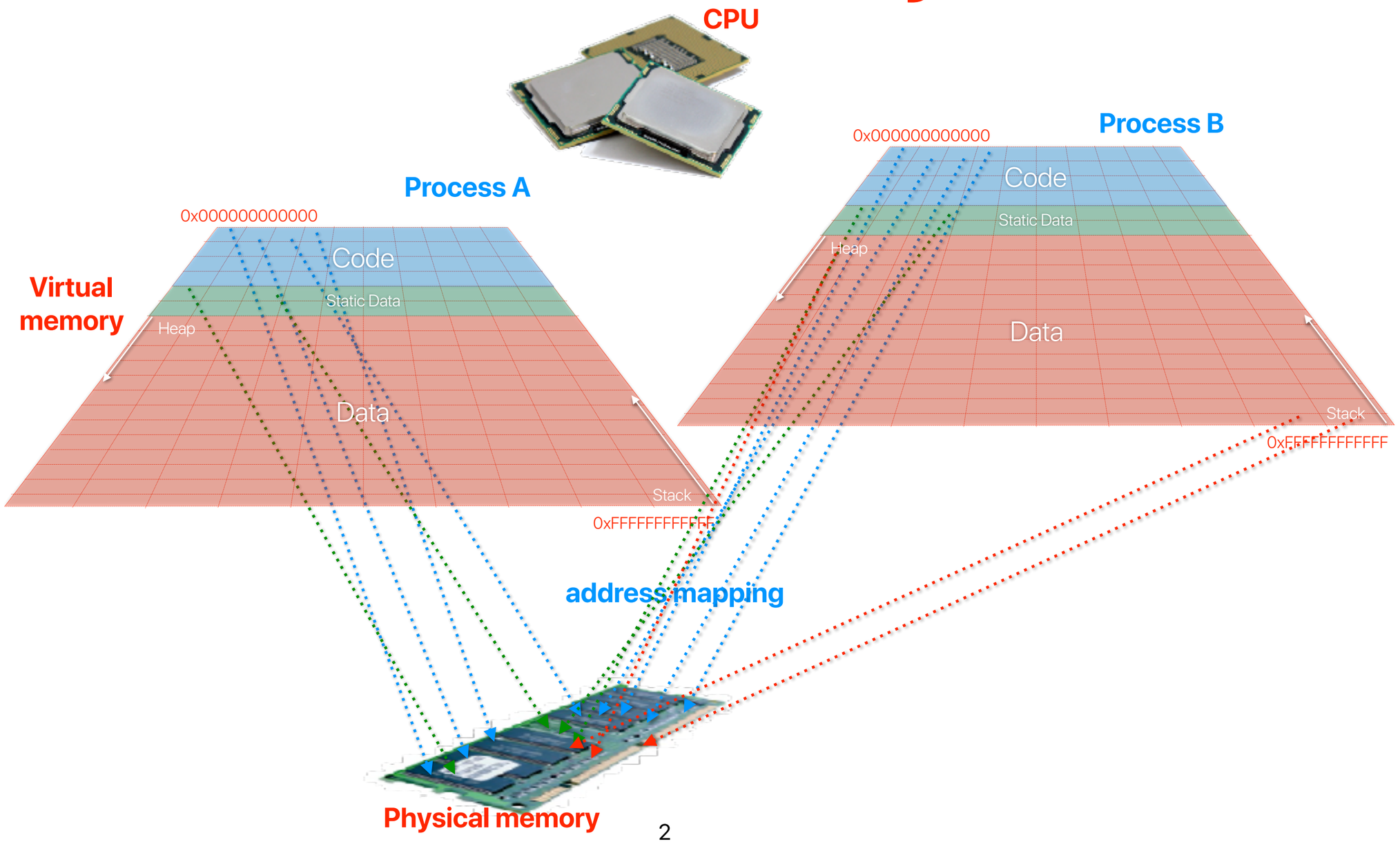
Virtual memory design in operating systems

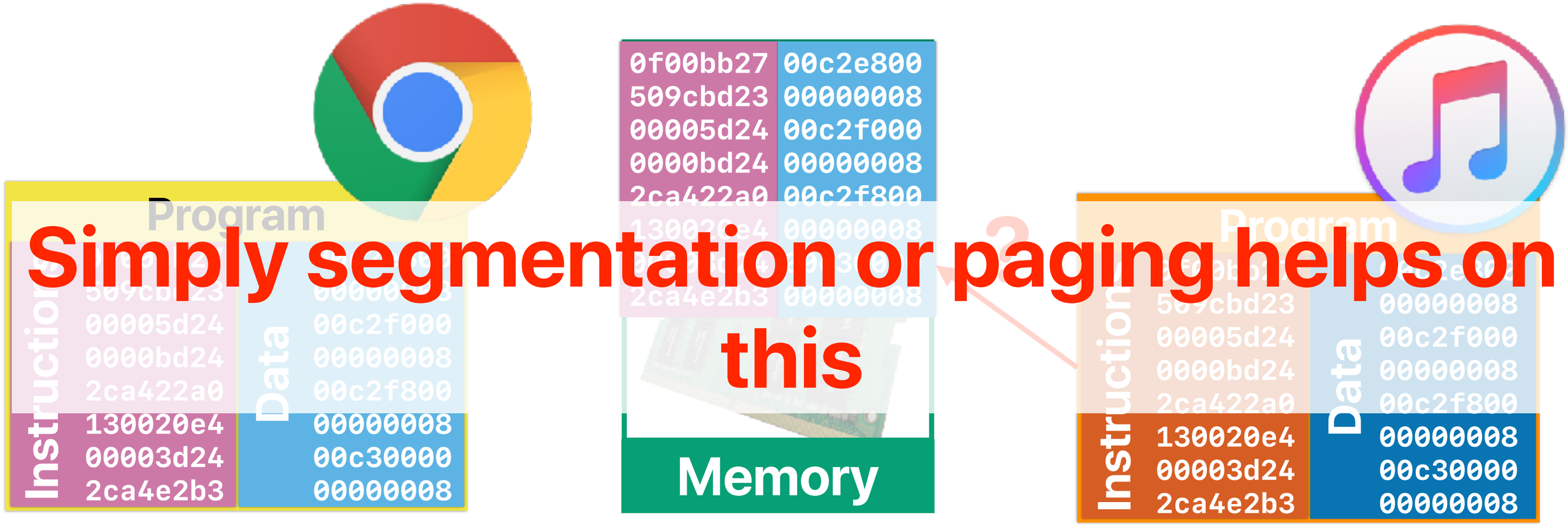
Hung-Wei Tseng

Virtual Memory



Recap: If we expose memory directly to the processor

What if both programs need to use memory?



