

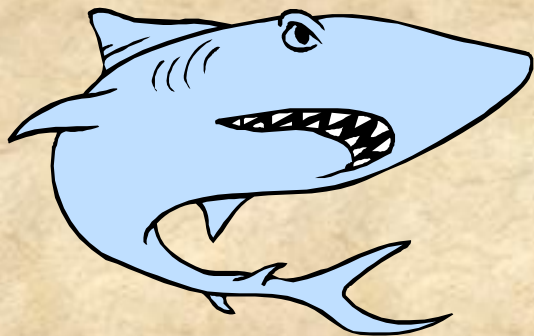
*Operating
Systems:
Internals
and Design
Principles*

Virtual Memory

- Implementation
 - Principle of locality
 - Trashing
 - Paging and multilevel paging
 - Segmentation and Segmented paging
 - Address translation
 - Inverted page table
 - Translation Lookaside Buffer (TLB)
 - Page size
 - Fetch policy, placement policy
 - Replacement policy & algorithms
 - Resident set size & management
 - Cleaning policy
 - Load control

Operating Systems: Internals and Design Principles

You're gonna need a bigger boat.



— Steven Spielberg,
JAWS, 1975

Hardware and Control Structures

- n Two characteristics fundamental to memory management:
 - 1) all memory references are logical addresses that are dynamically translated into physical addresses at run time
 - 2) a process may be broken up into a number of pieces that don't need to be contiguously located in main memory during execution
- n If these two characteristics are present, it is not necessary that all of the pages or segments of a process be in main memory during execution

Terminology

Virtual memory	A storage allocation scheme in which secondary memory can be addressed as though it were part of main memory. The addresses a program may use to reference memory are distinguished from the addresses the memory system uses to identify physical storage sites, and program-generated addresses are translated automatically to the corresponding machine addresses. The size of virtual storage is limited by the addressing scheme of the computer system and by the amount of secondary memory available and not by the actual number of main storage locations.
Virtual address	The address assigned to a location in virtual memory to allow that location to be accessed as though it were part of main memory.
Virtual address space	The virtual storage assigned to a process.
Address space	The range of memory addresses available to a process.
Real address	The address of a storage location in main memory.

Execution of a Process

- n Operating system brings into main memory a few pieces of the program
- n Resident set - portion of process that is in main memory
- n An interrupt (page fault interrupt) is generated when an address is needed that is not in main memory
- n Operating system places the process in a blocking state



I/O Implementation

- n Piece of process that contains the logical address is brought into main memory
- n Operating system issues a disk I/O Read request
- n Another process is dispatched to run while the disk I/O takes place
- n An interrupt (I/O interrupt) is issued when disk I/O is complete, which causes the operating system to place the affected process in the Ready state



Implications

- n More processes may be maintained in main memory
 - n Only load in some of the pieces of each process
 - n With so many processes in main memory, it is very likely that some process will be in the Ready state at any particular time
- n A process may be larger than all of main memory



Real and Virtual Memory

Real memory

- main memory, the actual RAM

Virtual memory

- all program memory on disk - some not uptodate
- some in real memory – all currently used areas
- allows for effective multiprogramming and relieves the user of tight constraints of main memory

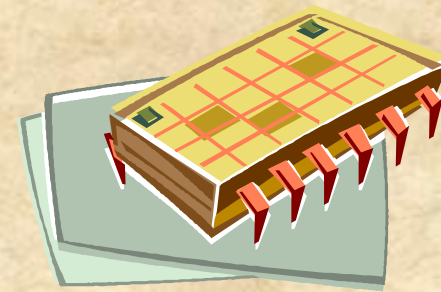


Table 8.2

Characteristics of

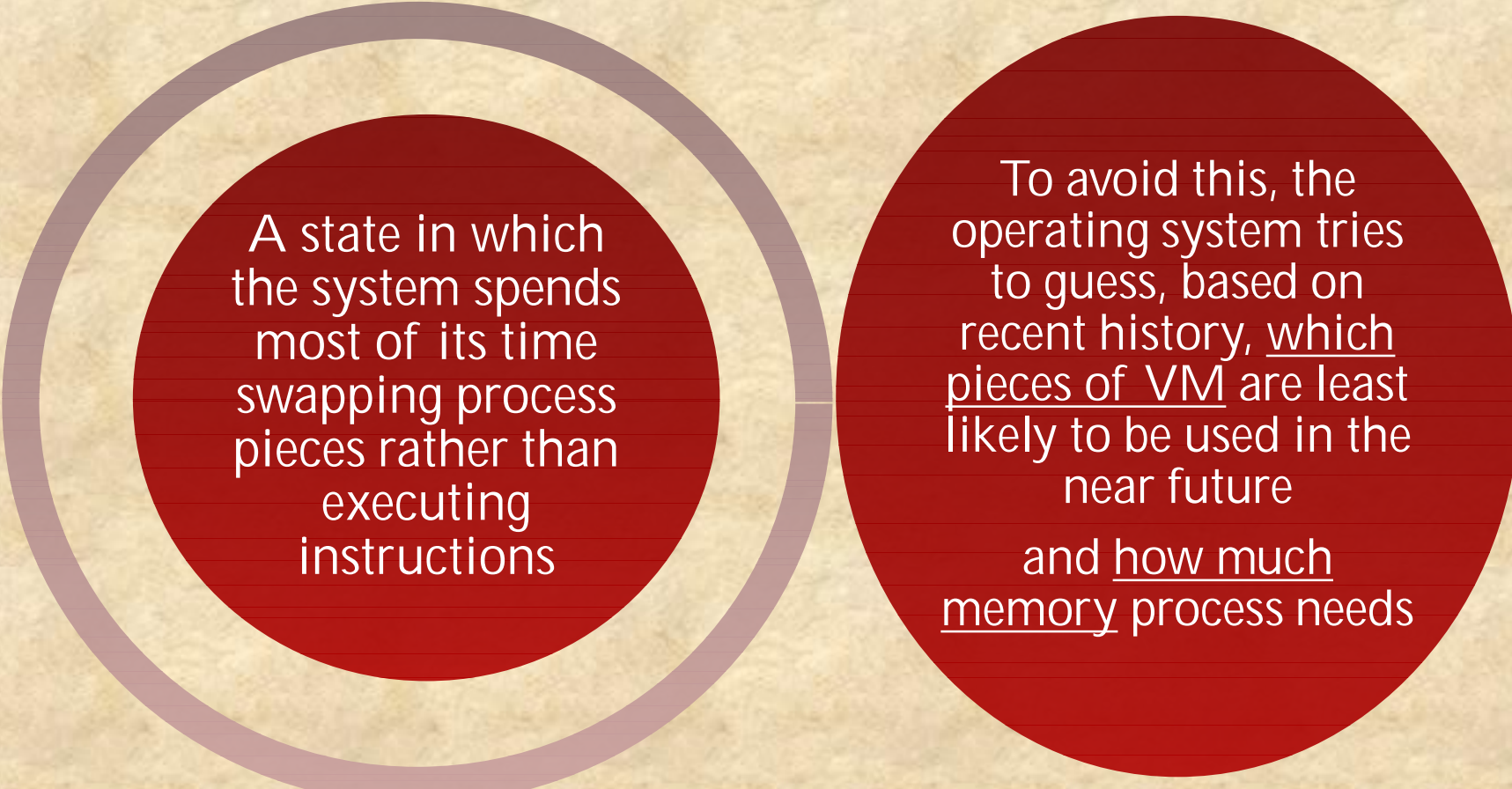
Paging and

Segmentation

Segmented paging?
Multi-level paging?

Simple Paging	Virtual Memory Paging	Simple Segmentation	Virtual Memory Segmentation
Main memory partitioned into small fixed-size chunks called frames	Main memory partitioned into small fixed-size chunks called frames	Main memory not partitioned	Main memory not partitioned
Program broken into pages by the compiler or memory management system	Program broken into pages by the compiler or memory management system	Program segments specified by the programmer to the compiler (i.e., the decision is made by the programmer)	Program segments specified by the programmer to the compiler (i.e., the decision is made by the programmer)
Internal fragmentation within frames	Internal fragmentation within frames	No internal fragmentation	No internal fragmentation
No external fragmentation	No external fragmentation	External fragmentation	<u>External fragmentation</u>
Operating system must maintain a page table for each process showing which frame each page occupies	Operating system must maintain a page table for each process showing which frame each page occupies	Operating system must maintain a segment table for each process showing the load address and length of each segment	Operating system must maintain a segment table for each process showing the load address and length of each segment
Operating system must maintain a free frame list	Operating system must maintain a free frame list	Operating system must maintain a list of free holes in main memory	Operating system must maintain a list of free holes in main memory
Processor uses page number, offset to calculate absolute address	Processor uses page number, offset to calculate absolute address	Processor uses segment number, offset to calculate absolute address	Processor uses segment number, offset to calculate absolute address
All the pages of a process must be in main memory for process to run, unless overlays are used	Not all pages of a process need be in main memory frames for the process to run. Pages may be read in as needed	All the segments of a process must be in main memory for process to run, unless overlays are used	Not all segments of a process need be in main memory for the process to run. Segments may be read in as needed
	Reading a page into main memory may require writing a page out to disk		Reading a segment into main memory may require writing one or more segments out to disk

Thrashing



A state in which the system spends most of its time swapping process pieces rather than executing instructions

The diagram features two large red circles on a light beige background. The left circle is nested within a larger, light purple circle. The right circle is positioned to the right of the left one. Both circles contain white text. The top of the slide has a yellow textured banner with the title 'Thrashing' in blue.

To avoid this, the operating system tries to guess, based on recent history, which pieces of VM are least likely to be used in the near future and how much memory process needs

Principle of Locality

- n Program and data references within a process tend to cluster
- n Only a few pieces of a process will be needed over a short period of time
- n Therefore it is possible to make intelligent guesses about which pieces will be needed in the future
- n Avoids thrashing



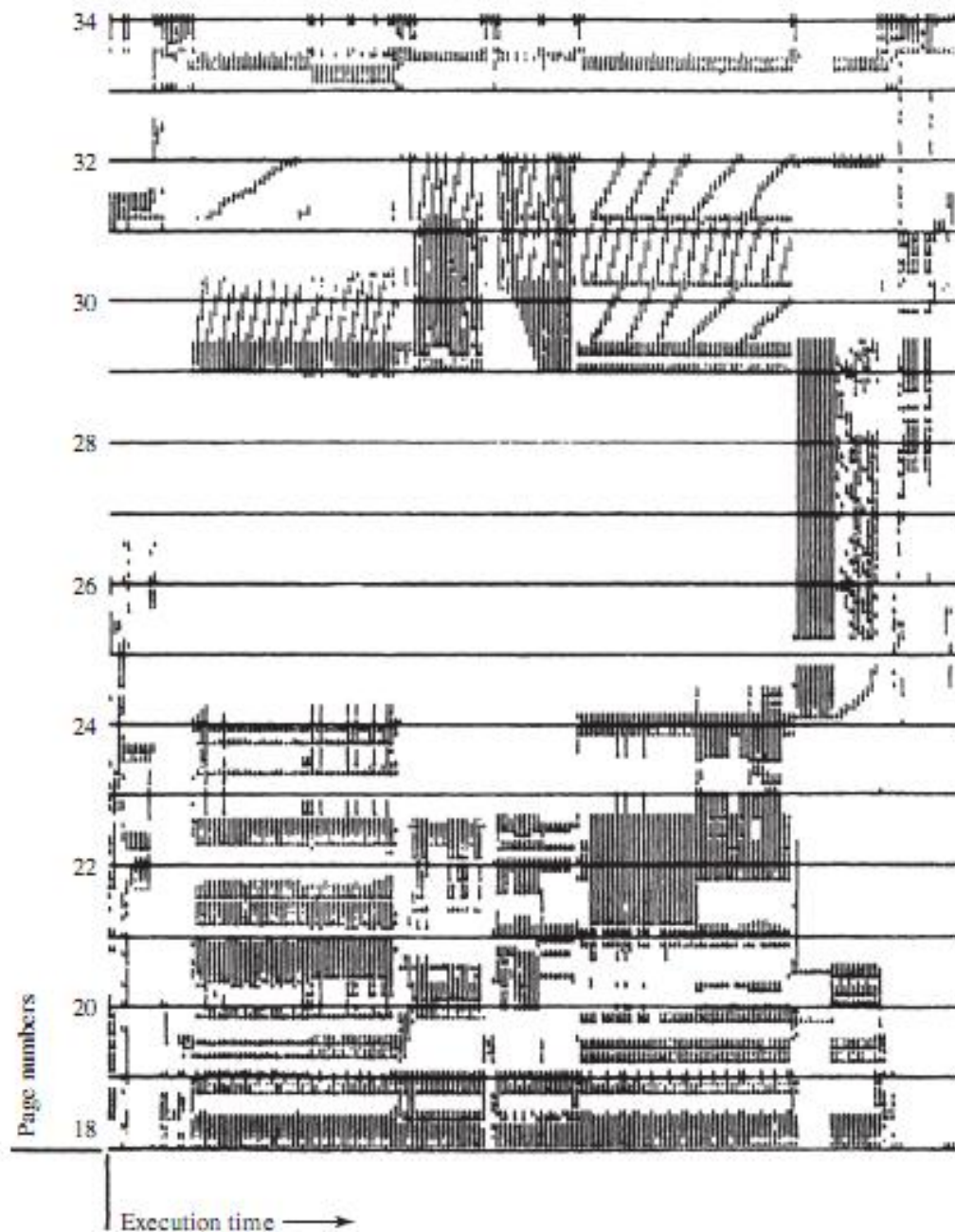


Figure 8.1 Paging Behavior

Copyright William Stallings & Teemu Kerola 2015

Paging Behavior

- n During the lifetime of the process, references are confined to a subset of pages
- n That subset will change over time
- n That subset will change from one execution to another

Support Needed for Virtual Memory

For virtual memory to be practical and effective:

- Hardware must support paging and segmentation
- Operating system must include software for managing the movement of pages and/or segments between secondary memory and main memory

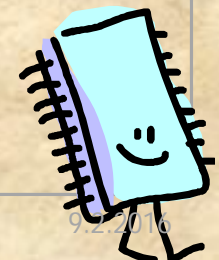
TLB

Paging

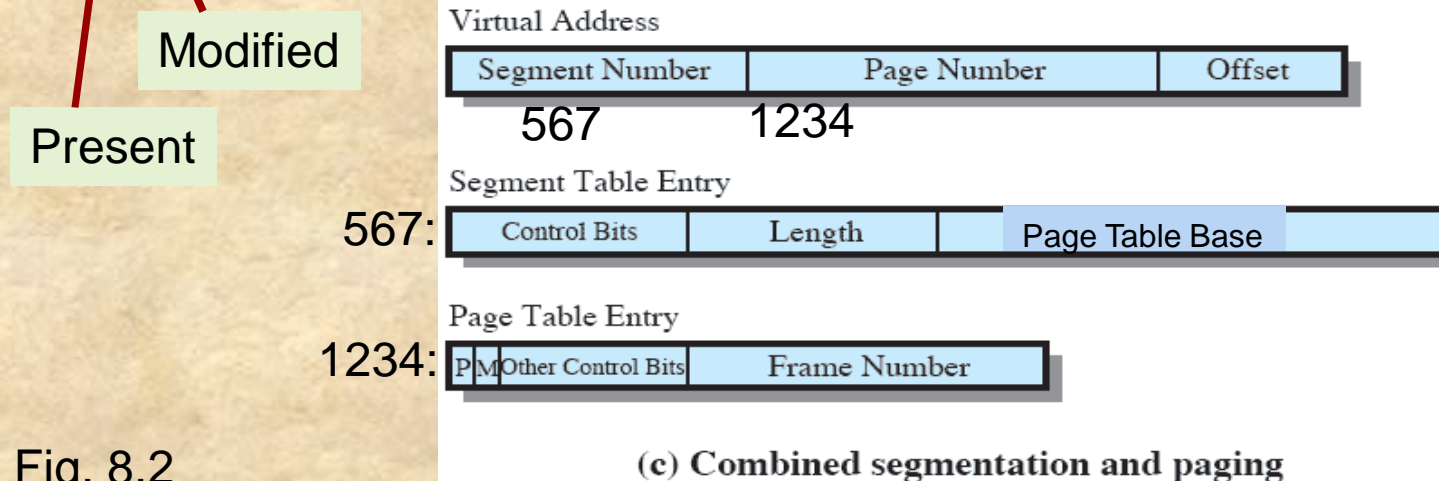
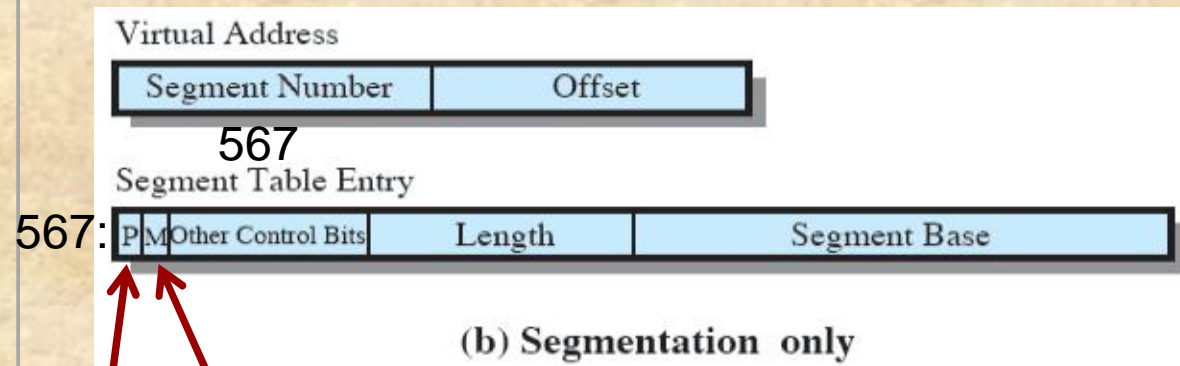
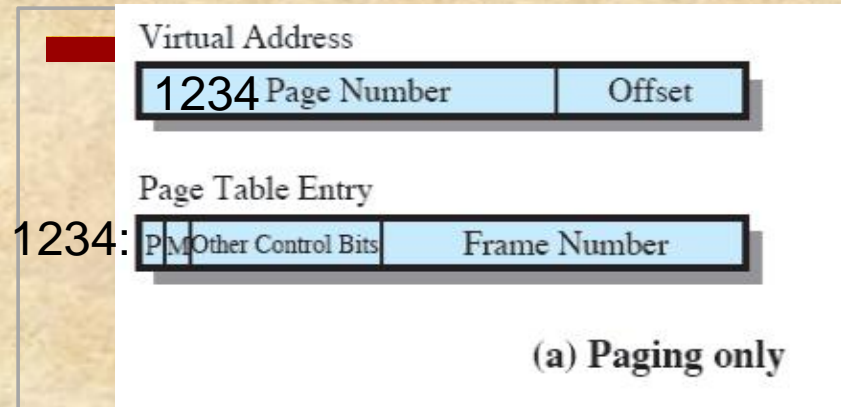
- n The term *virtual memory* is usually associated with systems that employ paging
- n Use of paging to achieve virtual memory was first reported for the Atlas computer (1962)
- n Each process has its own page table
 - n Each page table entry contains the frame number of the corresponding page in main memory



