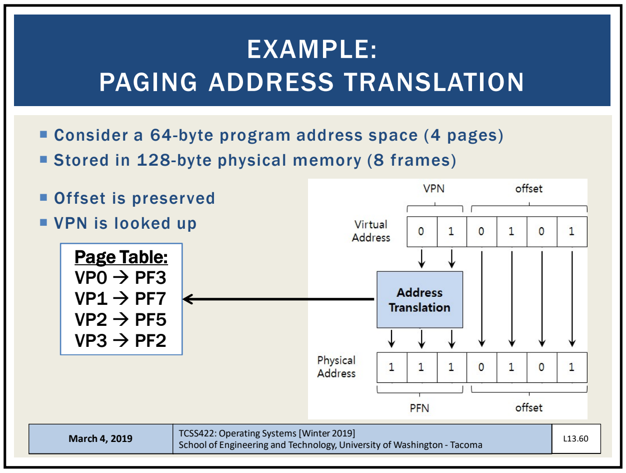


PAGING: ADDRESS TRANSLATION

- PAGE: Has two address components
  - VPN: Virtual Page Number
  - Offset: Offset within a Page

- Example:
  Page Size: 16-bytes, Address Space: 64-bytes

Here there are just four pages…



EXAMPLE: PAGING ADDRESS TRANSLATION

- Consider a 64-byte program address space (4 pages)
- Stored in 128-byte physical memory (8 frames)
- Offset is preserved
- VPN is looked up

Page Table:
VP0 → PF3
VP1 → PF7
VP2 → PF5
VP3 → PF2

PAGING DESIGN QUESTIONS

- (1) Where are page tables stored?
- (2) What are the typical contents of the page table?
- (3) How big are page tables?
- (4) Does paging make the system too slow?

(1) WHERE ARE PAGE TABLES STORED?

- Example:
  - Consider a 32-bit process address space (up to 4GB)
  - With 4 KB pages
  - 20 bits for VPN (2^{20} pages)
  - 12 bits for the page offset (2^{12} unique bytes in a page)
- Page tables for each process are stored in RAM
  - Support potential storage of 2^{20} translations = 1,048,576 pages per process
  - Each page has a page table entry size of 4 bytes

PAGE TABLE EXAMPLE

- With 2^{20} slots in our page table for a single process
- Each slot dereferences a VPN
- Provides physical frame number
- Each slot requires 4 bytes (32 bits)
  - 20 for the PFN on a 4GB system with 4KB pages
  - 12 for the offset which is preserved
  - (note we have no status bits, so this is unrealistically small)

VPN_0
VPN_1
VPN_2
…
…
$VPN_{1048576}$

- How much memory to store page table for 1 process?
  - 4,194,304 bytes (or 4MB) to index one process

NOW FOR AN ENTIRE OS

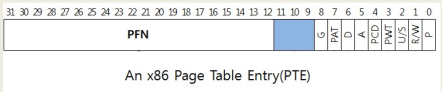
- If 4 MB is required to store one process
- Consider how much memory is required for an entire OS?
  - With for example 100 processes…
- Page table memory requirement is now 4MB x 100 = 400MB
- If computer has 4GB memory (maximum for 32-bits), the page table consumes 10% of memory

400 MB / 4000 GB

- Is this efficient?

(2) WHAT'S ACTUALLY IN THE PAGE TABLE

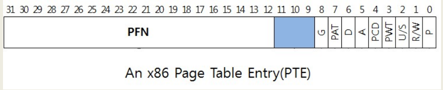
- Page table is data structure used to map virtual page numbers (VPN) to the physical address (Physical Frame Number PFN)
  - Linear page table → simple array
- Page-table entry
  - 32 bits for capturing state



