

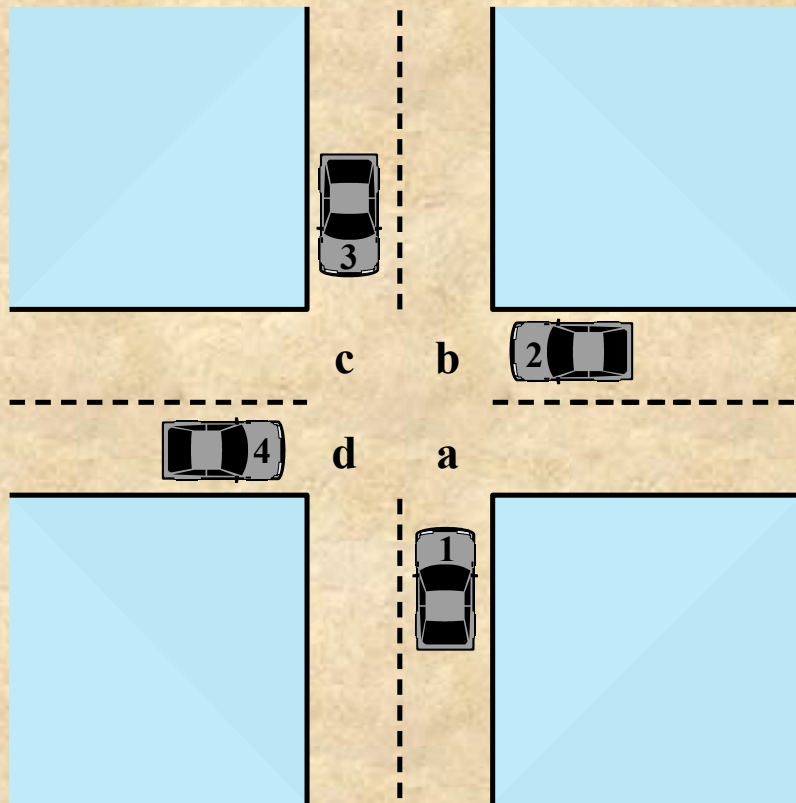
*Operating
Systems:
Internals
and Design
Principles*

Chapter 6
Concurrency:
Deadlock and
Starvation

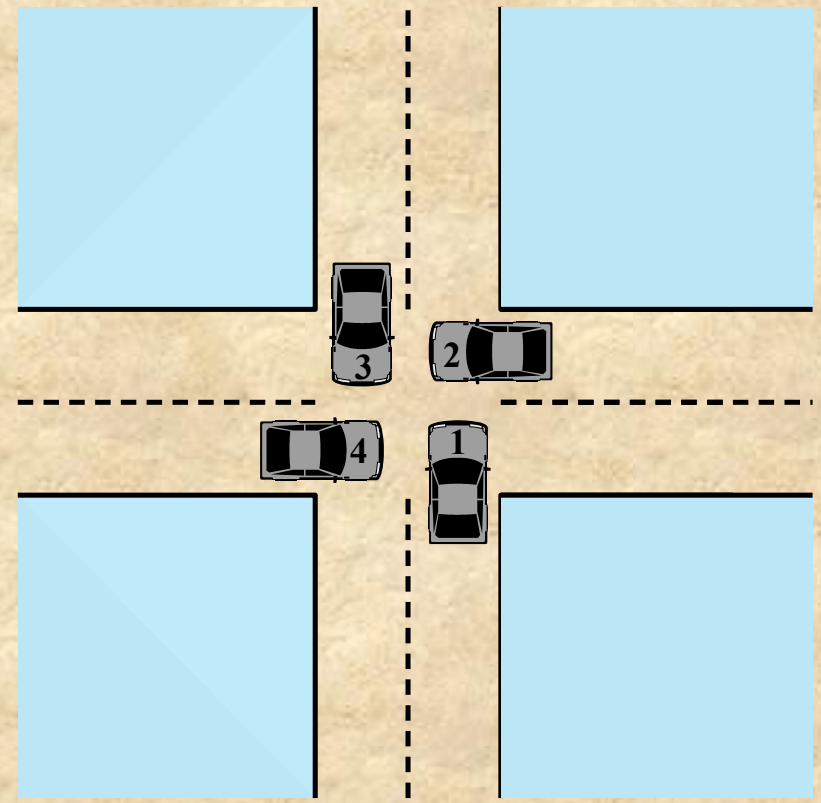
6.1 Principles of Deadlock

- The permanent blocking of a set of processes that either compete for system resources or communicate with each other
- A set of processes is deadlocked when each process in the set is blocked awaiting an event that can only be triggered by another blocked process in the set
- Permanent because none of the events can be triggered
- No efficient solution in the general case