Simply segmentation or paging helps on this

Recap: If we expose memory directly to the processor (I)

00c2f800
00000008
00c30000
00000008

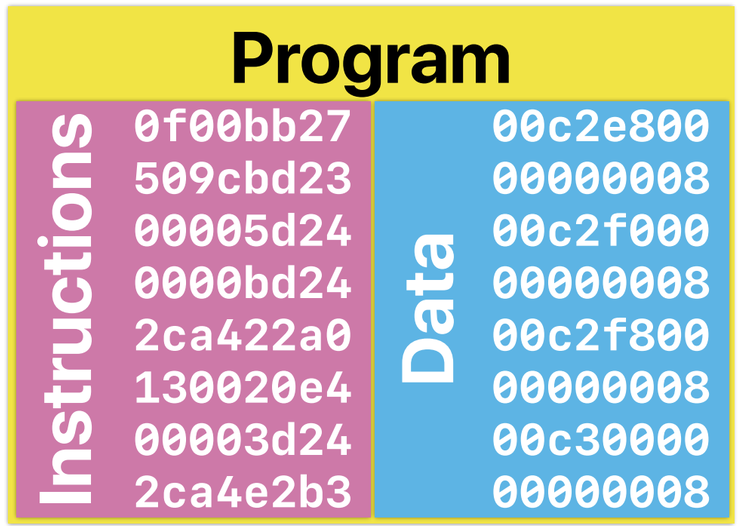
?

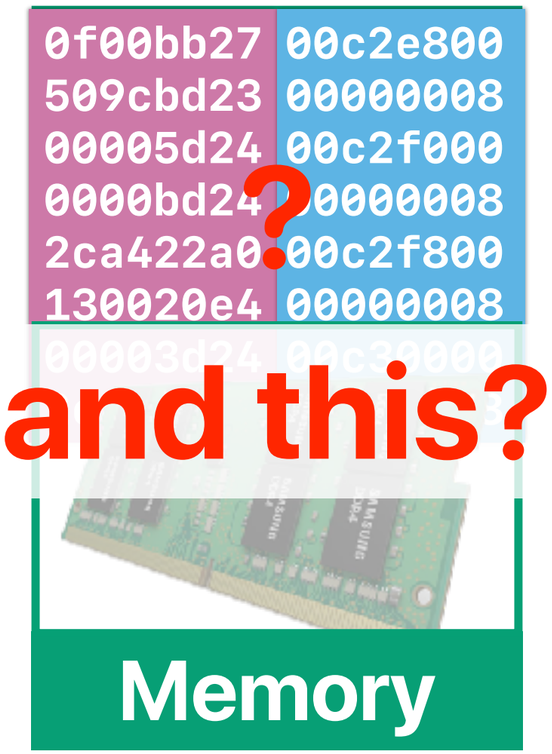
What if my program needs more memory?

But how about this?

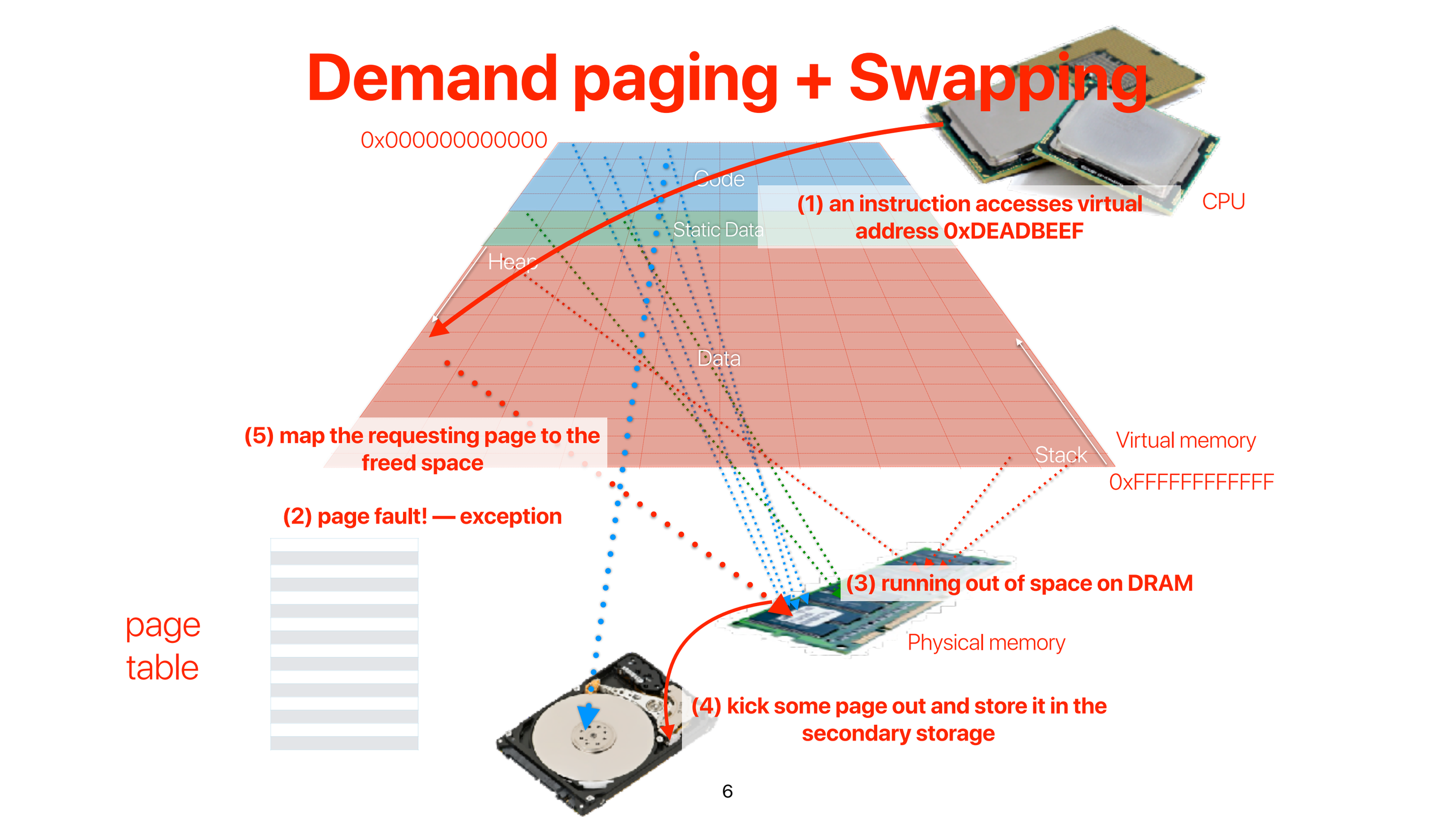
Recap: If we expose memory directly to the processor (II)

What if my program runs on a machine with a different memory size?





