Caches (continued)
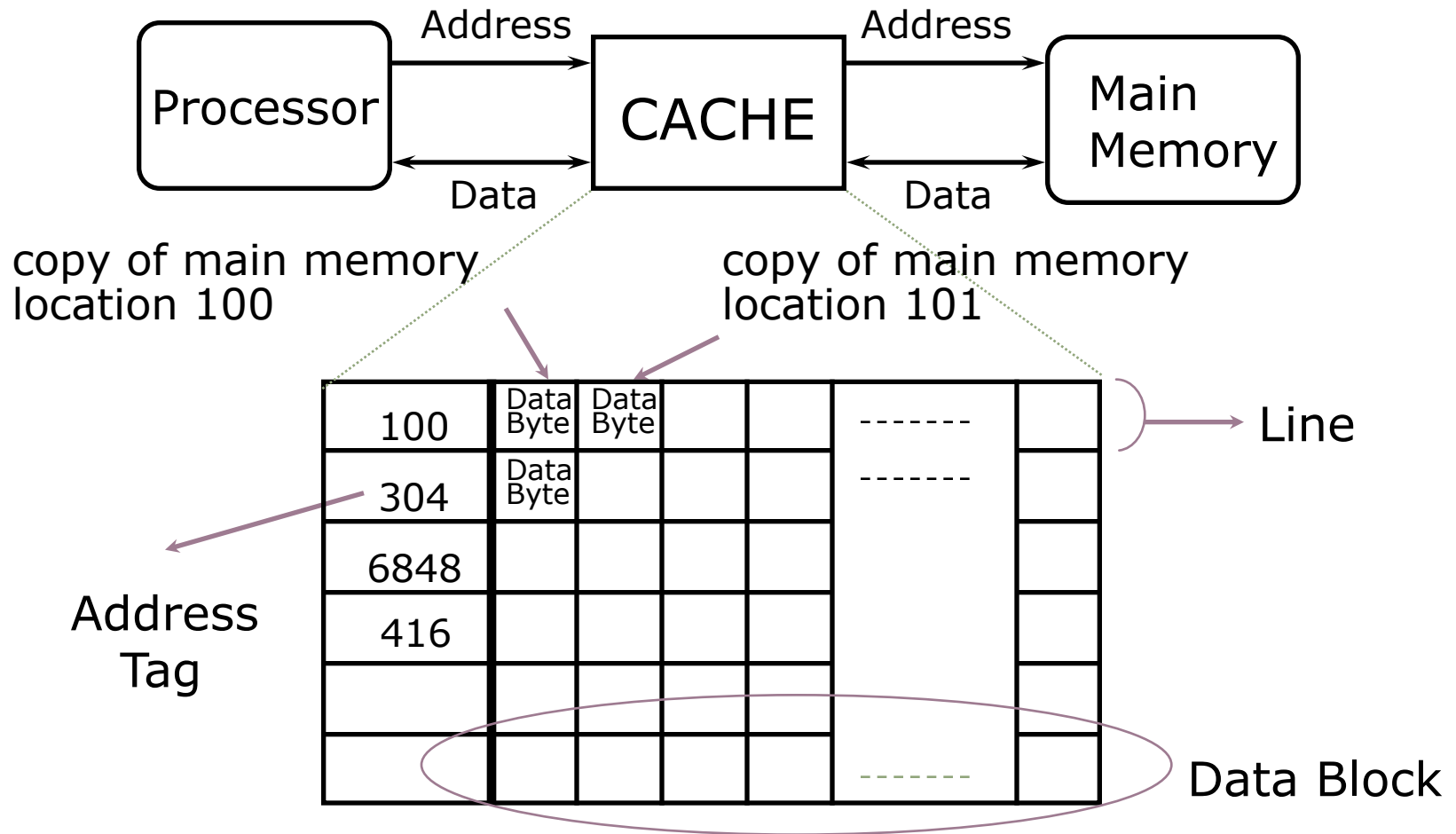
Mengjia Yan
Computer Science and Artificial Intelligence Laboratory
M.I.T.

Reminder: Inside a Cache

Processor → **Address** → CACHE → **Address** → Main Memory

Processor ← **Data** ← CACHE ← **Data** ← Main Memory

copy of main memory location 100

copy of main memory location 101

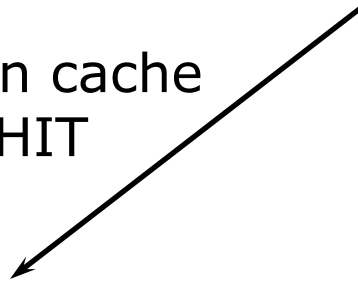| Address Tag	Data Block					
100	Data Byte	Data Byte			-------	
304	Data Byte				-------	
6848						
416						

→ Line

Data Block

Reminder: Cache Algorithm (Read)

Look at Processor Address, search cache tags to find match. Then either

Reminder: Cache Algorithm (Read)

Look at Processor Address, search cache tags to find match.
Then either

Found in cache
a.k.a. HIT

Return copy
of data from
cache

Reminder: Cache Algorithm (Read)

Look at Processor Address, search cache tags to find match. Then either

Found in cache
a.k.a. HIT

Not in cache
a.k.a. MISS

Return copy
of data from
cache

Read block of data from
Main Memory

Wait …

Return data to processor
and update cache

Reminder: Cache Algorithm (Read)

Look at Processor Address, search cache tags to find match. Then either

Found in cache
a.k.a. HIT

Not in cache
a.k.a. MISS

Return copy
of data from
cache

Read block of data from
Main Memory

Wait …

Return data to processor
and update cache

Which line do we replace?

Reminder: Cache Performance

Average memory access time (AMAT) =
Hit time + Miss rate x Miss penalty

Reminder: Cache Performance

Average memory access time (AMAT) =
            Hit time + Miss rate x Miss penalty

To improve performance:
- reduce the hit time
- reduce the miss rate (e.g., larger, better policy)
- reduce the miss penalty (e.g., L2 cache)

What is the simplest design strategy?

Reminder: Cache Performance

Average memory access time (AMAT) =
    Hit time + Miss rate x Miss penalty

To improve performance:
- reduce the hit time
- reduce the miss rate (e.g., larger, better policy)
- reduce the miss penalty (e.g., L2 cache)

What is the simplest design strategy?

Biggest cache that doesn't increase hit time past 1-2 cycles (approx. 16-64KB in modern technology)

[design issues more complex with out-of-order superscalar processors]

Causes for Cache Misses [Hill, 1989]

- *Compulsory:*

  First reference to a block *a.k.a.* cold start misses
  - misses that would occur even with infinite cache

- *Capacity:*

  cache is too small to hold all data the program needs
  - misses that would occur even under fully-associative placement & perfect replacement policy

- *Conflict:*

  misses from collisions due to block-placement strategy
  - misses that would not occur with full associativity

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache *
Compulsory misses			
Capacity misses			
Conflict misses			
Hit latency			
Miss latency			

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache *
Compulsory misses	═		
Capacity misses			
Conflict misses			
Hit latency			
Miss latency			

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	**Higher associativity cache**	**Larger block size cache ***
Compulsory misses	=		
Capacity misses	↓		
Conflict misses			
Hit latency			
Miss latency			

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache *
Compulsory misses	=		
Capacity misses	⬇		
Conflict misses	⬇		
Hit latency			
Miss latency			

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	**Higher associativity cache**	**Larger block size cache ***
Compulsory misses	=		
Capacity misses	⬇		
Conflict misses	⬇		
Hit latency	⬆		
Miss latency			

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	**Higher associativity cache**	**Larger block size cache ***
Compulsory misses	=		
Capacity misses	↓		
Conflict misses	↓		
Hit latency	↑		
Miss latency	=		

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	**Higher associativity cache**	**Larger block size cache ***
Compulsory misses	=	=	
Capacity misses	↓		
Conflict misses	↓		
Hit latency	↑		
Miss latency	=		

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache *
Compulsory misses	=	=	
Capacity misses	↓	=	
Conflict misses	↓		
Hit latency	↑		
Miss latency	=		

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache *
Compulsory misses	=	=	
Capacity misses	↓	=	
Conflict misses	↓	↓	
Hit latency	↑		
Miss latency	=		

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache *
Compulsory misses	=	=	
Capacity misses	↓	=	
Conflict misses	↓	↓	
Hit latency	↑	↑	
Miss latency	=		

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache *
Compulsory misses	=	=	
Capacity misses	↓	=	
Conflict misses	↓	↓	
Hit latency	↑	↑	
Miss latency	=	=	

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache *
Compulsory misses	=	=	⬇
Capacity misses	⬇	=	
Conflict misses	⬇	⬇	
Hit latency	⬆	⬆	
Miss latency	=	=	

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache *
Compulsory misses	=	=	⬇
Capacity misses	⬇	=	⬇
Conflict misses	⬇	⬇	
Hit latency	⬆	⬆	
Miss latency	=	=	

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache *
Compulsory misses	=	=	⬇
Capacity misses	⬇	=	⬇
Conflict misses	⬇	⬇	?
Hit latency	⬆	⬆	
Miss latency	=	=	

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache *
Compulsory misses	▬	▬	⬇
Capacity misses	⬇	▬	⬇
Conflict misses	⬇	⬇	?
Hit latency	⬆	⬆	▬
Miss latency	▬	▬	

* Assume substantial spatial locality

Effect of Cache Parameters on Performance

	Larger capacity cache	Higher associativity cache	Larger block size cache *
Compulsory misses	=	=	⬇
Capacity misses	⬇	=	⬇
Conflict misses	⬇	⬇	?
Hit latency	⬆	⬆	=
Miss latency	=	=	⬆⬇

* Assume substantial spatial locality

Replacement Policy

Which block from a set should be evicted?

Replacement Policy

Which block from a set should be evicted?

- Random

Replacement Policy

Which block from a set should be evicted?

- Random

Replacement Policy

Which block from a set should be evicted?

- Random

- Least Recently Used (LRU)
  - LRU cache state must be updated on every access
  - true implementation only feasible for small sets (2-way)
  - pseudo-LRU binary tree was often used for 4-8 way

Replacement Policy

Which block from a set should be evicted?

- Random

- Least Recently Used (LRU)
  - LRU cache state must be updated on every access
  - true implementation only feasible for small sets (2-way)
  - pseudo-LRU binary tree was often used for 4-8 way

- First In, First Out (FIFO) a.k.a. Round-Robin
  - used in highly associative caches

Replacement Policy

Which block from a set should be evicted?

- Random

- Least Recently Used (LRU)
  - LRU cache state must be updated on every access
  - true implementation only feasible for small sets (2-way)
  - pseudo-LRU binary tree was often used for 4-8 way

- First In, First Out (FIFO) a.k.a. Round-Robin
  - used in highly associative caches

- Not Least Recently Used (NLRU)
  - FIFO with exception for most recently used block or blocks

Replacement Policy

Which block from a set should be evicted?

- Random

- Least Recently Used (LRU)
    - LRU cache state must be updated on every access
    - true implementation only feasible for small sets (2-way)
    - pseudo-LRU binary tree was often used for 4-8 way

- First In, First Out (FIFO) a.k.a. Round-Robin
    - used in highly associative caches

- Not Least Recently Used (NLRU)
    - FIFO with exception for most recently used block or blocks

- One-bit LRU
    - Each way represented by a bit. Set on use, replace first unused.

Multiple replacement policies

Use the best replacement policy for a program

Cache

How do we decide
which policy to use?