University of Manchester Atlas, 1962



Memory Management Formats



P = present bit
M = Modified bit

Fig. 8.2

Address Translation

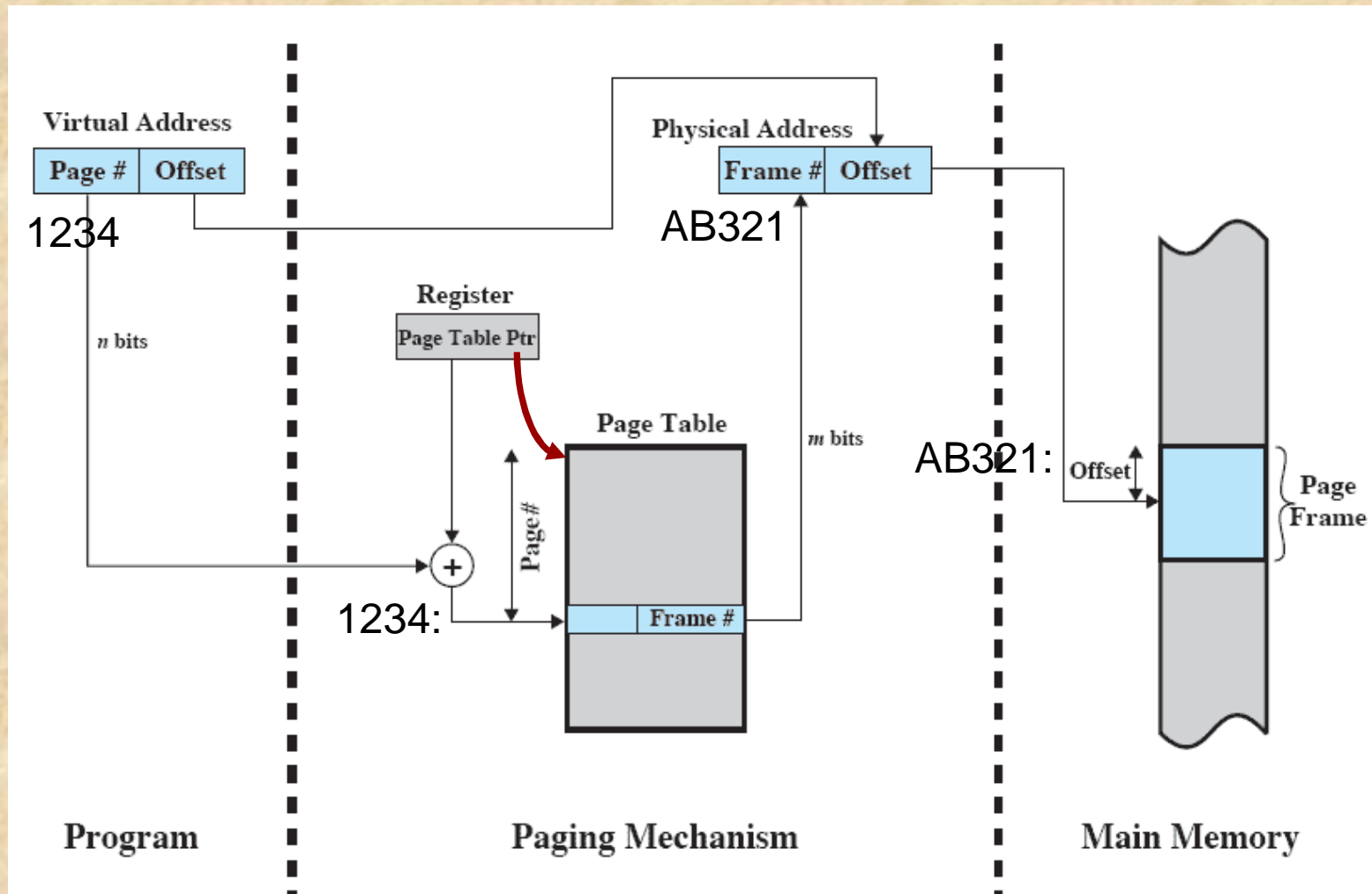
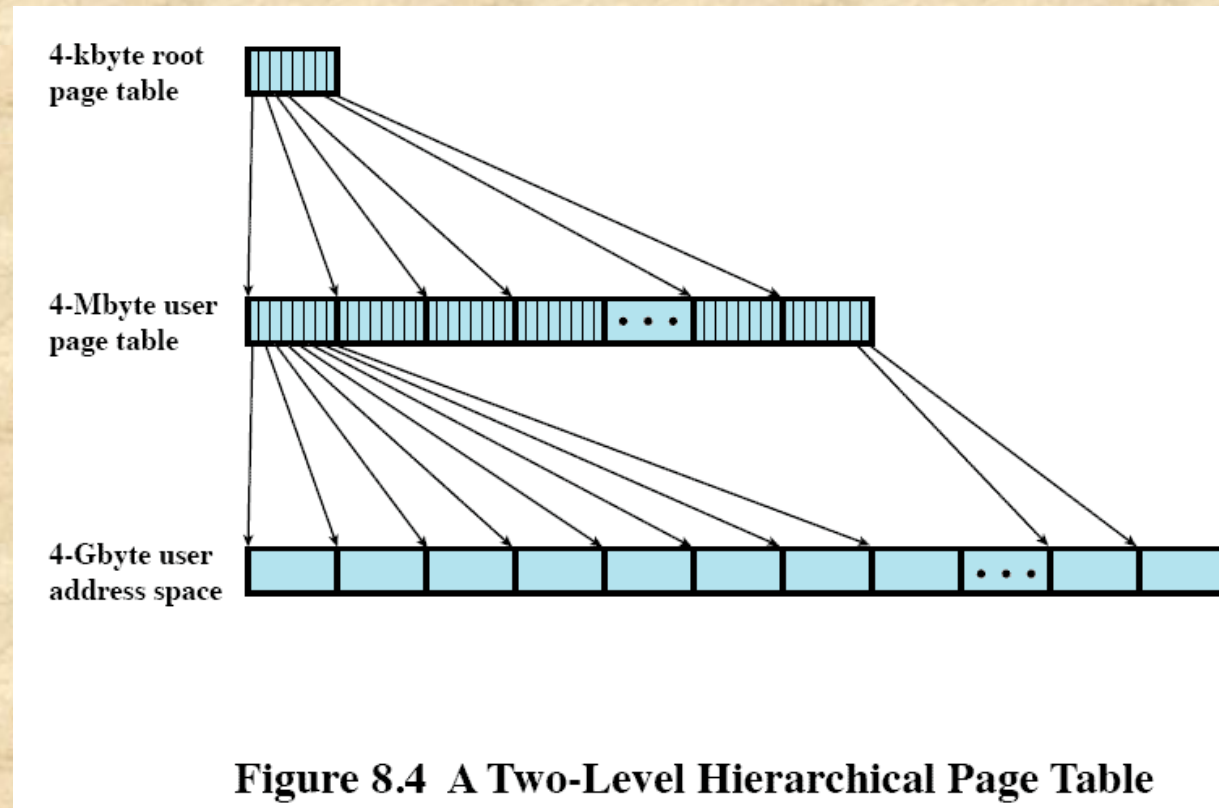


Figure 8.3 Address Translation in a Paging System

Two-Level Hierarchical Page Table



Address Translation for 2-level Page Table

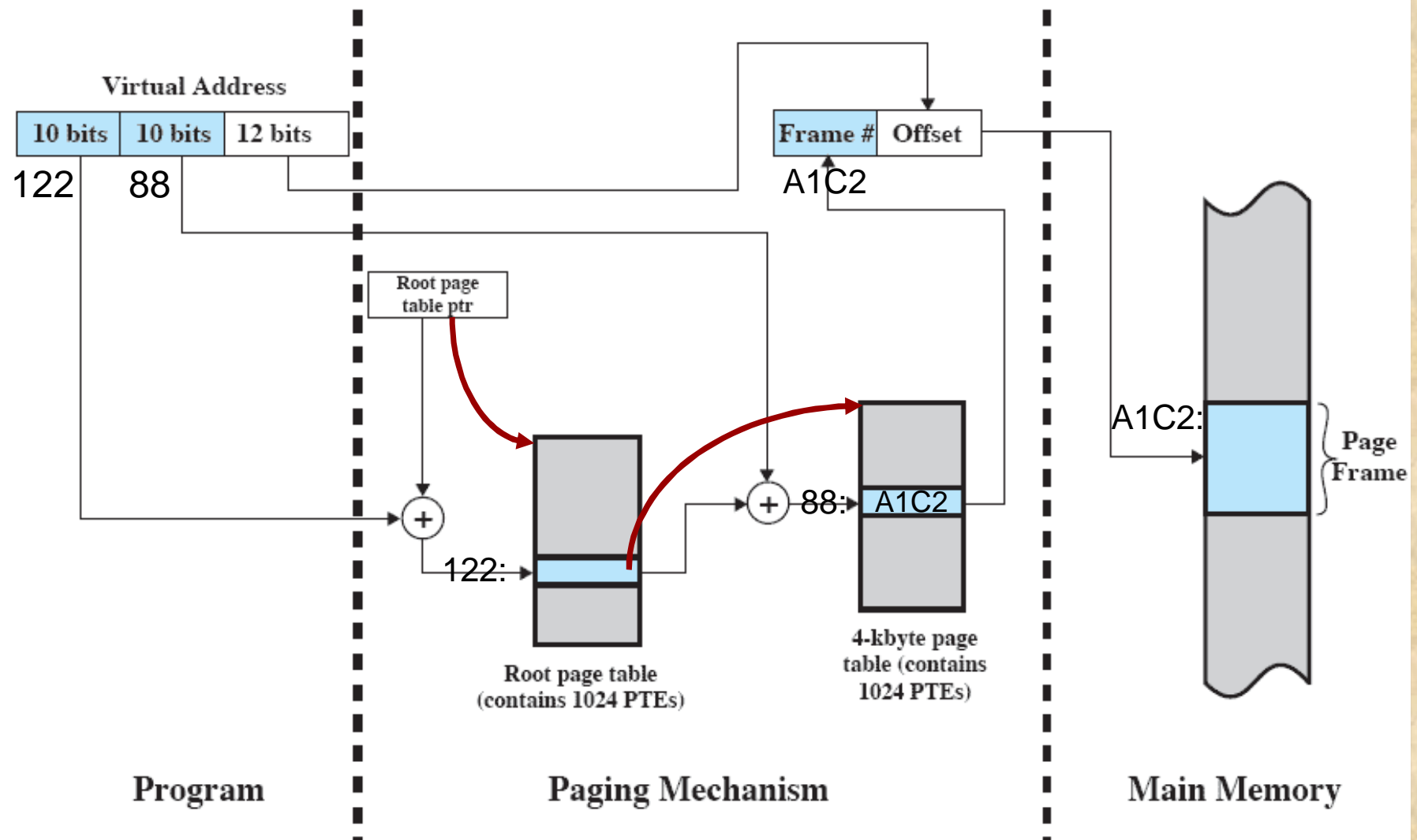
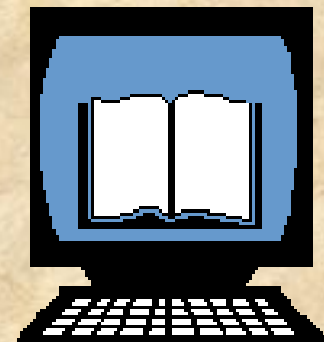


Figure 8.5 Address Translation in a Two-Level Paging System

Discuss

Inverted Page Table

- n Page number portion of a virtual address is mapped into a hash value
 - n hash value points to inverted page table
- n Fixed proportion of real memory is required for the tables regardless of the number of processes or virtual pages supported
- n Structure is called *inverted* because it indexes page table entries by frame number rather than by virtual page number