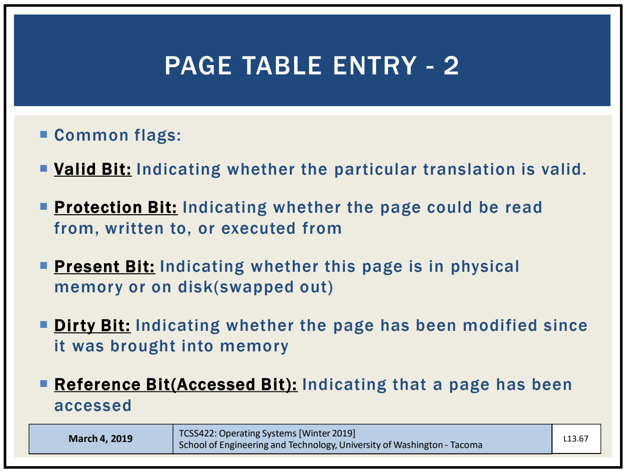
An x86 Page Table Entry(PTE)

PAGE TABLE ENTRY

- P: present
- R/W: read/write bit
- U/S: supervisor
- A: accessed bit
- D: dirty bit
- PFN: the page frame number



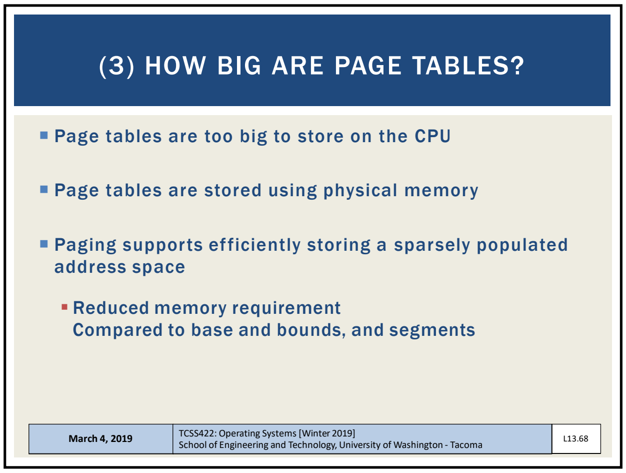
An x86 Page Table Entry(PTE)



PAGE TABLE ENTRY - 2

- Common flags:
- **Valid Bit:** Indicating whether the particular translation is valid.
- **Protection Bit:** Indicating whether the page could be read from, written to, or executed from
- **Present Bit:** Indicating whether this page is in physical memory or on disk(swapped out)
- **Dirty Bit:** Indicating whether the page has been modified since it was brought into memory
- **Reference Bit(Accessed Bit):** Indicating that a page has been accessed

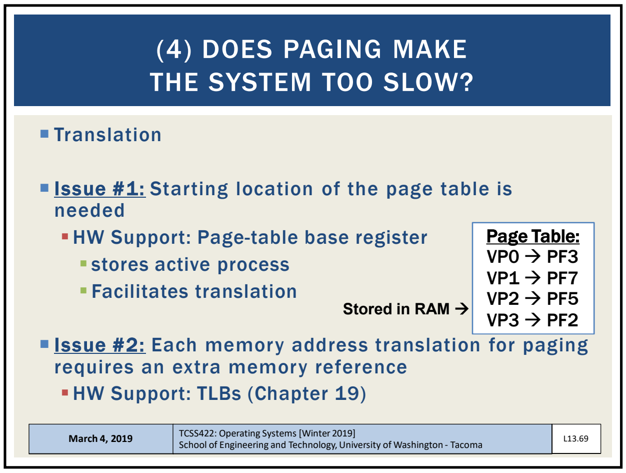
March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.67



(3) HOW BIG ARE PAGE TABLES?

- Page tables are too big to store on the CPU
- Page tables are stored using physical memory
- Paging supports efficiently storing a sparsely populated address space
  - Reduced memory requirement
    Compared to base and bounds, and segments

March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.68



(4) DOES PAGING MAKE THE SYSTEM TOO SLOW?