Sets

Policy A

Policy B

Miss
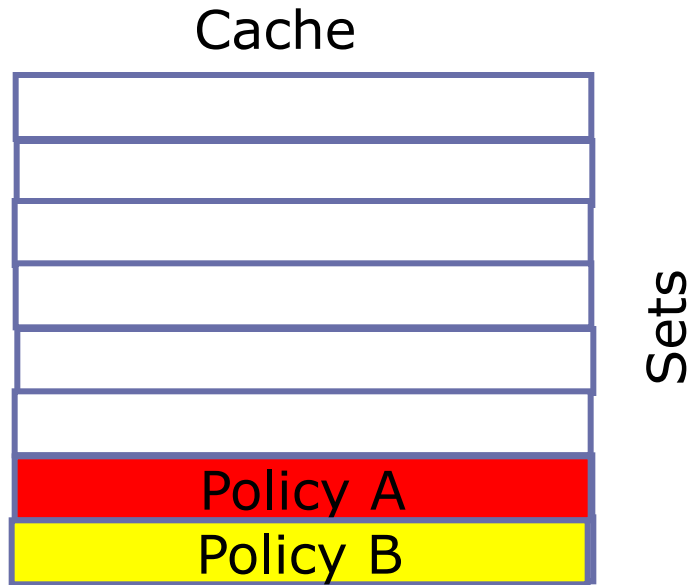
0: Policy A
1: Policy B

Multiple replacement policies
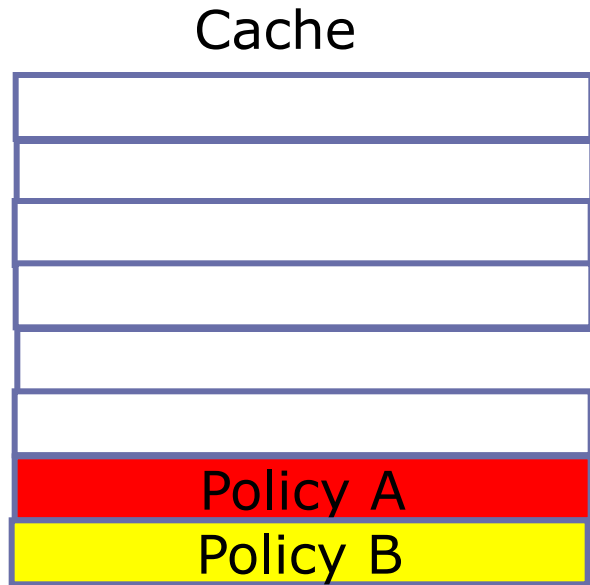
Use the best replacement policy for a program

Cache

How do we decide
which policy to use?

Sets

Policy A

Policy B

Miss

0: Policy A Missed
1: Policy B Missed

+1   -1

Counter

0: Policy A
1: Policy B

>0

Multilevel Caches

- A memory cannot be large and fast
- Add level of cache to reduce miss penalty
  - Each level can have longer latency than level above
  - So, increase sizes of cache at each level

```
CPU <---> L1 <---> L2 <---> DRAM
```

Multilevel Caches

- A memory cannot be large and fast
- Add level of cache to reduce miss penalty
  - Each level can have longer latency than level above
  - So, increase sizes of cache at each level

```
CPU  <->  L1  <->  L2  <->  DRAM
```
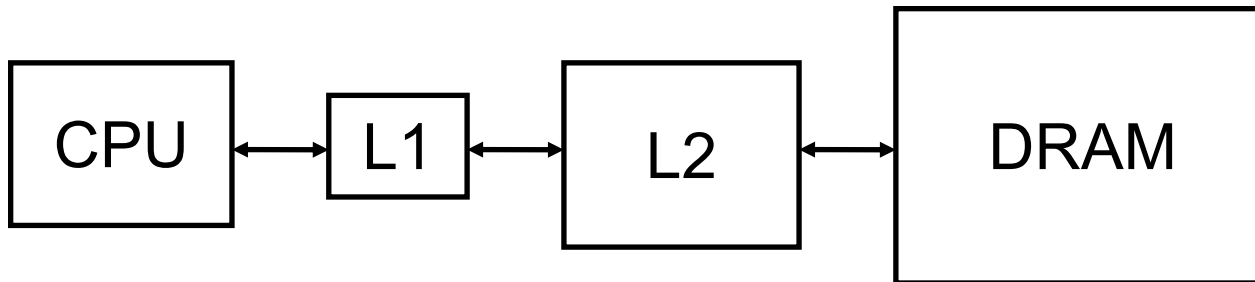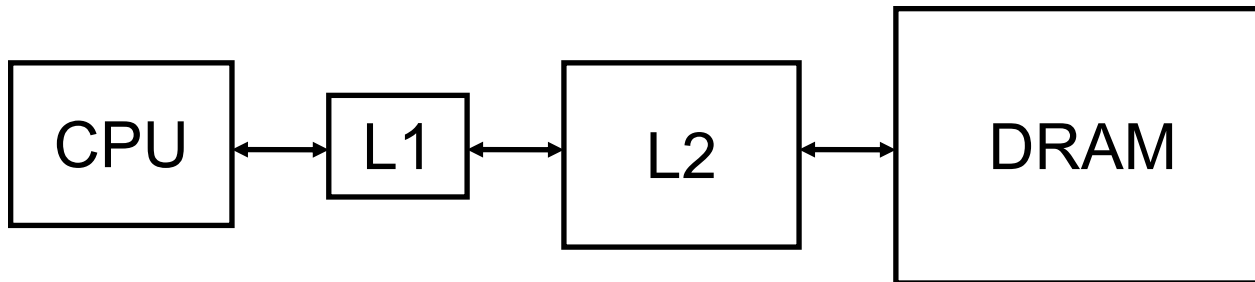
Metrics:

Local miss rate = misses in cache/ accesses to cache

Global miss rate = misses in cache / CPU memory accesses

Misses per instruction (MPI) = misses in cache / number of instructions

Typical memory hierarchies
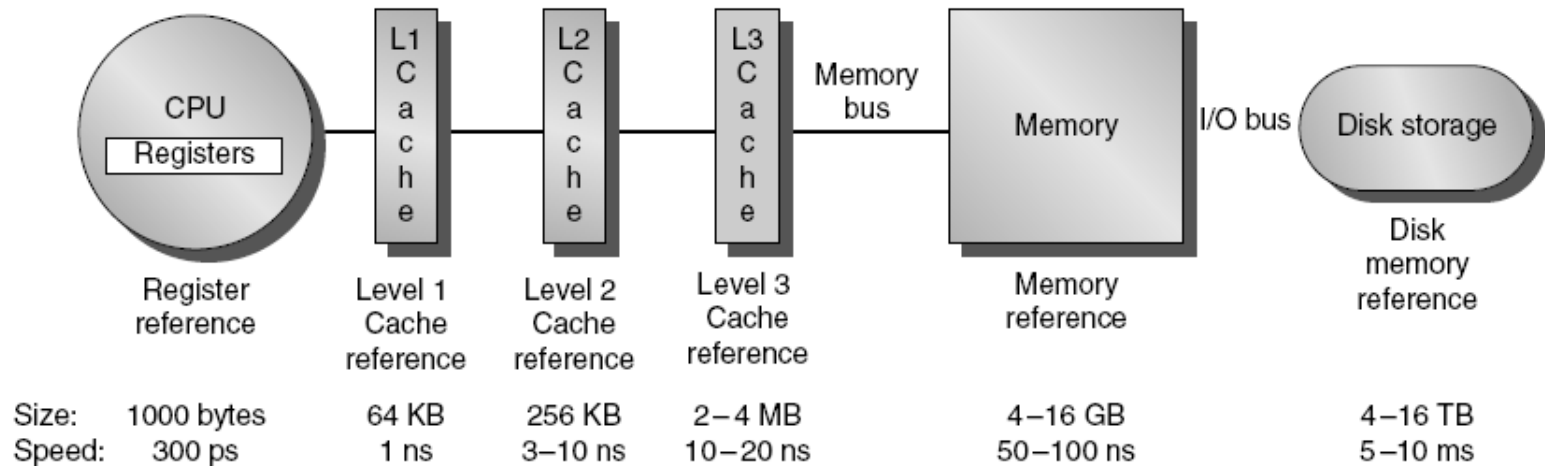
(a) Memory hierarchy for server

	Register reference	Level 1 Cache reference	Level 2 Cache reference	Level 3 Cache reference	Memory reference	Disk memory reference
Size:	1000 bytes	64 KB	256 KB	2–4 MB	4–16 GB	4–16 TB
Speed:	300 ps	1 ns	3–10 ns	10–20 ns	50–100 ns	5–10 ms
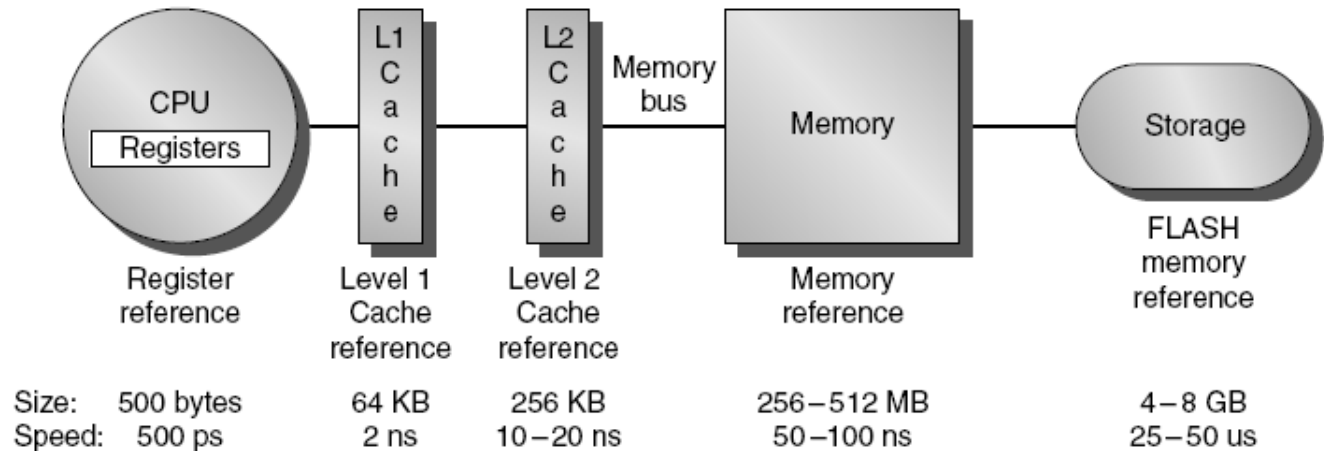
(b) Memory hierarchy for a personal mobile device

	Register reference	Level 1 Cache reference	Level 2 Cache reference	Memory reference	FLASH memory reference
Size:	500 bytes	64 KB	256 KB	256–512 MB	4–8 GB
Speed:	500 ps	2 ns	10–20 ns	50–100 ns	25–50 us

Block-level Optimizations

- Tags are too large, i.e., too much overhead

Block-level Optimizations

- Tags are too large, i.e., too much overhead
  - Simple solution: Larger blocks, but miss penalty could be large.

Block-level Optimizations

- ## Tags are too large, i.e., too much overhead
  - Simple solution: Larger blocks, but miss penalty could be large.