Inverted Page Table

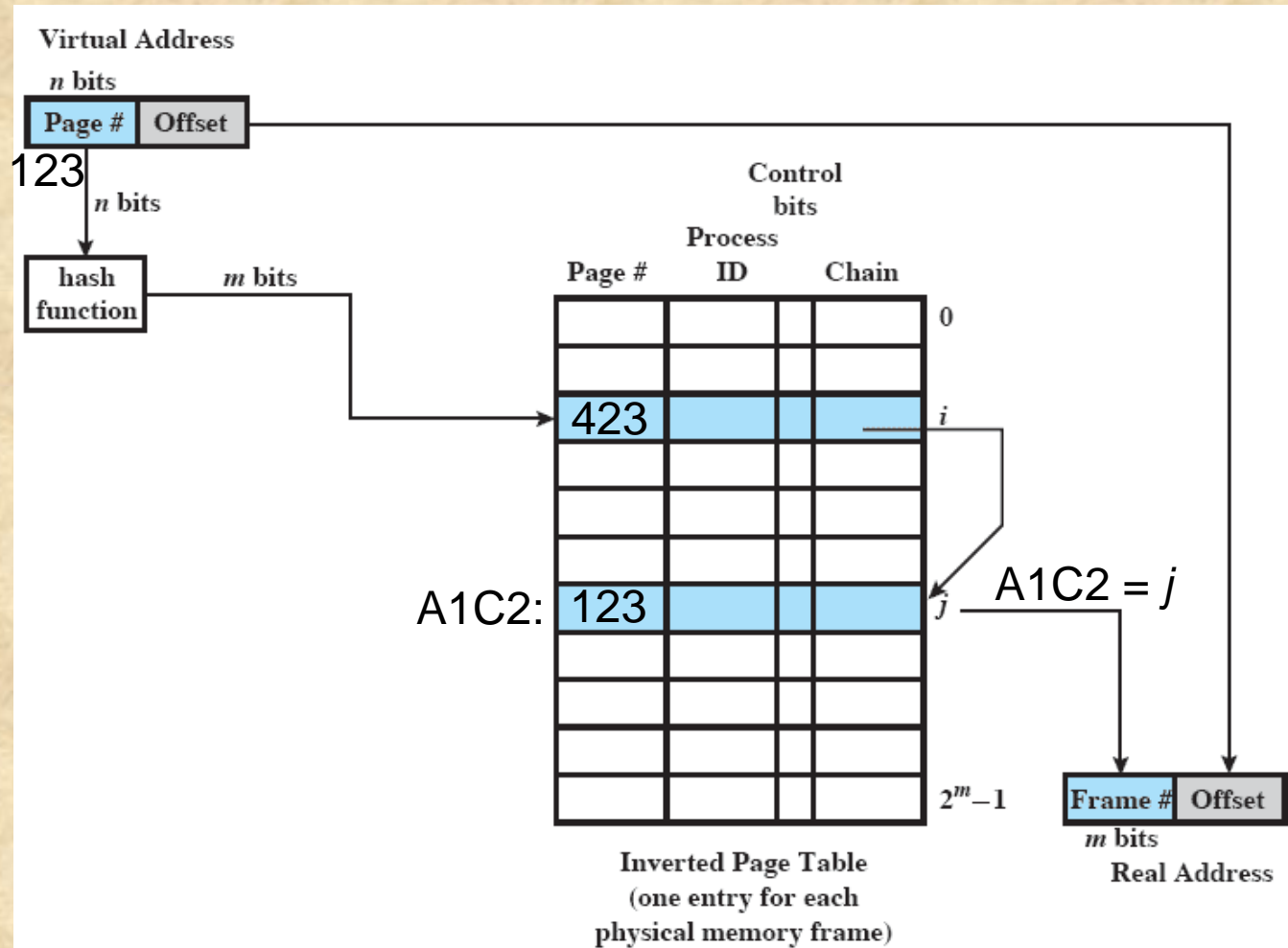
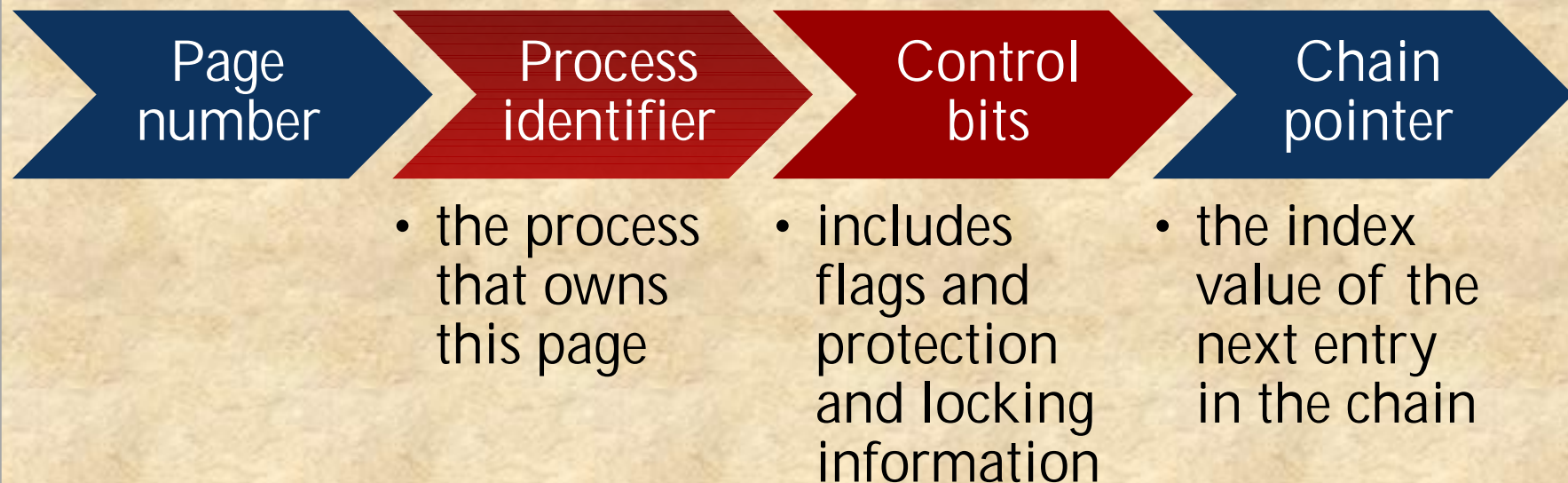


Figure 8.6 Inverted Page Table Structure

Discuss

Inverted Page Table

Each entry in the page table includes:



Translation Lookaside Buffer (TLB)

- n Each virtual memory reference can cause two (or many) physical memory accesses:
 - n one (or more) to fetch the page table entry
 - n one to load/store the data
- n To overcome the effect of doubling the memory access time, most virtual memory schemes make use of a special high-speed cache called a *translation lookaside buffer (TLB)*
 - n *Hw-assistance*

Use of a TLB

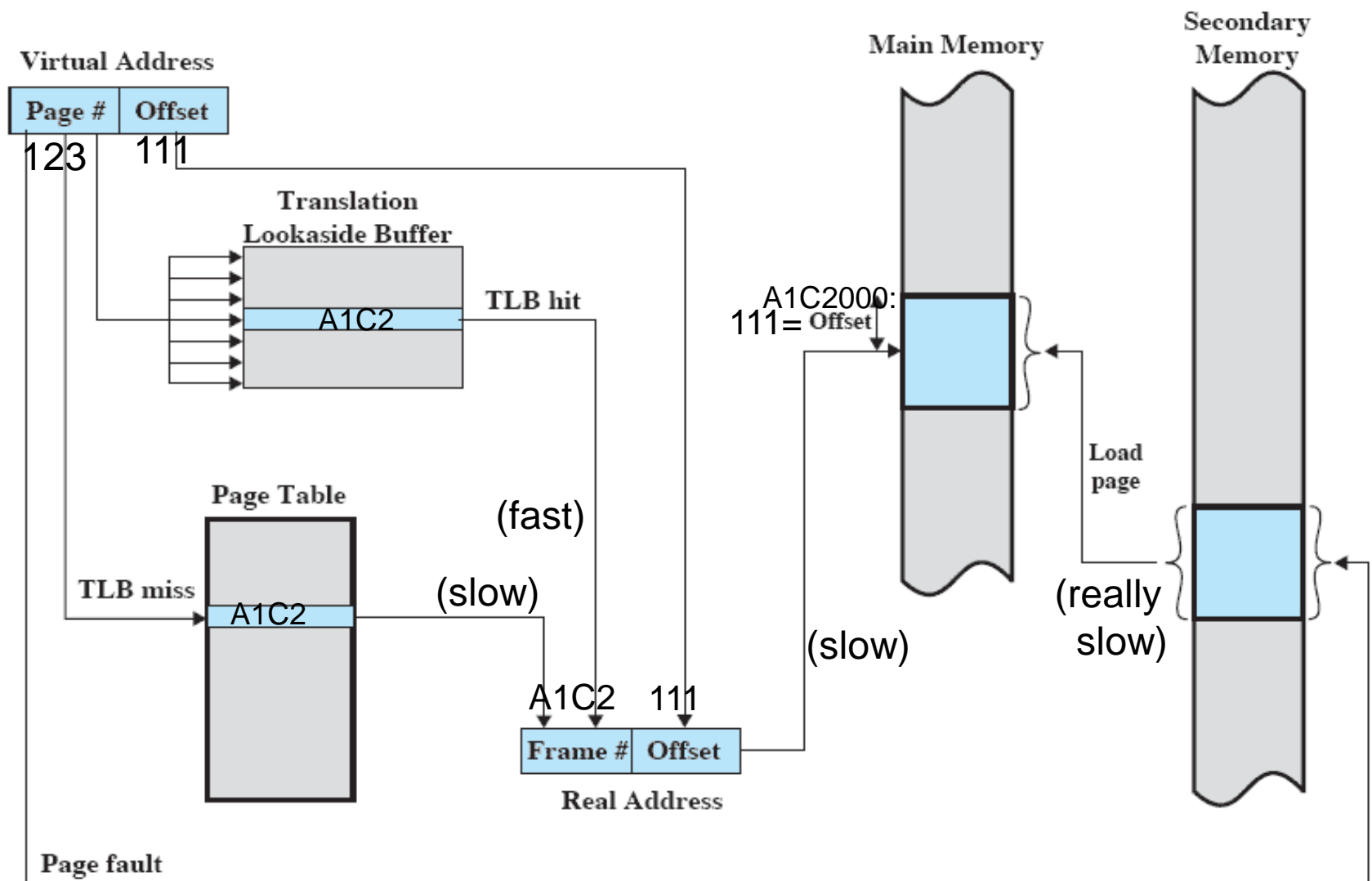


Figure 8.7 Use of a Translation Lookaside Buffer

TLB Operation

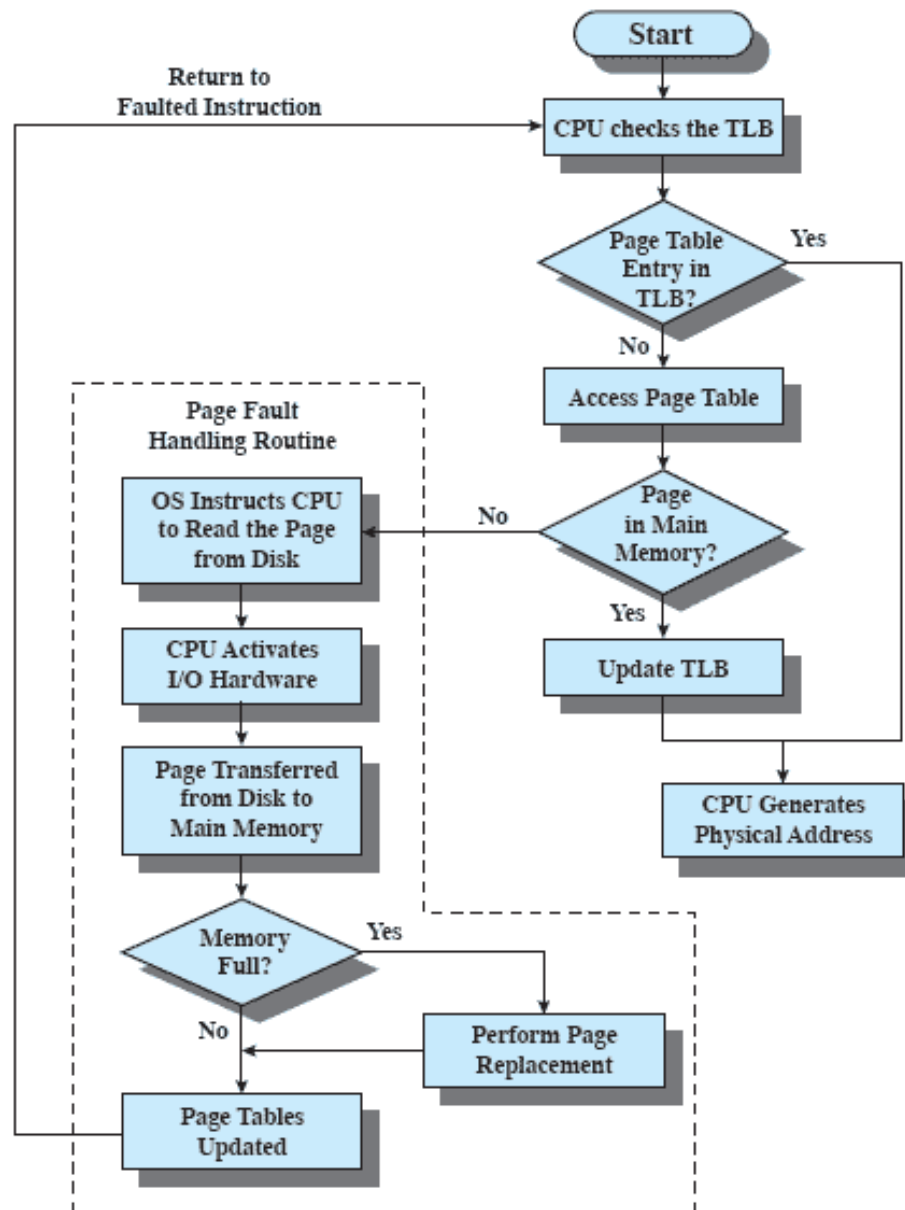
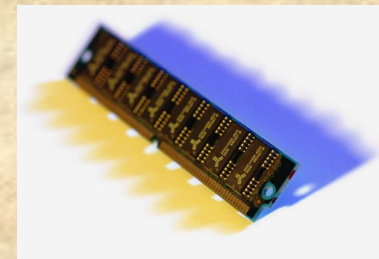


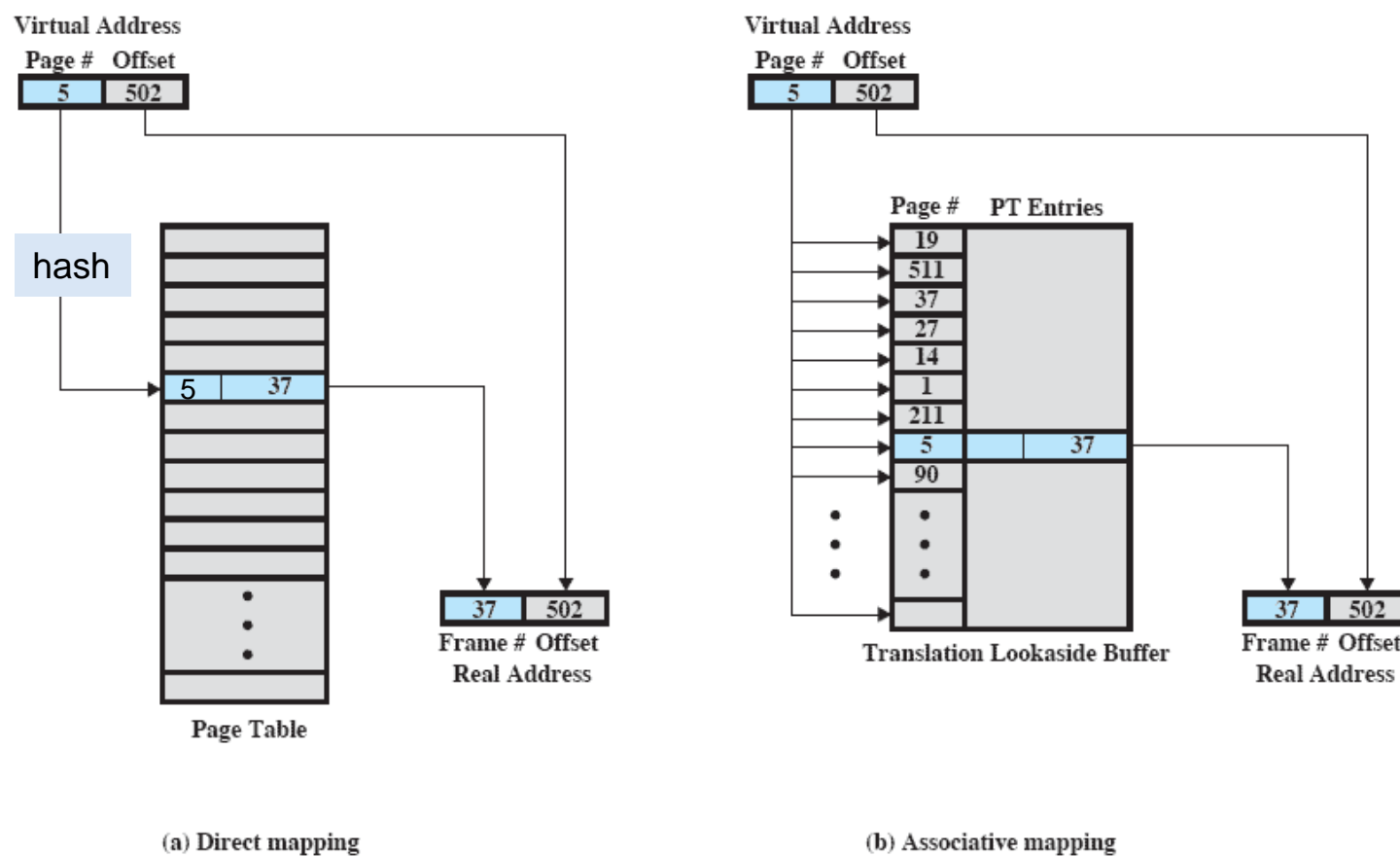
Figure 8.8 Operation of Paging and Translation Lookaside Buffer (TLB) [FURH87]

Associative Mapping

- n The TLB only contains some of the page table entries so we cannot simply index into the TLB based on page number
- n each TLB entry must include the page number as well as the complete page table entry
- n The processor is equipped with hardware that allows it to interrogate simultaneously a number of TLB entries to determine if there is a match on page number



Direct Versus Associative Lookup



Just one

(c) Set associative mapping (set size 2? 4?)