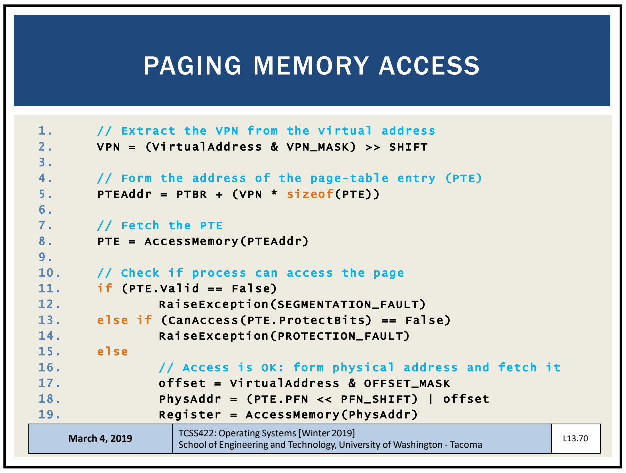
- Translation
- **Issue #1:** Starting location of the page table is needed
  - HW Support: Page-table base register
    - stores active process
    - Facilitates translation

Stored in RAM →

Page Table:
VP0 → PF3
VP1 → PF7
VP2 → PF5
VP3 → PF2

- **Issue #2:** Each memory address translation for paging requires an extra memory reference
  - HW Support: TLBs (Chapter 19)

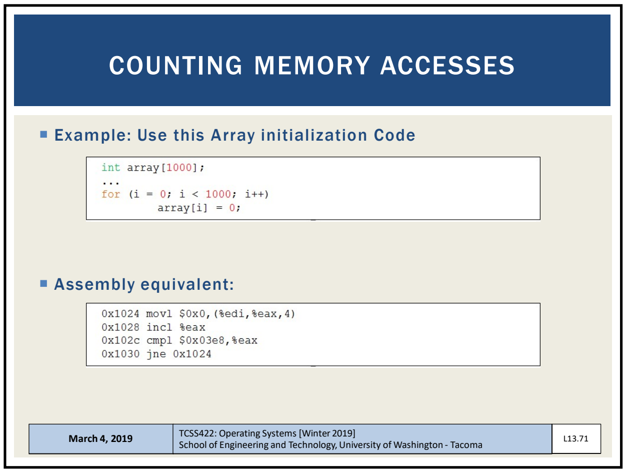
March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.69



PAGING MEMORY ACCESS

```
// Extract the VPN from the virtual address
VPN = (VirtualAddress & VPN_MASK) >> SHIFT

// Form the address of the page-table entry (PTE)
PTEAddr = PTBR + (VPN * sizeof(PTE))

// Fetch the PTE
PTE = AccessMemory(PTEAddr)

// Check if process can access the page
if (PTE.Valid == False)
        RaiseException(SEGMENTATION_FAULT)
else if (CanAccess(PTE.ProtectBits) == False)
        RaiseException(PROTECTION_FAULT)
else
        // Access is OK: form physical address and fetch it
        offset = VirtualAddress & OFFSET_MASK
        PhysAddr = (PTE.PFN << PFN_SHIFT) | offset
        Register = AccessMemory(PhysAddr)
```

March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.70



COUNTING MEMORY ACCESSES

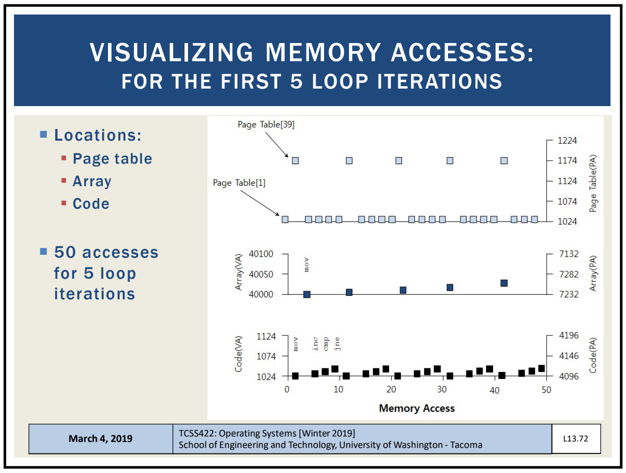
- Example: Use this Array initialization Code

```
int array[1000];
...
for (i = 0; i < 1000; i++)
        array[i] = 0;
```

- Assembly equivalent:

```
0x1024 movl $0x0,(%edi,%eax,4)
0x1028 incl %eax
0x102c cmpl $0x03e8,%eax
0x1030 jne 0x1024
```

March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.71



VISUALIZING MEMORY ACCESSES: FOR THE FIRST 5 LOOP ITERATIONS

- Locations:
  - Page table
  - Array
  - Code
- 50 accesses for 5 loop iterations

March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.72

PAGING SYSTEM EXAMPLE

- Consider a 4GB Computer:
- With a 4096-byte page size (4KB)
- How many pages would fit in physical memory?