- ## Sub-block placement (aka sector cache)
  - A valid bit added to units smaller than the full block, called sub-blocks
  - Only read a sub-block on a miss
  - *If a tag matches, is the sub-block in the cache?*

100	1	1	1	1
300	1	1	0	0
204	0	1	0	1

Victim Caches (HP 7200)

CPU
RF

L1 Data Cache

Unified L2 Cache

Evicted data from L1

Hit data (miss in L1)

Victim Cache FA, 4 blocks

where ?
Evicted data from VC

Victim Caches (HP 7200)

CPU — RF

L1 Data Cache

Unified L2 Cache

Evicted data from L1

Victim Cache
FA, 4 blocks

Hit data (miss in L1)

where ?
Evicted data from VC
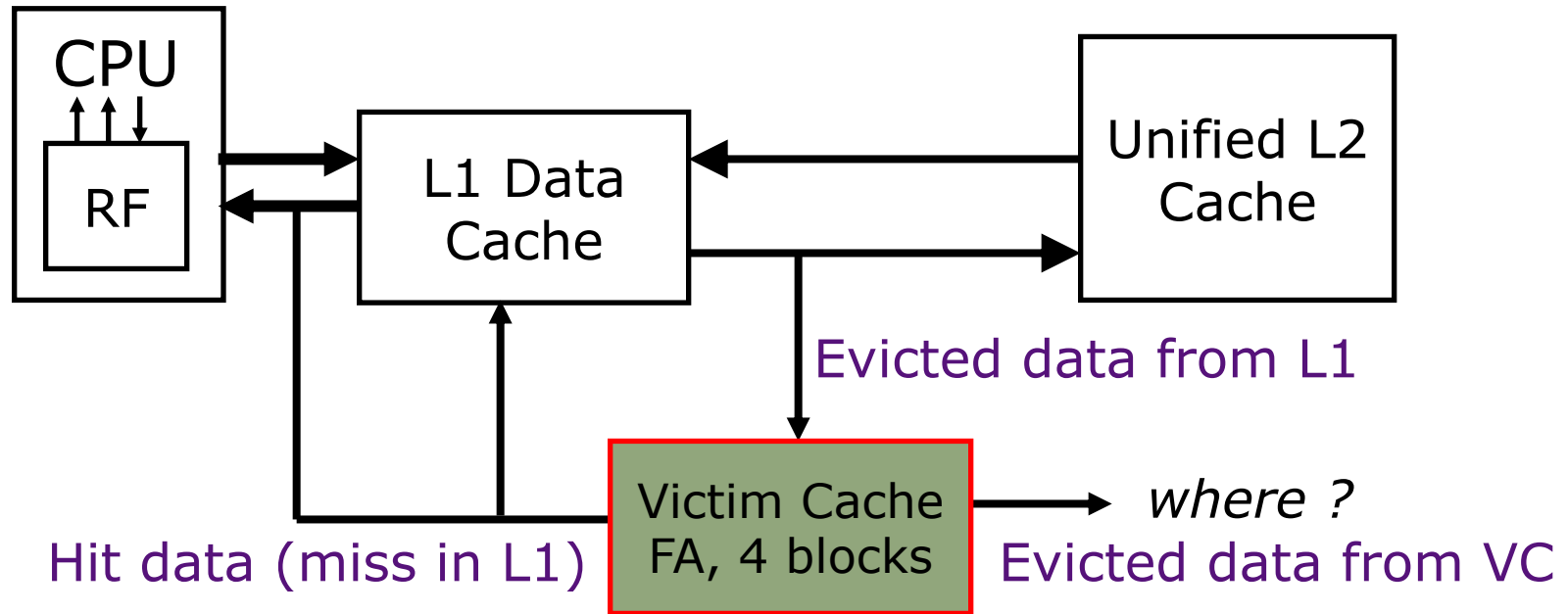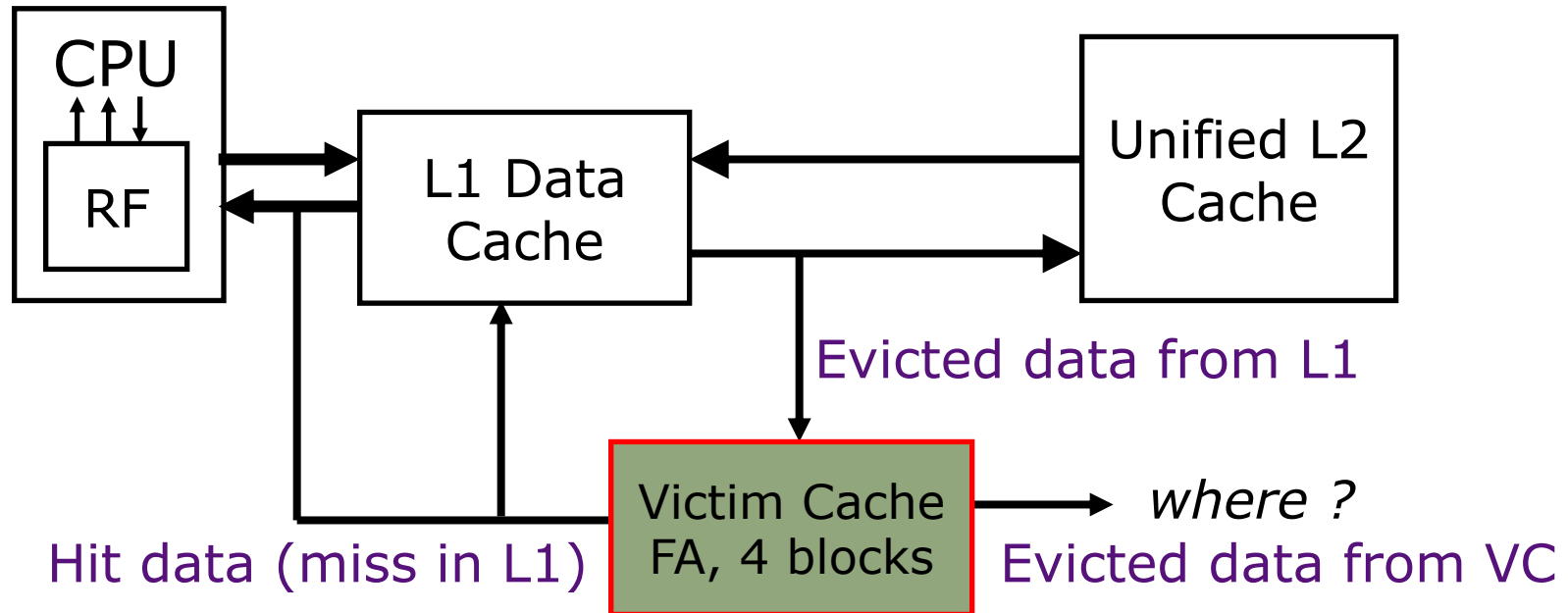
Victim cache is a small associative back up cache, added to a direct mapped cache, which holds recently evicted lines
• First look up in direct mapped cache
• If miss, look in victim cache
• If hit in victim cache, swap hit line with line now evicted from L1
• If miss in victim cache, L1 victim -> VC, VC victim->?
Fast hit time of direct mapped but with reduced conflict misses
-> Nowadays, more general, L4 in Intel Haswell, L3 in IBM Power5

Inclusion Policy

- ## Inclusive multilevel cache:
  - Inner cache holds copies of data in outer cache
  - On miss, line inserted in inner and outer cache; replacement in outer cache invalidates line in inner cache
  - External accesses need only check outer cache
  - Commonly used (e.g., Intel CPUs up to Broadwell)

- ## Non-inclusive multilevel caches:
  - Inner cache may hold data not in outer cache
  - Replacement in outer cache doesn't invalidate line in inner cache
  - Used in Intel Skylake, ARM

- ## Exclusive multilevel caches:
  - Inner cache and outer cache hold different data
  - Swap lines between inner/outer caches on miss
  - Used in AMD processors

Inclusion Policy

- Inclusive multilevel cache:
  - Inner cache holds copies of data in outer cache
  - On miss, line inserted in inner and outer cache; replacement in outer cache invalidates line in inner cache
  - External accesses need only check outer cache
  - Commonly used (e.g., Intel CPUs up to Broadwell)

- Non-inclusive multilevel caches:
  - Inner cache may hold data not in outer cache
  - Replacement in outer cache doesn't invalidate line in inner cache
  - Used in Intel Skylake, ARM

- Exclusive multilevel caches:
  - Inner cache and outer cache hold different data
  - Swap lines between inner/outer caches on miss
  - Used in AMD processors

Why choose one type or the other?

Memory Management:
From Absolute Addresses to Demand Paging

Mengjia Yan
Computer Science and Artificial Intelligence Laboratory
M.I.T.

Memory Management

- The Fifties
  - Absolute Addresses
  - Dynamic address translation

- The Sixties
  - Atlas and Demand Paging
  - Paged memory systems and TLBs

- Modern Virtual Memory Systems

Names for Memory Locations

machine
language
address
→ **ISA** →
virtual
address
→ **Address Mapping** →
physical
address
→ **Physical Memory (DRAM)**

Names for Memory Locations

machine language address → **ISA** → virtual address → **Address Mapping** → physical address → **Physical Memory (DRAM)**
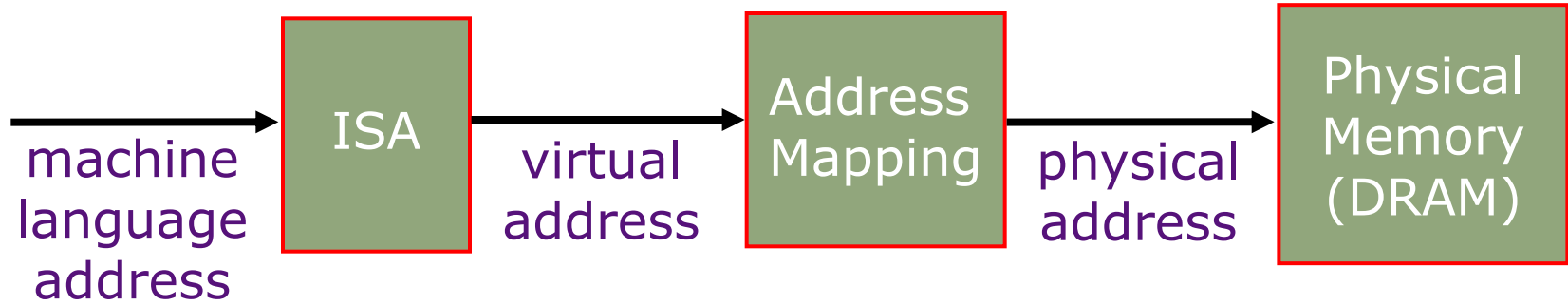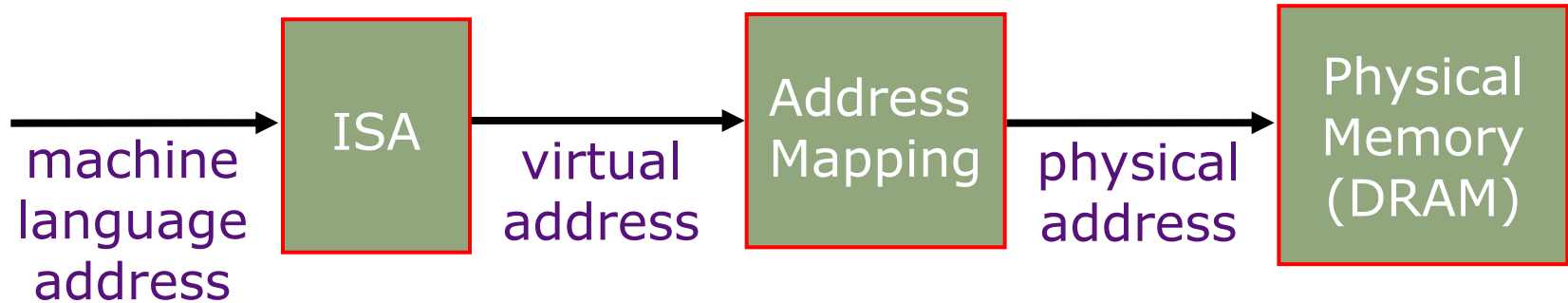
- ## Machine language address
  - as specified in machine code

- ## Virtual address
  - ISA specifies translation of machine code address into virtual address of program variable (sometimes called *effective* address)

- ## Physical address
  - Operating system specifies mapping of virtual address into name for a physical memory location

Absolute Addresses

EDSAC, early 50's

virtual address = physical memory address

Absolute Addresses

EDSAC, early 50's

virtual address = physical memory address

- Only one program ran at a time, with unrestricted access to entire machine (RAM + I/O devices)
- Addresses in a program depended upon where the program was to be loaded in memory

Absolute Addresses

EDSAC, early 50's

virtual address = physical memory address

- Only one program ran at a time, with unrestricted access to entire machine (RAM + I/O devices)
- Addresses in a program depended upon where the program was to be loaded in memory
- *But* it was more convenient for programmers to write location-independent subroutines

Absolute Addresses

EDSAC, early 50's

> virtual address = physical memory address

- Only one program ran at a time, with unrestricted access to entire machine (RAM + I/O devices)
- Addresses in a program depended upon where the program was to be loaded in memory
- *But* it was more convenient for programmers to write location-independent subroutines

How could location independence be achieved?