TLB and Cache

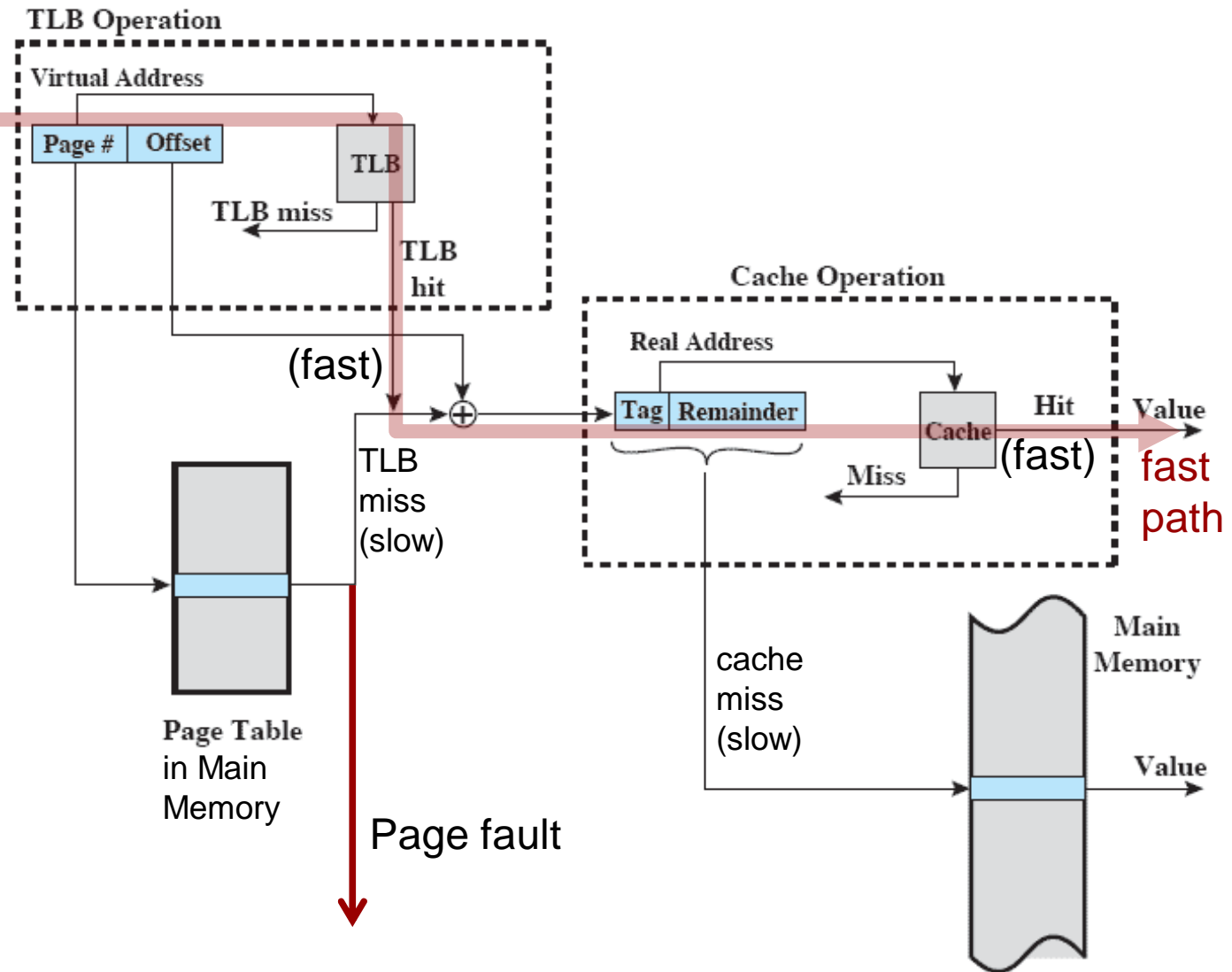
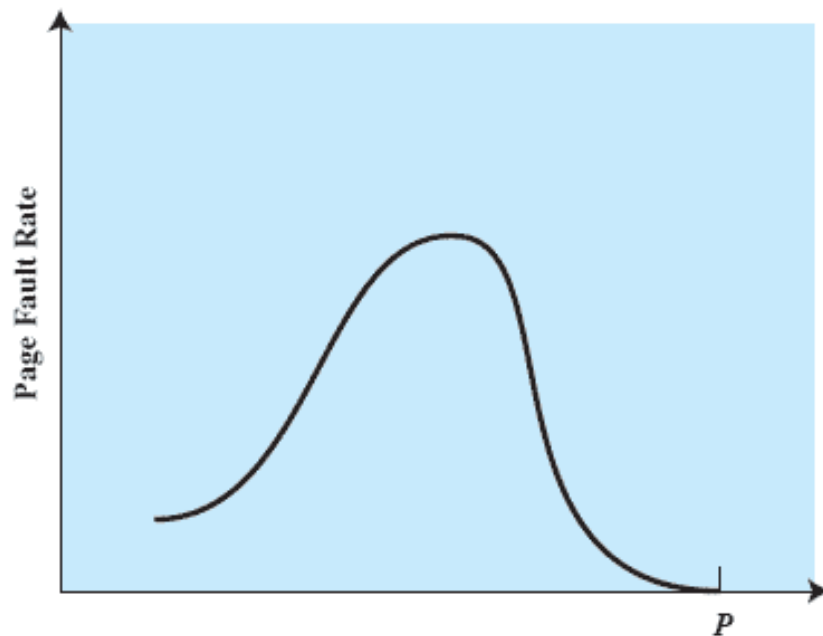


Figure 8.10 Translation Lookaside Buffer and Cache Operation

Page Size

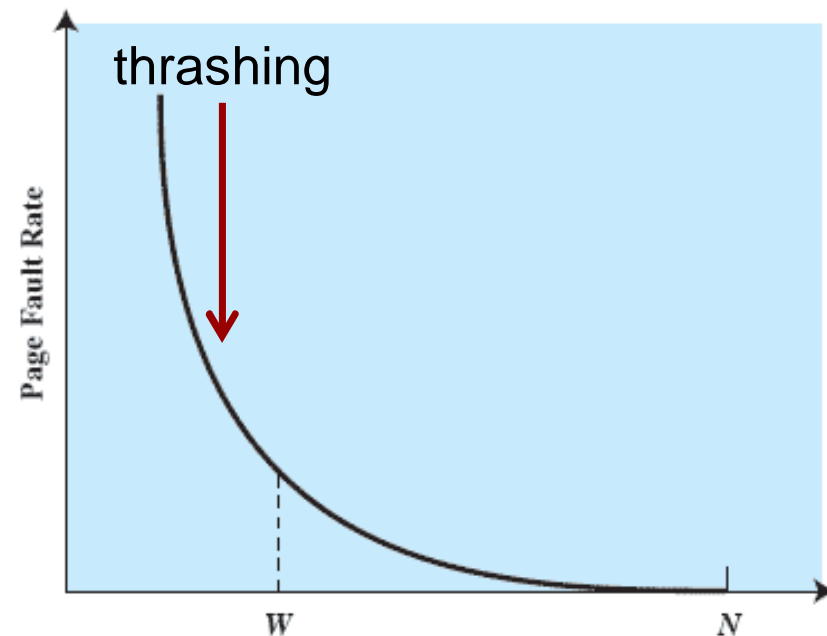
- n The smaller the page size, the lesser the amount of internal fragmentation
- n however, more pages are required per process
- n more pages per process means larger page tables
- n for large programs in a heavily multiprogrammed environment some portion of the page tables of active processes must be in virtual memory instead of main memory
- n the physical characteristics of most secondary-memory devices favor a larger page size for more efficient block transfer of data

Paging Behavior of a Program



(a) Page Size

P = size of entire process
 W = working set size
 N = total number of pages in process



(b) Number of Page Frames Allocated
(allocated memory size)

Large pages good for spatial locality
Many pages good for temporal locality

Figure 8.11 Typical Paging Behavior of a Program

Example: Page Sizes

Computer	Page Size
Atlas	512 48-bit words
Honeywell-Multics	1024 36-bit words
IBM 370/XA and 370/ESA	4 Kbytes
VAX family	512 bytes
IBM AS/400	512 bytes
DEC Alpha	8 Kbytes
MIPS	4 Kbytes to 16 Mbytes
UltraSPARC	8 Kbytes to 4 Mbytes
Pentium	4 Kbytes or 4 Mbytes
IBM POWER	4 Kbytes
Itanium	4 Kbytes to 256 Mbytes

Page Size

The design issue of page size is related to the size of physical main memory and program size



main memory is getting larger and address space used by applications is also growing



most obvious on personal computers where applications are becoming increasingly complex

- n Contemporary programming techniques used in large programs tend to decrease the locality of references within a process

Segmentation

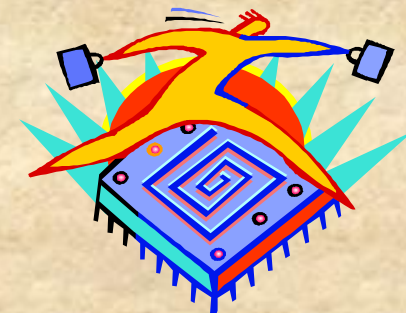
- n Segmentation allows the programmer to view memory as consisting of multiple address spaces or segments

Advantages:

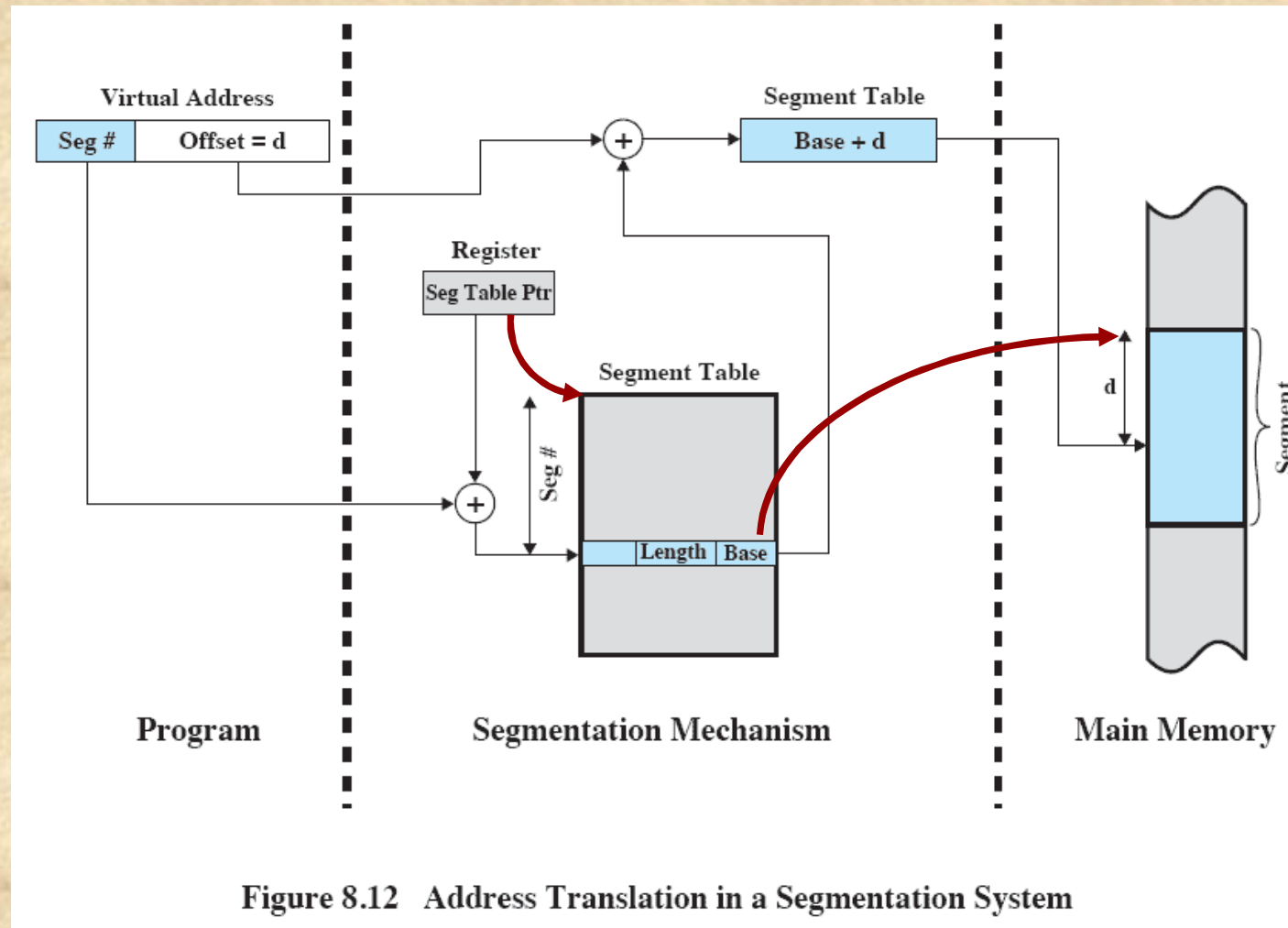
- Simplifies handling of growing data structures
- Allows programs to be altered and recompiled independently
- Lends itself to sharing data among processes
- Lends itself to protection

Segment Organization

- n Each segment table entry contains the starting address of the corresponding segment in main memory and the length of the segment
- n A bit is needed to determine if the segment is already in main memory
- n Another bit is needed to determine if the segment has been modified since it was loaded in main memory



Address Translation



Combined Paging and Segmentation

In a combined paging/segmentation system a user's address space is broken up into a number of segments. Each segment is broken up into a number of fixed-sized pages which are equal in length to a main memory frame

Segmentation is visible to the programmer

Paging is transparent to the programmer

Segmented Paging Address Translation

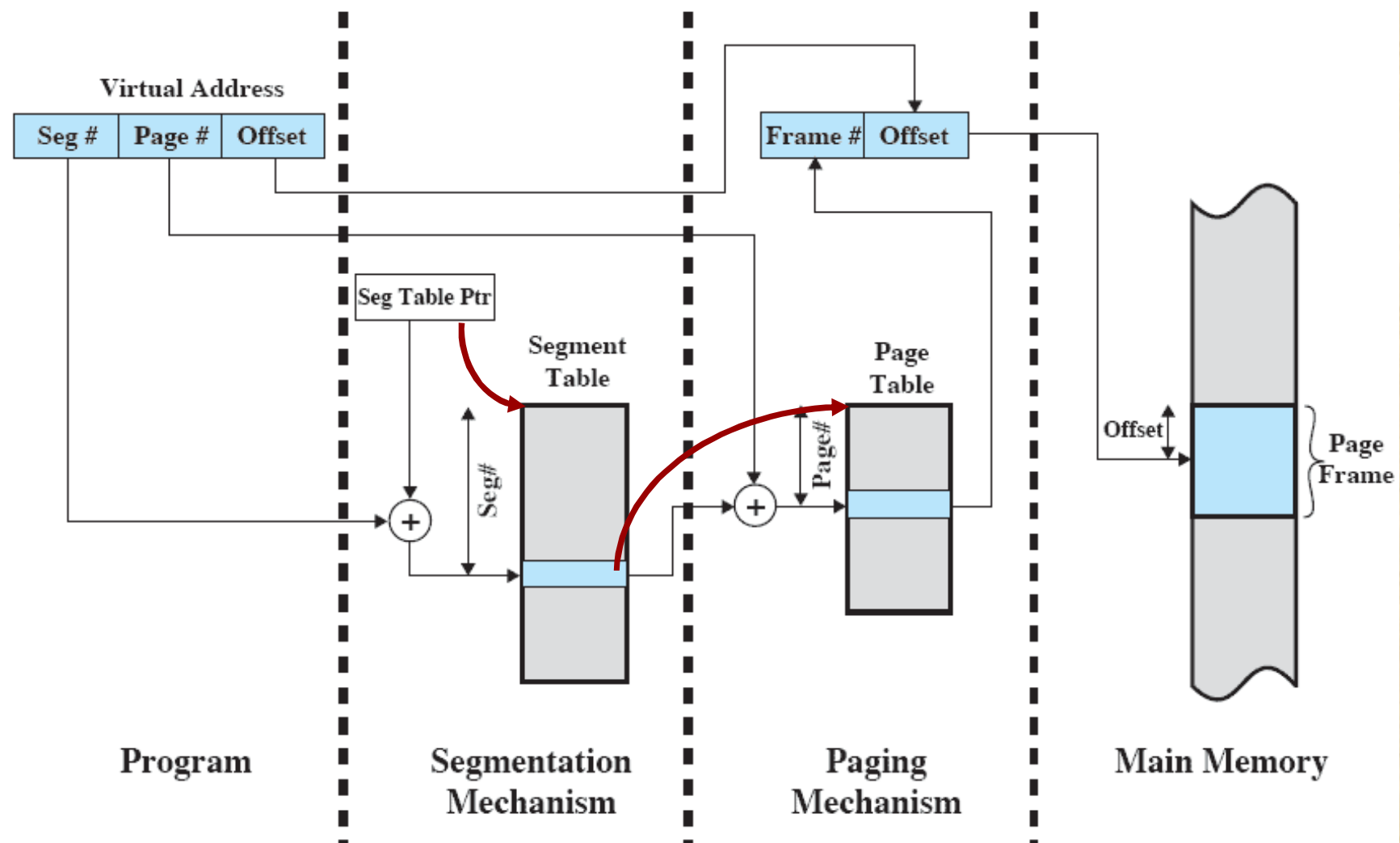


Figure 8.13 Address Translation in a Segmentation/Paging System

Combined Segmentation and Paging

Virtual Address



Segment Table Entry



Page Table Entry

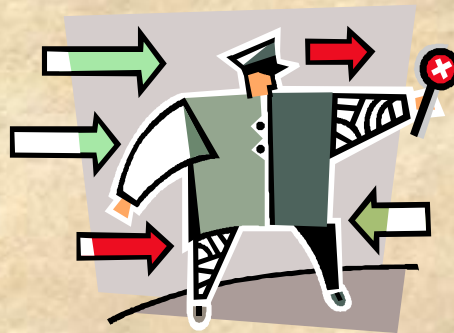


P = present bit
M = Modified bit

(c) Combined segmentation and paging

Protection and Sharing

- n Segmentation lends itself to the implementation of protection and sharing policies
- n Each entry has a base address and length so inadvertent memory access can be controlled
- n Sharing can be achieved by segments referencing multiple processes