- Now consider a page table:
- For the page table entry, how many bits are required for the VPN?
- If we assume the use of 4-byte (32 bit) page table entries, how many bits are available for status bits?
- How much space does this page table require? Page Table Entries x Number of pages
- How many page tables (for user processes) would fill the entire 4GB of memory?

CHAPTER 19: TRANSLATION LOOKASIDE BUFFER (TLB)

OBJECTIVES

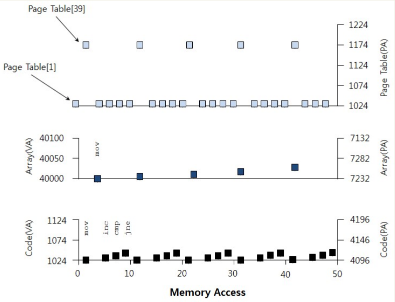
- Chapter 19
  - TLB Algorithm
  - TLB Tradeoffs
  - TLB Context Switch

TRANSLATION LOOKASIDE BUFFER

- Legacy name…
- Better name, “Address Translation Cache”
- TLB is an on CPU cache of address translations
  - virtual → physical memory

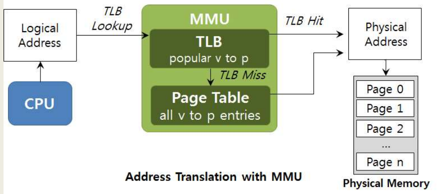
TRANSLATION LOOKASIDE BUFFER - 2

- Goal: Reduce access to the page tables
- Example: 50 RAM accesses for first 5 for-loop iterations
- Move lookups from RAM to TLB by caching page table entries

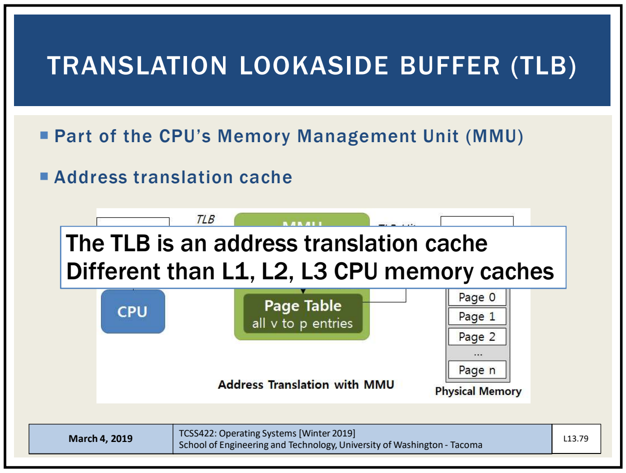


TRANSLATION LOOKASIDE BUFFER (TLB)

- Part of the CPU’s Memory Management Unit (MMU)
- Address translation cache



Address Translation with MMU



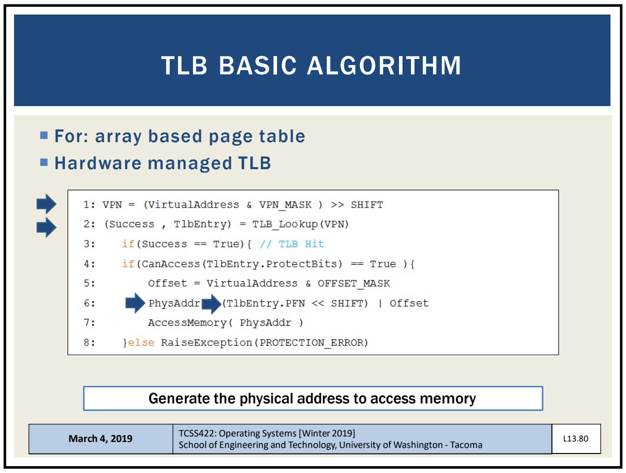
TRANSLATION LOOKASIDE BUFFER (TLB)