Absolute Addresses

EDSAC, early 50's

> virtual address = physical memory address

- Only one program ran at a time, with unrestricted access to entire machine (RAM + I/O devices)
- Addresses in a program depended upon where the program was to be loaded in memory
- *But* it was more convenient for programmers to write location-independent subroutines

How could location independence be achieved?

Linker and/or loader modify addresses of subroutines and callers when building a program memory image

Multiprogramming

Motivation

In the early machines, I/O operations were slow and each word transferred involved the CPU

Higher throughput if CPU and I/O of 2 or more programs were overlapped. *How?*
$\Rightarrow$ *multiprogramming*

prog1

prog2

Physical Memory

Multiprogramming

Motivation

In the early machines, I/O operations were slow and each word transferred involved the CPU

Higher throughput if CPU and I/O of 2 or more programs were overlapped. *How?*
⇒ *multiprogramming*

Location-independent programs

Programming and storage management ease
⇒ need for a *base register*

prog1

prog2

Physical Memory

Multiprogramming

Motivation

In the early machines, I/O operations were slow and each word transferred involved the CPU

Higher throughput if CPU and I/O of 2 or more programs were overlapped. *How?*
  $\Rightarrow$ *multiprogramming*

Location-independent programs

Programming and storage management ease
  $\Rightarrow$ need for a *base register*

Protection

Independent programs should not affect each other inadvertently

  $\Rightarrow$ need for a *bound register*

Physical Memory

prog1

prog2
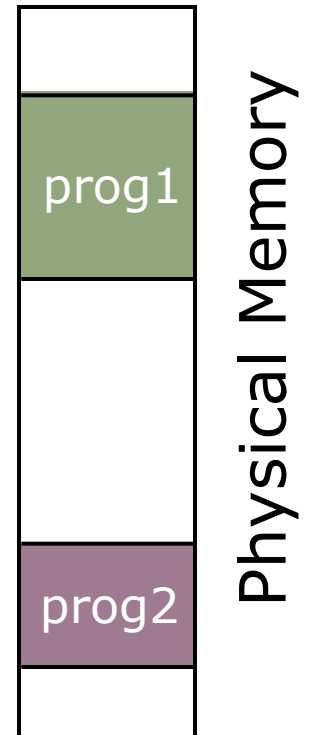
Simple Base and Bound Translation
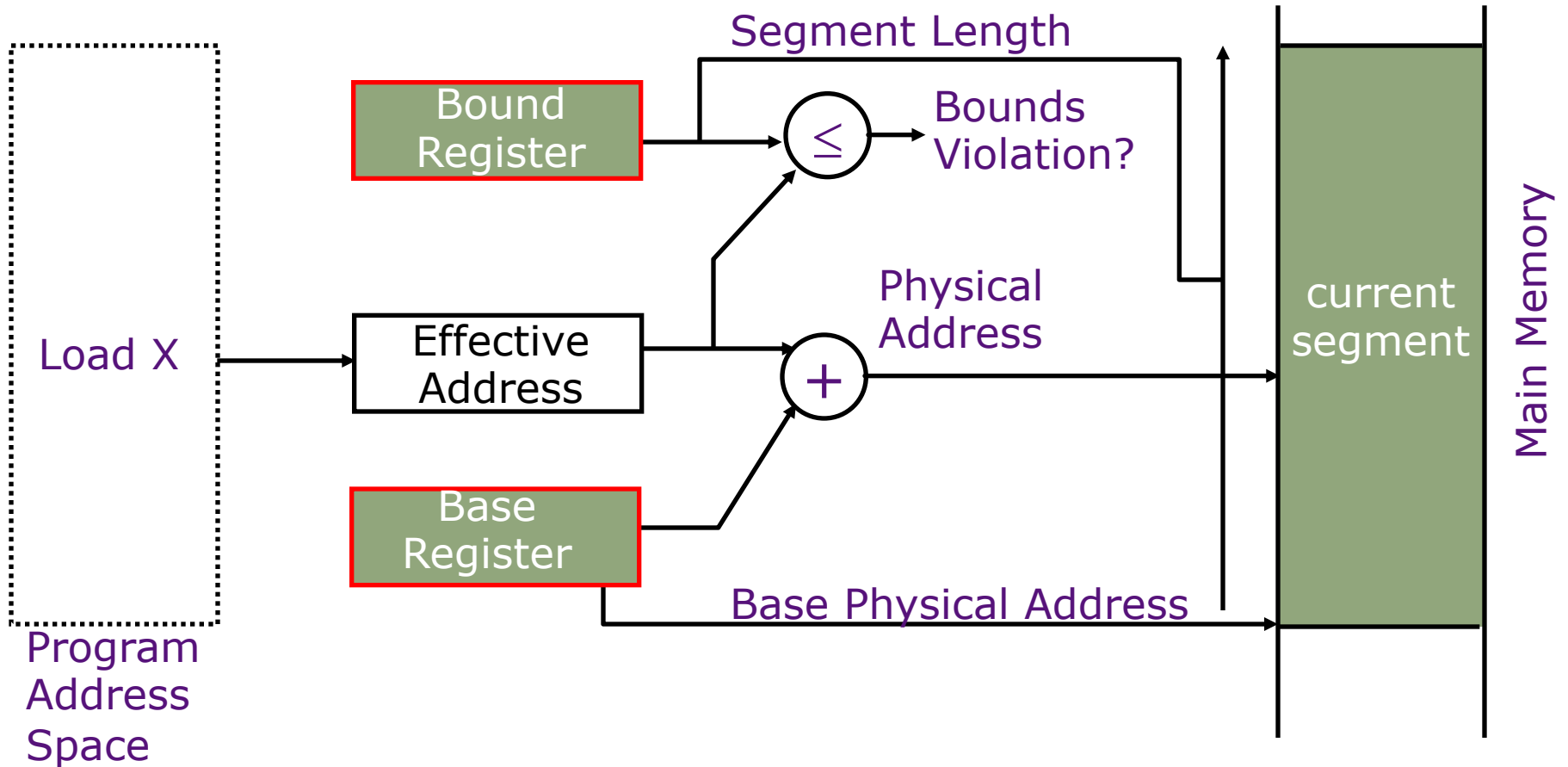
Simple Base and Bound Translation
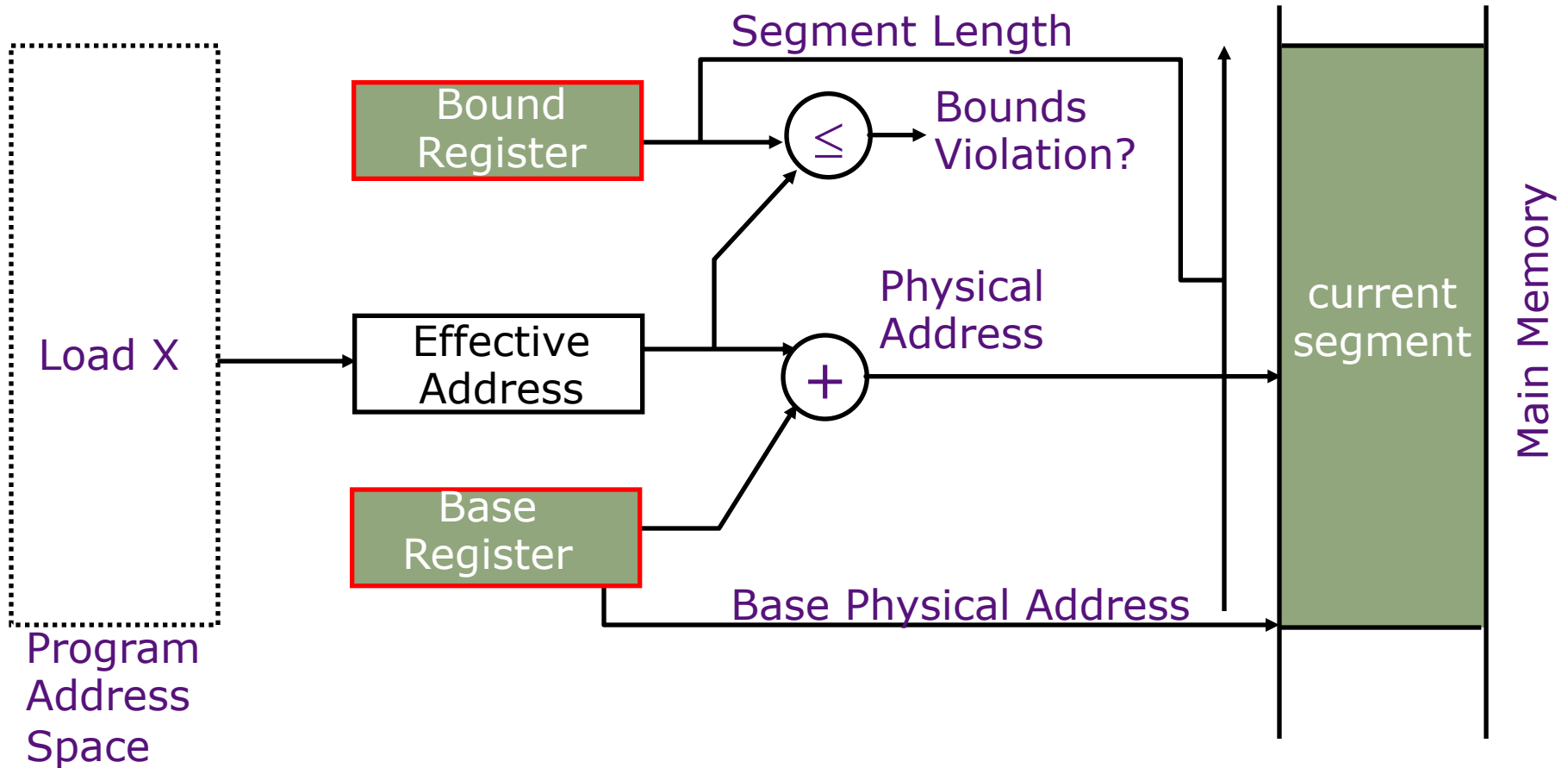
Base and bounds registers are visible/accessible only when processor is running in *supervisor mode*

Separate Areas for Code and Data

Program Address Space

Load X

Data Bound Register

Effective Addr Register

Data Base Register

Code Bound Register

Program Counter

Code Base Register

≤ Bounds Violation?

+

≤ Bounds Violation?