Protection Relationships

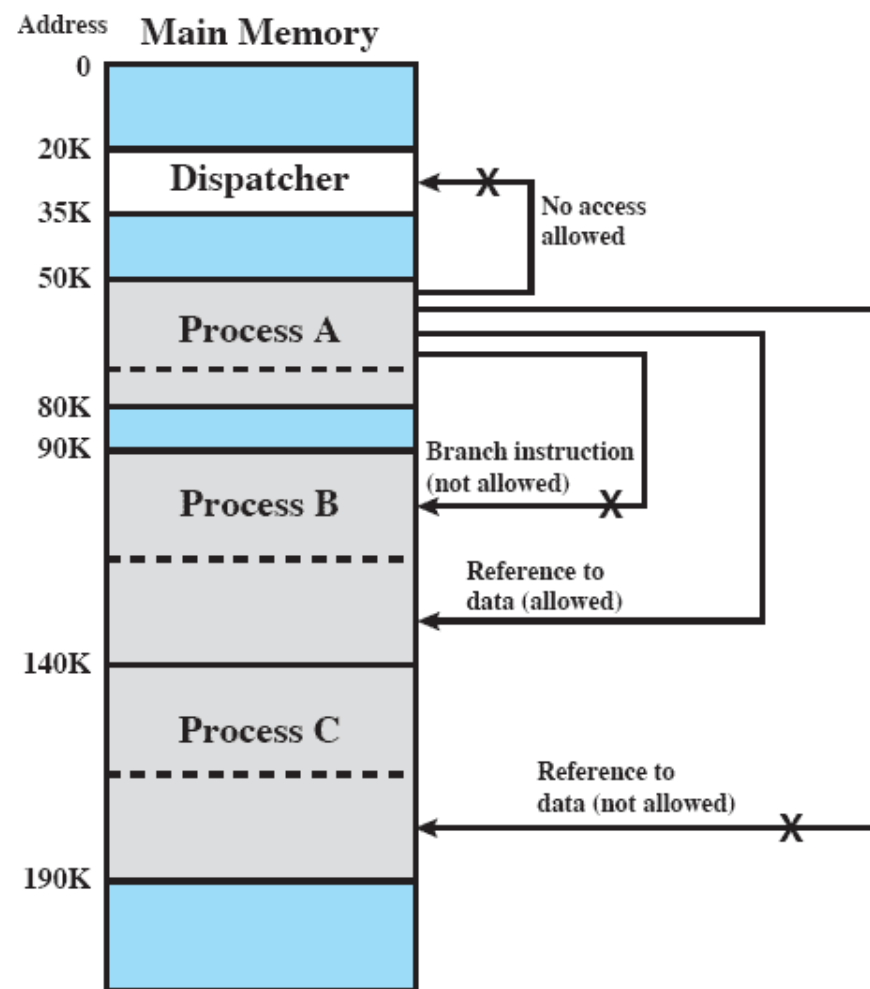


Figure 8.14 Protection Relationships Between Segments

Operating System Software

The design of the memory management portion of an operating system depends on three fundamental areas of choice:

- whether or not to use virtual memory techniques
- the use of paging or segmentation or both
- the algorithms employed for various aspects of memory management

Policies for Virtual Memory

n Key issue: Performance

§ minimize page faults

Fetch Policy Demand paging Prepaging	noutopolitiikka tarvesivutus	Resident Set Management Resident set size Fixed Variable Replacement Scope Global Local	käyttäjöjoukon koko
Placement Policy	sijoituspolitiikka		
Replacement Policy Basic Algorithms Optimal Least recently used (LRU) First-in-first-out (FIFO) Clock Page Buffering	poistopolitiikka	Cleaning Policy Demand Precleaning	puhdistuspolitiikka levylle kirjoitus politiikka
		Load Control Degree of multiprogramming	Moniajoasteen säätely

Fetch Policy

noutopolitiikka

n Determines when a page should be brought into memory



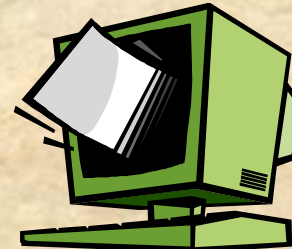
Two main types:

Demand
Paging

Prepaging

Demand Paging

- n Demand Paging
 - n only brings pages into main memory when a reference is made to a location on the page
 - n many page faults when process is first started
 - n principle of locality suggests that as more and more pages are brought in, most future references will be to pages that have recently been brought in, and page faults should drop to a very low level



Prepaging

- n Prepaging
 - n pages other than the one demanded by a page fault are brought in
 - n exploits the characteristics of most secondary memory devices
 - n if pages of a process are stored contiguously in secondary memory it is more efficient to bring in a number of pages at one time
 - n ineffective if extra pages are not referenced
 - n should not be confused with “swapping”

Placement Policy

- n Determines where in real memory a process piece is to reside
- n Important design issue in a segmentation system
- n With paging or combined paging with segmentation placing is irrelevant because hardware performs functions with equal efficiency
- n For NUMA systems an automatic placement strategy is desirable

Replacement Policy

- n Deals with the selection of a page in main memory to be replaced when a new page must be brought in
 - n Objective is that the page that is removed be the page least likely to be referenced in the near future
- n The more elaborate the replacement policy, the greater the hardware and software overhead to implement it

Frame Locking

- § When a frame is locked the page currently stored in that frame may not be replaced
 - § Kernel of the OS as well as key control structures are held in locked frames
 - § I/O buffers and time-critical areas may be locked into main memory frames
 - § Locking is achieved by associating a lock bit with each frame (HW assistance)



Replacement Policy

Basic Algorithms

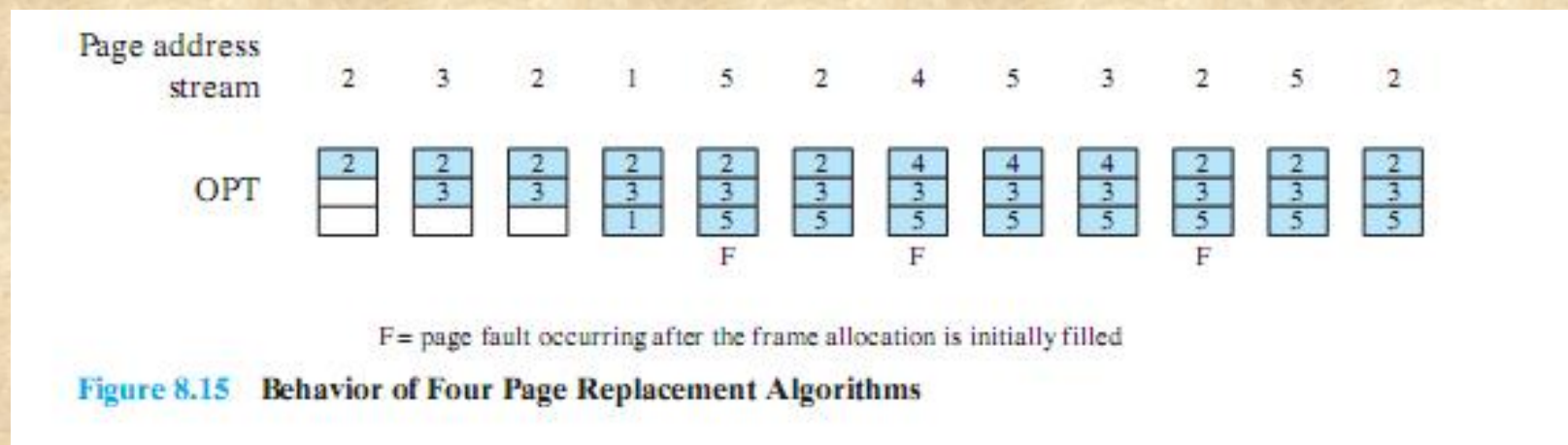


Algorithms used for the selection of a page to replace:

- Optimal
- Least recently used (LRU)
- First-in-first-out (FIFO)
- Clock
- Random (!)

Optimal Policy

- § Selects the page for which the time to the next reference is the longest
- § Produces three page faults after the frame allocation has been filled



Least Recently Used (LRU)

- n Replaces the page that has not been referenced for the longest time
- n By the principle of locality, this should be the page least likely to be referenced in the near future
- n Difficult to implement
 - n One approach is to tag each page with the time of last reference (which requires a great deal of overhead)



LRU Example

Page address stream	2	3	2	1	5	2	4	5	3	2	5	2																																				
LRU	<table><tr><td>2</td></tr><tr><td></td></tr><tr><td></td></tr></table>	2			<table><tr><td>2</td></tr><tr><td>3</td></tr><tr><td></td></tr></table>	2	3		<table><tr><td>2</td></tr><tr><td>3</td></tr><tr><td></td></tr></table>	2	3		<table><tr><td>2</td></tr><tr><td>3</td></tr><tr><td>1</td></tr></table>	2	3	1	<table><tr><td>2</td></tr><tr><td>5</td></tr><tr><td>1</td></tr></table>	2	5	1	<table><tr><td>2</td></tr><tr><td>5</td></tr><tr><td>1</td></tr></table>	2	5	1	<table><tr><td>2</td></tr><tr><td>5</td></tr><tr><td>4</td></tr></table>	2	5	4	<table><tr><td>2</td></tr><tr><td>5</td></tr><tr><td>4</td></tr></table>	2	5	4	<table><tr><td>3</td></tr><tr><td>5</td></tr><tr><td>4</td></tr></table>	3	5	4	<table><tr><td>3</td></tr><tr><td>5</td></tr><tr><td>2</td></tr></table>	3	5	2	<table><tr><td>3</td></tr><tr><td>5</td></tr><tr><td>2</td></tr></table>	3	5	2	<table><tr><td>3</td></tr><tr><td>5</td></tr><tr><td>2</td></tr></table>	3	5	2
2																																																
2																																																
3																																																
2																																																
3																																																
2																																																
3																																																
1																																																
2																																																
5																																																
1																																																
2																																																
5																																																
1																																																
2																																																
5																																																
4																																																
2																																																
5																																																
4																																																
3																																																
5																																																
4																																																
3																																																
5																																																
2																																																
3																																																
5																																																
2																																																
3																																																
5																																																
2																																																
				F		F			F	F																																						

F = page fault occurring after the frame allocation is initially filled

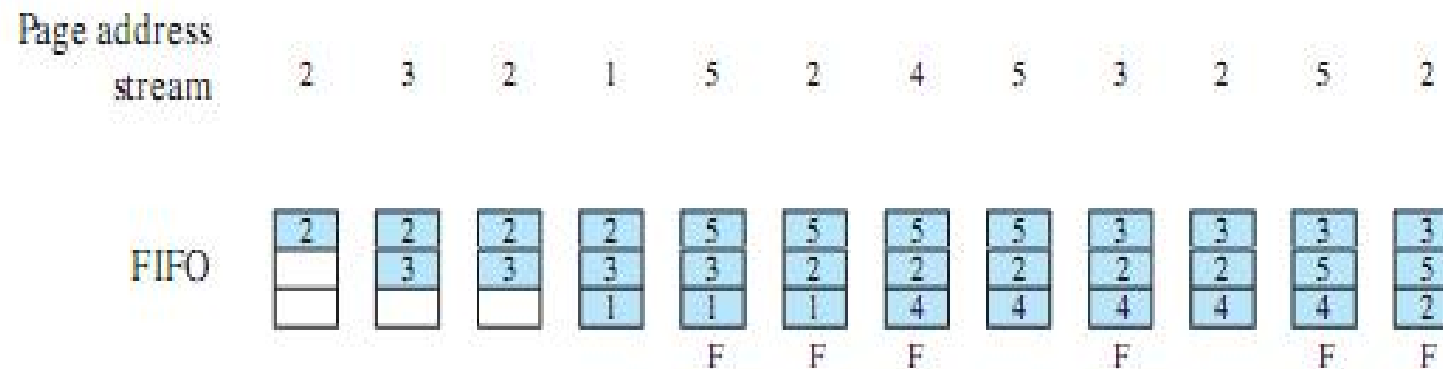
Figure 8.15 Behavior of Four Page Replacement Algorithms

First-in-First-out (FIFO)

- n Treats page frames allocated to a process as a circular buffer
- n Pages are removed in round-robin style
 - § Simple replacement policy to implement
- n Page that has been in memory the longest is replaced



FIFO Example

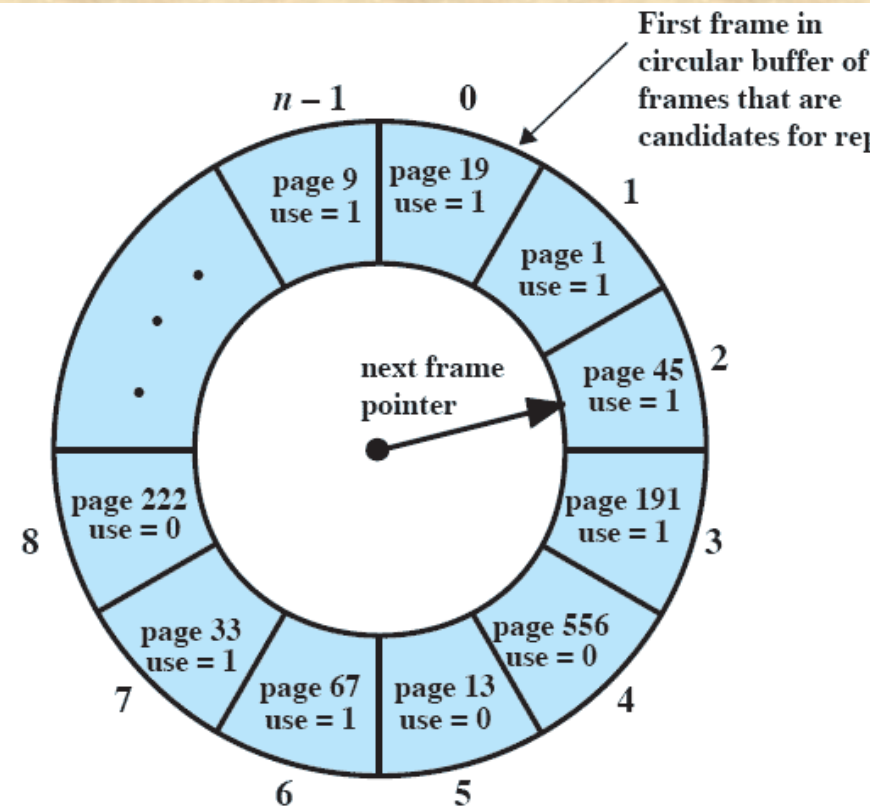


F = page fault occurring after the frame allocation is initially filled

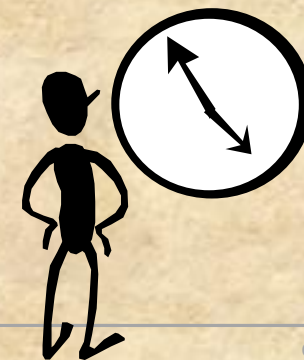
Figure 8.15 Behavior of Four Page Replacement Algorithms

Clock

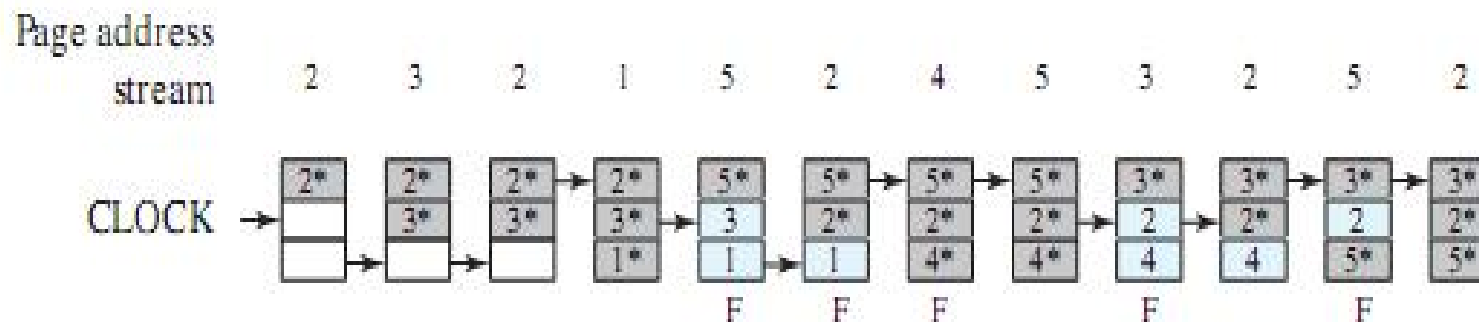
- Requires the association of an additional bit with each frame
 - Referred to as the *use* bit
- When a page is first loaded in memory or referenced, the use bit is set to 1
- The set of frames is considered to be a circular buffer
- Any frame with a use bit of 1 is passed over by the algorithm
- Page frames visualized as laid out in a circle