- Part of the CPU's Memory Management Unit (MMU)
- Address translation cache

TLB
MMU
CPU
Page Table
all v to p entries
Page 0
Page 1
Page 2
...
Page n
Physical Memory

The TLB is an address translation cache
Different than L1, L2, L3 CPU memory caches

Address Translation with MMU

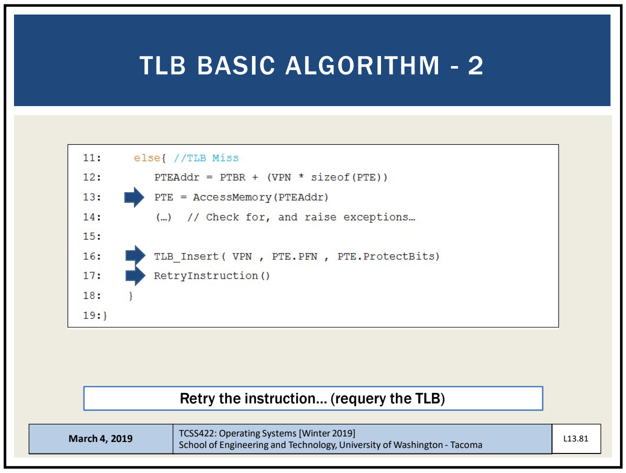


TLB BASIC ALGORITHM

- For: array based page table
- Hardware managed TLB

```
VPN = (VirtualAddress & VPN_MASK ) >> SHIFT
(Success , TlbEntry) = TLB_Lookup(VPN)
   if(Success == True){ // TLB Hit
   if(CanAccess(TlbEntry.ProtectBits) == True ){
       Offset = VirtualAddress & OFFSET_MASK
       PhysAddr = (TlbEntry.PFN << SHIFT) | Offset
       AccessMemory( PhysAddr )
   }else RaiseException(PROTECTION_ERROR)
```

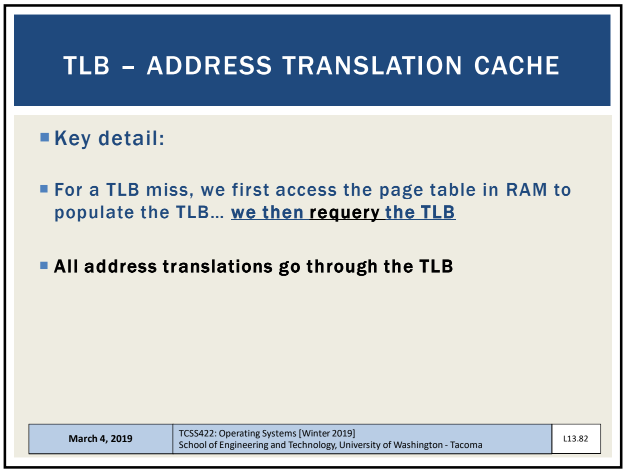
Generate the physical address to access memory



TLB BASIC ALGORITHM - 2

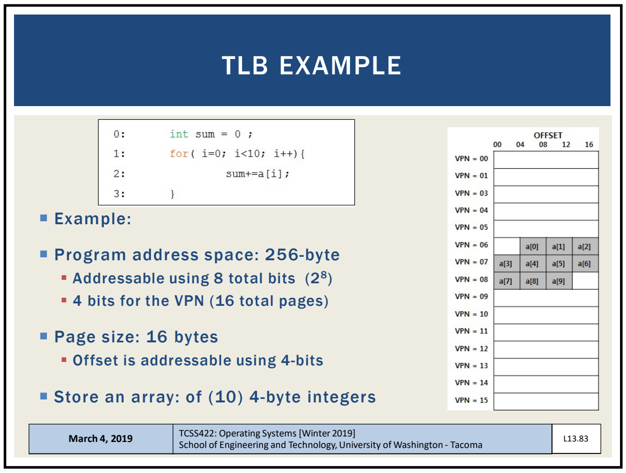
```
    else{ //TLB Miss
        PTEAddr = PTBR + (VPN * sizeof(PTE))
        PTE = AccessMemory(PTEAddr)
        (…)  // Check for, and raise exceptions…

        TLB_Insert( VPN , PTE.PFN , PTE.ProtectBits)
        RetryInstruction()
    }
}
```