+

data segment

code segment

Main Memory
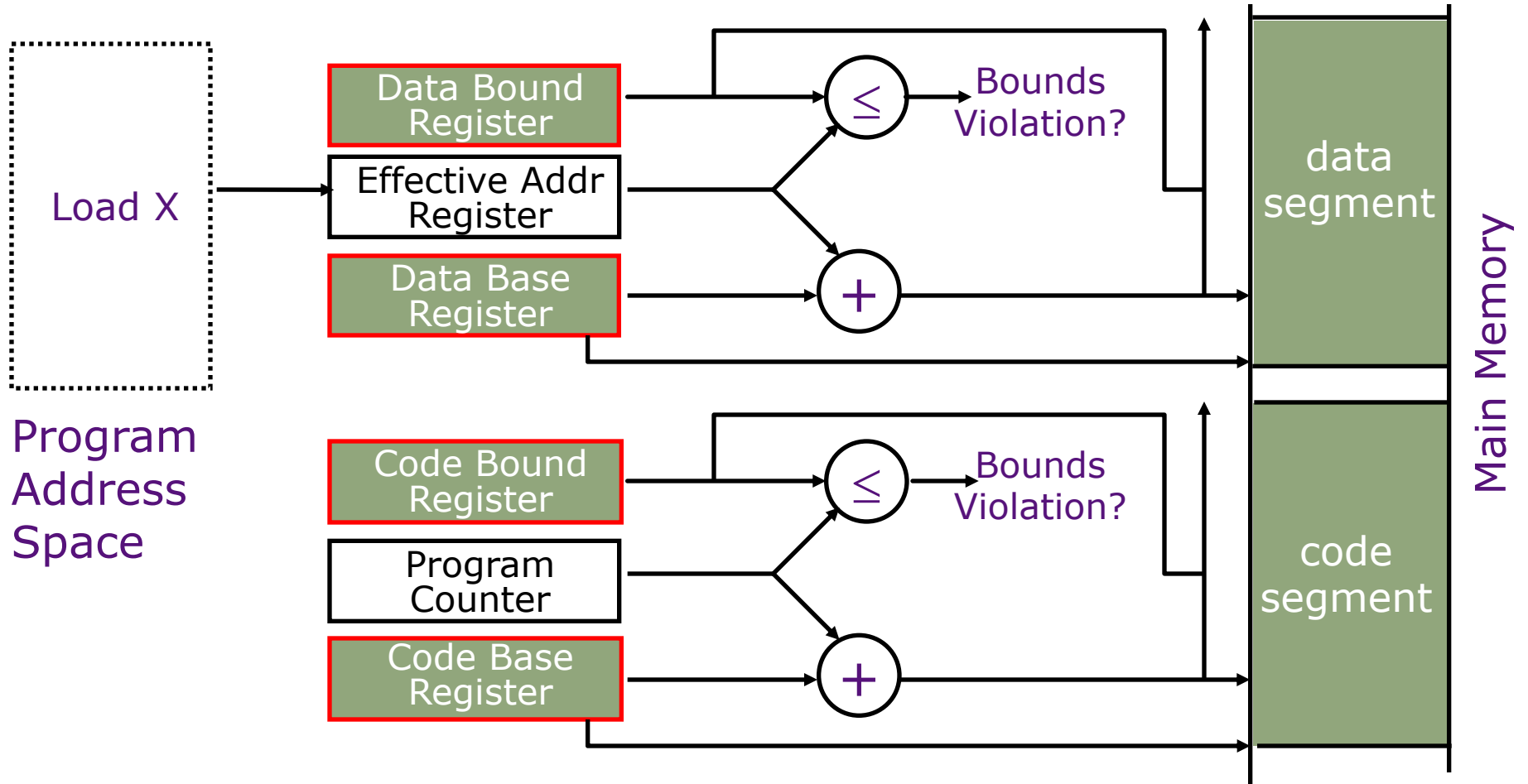
What is an advantage of this separation?
(Scheme used on all Cray vector supercomputers prior to X1, 2002)

Memory Fragmentation

free

OS
Space

user 1 — 16K

user 2 — 24K

24K

user 3 — 32K

24K

Memory Fragmentation

free

Users 4 & 5
arrive

OS
Space

user 1	16K
user 2	24K
	24K
user 3	32K
	24K

OS
Space

user 1	16K
user 2	24K
user 4	16K
	8K
user 3	32K
user 5	24K
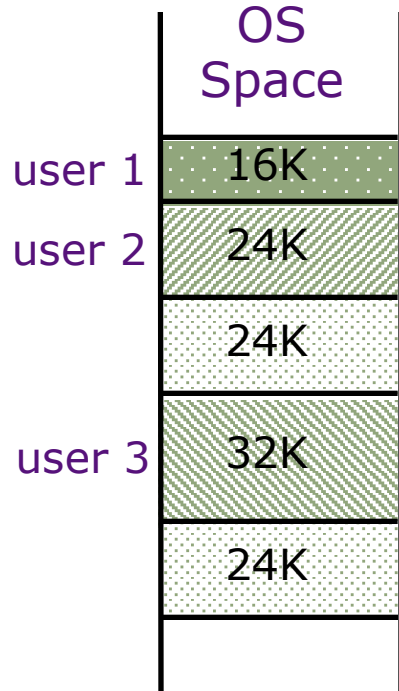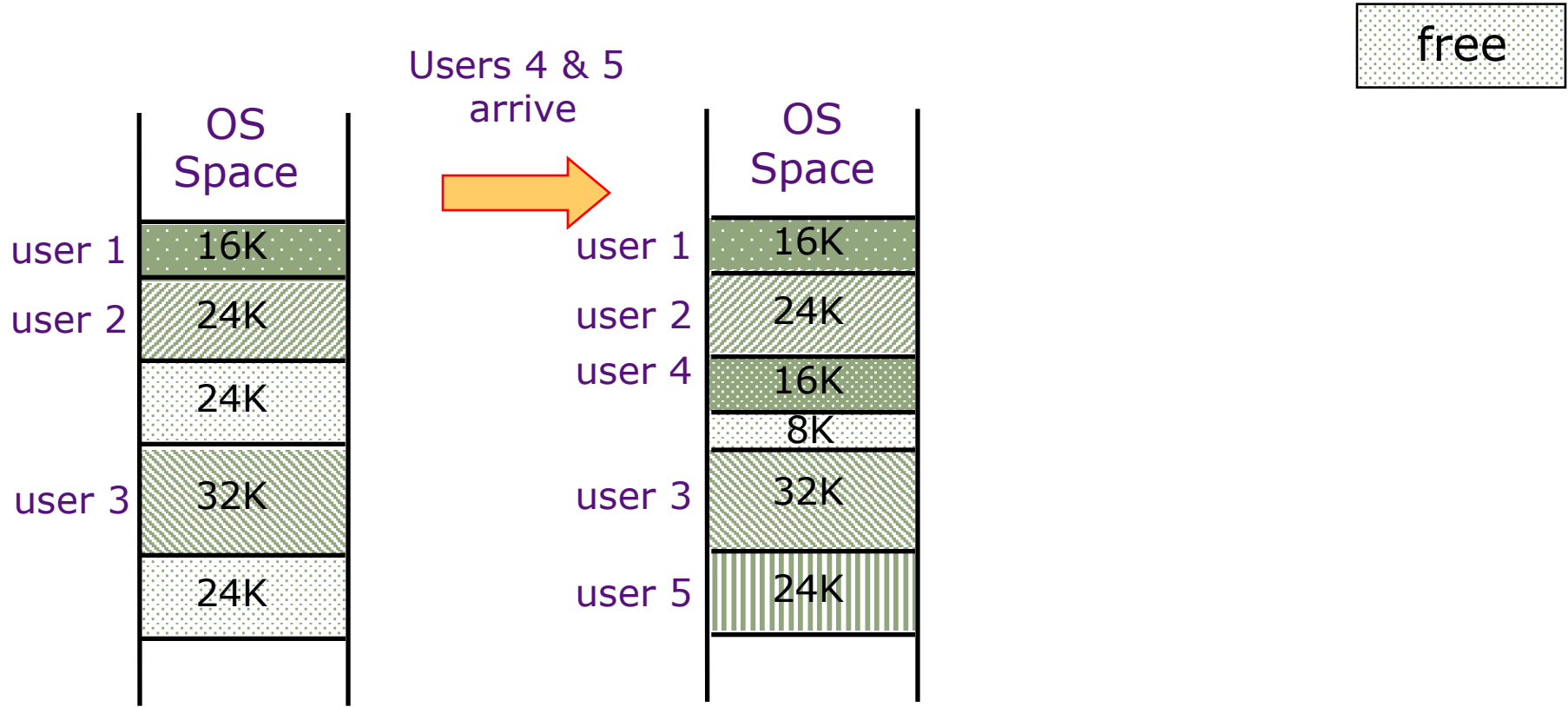
Memory Fragmentation

Memory Fragmentation
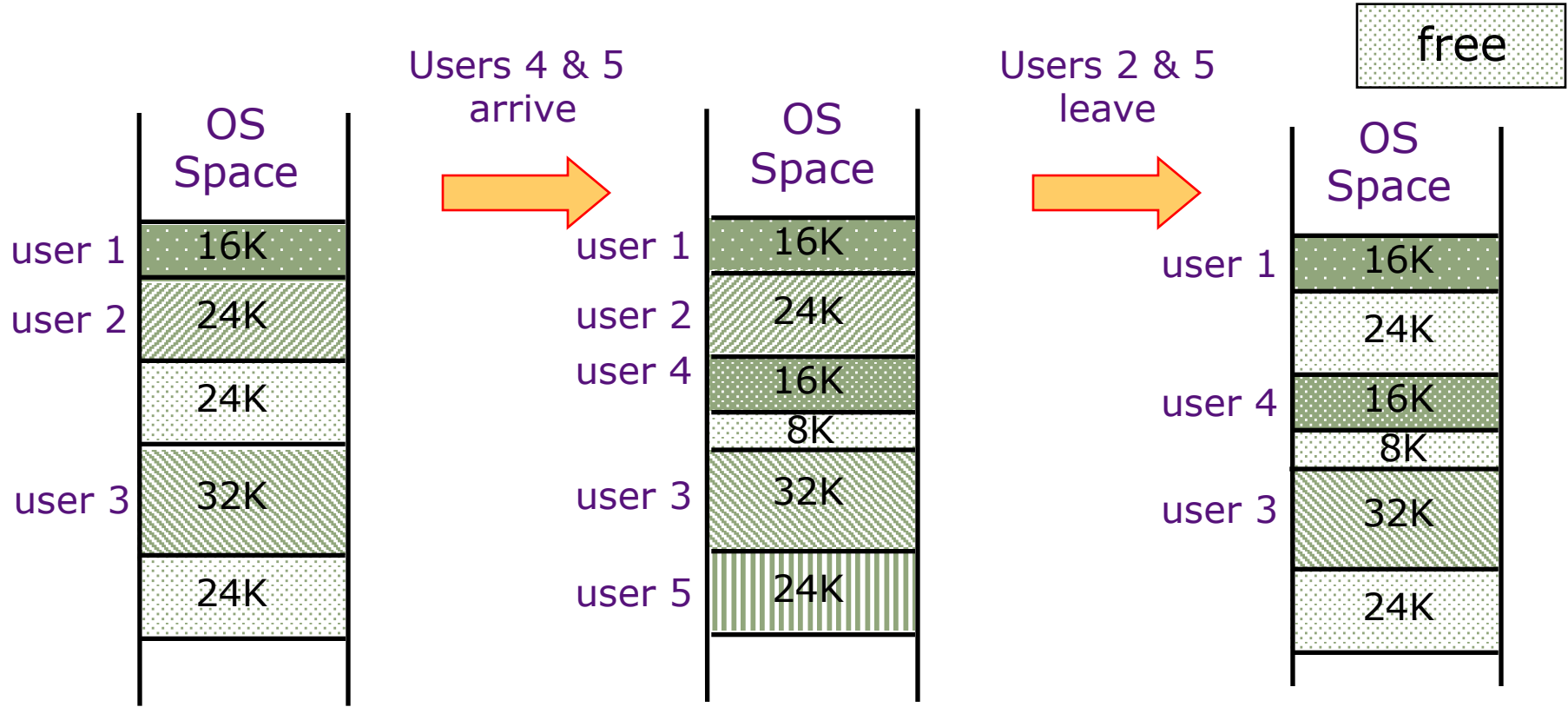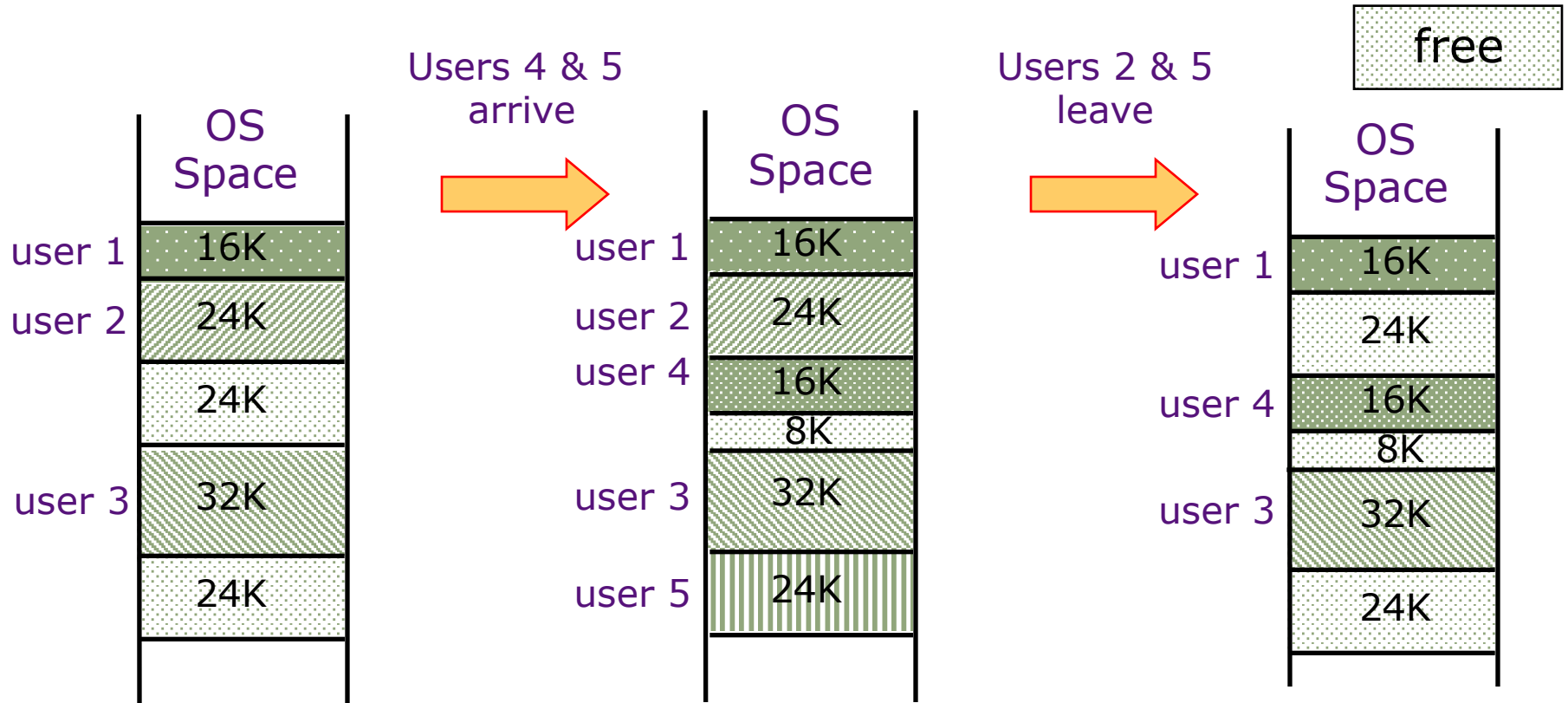
As users come and go, the storage is "fragmented".