(a) State of buffer just prior to a page replacement



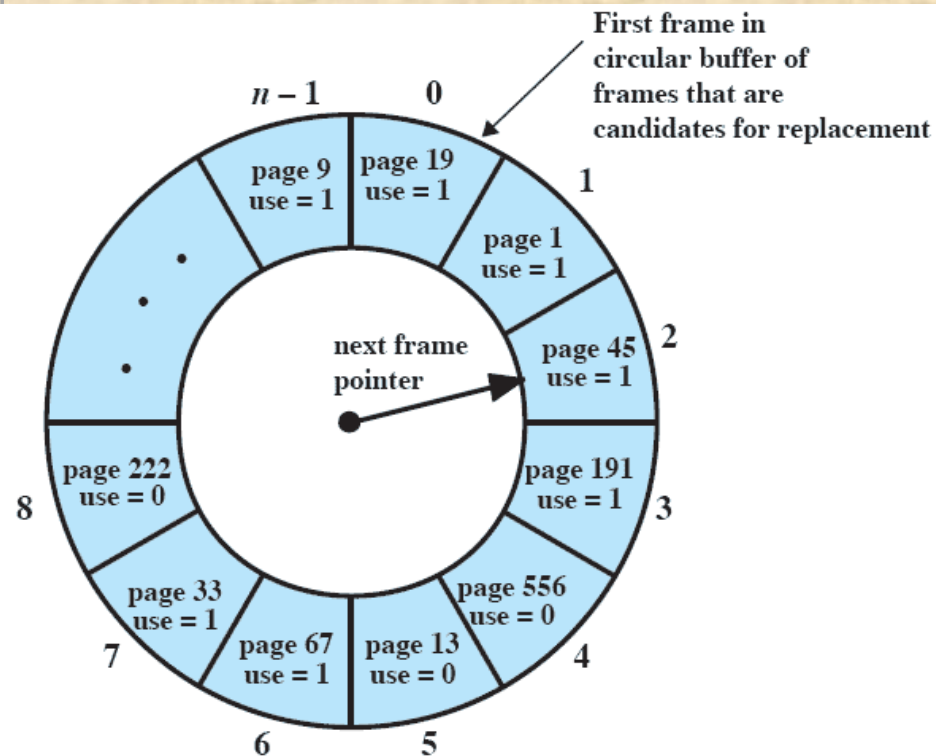
Clock Policy Example



F = page fault occurring after the frame allocation is initially filled

Figure 8.15 Behavior of Four Page Replacement Algorithms

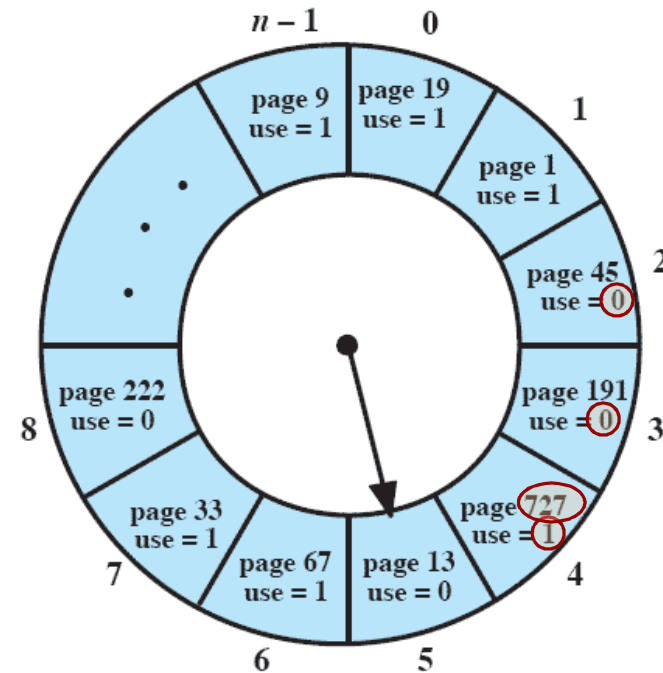
Clock Policy



(a) State of buffer just prior to a page replacement

Reference to page 727

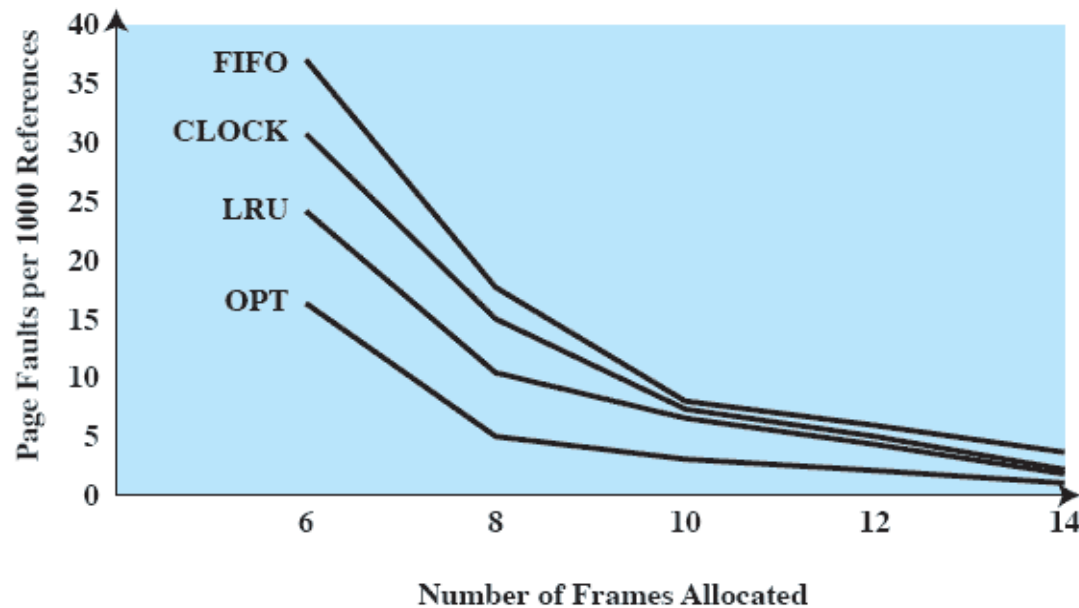
- Page fault
- All frames at use
- Which page is replaced?



(b) State of buffer just after the next page replacement

Figure 8.16 Example of Clock Policy Operation

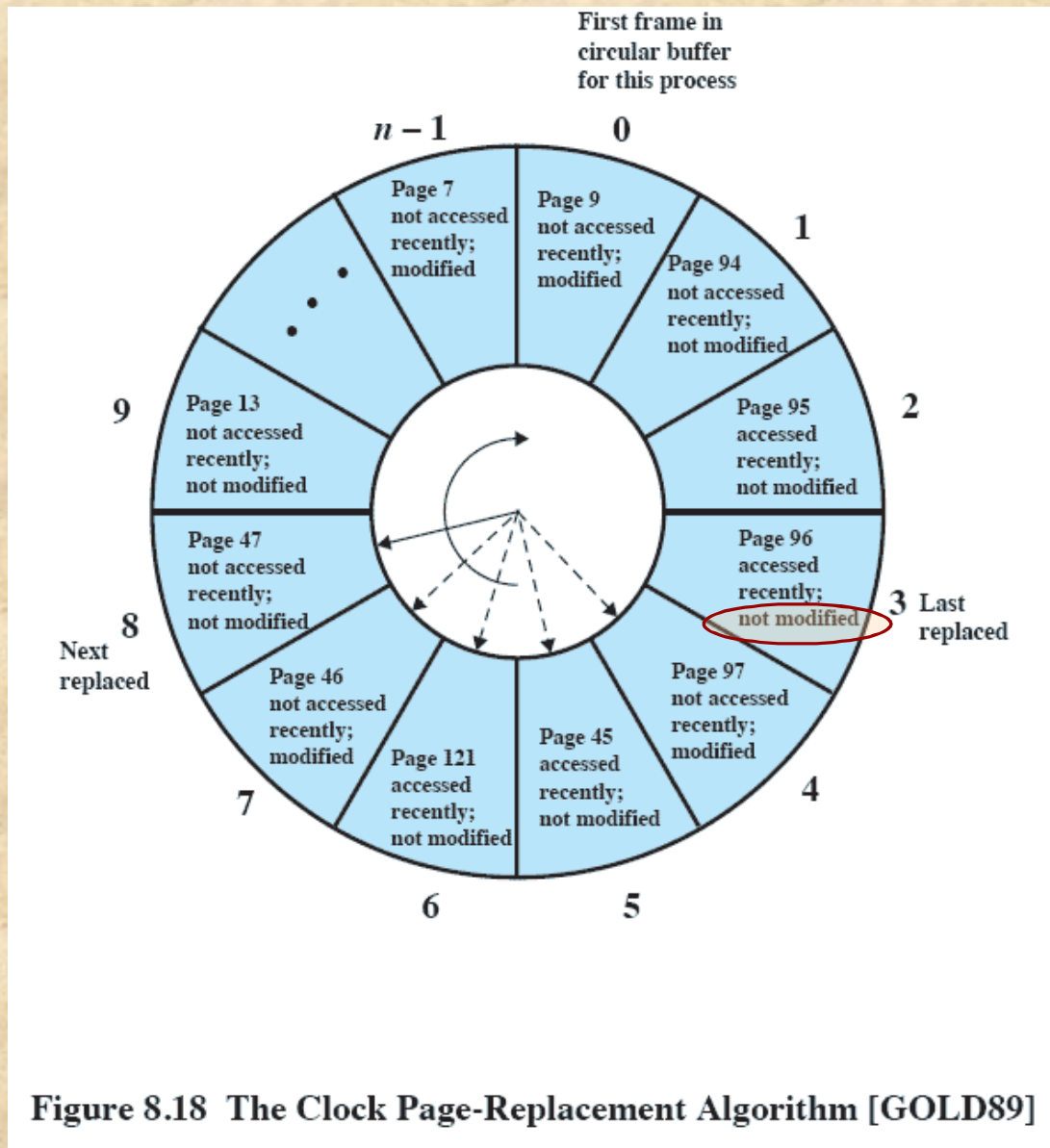
Comparison of Algorithms



Baer 1980 !!

Figure 8.17 Comparison of Fixed-Allocation, Local Page Replacement Algorithms

Clock Policy with Modified Bit



Combined Examples

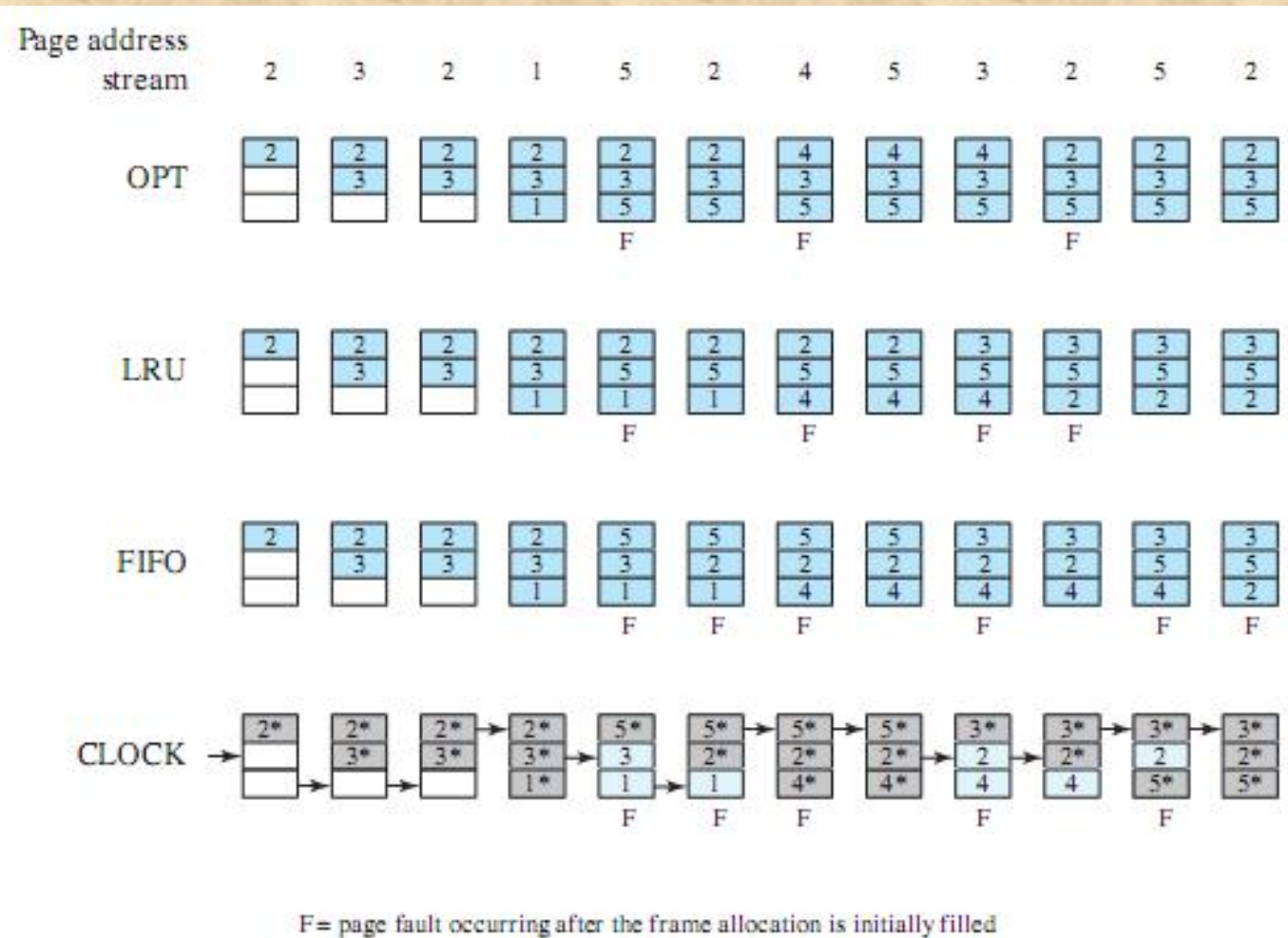
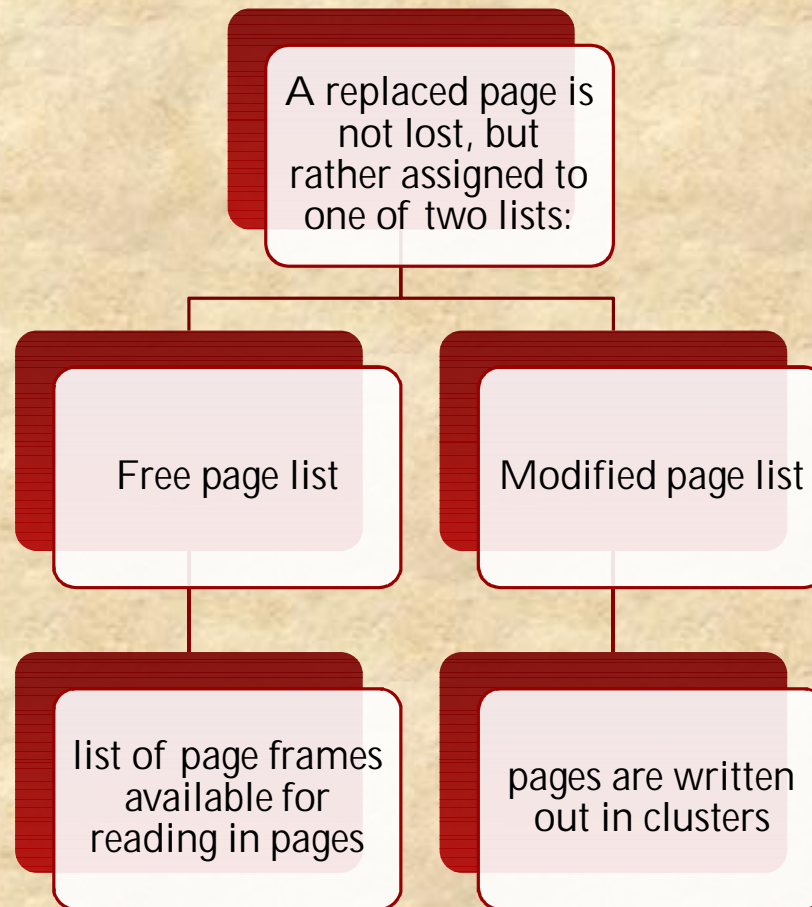