(a) Deadlock possible



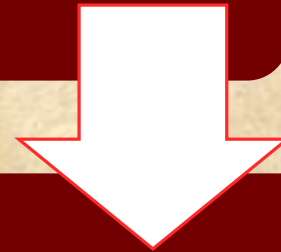
(b) Deadlock

Figure 6.1 Illustration of Deadlock

Resource Categories

Reusable

- can be safely used by only one process at a time and is not depleted by that use
- processors, I/O channels, main and secondary memory, devices, and data structures such as files, databases, and semaphores



Consumable

- one that can be created (produced) and destroyed (consumed)
 - interrupts, signals, messages, and information in I/O buffers

Example 1: File & Tape Drive Request

Process P

Process Q

Step	Action
p ₀	Request (D)
p ₁	Lock (D)
p ₂	Request (T)
p ₃	Lock (T)
p ₄	Perform function
p ₅	Unlock (D)
p ₆	Unlock (T)

Step	Action
q ₀	Request (T)
q ₁	Lock (T)
q ₂	Request (D)
q ₃	Lock (D)
q ₄	Perform function
q ₅	Unlock (T)
q ₆	Unlock (D)

Figure 6.4
Example of Two Processes Competing for Reusable Resources

Example 2: Memory Request

- Suppose that the space is available for allocation of 200Kbytes, and the following sequence of events occur:



- Deadlock occurs if both processes progress to their second request

Consumable Resources Deadlock

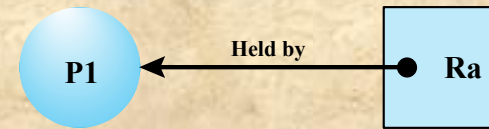
- Consider a pair of processes, in which each process attempts to receive a message from the other process and then send a message to the other process:

P1	P2
...	...
Receive (P2);	Receive (P1);
...	...
Send (P2, M1);	Send (P1, M2);

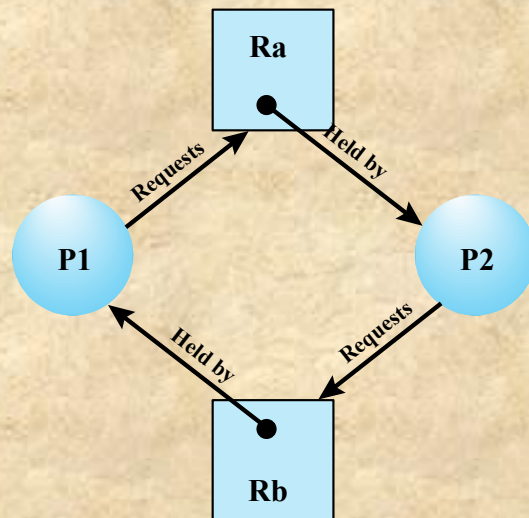
- Deadlock occurs if the Receive is blocking