Memory Fragmentation
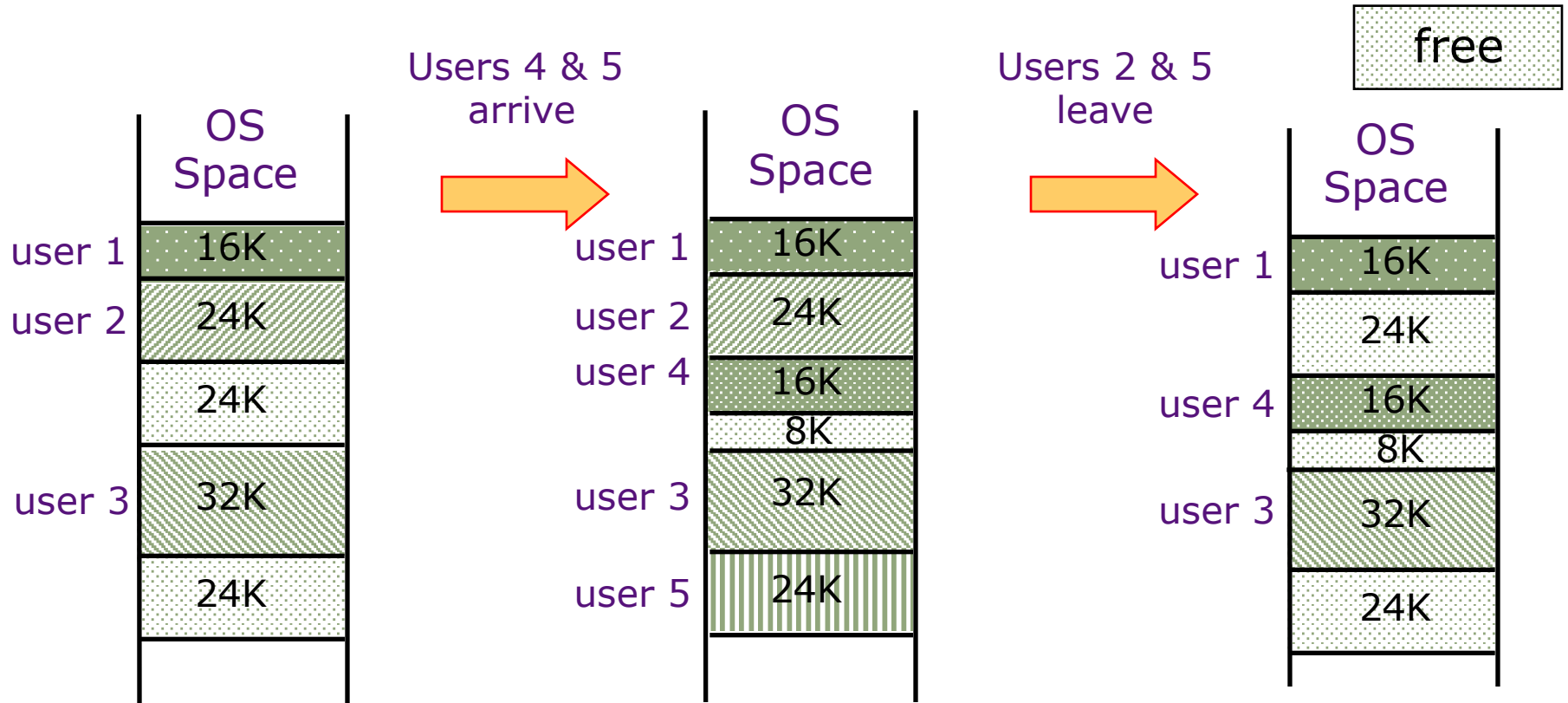
As users come and go, the storage is "fragmented". Therefore, at some stage programs have to be moved around to compact the storage.

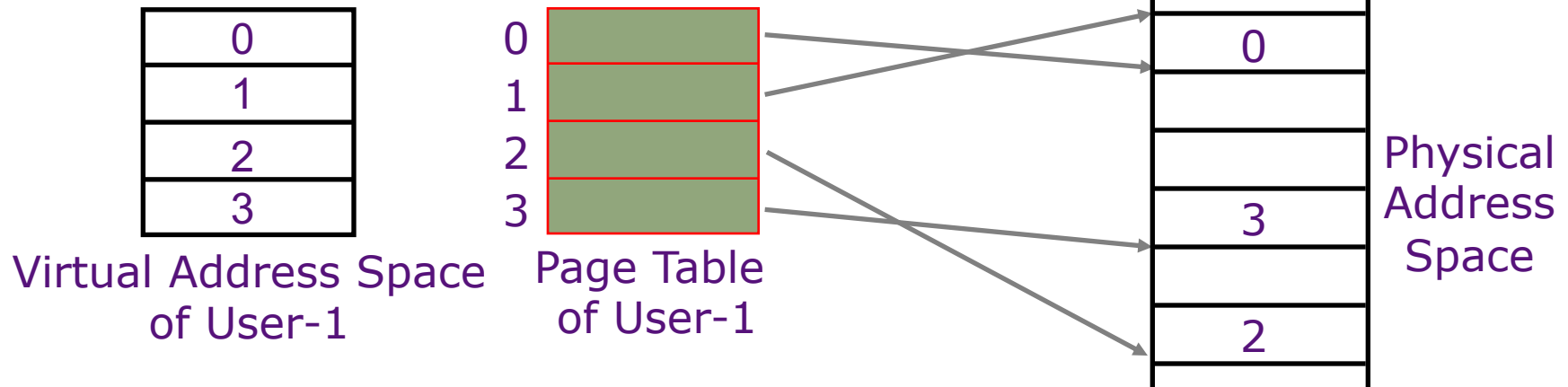# Paged Memory Systems
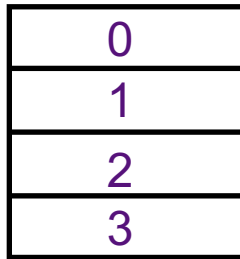
Virtual Address Space of User-1
0
1
2
3
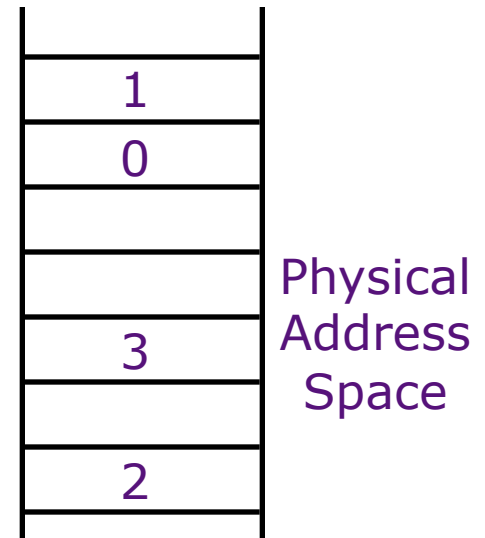
Physical Address Space

1
0
3
2

Paged Memory Systems

- Processor-generated address can be interpreted as a pair <page number, offset>

page number	offset

- A page table contains the physical address of the base of each page

Virtual Address Space
of User-1

Page Table
of User-1

Physical
Address
Space

Paged Memory Systems

- Processor-generated address can be interpreted as a pair <page number, offset>

page number	offset

- A page table contains the physical address of the base of each page

Virtual Address Space of User-1

Page Table of User-1
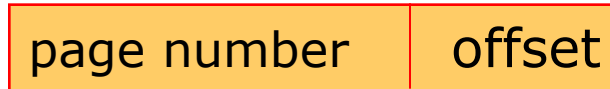
Physical Address Space
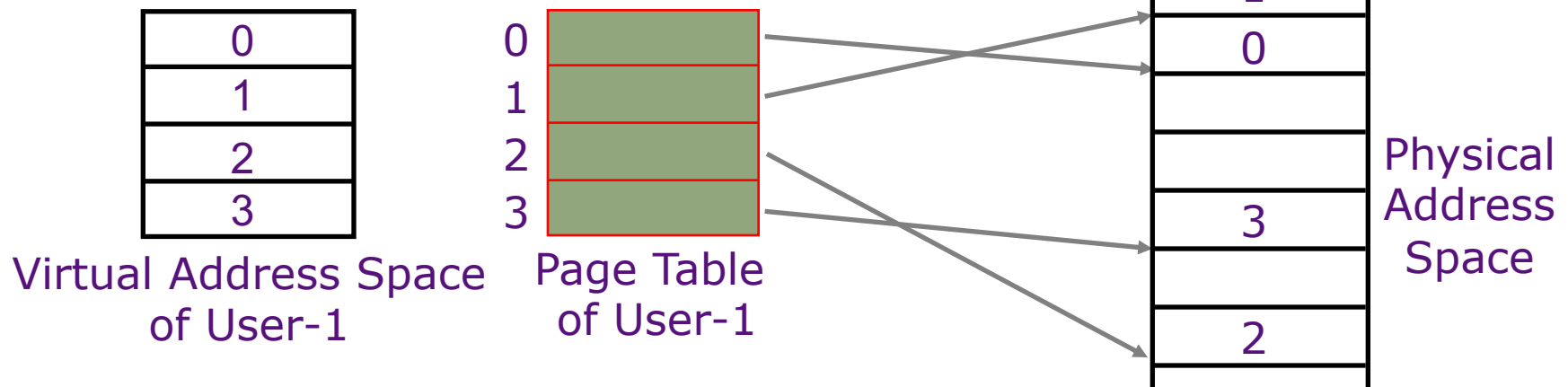
*Page tables make it possible to store the pages of a program **non-contiguously**.*

A Problem in Early Sixties

- There were many applications whose data could not fit in the main memory, e.g., payroll

  Paged memory system reduced fragmentation but still required the whole program to be resident in the main memory

- Programmers moved the data back and forth from the secondary store by *overlaying* it repeatedly on the primary store

  tricky programming!