Retry the instruction… (requery the TLB)



TLB – ADDRESS TRANSLATION CACHE

- Key detail:
- For a TLB miss, we first access the page table in RAM to populate the TLB… **we then requery the TLB**
- **All address translations go through the TLB**



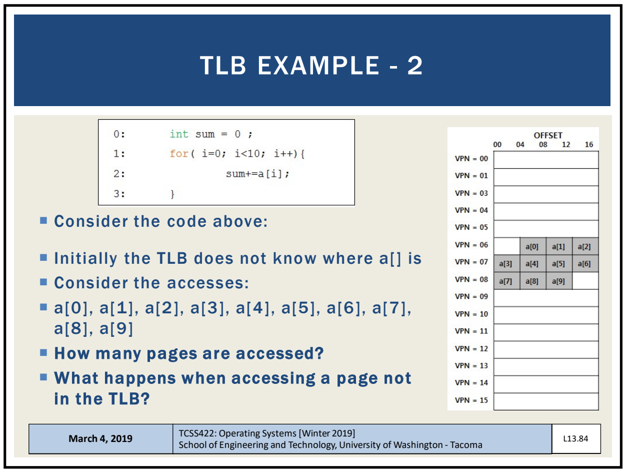
TLB EXAMPLE

```
    int sum = 0 ;
    for( i=0; i<10; i++){
            sum+=a[i];
    }
```

- Example:
- Program address space: 256-byte
  - Addressable using 8 total bits (2^8)
  - 4 bits for the VPN (16 total pages)
- Page size: 16 bytes
  - Offset is addressable using 4-bits
- Store an array: of (10) 4-byte integers

OFFSET

	00	04	08	12	16
VPN = 00					
VPN = 01					
VPN = 03					
VPN = 04					
VPN = 05					
VPN = 06		a[0]	a[1]	a[2]	
VPN = 07	a[3]	a[4]	a[5]	a[6]	
VPN = 08	a[7]	a[8]	a[9]		
VPN = 09					
VPN = 10					
VPN = 11					
VPN = 12					
VPN = 13					
VPN = 14					
VPN = 15					



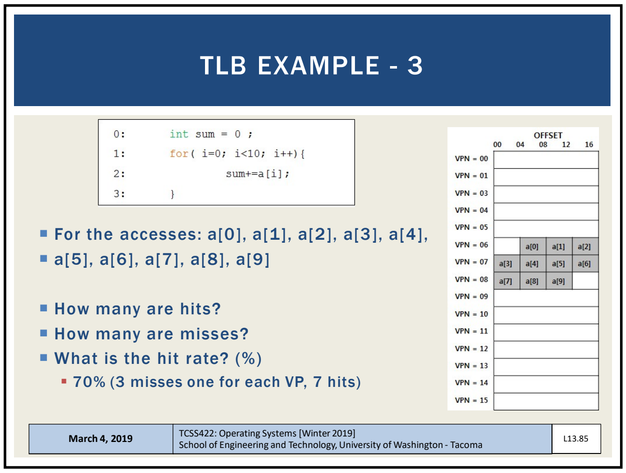
TLB EXAMPLE - 2

```
    int sum = 0 ;
    for( i=0; i<10; i++){
            sum+=a[i];
    }
```

- Consider the code above:
- Initially the TLB does not know where a[] is
- Consider the accesses:
- a[0], a[1], a[2], a[3], a[4], a[5], a[6], a[7], a[8], a[9]
- **How many pages are accessed?**
- **What happens when accessing a page not in the TLB?**