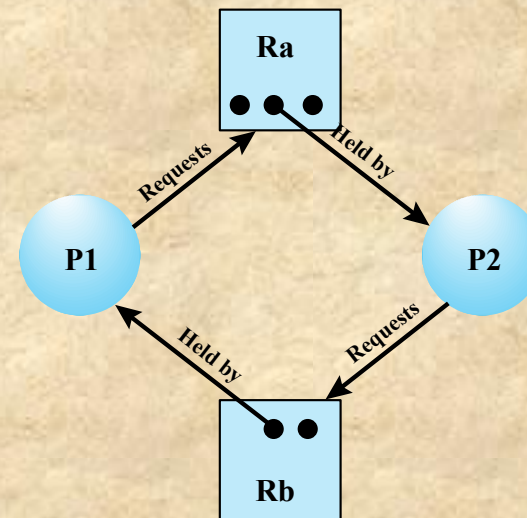
(a) Resource is requested



(b) Resource is held



(c) Circular wait



(d) No deadlock

Figure 6.5 Examples of Resource Allocation Graphs

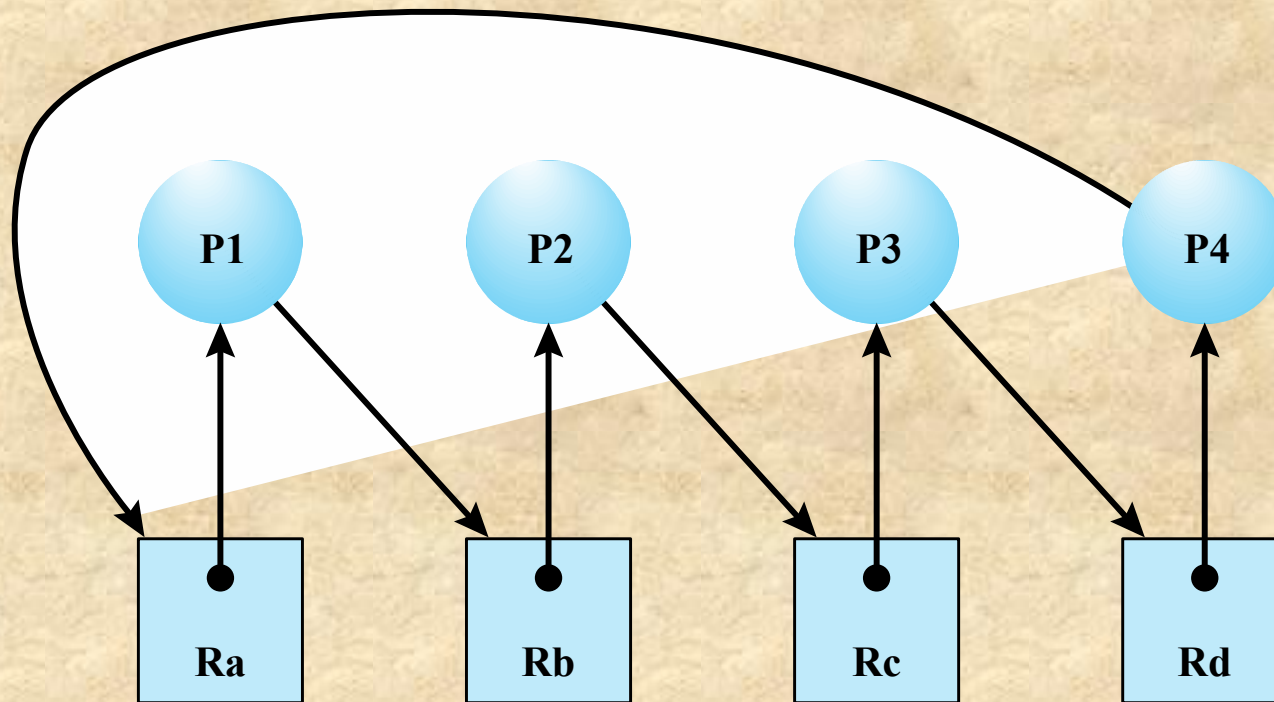


Figure 6.6 Resource Allocation Graph for Figure 6.1b

Conditions for Deadlock

▪ Necessary conditions

Mutual Exclusion

- only one process may use a resource at a time

Hold-and-Wait

- a process may hold allocated resources while awaiting assignment of others

No Pre-emption

- no resource can be forcibly removed from a process holding it

Circular Wait

- a closed chain of processes exists, such that each process holds at least one resource needed by the next process in the chain

▪ Sufficient condition

Dealing with Deadlock

- Three general approaches exist for dealing with deadlock:

Prevent Deadlock

- adopt a policy that eliminates one of the conditions

Avoid Deadlock

- make the appropriate dynamic choices based on the current state of resource allocation

Detect Deadlock

- attempt to detect the presence of deadlock and take action to recover

6.2 Deadlock Prevention

- Design a system in such a way that the possibility of deadlock is excluded
- Two main methods:
 - Indirect
 - prevent the occurrence of one of the three necessary conditions
 - Direct
 - prevent the occurrence of a circular wait



Deadlock Prevention Condition

Mutual Exclusion

if access to a resource requires mutual exclusion then it must be supported by the OS

Hold and Wait

require that a process request all of its required resources at one time and blocking the process until all requests can be granted simultaneously