Manual Overlays

- Assume an instruction can address all the storage on the drum

- *Method 1:* programmer keeps track of addresses in the main memory and initiates an I/O transfer when required
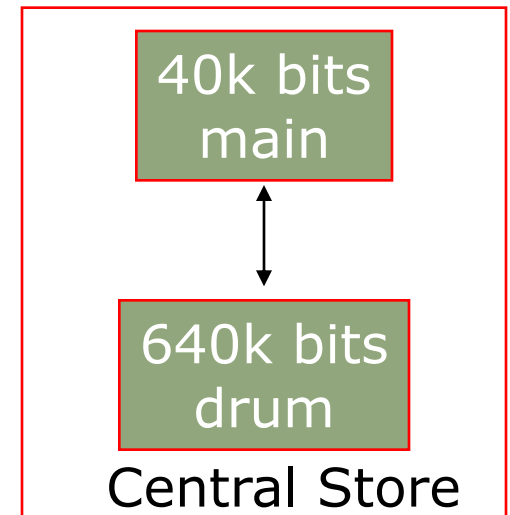
40k bits main

640k bits drum

Central Store

Ferranti Mercury 1956

MIT 6.5900 Fall 2023

Manual Overlays

- Assume an instruction can address all the storage on the drum

- *Method 1:* programmer keeps track of addresses in the main memory and initiates an I/O transfer when required

- *Method 2:* automatic initiation of I/O transfers by software address translation
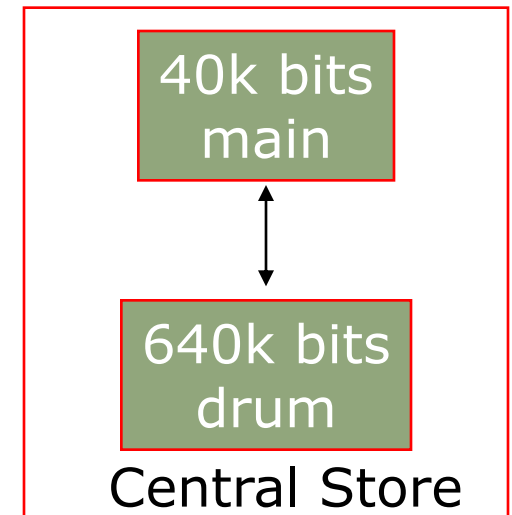
  Brooker's interpretive coding, 1960

40k bits main

640k bits drum

Central Store

Ferranti Mercury 1956

Manual Overlays

- Assume an instruction can address all the storage on the drum

- *Method 1:* programmer keeps track of addresses in the main memory and initiates an I/O transfer when required

- *Method 2:* automatic initiation of I/O transfers by software address translation

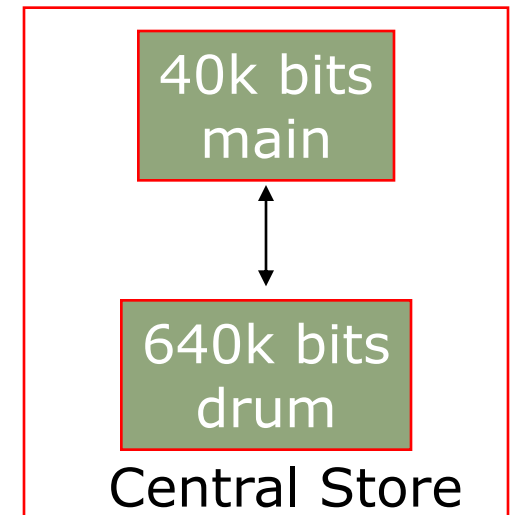 *Brooker's interpretive coding, 1960*

Problems?

40k bits main	
↕	
640k bits drum	

Central Store

Ferranti Mercury 1956

Manual Overlays

- Assume an instruction can address all the storage on the drum

- *Method 1:* programmer keeps track of addresses in the main memory and initiates an I/O transfer when required

- *Method 2:* automatic initiation of I/O transfers by software address translation

  Brooker's interpretive coding, 1960

40k bits main

↕

640k bits drum

Central Store

Ferranti Mercury 1956

Problems? | Method1: Difficult, error prone |

Manual Overlays

- Assume an instruction can address all the storage on the drum

- *Method 1:* programmer keeps track of addresses in the main memory and initiates an I/O transfer when required

- *Method 2:* automatic initiation of I/O transfers by software address translation

  Brooker's interpretive coding, 1960

40k bits main

640k bits drum
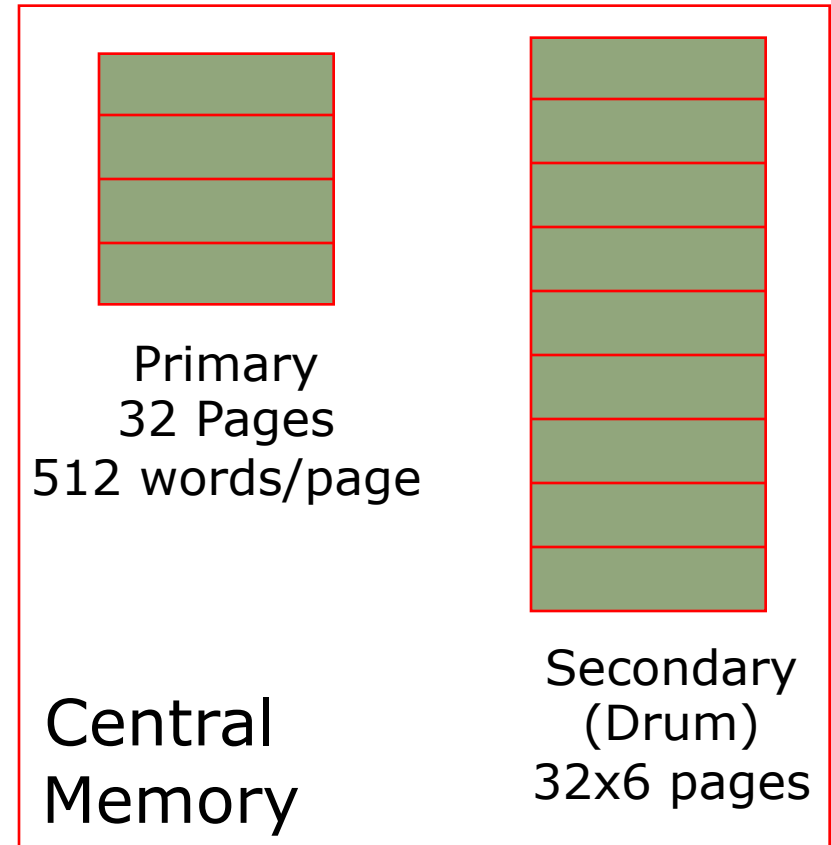
Central Store
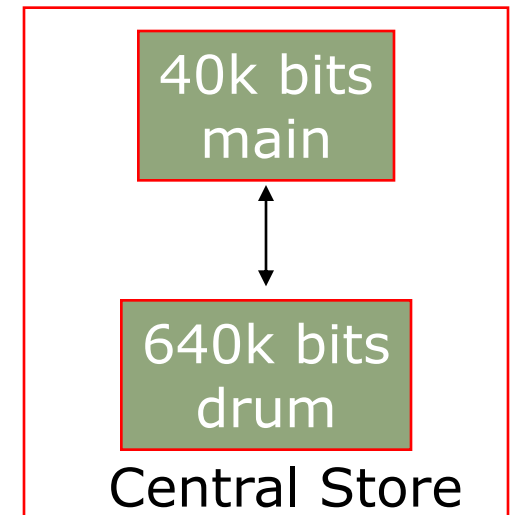
Ferranti Mercury 1956

Problems?

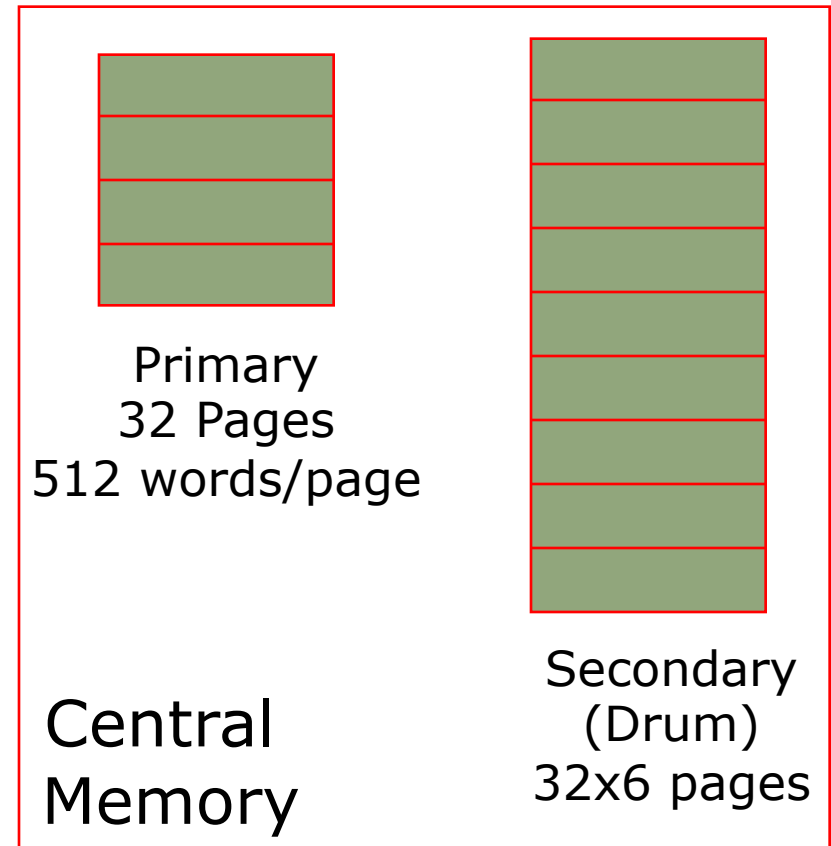Method1: Difficult, error prone
Method2: Inefficient

Demand Paging in Atlas (1962)

"A page from secondary storage is brought into the primary storage whenever it is (*implicitly*) demanded by the processor."

Tom Kilburn

Primary
32 Pages
512 words/page

Secondary
(Drum)
32x6 pages

Central Memory

Demand Paging in Atlas (1962)