Figure 8.15 Behavior of Four Page Replacement Algorithms

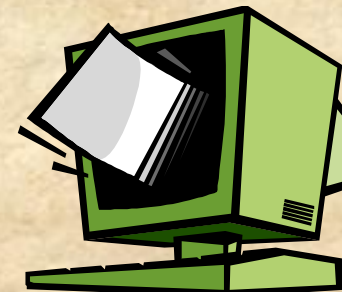
Page Buffering

- Improves paging performance and allows the use of a simpler page replacement policy



Replacement Policy and Cache Size

- n With large caches, replacement of pages can have a performance impact
- n If the page frame selected for replacement is in the cache, that cache block is lost as well as the page that it holds
- n In systems using page buffering, cache performance can be improved with a policy for page placement in the page buffer
- n Most operating systems place pages by selecting an arbitrary page frame from the page buffer



Resident Set Management

- n The OS must decide how many pages to bring into main memory
- n The smaller the amount of memory allocated to each process, the more processes can reside in memory
- n Small number of pages loaded increases page faults
- n Beyond a certain size, further allocations of pages will not effect the page fault rate

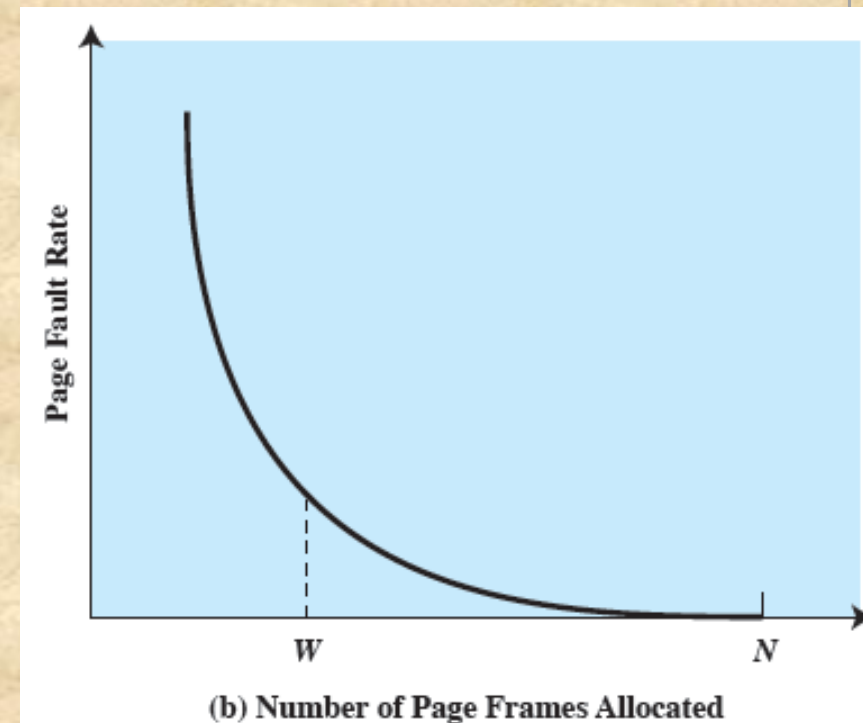


Fig. 8.11

Resident Set Size

Fixed-allocation

- n Gives a process a fixed number of frames in main memory within which to execute
- n When a page fault occurs, one of the pages of that process must be replaced

"fixed partition"?

Variable-allocation

- n Allows the number of page frames allocated to a process to be varied over the lifetime of the process

"run time dynamic partition"?

Replacement Scope

- n The scope of a replacement strategy can be categorized as *global* or *local*
 - n Both types are activated by a page fault when there are no free page frames

Local

- Chooses only among the resident pages of the process that generated the page fault

Global

- Considers all unlocked pages in main memory

Resident Set Management Summary

	Local Replacement	Global Replacement
Fixed Allocation	<ul style="list-style-type: none">•Number of frames allocated to a process is fixed.•Page to be replaced is chosen from among the frames allocated to that process.	<ul style="list-style-type: none">•Not possible.
Variable Allocation	<ul style="list-style-type: none">•The number of frames allocated to a process may be changed from time to time to maintain the working set of the process.•Page to be replaced is chosen from among the frames allocated to that process.	<ul style="list-style-type: none">•Page to be replaced is chosen from all available frames in main memory; this causes the size of the resident set of processes to vary.

Table 8.5

Fixed Allocation, Local Scope

- n Necessary to decide ahead of time the amount of allocation to give a process
- n If allocation is too small, there will be a high page fault rate

If allocation is too large, there will be too few programs in main memory

- increased processor idle time
- increased time spent in swapping

Variable Allocation

Global Scope

- n Easiest to implement
 - n Adopted in a number of operating systems
- n OS maintains a list of free frames
- n Free frame is added to resident set of process when a page fault occurs
- n If no frames are available the OS must choose a page currently in memory
- n One way to counter potential problems is to use page buffering

Variable Allocation Local Scope

- n When a new process is loaded into main memory, allocate to it a certain number of page frames as its resident set
- n When a page fault occurs, select the page to replace from among the resident set of the process that suffers the fault
- n Reevaluate the allocation provided to the process and increase or decrease it to improve overall performance



Variable Allocation Local Scope

- n Decision to increase or decrease a resident set size is based on the assessment of the likely future demands of active processes

Key elements:

- criteria used to determine resident set size
- the timing of changes

Variable Allocation, Local Scope: Working Set Strategy

Fig. 8.19

Sequence of Page References	Window Size, Δ			
	2	3	4	5
24	24	24	24	24
15	24 15	24 15	24 15	24 15
18	15 18	24 15 18	24 15 18	24 15 18
23	18 23	15 18 23	24 15 18 23	24 15 18 23
24	23 24	18 23 24	•	•
17	24 17	23 24 17	18 23 24 17	15 18 23 24 17
18	17 18	24 17 18	•	18 23 24 17
24	18 24	•	24 17 18	•
18	•	18 24	•	24 17 18
17	18 17	24 18 17	•	•
17	17	18 17	•	•
15	17 15	17 15	18 17 15	24 18 17 15
24	15 24	17 15 24	17 15 24	•
17	24 17	•	•	17 15 24
24	•	24 17	•	•
18	24 18	17 24 18	17 24 18	15 17 24 18

n Working set of process is defined by window size

n Working set = Pages referenced in window

n Periodically remove pages not in working set

Variable Allocation, Local Scope: Page Fault Frequency (PFF)

- n Requires a use bit to be associated with each page in memory
- n Bit is set to 1 when that page is accessed
- n When a page fault occurs, the OS notes the virtual time since the last page fault for that process
 - n Short time, increase working set size (by one, this new referenced page)
 - n Long time, decrease working set size (free all pages with use bit 0)
 - n Reset all use bits to 0
- n Does not perform well during the transient periods when there is a shift to a new locality

Variable-interval Sampled Working Set (VSWS)

- n Evaluates the working set of a process at sampling instances based on elapsed virtual time
- n Driven by three parameters:

the minimum
duration of the
sampling
interval

the maximum
duration of the
sampling
interval

the number of
page faults that
are allowed to
occur between
sampling
instances

Cleaning Policy

- n Concerned with determining when a modified page should be written out to secondary memory

Demand Cleaning

a page is written out to secondary memory only when it has been selected for replacement

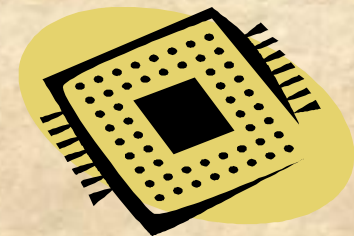


Precleaning

allows the writing of pages in batches

Load Control

- n Determines the number of processes that will be resident in main memory
 - n *multiprogramming* level
- n Critical in effective memory management
- n Too few processes, many occasions when all processes will be blocked and much time will be spent in swapping
- n Too many processes will lead to thrashing



Multiprogramming

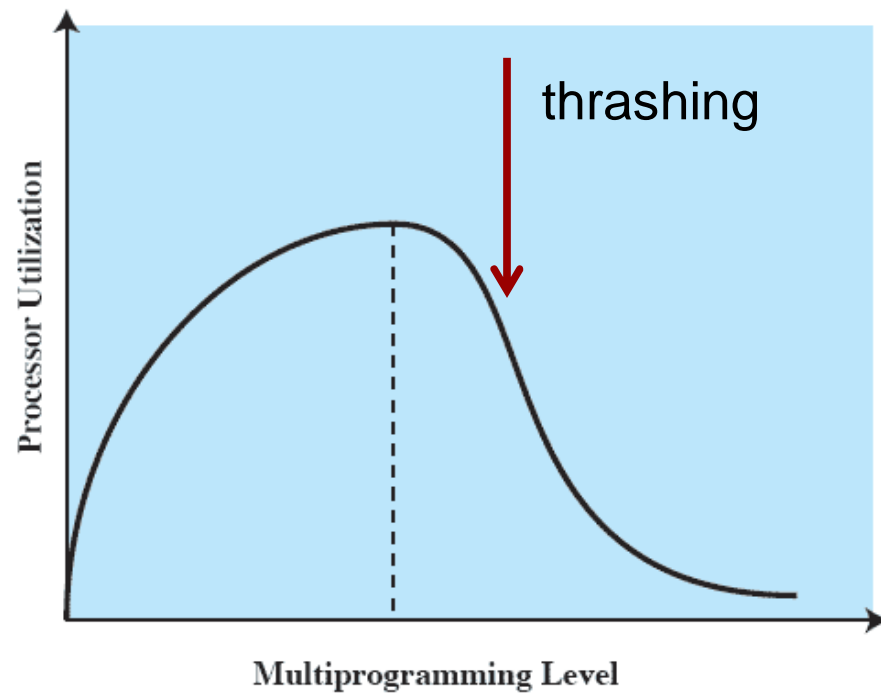


Figure 8.21 Multiprogramming Effects

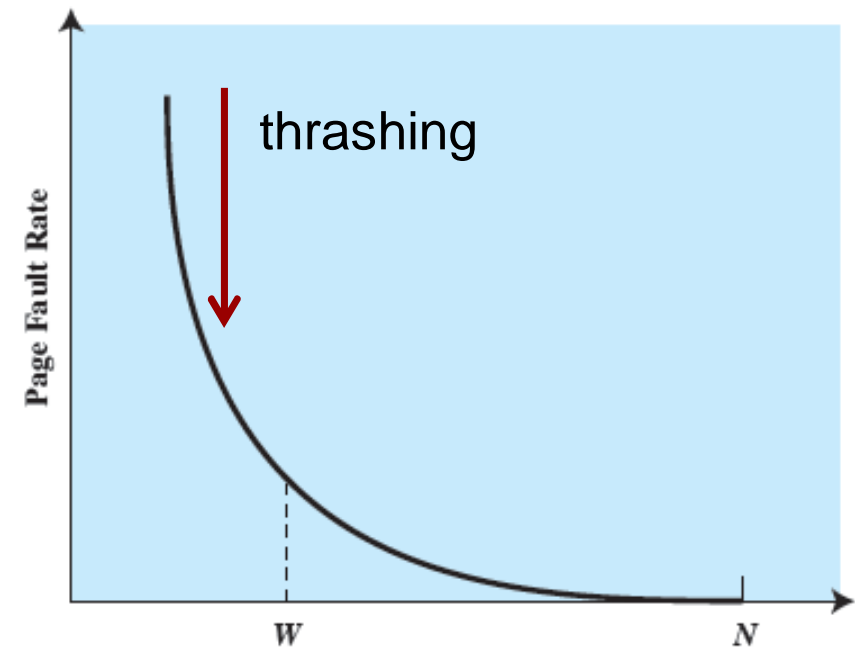


Fig. 8.11 (b) Number of Page Frames Allocated

Process Suspension

- n If the degree of multiprogramming is to be reduced, one or more of the currently resident processes must be swapped out

Six possibilities exist:

- Lowest-priority process
- Faulting process
- Last process activated
- Process with the smallest resident set
- Largest process
- Process with the largest remaining execution window

Unix

- n Intended to be machine independent so its memory management schemes will vary
- n Early Unix: variable partitioning with no virtual memory scheme
- n Current implementations of UNIX and Solaris make use of paged virtual memory

SVR4 and Solaris use two separate schemes:

- paging system
- kernel memory allocator

Paging System and Kernel Memory Allocator

Paging system



provides a virtual memory capability that allocates page frames in main memory to processes



allocates page frames to disk block buffers

Kernel Memory Allocator



allocates memory for the kernel

UNIX SVR4 Memory Management Formats

Page frame number	Age	Copy on write	Mod-ify	Refer-ence	Valid	Pro-tect
-------------------	-----	---------------	---------	------------	-------	----------

(a) Page table entry

Swap device number	Device block number	Type of storage
--------------------	---------------------	-----------------

(b) Disk block descriptor

Page state	Reference count	Logical device	Block number	Pfdata pointer
------------	-----------------	----------------	--------------	----------------

(c) Page frame data table entry

Reference count	Page/storage unit number
-----------------	--------------------------

(d) Swap-use table entry

Figure 8.22 UNIX SVR4 Memory Management Formats

Table 8.6

UNIX SVR4 Memory Management Parameters (page 1 of 2)

Page Table Entry

Page frame number

Refers to frame in real memory.

Age

Indicates how long the page has been in memory without being referenced. The length and contents of this field are processor dependent.

Copy on write

Set when more than one process shares a page. If one of the processes writes into the page, a separate copy of the page must first be made for all other processes that share the page. This feature allows the copy operation to be deferred until necessary and avoided in cases where it turns out not to be necessary.

Modify

Indicates page has been modified.

Reference

Indicates page has been referenced. This bit is set to 0 when the page is first loaded and may be periodically reset by the page replacement algorithm.

Valid

Indicates page is in main memory.

Protect

Indicates whether write operation is allowed.

Disk Block Descriptor

Swap device number

Logical device number of the secondary device that holds the corresponding page. This allows more than one device to be used for swapping.

Device block number

Block location of page on swap device.

Type of storage

Storage may be swap unit or executable file. In the latter case, there is an indication as to whether the virtual memory to be allocated should be cleared first.

Table 8.6

UNIX SVR4 Memory Management Parameters (page 2 of 2)

Page Frame Data Table Entry

Page state

Indicates whether this frame is available or has an associated page. In the latter case, the status of the page is specified: on swap device, in executable file, or DMA in progress.

Reference count

Number of processes that reference the page.

Logical device

Logical device that contains a copy of the page.

Block number

Block location of the page copy on the logical device.

Pfdata pointer

Pointer to other pfdata table entries on a list of free pages and on a hash queue of pages.

Swap-Use Table Entry

Reference count

Number of page table entries that point to a page on the swap device.

Page/storage unit number

Page identifier on storage unit.