Demand paging + Swapping



The mechanism: demand paging + swapping

- Divide physical & virtual memory spaces into fix-sized units — pages
- Allocate a physical memory page whenever the virtual memory page containing your data is absent
- In case if we are running out of physical memory —
  - Reserve space on disks
    - Disks are slow: the access time for HDDs is around 10 ms, the access time for SSDs is around 30us - 1 ms
    - Disks are orders of magnitude larger than main memory
  - When you need to make rooms in the physical main memory, allocate a page in the swap space and put the content of the evicted page there
  - When you need to reference a page in the swap space, make a room in the physical main memory and swap the disk space with the evicted page

Latency Numbers Every Programmer Should Know

Operations	Latency (ns)	Latency (us)	Latency (ms)	
L1 cache reference	0.5 ns			~ 1 CPU cycle
Branch mispredict	5 ns			
L2 cache reference	7 ns			14x L1 cache
Mutex lock/unlock	25 ns			
Main memory reference	100 ns			20x L2 cache, 200x L1 cache
Compress 1K bytes with Zippy	3,000 ns	3 us		
Send 1K bytes over 1 Gbps network	10,000 ns	10 us		
Read 4K randomly from SSD*	150,000 ns	150 us		~1GB/sec SSD
Read 1 MB sequentially from memory	250,000 ns	250 us		
Round trip within same datacenter	500,000 ns	500 us		
Read 1 MB sequentially from SSD*	1,000,000 ns	1,000 us	1 ms	~1GB/sec SSD, 4X memory
Disk seek	10,000,000 ns	10,000 us	10 ms	20x datacenter roundtrip
Read 1 MB sequentially from disk	20,000,000 ns	20,000 us	20 ms	80x memory, 20X SSD
Send packet CA-Netherlands-CA	150,000,000 ns	150,000 us	150 ms	

Page replacement policy

- Goal: Identify page to remove that will avoid future page faults (i.e. utilize locality as much as possible)
- Implementation Goal: Minimize the amount of software and hardware overhead
  - Example:
    - Memory (i.e. RAM) access time: 100ns
    - Disk access time: 10ms
    - P_f: probability of a page fault
    - Effective Access Time = $10^{-7} + P_f * 10^{-3}$
  - When $P_f = 0.001$:
    Effective Access Time = 10,100ns
  - Takeaway: Disk access tolerable only when it is extremely rare

Outline

- VAX/VMS Design
- Mach VM

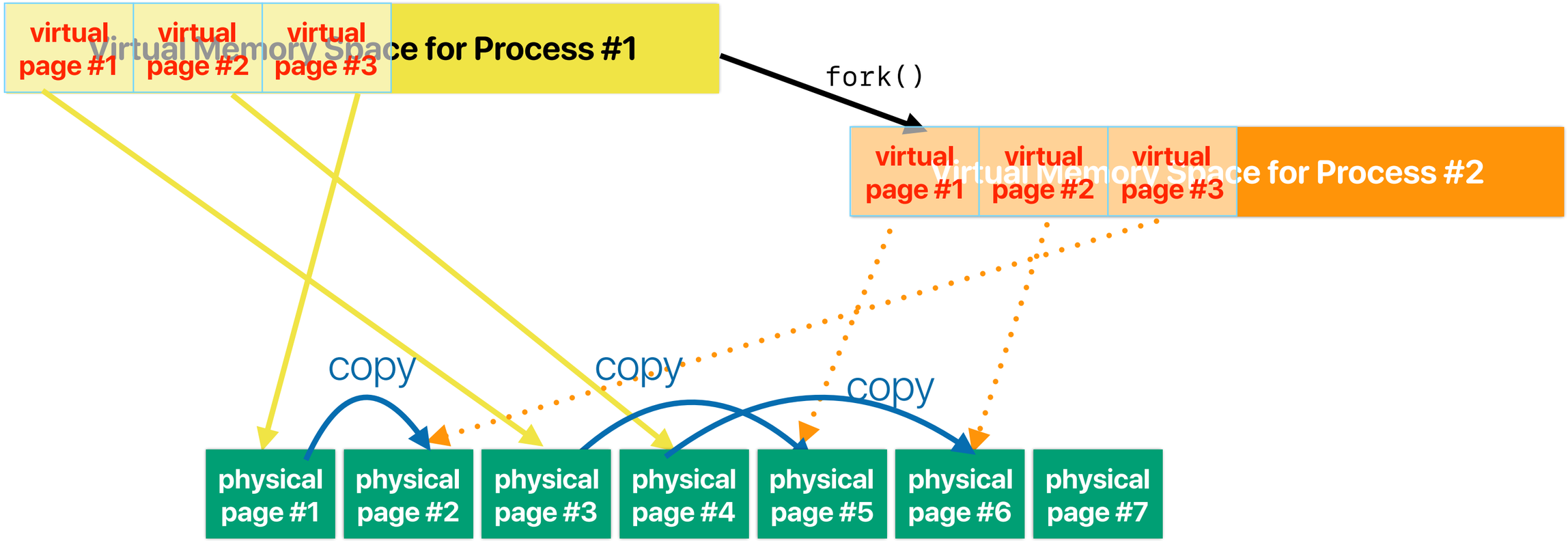
Virtual Memory Management in the VAX/VMS Operating System