"A page from secondary storage is brought into the primary storage whenever it is (*implicitly*) demanded by the processor."
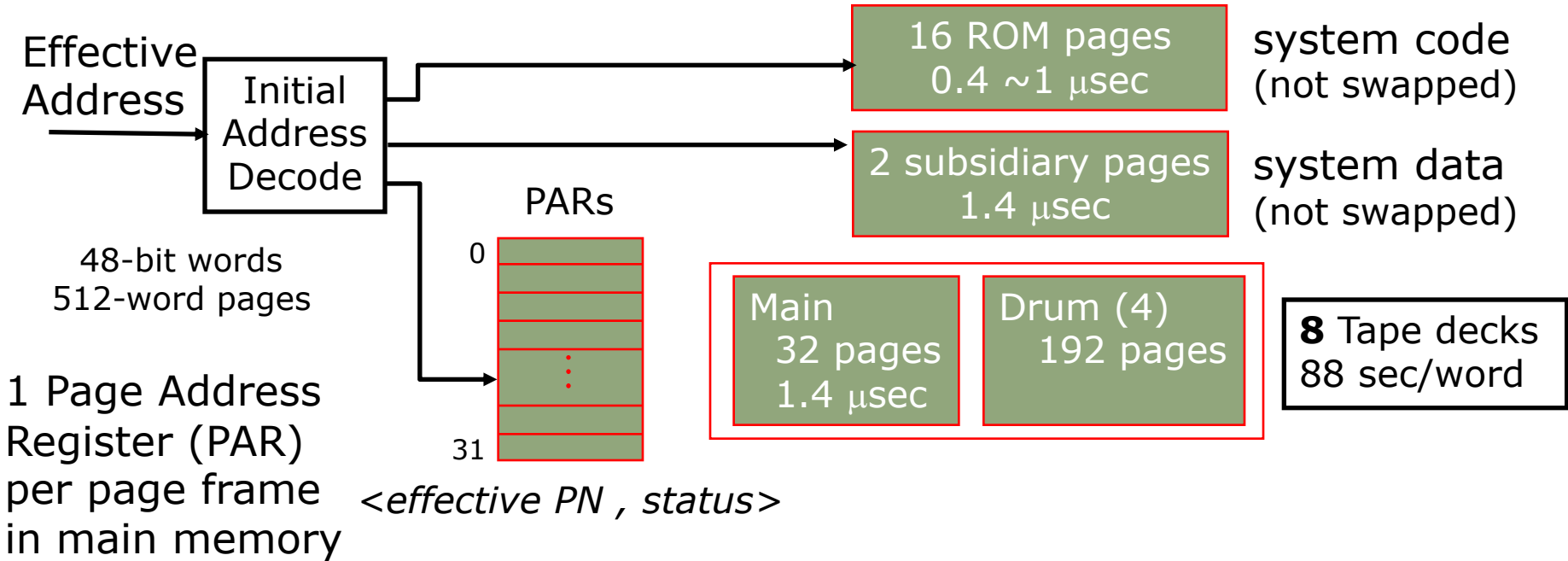
Tom Kilburn
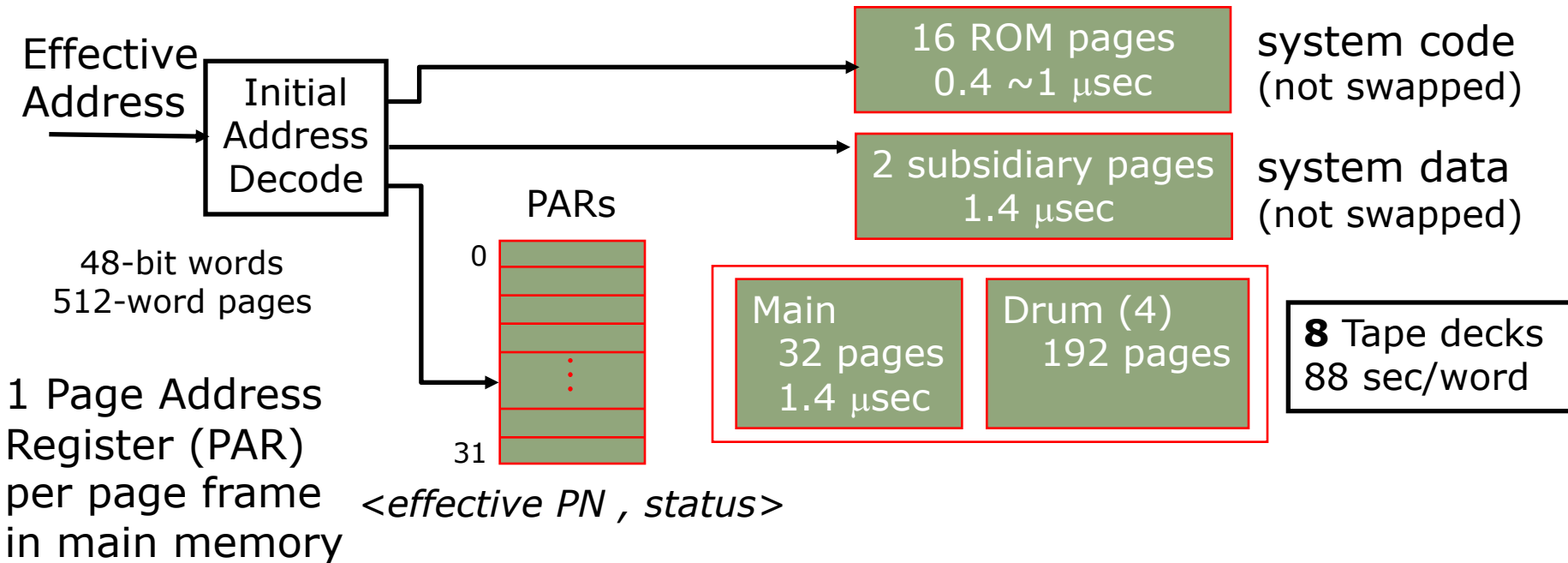
Primary memory as a *cache* for secondary memory

User sees the storage size of the secondary storage, since data transfer happens automatically

Primary
32 Pages
512 words/page

Secondary
(Drum)
32x6 pages

Central
Memory

Hardware Organization of Atlas

Effective Address → Initial Address Decode

48-bit words
512-word pages

1 Page Address Register (PAR) per page frame in main memory

PARs

0
...
31

<effective PN , status>

16 ROM pages
0.4 ~1 μsec
system code (not swapped)

2 subsidiary pages
1.4 μsec
system data (not swapped)

Main
32 pages
1.4 μsec

Drum (4)
192 pages

8 Tape decks
88 sec/word

Hardware Organization of Atlas

Effective Address → Initial Address Decode

48-bit words
512-word pages

1 Page Address Register (PAR) per page frame in main memory

PARs

0
:
31

<effective PN , status>

16 ROM pages
0.4 ~1 μsec

system code
(not swapped)

2 subsidiary pages
1.4 μsec

system data
(not swapped)

Main
32 pages
1.4 μsec

Drum (4)
192 pages

8 Tape decks
88 sec/word

Compare the effective page address against all 32 PARs
        match   ⇒ normal access
        no match   ⇒ *page fault*
   save the state of the partially executed instruction

Atlas Demand Paging Scheme

- ## On a page fault:

  - Input transfer into a free page is initiated

  - The Page Address Register (PAR) is updated

  - If no free page is left, a *page is selected to be replaced* (based on usage)

  - The replaced page is written on the drum
    - to minimize the drum latency effect, the first empty page on the drum was selected

  - The *page table is updated* to point to the new location of the page on the drum

Caching vs. Demand Paging
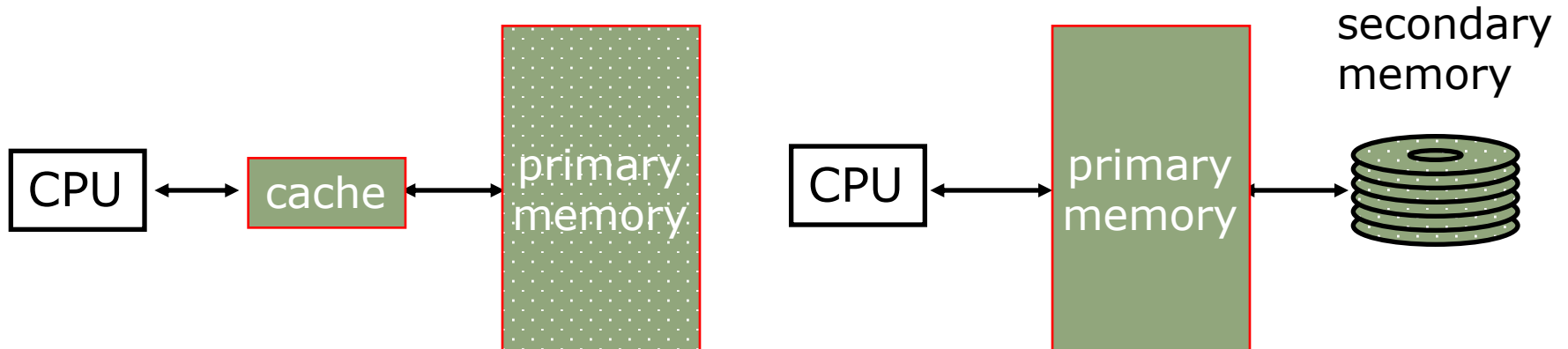
Caching vs. Demand Paging
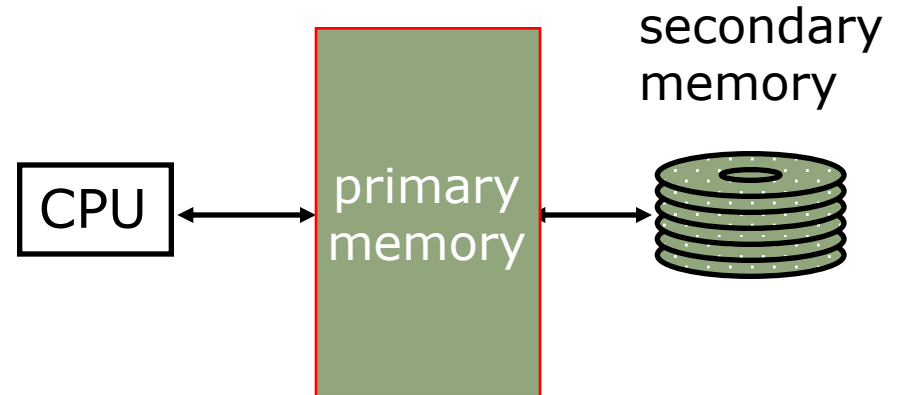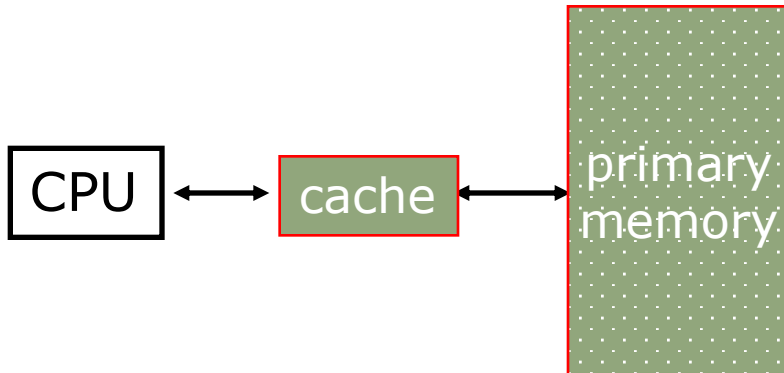
Caching

cache entry

cache block (~32 bytes)

cache miss rate (1% to 20%)

cache hit (~1 cycle)

cache miss (~100 cycles)

a miss is handled
    in *hardware*

Demand paging

page frame

page (~4K bytes)

page miss rate (<0.001%)

page hit (~100 cycles)

page miss (~5M cycles)

a miss is handled
    mostly in *software*

Modern Virtual Memory Systems
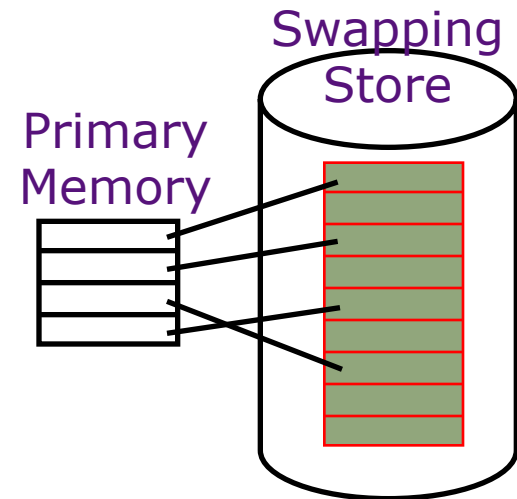*Illusion of a large, private, uniform store*

Protection & Privacy
several users, each with their private address space and one or more shared address spaces
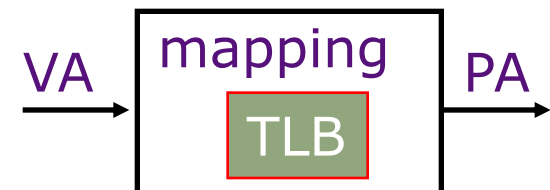page table $\equiv$ name space

OS

user$_i$
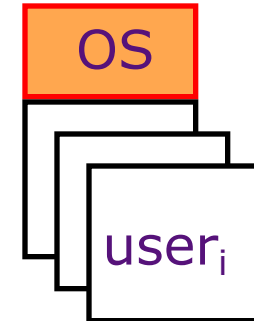
Demand Paging
Provides the ability to run programs larger than the primary memory

Hides differences in machine configurations

Swapping Store

Primary Memory

The price is address translation on each memory reference

VA → mapping TLB → PA

Next lecture:

Modern Virtual Memory Systems