OFFSET

	00	04	08	12	16
VPN = 00					
VPN = 01					
VPN = 03					
VPN = 04					
VPN = 05					
VPN = 06		a[0]	a[1]	a[2]	
VPN = 07	a[3]	a[4]	a[5]	a[6]	
VPN = 08	a[7]	a[8]	a[9]		
VPN = 09					
VPN = 10					
VPN = 11					
VPN = 12					
VPN = 13					
VPN = 14					
VPN = 15					



TLB EXAMPLE - 3

```
0:      int sum = 0 ;
1:      for( i=0; i<10; i++){
2:              sum+=a[i];
3:      }
```

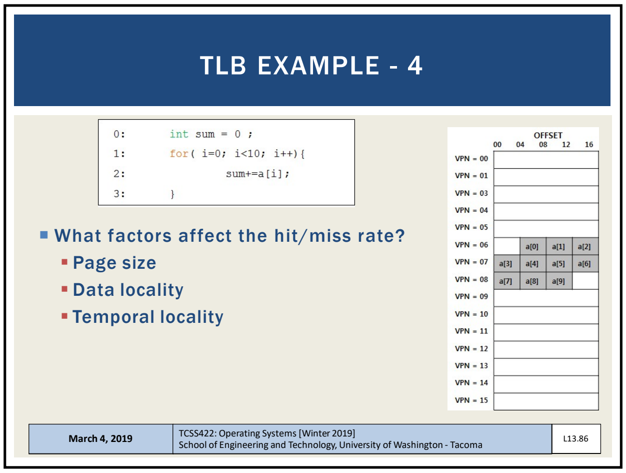
- For the accesses: a[0], a[1], a[2], a[3], a[4],
- a[5], a[6], a[7], a[8], a[9]

- How many are hits?
- How many are misses?
- What is the hit rate? (%)
  - 70% (3 misses one for each VP, 7 hits)

OFFSET

	00	04	08	12	16
VPN = 00					
VPN = 01					
VPN = 03					
VPN = 04					
VPN = 05					
VPN = 06		a[0]	a[1]	a[2]	
VPN = 07	a[3]	a[4]	a[5]	a[6]	
VPN = 08	a[7]	a[8]	a[9]		
VPN = 09					
VPN = 10					
VPN = 11					
VPN = 12					
VPN = 13					
VPN = 14					
VPN = 15					

March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.85



TLB EXAMPLE - 4

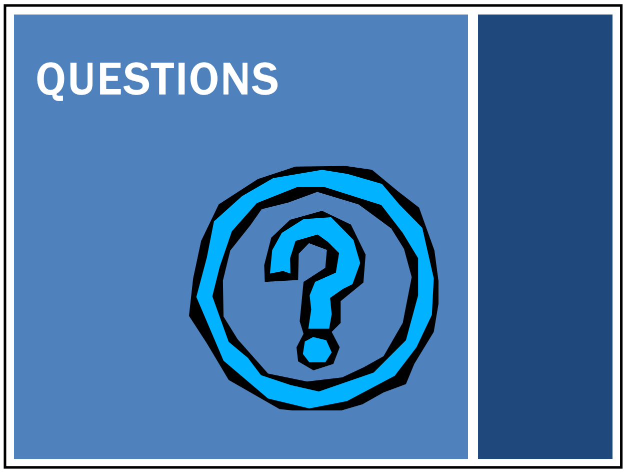
```
0:      int sum = 0 ;
1:      for( i=0; i<10; i++){
2:              sum+=a[i];
3:      }
```

- What factors affect the hit/miss rate?
  - Page size
  - Data locality
  - Temporal locality

OFFSET

	00	04	08	12	16
VPN = 00					
VPN = 01					
VPN = 03					
VPN = 04					
VPN = 05					
VPN = 06		a[0]	a[1]	a[2]	
VPN = 07	a[3]	a[4]	a[5]	a[6]	
VPN = 08	a[7]	a[8]	a[9]		
VPN = 09					
VPN = 10					
VPN = 11					
VPN = 12					
VPN = 13					
VPN = 14					
VPN = 15					

March 4, 2019 | TCSS422: Operating Systems [Winter 2019]
School of Engineering and Technology, University of Washington - Tacoma | L13.86



QUESTIONS