H. M. Levy and P. H. Lipman
Digital Equipment Corporation

The “Why” behind VAX/VMS VM

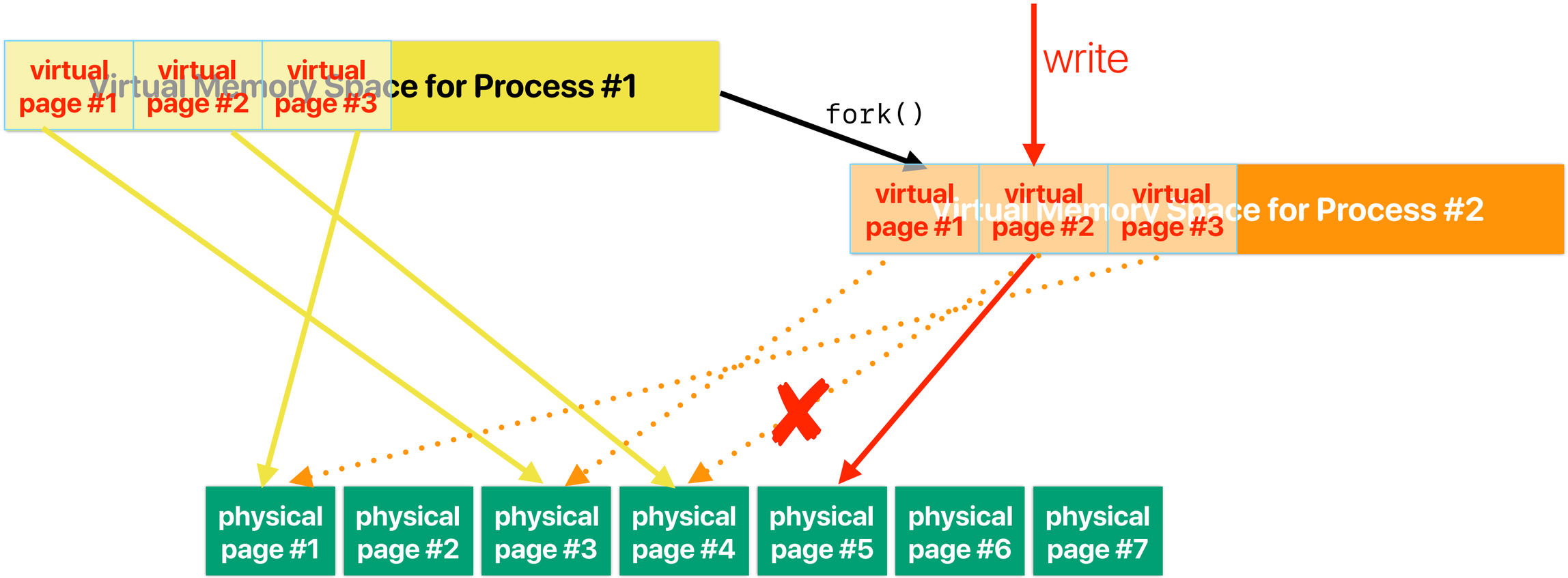
- The system needs to execute various types of applications efficiently
- The system runs on different types of hardware
- As a result, the memory management system has to be capable of adjusting the changing demands characteristic of time sharing while allowing predictable performance required by real-time and batch processes

What happens on a `fork`?



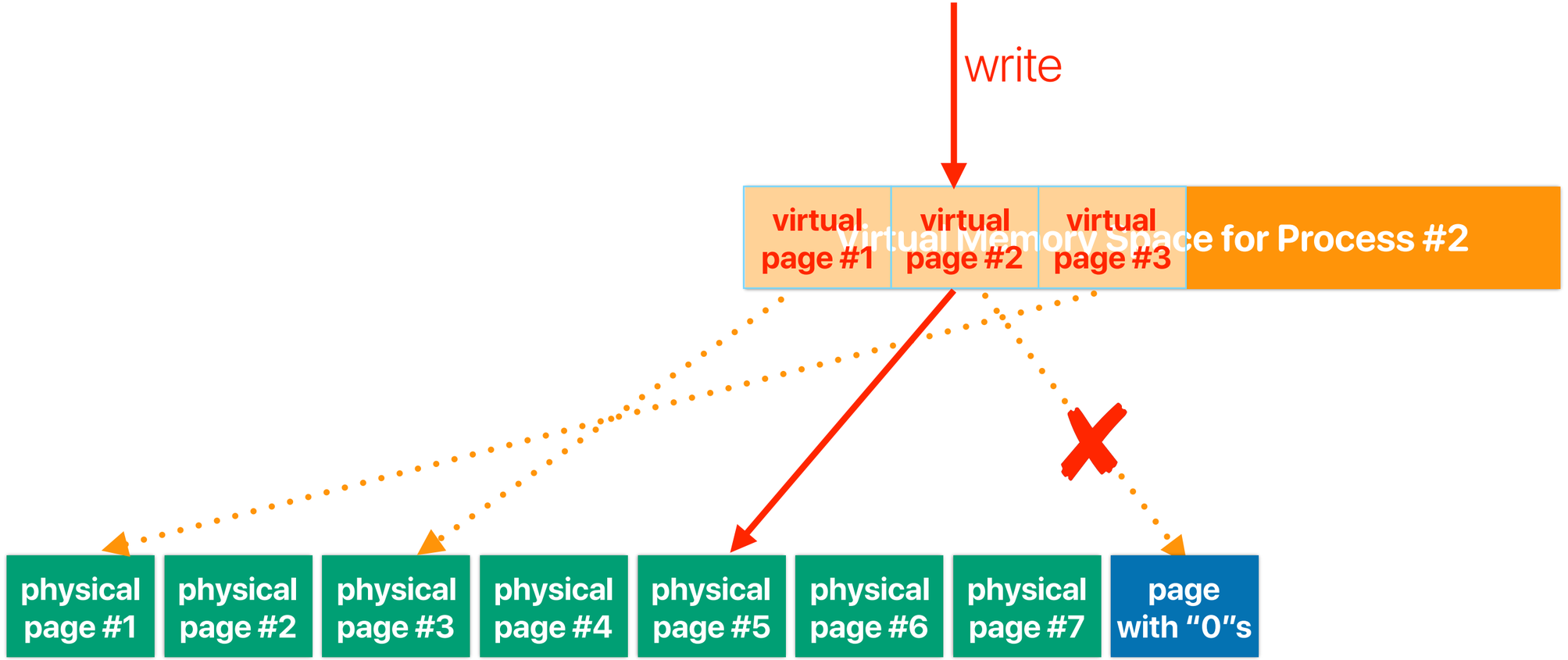
- **Copy the page content to different locations before the new process can start**

Copy-on-write



- The modified bit of a writable page will be set when it's loaded from the executable file
- The process eventually will have its own copy of that page

Demand zero



- The linker does not embed the pages with all 0s in the compiled program
- When page fault occurs, allocate a physical page fills with zeros
- Set the modified bit so that the page can be written back