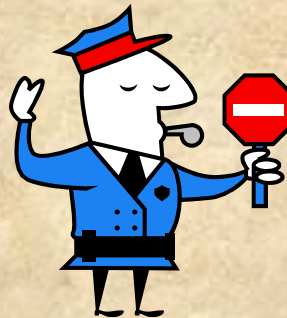
Deadlock Prevention Condition

- No Preemption
 - if a process holding certain resources is denied a further request, that process must release its original resources and request them again
 - OS may preempt the second process and require it to release its resources
- Circular Wait
 - define a linear ordering of resource types
- These lead to inefficient use of resources and inefficient execution of processes



6.3 Deadlock Avoidance

- A decision is made dynamically whether the current resource allocation request will, if granted, potentially lead to a deadlock
- Requires knowledge of future process requests



Two Approaches to Deadlock Avoidance

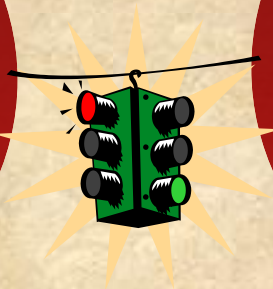
Deadlock Avoidance

Resource Allocation Denial

- do not grant an incremental resource request to a process if this allocation might lead to deadlock

Process Initiation Denial

- do not start a process if its demands might lead to deadlock



Resource Initiation Denial

- Consider a system of n processes and m different types of resources
- Define two vectors and two matrices:
 - *Resource* = R = Total number of each resource in the system
 - *Available* = V = Total amount of each resource not allocated to any process
 - *Claim* = C (C_{ij} = requirement of P_i for resource R_j)
 - *Allocation* = A (C_{ij} = current allocation to P_i to R_j)
- Start a new process P_{n+1} only if

$$R_j \geq C_{(n+1)j} + \sum_{i=1}^n C_{ij} \text{ for all } j$$



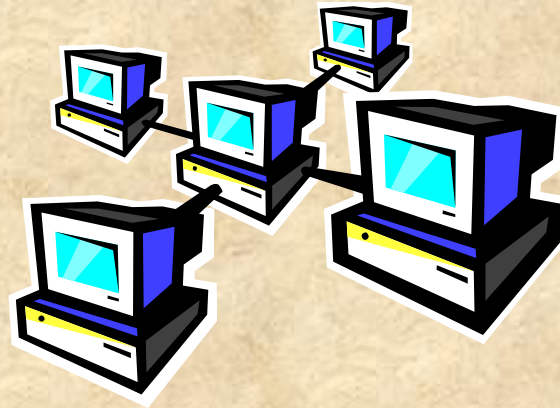
Resource Allocation Denial

- Referred to as the *banker's algorithm* (Figure 6.9 on pp. 306)
- The *state* of the system reflects the current allocation of resources to processes
- A *safe state* is one in which there is at least one sequence of resource allocations to processes that does not result in a deadlock
- An *unsafe state* is a state that is not safe




Deadlock Avoidance Advantages

- It is not necessary to preempt and rollback processes, as in deadlock detection
- It is less restrictive and allow more concurrency than deadlock prevention



Deadlock Avoidance Restrictions

- 
- The maximum resource requirement for each process must be stated in advance
 - The processes under consideration must be independent and with no synchronization requirements
 - There must be a fixed number of resources to allocate
 - No process may exit while holding resources

6.4 Deadlock Detection

Deadlock prevention strategies are very conservative

- limit access to resources by imposing restrictions on processes

Deadlock detection strategies do the opposite

- resource requests are granted whenever possible

Deadline Detection Algorithms

- A check for deadlock can be made as frequently as each resource request or, less frequently, depending on how likely it is for a deadlock to occur



Advantages:

- it leads to early detection
- the algorithm is relatively simple due to incremental changes to the state of the system



Disadvantage

- frequent checks consume considerable processor time

Recovery Strategies

- Possible approaches (in order of increasing sophistication)
 - Abort all deadlocked processes
 - Back up each deadlocked process to some previously defined checkpoint and restart all processes
 - Successively abort deadlocked processes until deadlock no longer exists
 - Successively preempt resources until deadlock no longer exists