Page Replacement

- n The page frame data table is used for page replacement
- n Pointers are used to create lists within the table
 - n all available frames are linked together in a list of free frames available for bringing in pages
 - n when the number of available frames drops below a certain threshold, the kernel will steal a number of frames to compensate



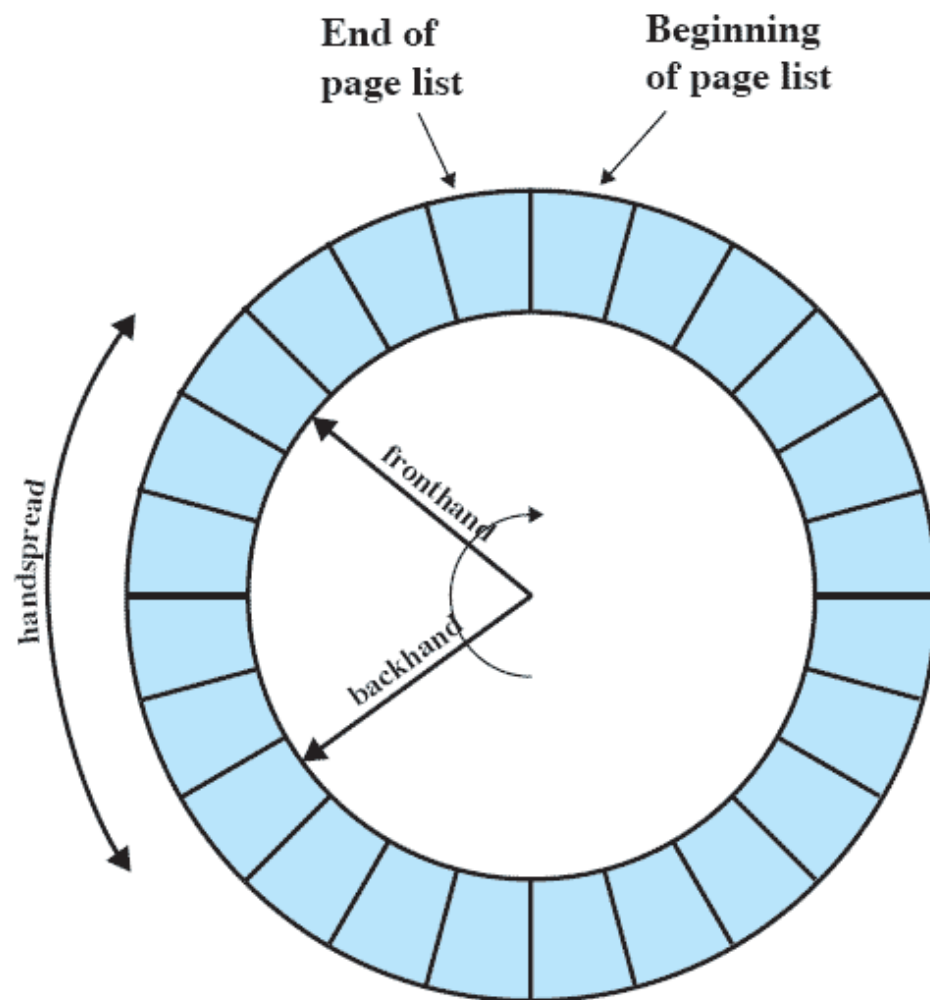


Figure 8.23 Two-Handed Clock Page-Replacement Algorithm

“Two Handed”
Clock
Page
Replacement

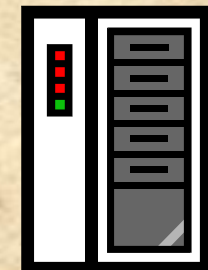
Kernel Memory Allocator

- n The kernel generates and destroys small tables and buffers frequently during the course of execution, each of which requires dynamic memory allocation.
- n Most of these blocks are significantly smaller than typical pages (therefore paging would be inefficient)
- n Allocations and free operations must be made as fast as possible



Lazy Buddy

- n Technique adopted for SVR4
- n UNIX often exhibits steady-state behavior in kernel memory demand
 - n i.e. the amount of demand for blocks of a particular size varies slowly in time
- n Defers coalescing until it seems likely that it is needed, and then coalesces as many blocks as possible



Lazy Buddy System Algorithm

Initial value of D_i is 0

After an operation, the value of D_i is updated as follows

(I) if the next operation is a block allocate request:

if there is any free block, select one to allocate

if the selected block is locally free

then $D_i := D_i + 2$

else $D_i := D_i + 1$

otherwise

first get two blocks by splitting a larger one into two (recursive operation)

allocate one and mark the other locally free

D_i remains unchanged (but D may change for other block sizes because of the recursive call)

(II) if the next operation is a block free request

Case $D_i \geq 2$

mark it locally free and free it locally

$D_i := D_i - 2$

Case $D_i = 1$

mark it globally free and free it globally; coalesce if possible

$D_i := 0$

Case $D_i = 0$

mark it globally free and free it globally; coalesce if possible

select one locally free block of size 2^i and free it globally; coalesce if possible

$D_i := 0$

Figure 8.24 Lazy Buddy System Algorithm

Linux Memory Management

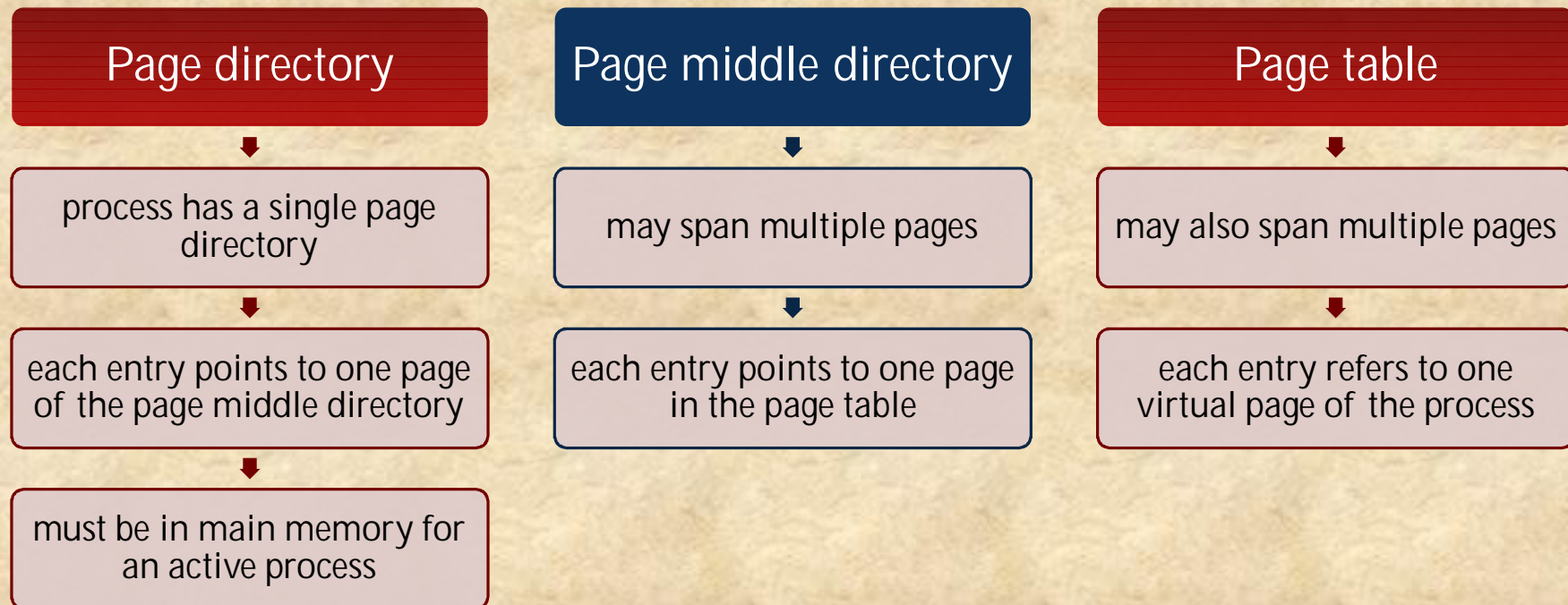
- n Shares many characteristics with Unix
- n Is quite complex

Two main
aspects

- process virtual memory
- kernel memory allocation

Linux Virtual Memory

n Three level page table structure:



Address Translation

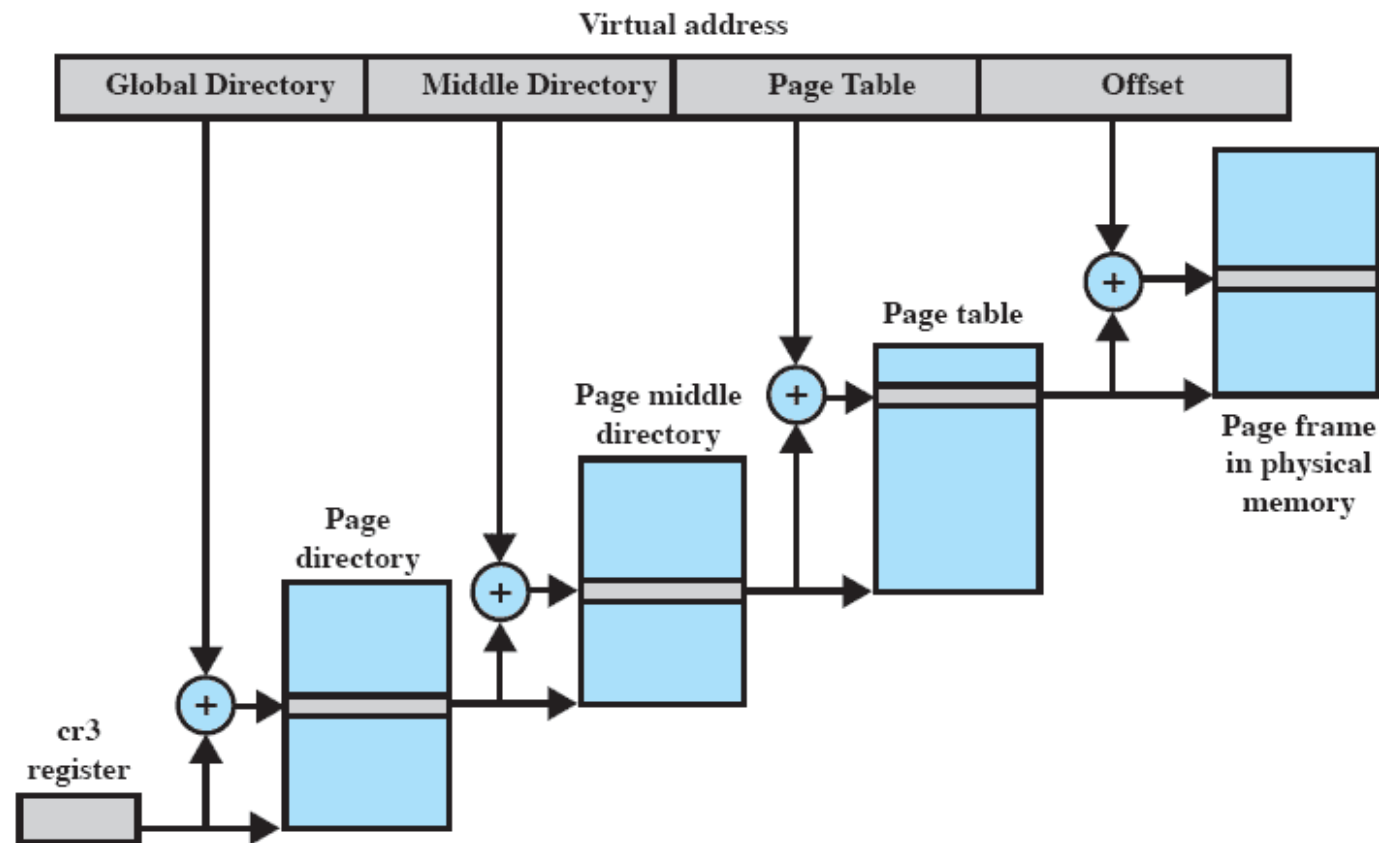


Figure 8.25 Address Translation in Linux Virtual Memory Scheme

Linux Page Replacement

- n Based on the clock algorithm
- n The use bit is replaced with an 8-bit age variable
 - n incremented each time the page is accessed
- n Periodically decrements the age bits
 - n a page with an age of 0 is an “old” page that has not been referenced in some time and is the best candidate for replacement
- n A form of least frequently used policy

Kernel Memory Allocation

- n Kernel memory capability manages physical main memory page frames
- n primary function is to allocate and deallocate frames for particular uses

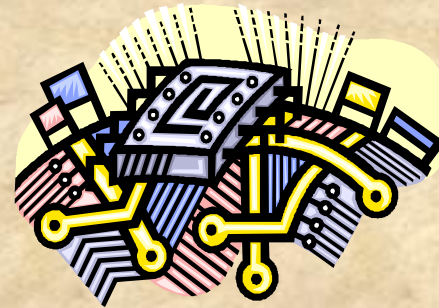
Possible owners of a frame include:

- user-space processes
- dynamically allocated kernel data
- static kernel code
- page cache

- n A buddy algorithm is used so that memory for the kernel can be allocated and deallocated in units of one or more pages
- n Page allocator alone would be inefficient because the kernel requires small short-term memory chunks in odd sizes
- n Slab allocation
 - n used by Linux to accommodate small kernel chunks of memory

Windows Memory Management

- n Virtual memory manager controls how memory is allocated and how paging is performed
- n Designed to operate over a variety of platforms
- n Uses page sizes ranging from 4 Kbytes to 64 Kbytes

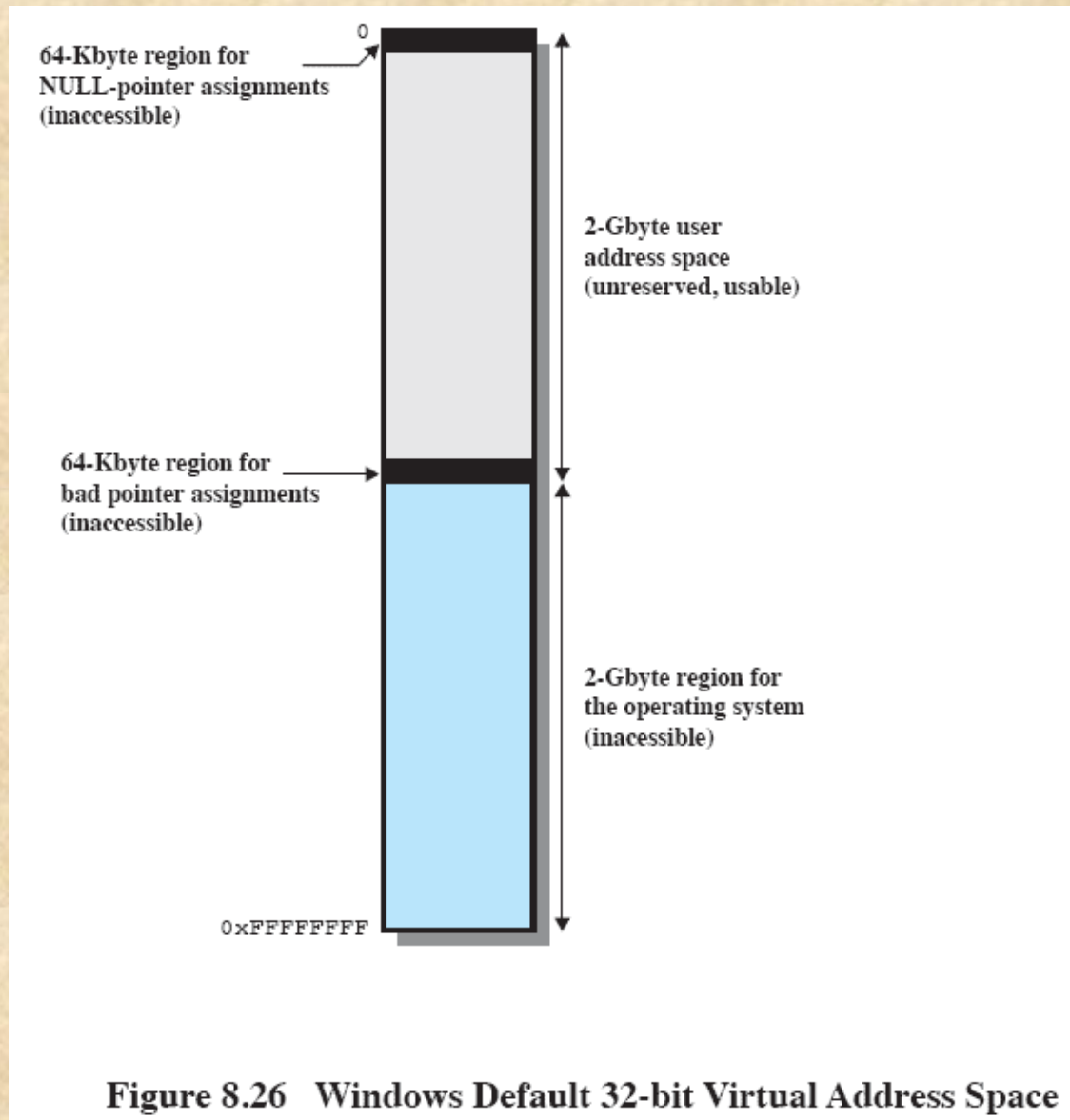


Windows Virtual Address Map

- n On 32 bit platforms each user process sees a separate 32 bit address space allowing 4 Gbytes of virtual memory per process
 - § by default half is reserved for the OS
- n Large memory intensive applications run more effectively using 64-bit Windows
- n Most modern PCs use the AMD64 processor architecture which is capable of running as either a 32-bit or 64-bit system



32-Bit Windows Address Space



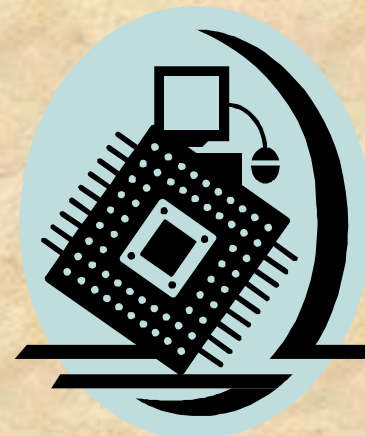
Windows Paging

- n On creation, a process can make use of the entire user space of almost 2 Gbytes
- n This space is divided into fixed-size pages managed in contiguous regions allocated on 64 Kbyte boundaries
- n Regions may be in one of three states:



Resident Set Management System

- n Windows uses variable allocation, local scope
- n When activated, a process is assigned a data structure to manage its working set
- n Working sets of active processes are adjusted depending on the availability of main memory



Summary

- n Desirable to:
 - n Maintain as many processes in main memory as possible
 - n Free programmers from size restrictions in program development
- n With virtual memory:
 - n All address references are logical references that are translated at run time to real addresses
 - n Locality makes it work.
 - n Two approaches are paging and segmentation
 - n Management scheme requires both hardware and software support: TLB + different VM policies