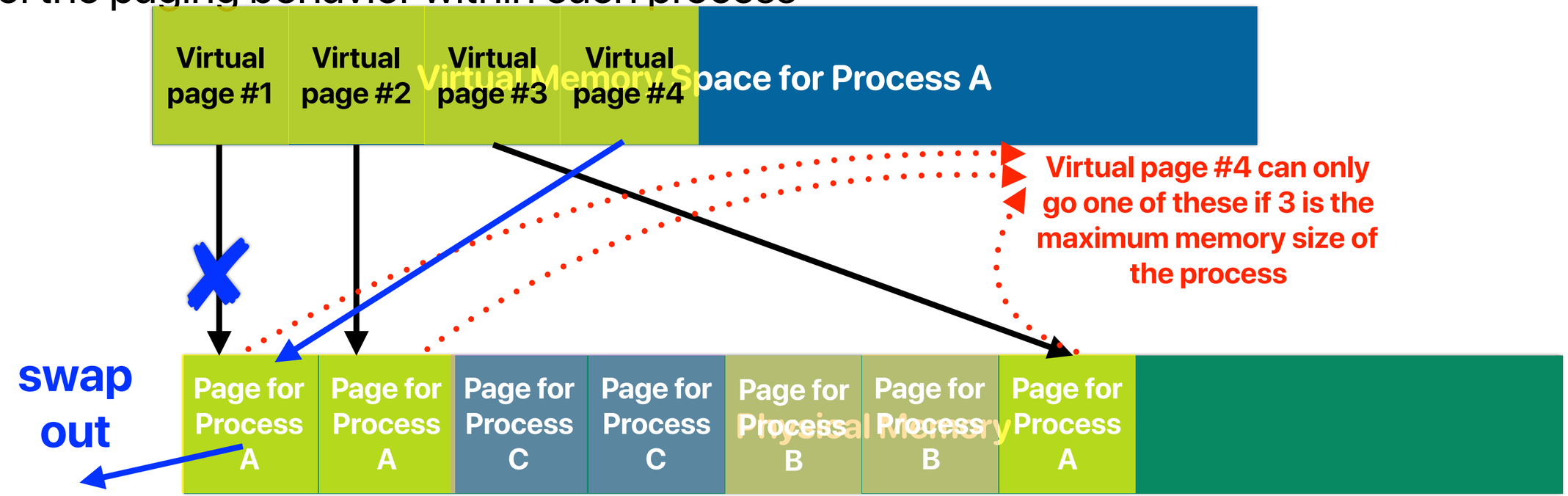
What VAX/VMS proposed to achieve these goals?

- Considering the optimization goals and the proposed VAX/VMS mechanisms, which of the following combinations is incorrect?

Goal		Optimization	
A	Process startup cost	W	Demand-zero & copy-on-refernce
B	Process performance interference	X	Process-local replacement
C	Page table lookup overhead	Y	Page clustering
D	Paging load on disks	Z	Page caching

Local page replacement policy

- Each process has a maximum size of memory
- When the process exceeds the maximum size, replaces from its own set of memory pages
- Control the paging behavior within each process



What's the policy? **FIFO!** **Low overhead!**

What VAX/VMS proposed to achieve these goals?

- Considering the optimization goals and the proposed VAX/VMS mechanisms, which of the following combinations is incorrect?

Goal		Optimization	
A	Process startup cost	W	Demand-zero & copy-on-refernce
B	Process performance interference	X	Process-local replacement
C	Page table lookup overhead	Y	Page clustering
D	Paging load on disks	Z	Page caching

Page clustering

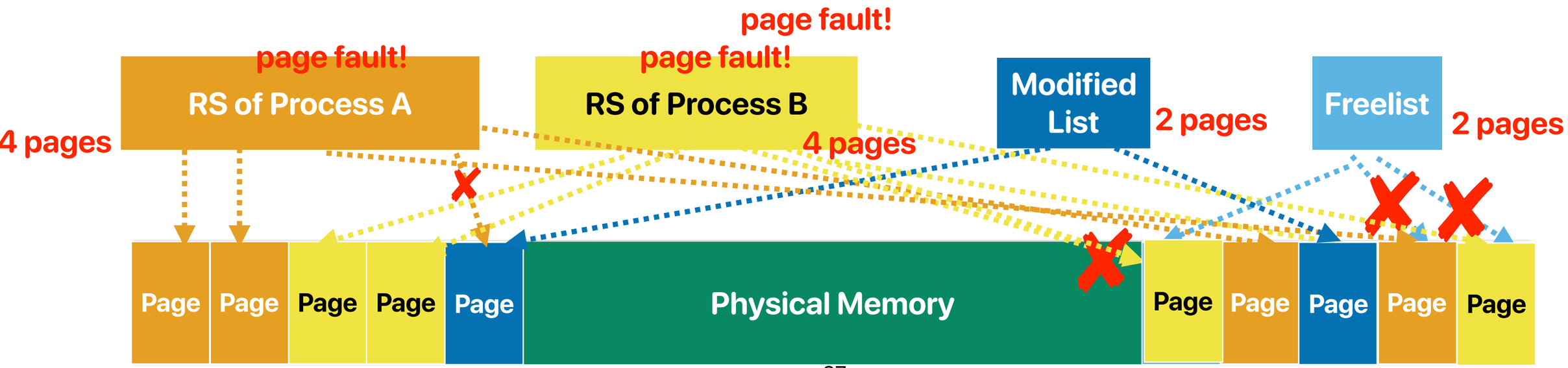
- Read or write a cluster of pages that are both consecutive in virtual memory and the disk
- Combining consecutive writes into single writes

Latency Numbers Every Programmer Should Know

Operations	Latency (ns)	Latency (us)	Latency (ms)	
L1 cache reference	0.5 ns			~ 1 CPU cycle
Branch mispredict	5 ns			
L2 cache reference	7 ns			14x L1 cache
Mutex lock/unlock	25 ns			
Main memory reference	100 ns			20x L2 cache, 200x L1 cach
Compress 1K bytes with Zippy	3,000 ns	3 us		
Send 1K bytes over 1 Gbps network	10,000 ns	10 us		
Read 4K randomly from SSD*	150,000 ns	150 us		~1GB/sec SSD
Read 1 MB sequentially from memory	250,000 ns	250 us		
Round trip within same datacenter	500,000 ns	500 us		
Read 1 MB sequentially from SSD*	1,000,000 ns	1,000 us	1 ms	~1GB/sec SSD, 4X memory
Disk seek for a 512B sector	10,000,000 ns	10,000 us	10 ms	20x datacenter roundtrip
Read 1 MB sequentially from disk	20,000,000 ns	20,000 us	20 ms	80x memory, 20X SSD
Send packet CA-Netherlands-CA	150,000,000 ns	150,000 us	150 ms	