Approach	Resource Allocation Policy	Different Schemes	Major Advantages	Major Disadvantages
Prevention	Conservative; undercommits resources	Requesting all resources at once	<ul style="list-style-type: none"> •Works well for processes that perform a single burst of activity •No preemption necessary 	<ul style="list-style-type: none"> •Inefficient •Delays process initiation •Future resource requirements must be known by processes
		Preemption	<ul style="list-style-type: none"> •Convenient when applied to resources whose state can be saved and restored easily 	<ul style="list-style-type: none"> •Preempts more often than necessary
		Resource ordering	<ul style="list-style-type: none"> •Feasible to enforce via compile-time checks •Needs no run-time computation since problem is solved in system design 	<ul style="list-style-type: none"> •Disallows incremental resource requests
Avoidance	Midway between that of detection and prevention	Manipulate to find at least one safe path	<ul style="list-style-type: none"> •No preemption necessary 	<ul style="list-style-type: none"> •Future resource requirements must be known by OS •Processes can be blocked for long periods
Detection	Very liberal; requested resources are granted where possible	Invoke periodically to test for deadlock	<ul style="list-style-type: none"> •Never delays process initiation •Facilitates online handling 	<ul style="list-style-type: none"> •Inherent preemption losses

Table 6.1

**Summary of
Deadlock
Detection,
Prevention, and
Avoidance
Approaches for
Operating
Systems
[ISLO80]**

6.6 Dining Philosophers Problem

- No two philosophers can use the same fork at the same time (mutual exclusion)
- No philosopher must starve to death (avoid deadlock and starvation)

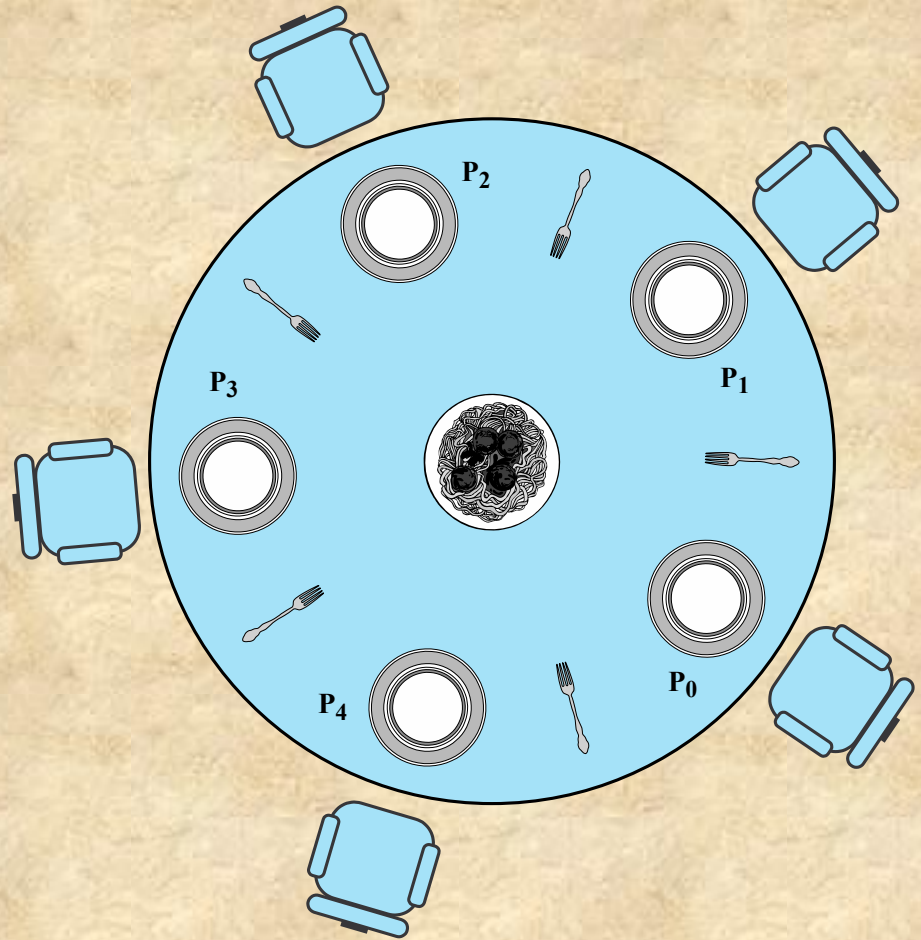


Figure 6.11 Dining Arrangement for Philosophers

```

/* program      diningphilosophers */
semaphore fork [5] = {1};
int i;
void philosopher (int i)
{
    while (true) {
        think();
        wait (fork[i]);
        wait (fork [(i+1) mod 5]);
        eat();
        signal(fork [(i+1) mod 5]);
        signal(fork[i]);
    }
}
void main()
{
    parbegin (philosopher (0), philosopher (1), philosopher
(2),
            philosopher (3), philosopher (4));
}