Page caching to cover the performance loss

- Evicted pages will be put into one of the lists in DRAM
  - Free list: clean pages
  - Modified list: dirty pages — needs to copy data to the disk
- Page fault to any of the page in the lists will bring the page back
  - Reduces the demand of accessing disks



Page caching

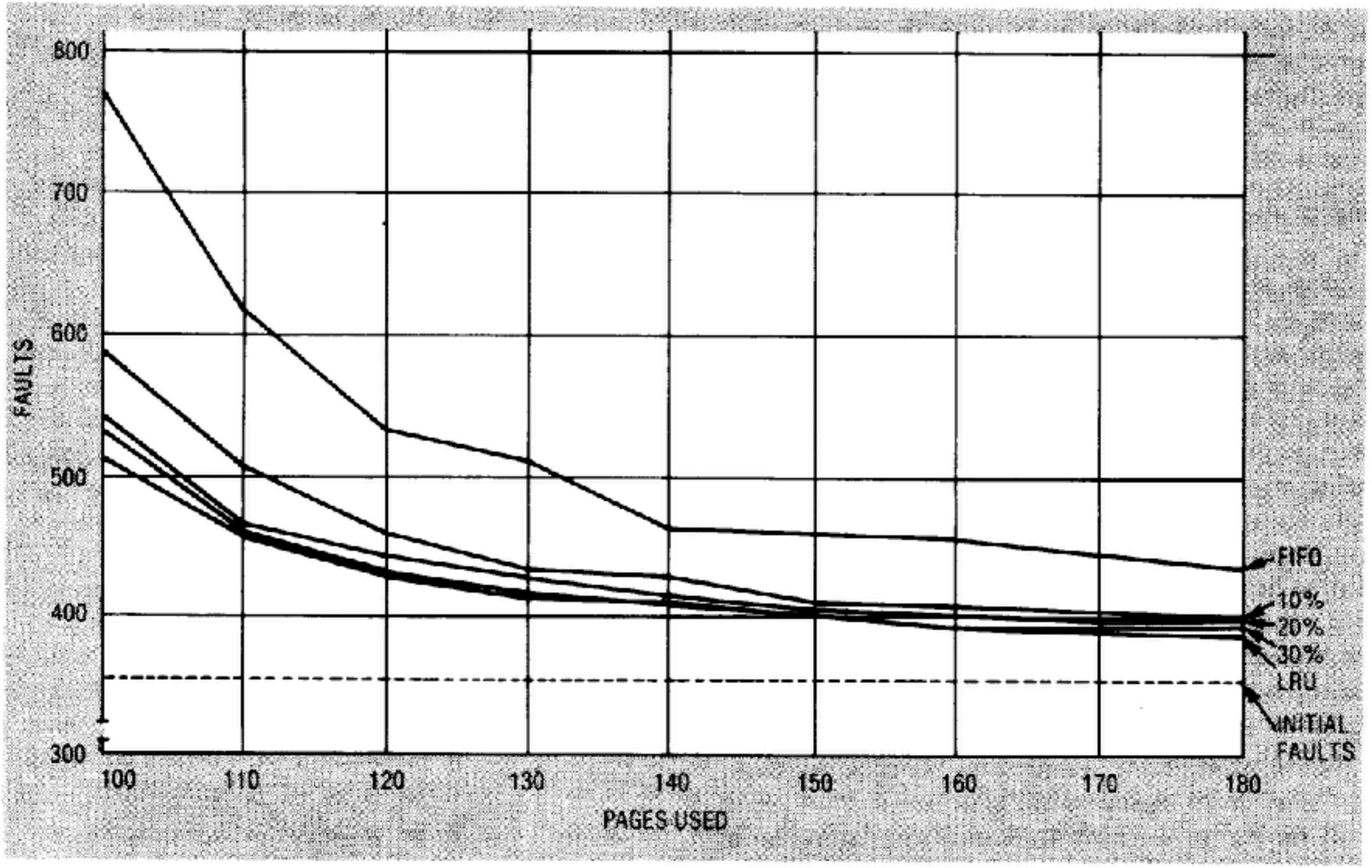
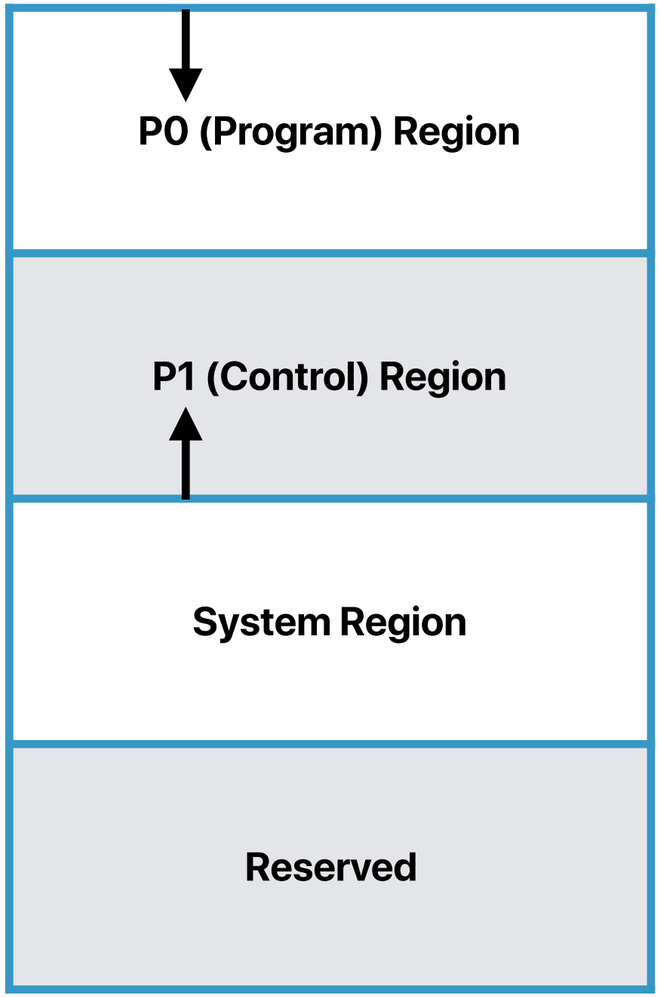


Figure 3. Faults vs. memory usage in Fortran compilation.

Process memory layout



Code
Heap

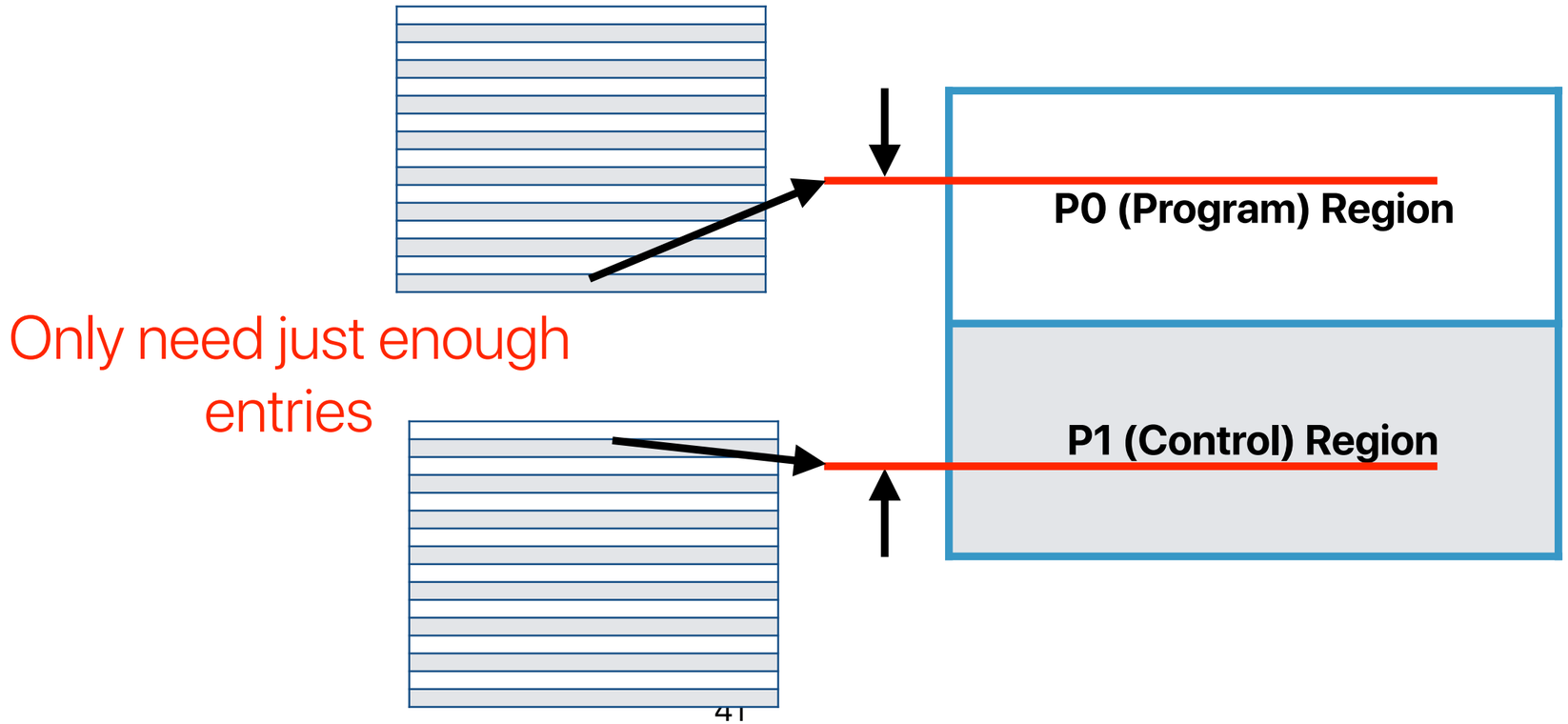
Stack
Other data

System: software vectors, hardware data structures, executive data, executive procedures, record management, dynamic storage

The VAX/VMS allows the OS code to access user-space memory

Why segmented layout?

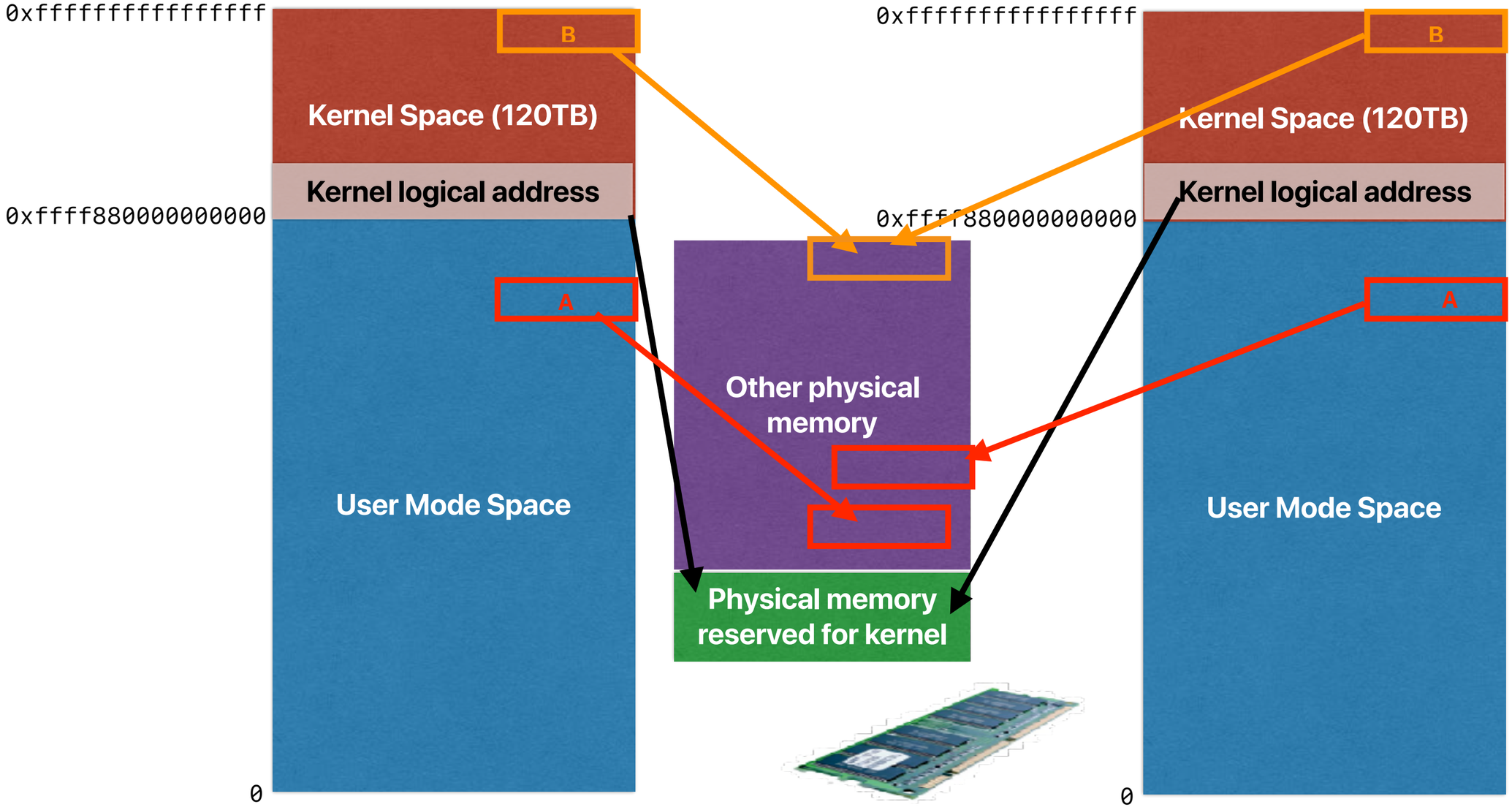
- Each segment has its own page table
- Entries between stack and heap boundaries do not need to be allocated — reduce the size of page table