```

**Figure 6.12 A First Solution to the Dining Philosophers Problem
(that may lead to deadlock)**

```

/* program diningphilosophers */
semaphore fork[5] = {1};
semaphore room = {4};
int i;
void philosopher (int i)
{
    while (true) {
        think();
        wait (room);
        wait (fork[i]);
        wait (fork [(i+1) mod 5]);
        eat();
        signal (fork [(i+1) mod 5]);
        signal (fork[i]);
        signal (room);
    }
}
void main()
{
    parbegin (philosopher (0), philosopher (1), philosopher (2),
              philosopher (3), philosopher (4));
}

```

Figure 6.13 A Second Solution to the Dining Philosophers Problem


```

monitor dining_controller;
cond ForkReady[5];          /* condition variable for synchronization */
boolean fork[5] = {true};    /* availability status of each fork */

void get_forks(int pid)      /* pid is the philosopher id number */
{
    int left = pid;
    int right = (++pid) % 5;
    /*grant the left fork*/
    if (!fork[left])
        cwait(ForkReady[left]);          /* queue on condition variable */
    fork[left] = false;
    /*grant the right fork*/
    if (!fork[right])
        cwait(ForkReady[right]);         /* queue on condition variable */
    fork[right] = false;
}

void release_forks(int pid)
{
    int left = pid;
    int right = (++pid) % 5;
    /*release the left fork*/
    if (empty(ForkReady[left])           /*no one is waiting for this fork */
        fork[left] = true;
    else                                /* awaken a process waiting on this fork */
        csignal(ForkReady[left]);
    /*release the right fork*/
    if (empty(ForkReady[right])          /*no one is waiting for this fork */
        fork[right] = true;
    else                                /* awaken a process waiting on this fork */
        csignal(ForkReady[right]);
}

```

```

void philosopher[k=0 to 4]      /* the five philosopher clients */
{
    while (true) {
        <think>;
        get_forks(k);            /* client requests two forks via monitor */
        <eat spaghetti>;
        release_forks(k);        /* client releases forks via the monitor */
    }
}

```

Figure 6.14

A Solution to the Dining Philosophers Problem Using a Monitor

6.7 UNIX Concurrency Mechanisms

- UNIX provides a variety of mechanisms for interprocessor communication and synchronization including:

Pipes

Messages

Shared
memory

Semaphores

Signals

Pipes

- Circular buffers allowing two processes to communicate on the producer-consumer model
 - A first-in-first-out queue, written by one process and read by another
 - When a pipe is created, it is given a fixed size in bytes
 - The OS enforces mutual exclusion for a pipe

Two types:

- Named
- Unnamed

Messages



- A block of bytes with an accompanying type
- UNIX provides *msgsnd* and *msgrcv* system calls for processes to engage in message passing
- Associated with each process is a message queue, which functions like a mailbox
- The receiver can either retrieve messages in first-in-first-out order or by type
 - If a process attempts to read a message of a certain type and fails because no message of that type is present, the process is not blocked

Shared Memory

- The fastest form of interprocess communication
- A common block of virtual memory shared by multiple processes
- Permission is read-only or read-write for a process
- Mutual exclusion constraints are not part of the shared-memory facility but must be provided by the processes using the shared memory

Semaphores

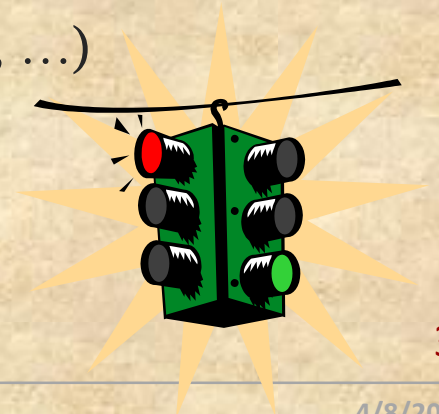
- A generalization of the `semWait` and `semSignal` primitives
 - no other process may access the semaphore until all operations have completed (i.e., atomic)

Consists of the following elements:

- current value of the semaphore
- process ID of the last process to operate on the semaphore
- number of processes waiting for the semaphore value to be greater than its current value
- number of processes waiting for the semaphore value to be zero