The impact of VAX/VMS

- VAX is popular in universities and UNIX is later ported to VAX — a popular OS research platform
- Affect the UNIX virtual memory design
- Affect the Windows virtual memory design

64-bit Linux process memory layout



Machine-Independent Virtual Memory Management for Paged Uniprocessor and Multiprocessor Architectures

Richard Rashid, Avadis Tevanian, Michael Young, David Golub, Robert Baron, David Black, William Bolosky, and Jonathan Chew
Carnegie-Mellon University, NeXT, University of Rochester

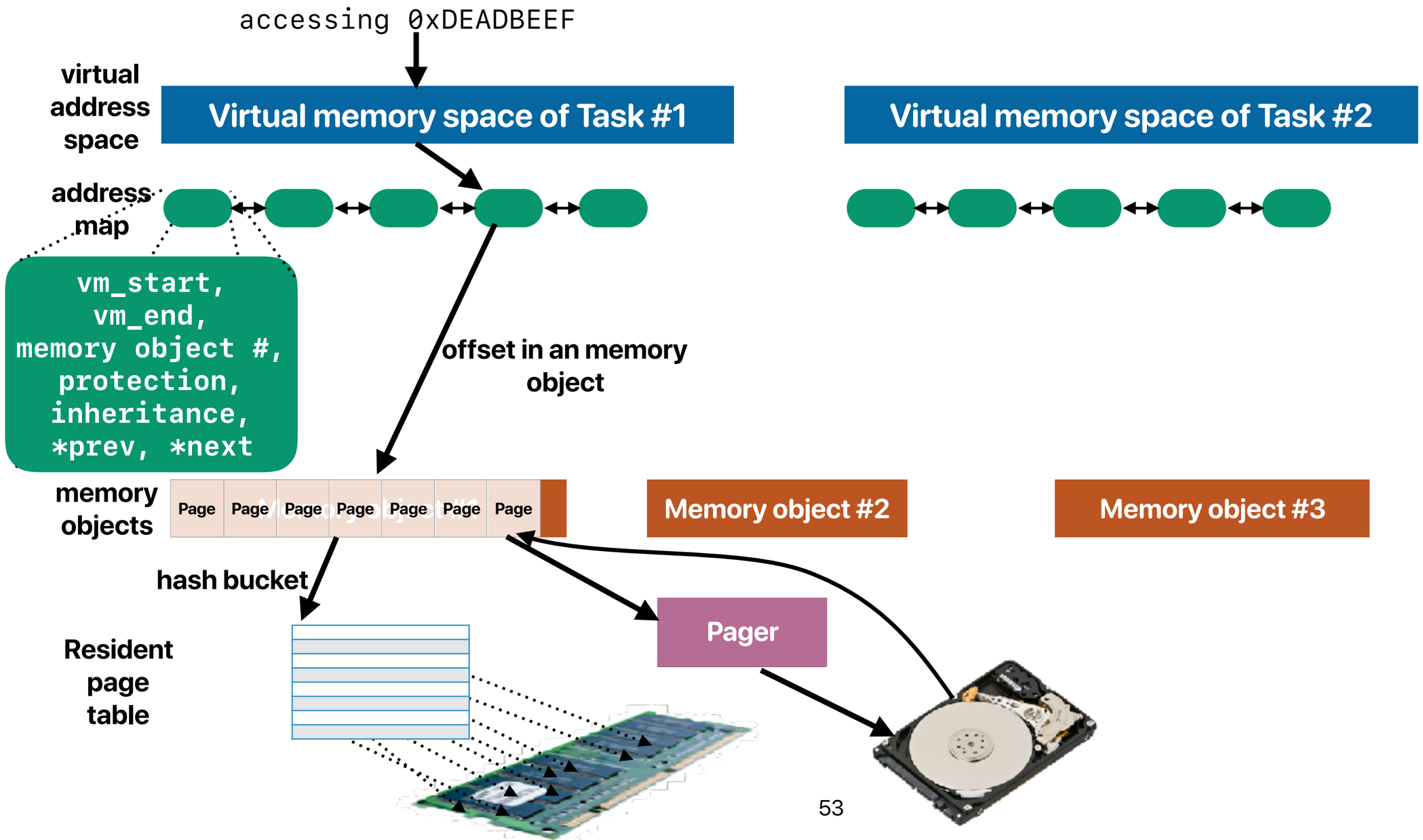
Mach abstractions

- Task: process in UNIX
- Thread: the basic scheduling identity
- Port: message queues protected by the kernel
- Message: data objects for inter-thread communication
- Memory object: data mapped into the address space of a task/process

What Mach VM proposed?

- Machine-independent virtual memory design by maintaining all VM state in a machine-independent module
- Treat hardware page tables/TLBs as caches of machine-independent information

Overview of Mach's VM



Where is pmap?

- Pmap is just a **cache** of virtual to physical address mapping
- It accelerates address translation by caching the address mapping, but not required
- As a result, it can be a small as several KBs

The impact of Mach VM

- MacOS X's virtual memory resembles the Mach VM design
  - Why?

Announcement

- Reading quiz due next Tuesday
- Project due 3/3
- Check your grades on iLearn
- Use office hours to discuss projects