Signals

- A software mechanism that informs a process of the occurrence of asynchronous events
 - similar to a hardware interrupt, but does not employ priorities with no particular ordering
- A signal is delivered by updating a field (*p_cursig*) in the process table (*proc* structure) for the process to which the signal is being sent
- A process may respond to a signal by:
 - performing some default action (e.g., abort, exit, ...)
 - executing a signal-handler function
 - ignoring the signal



Value	Name	Description
01	SIGHUP	Hang up; sent to process when kernel assumes that the user of that process is doing no useful work
02	SIGINT	Interrupt
03	SIGQUIT	Quit; sent by user to induce halting of process and production of core dump
04	SIGILL	Illegal instruction
05	SIGTRAP	Trace trap; triggers the execution of code for process tracing
06	SIGIOT	IOT instruction
07	SIGEMT	EMT instruction
08	SIGFPE	Floating-point exception
09	SIGKILL	Kill; terminate process
10	SIGBUS	Bus error
11	SIGSEGV	Segmentation violation; process attempts to access location outside its virtual address space
12	SIGSYS	Bad argument to system call
13	SIGPIPE	Write on a pipe that has no readers attached to it
14	SIGALRM	Alarm clock; issued when a process wishes to receive a signal after a period of time
15	SIGTERM	Software termination
16	SIGUSR1	User-defined signal 1
17	SIGUSR2	User-defined signal 2
18	SIGCHLD	Death of a child
19	SIGPWR	Power failure

Table 6.2

UNIX Signals

(Table can be found on page 316 in textbook)**35**

6.8 Linux Kernel Concurrency Mechanism

- Includes all the mechanisms found in UNIX plus:



Atomic Operations

- Atomic operations execute without interruption and without interference to avoid simple race conditions
- Two types:



Integer Operations

operate on an integer variable

typically used to implement counters

Bitmap Operations

operate on one of a sequence of bits at an arbitrary memory location indicated by a pointer variable

Atomic Integer Operations	
ATOMIC_INIT (int i)	At declaration: initialize an atomic t to i
int atomic_read(atomic_t *v)	Read integer value of v
void atomic_set(atomic_t *v, int i)	Set the value of v to integer i
void atomic_add(int i, atomic_t *v)	Add i to v
void atomic_sub(int i, atomic_t *v)	Subtract i from v
void atomic_inc(atomic_t *v)	Add 1 to v
void atomic_dec(atomic_t *v)	Subtract 1 from v
int atomic_sub_and_test(int i, atomic_t *v)	Subtract i from v; return 1 if the result is zero; return 0 otherwise
int atomic_add_negative(int i, atomic_t *v)	Add i to v; return 1 if the result is negative; return 0 otherwise (used for implementing semaphores)
int atomic_dec_and_test(atomic_t *v)	Subtract 1 from v; return 1 if the result is zero; return 0 otherwise
int atomic_inc_and_test(atomic_t *v)	Add 1 to v; return 1 if the result is zero; return 0 otherwise
Atomic Bitmap Operations	
void set_bit(int nr, void *addr)	Set bit nr in the bitmap pointed to by addr
void clear_bit(int nr, void *addr)	Clear bit nr in the bitmap pointed to by addr
void change_bit(int nr, void *addr)	Invert bit nr in the bitmap pointed to by addr
int test_and_set_bit(int nr, void *addr)	Set bit nr in the bitmap pointed to by addr; return the old bit value
int test_and_clear_bit(int nr, void *addr)	Clear bit nr in the bitmap pointed to by addr; return the old bit value
int test_and_change_bit(int nr, void *addr)	Invert bit nr in the bitmap pointed to by addr; return the old bit value
int test_bit(int nr, void *addr)	Return the value of bit nr in the bitmap pointed to by addr

Table 6.3

Linux

Atomic

Operations

(Table can be found on page 317 in textbook)

Spinlocks

- Most common technique for protecting a critical section in Linux
- Can only be acquired by one thread at a time
 - any other thread will keep trying (spinning) until it can acquire the lock
- Built on an integer location in memory that is checked by each thread before it enters its critical section
- Effective in situations where the wait time for acquiring a lock is expected to be very short, say on the order of less than two context switches
- Disadvantage:
 - locked-out threads continue to execute in a busy-waiting mode

<code>void spin_lock(spinlock_t *lock)</code>	Acquires the specified lock, spinning if needed until it is available
<code>void spin_lock_irq(spinlock_t *lock)</code>	Like spin lock, but also disables interrupts on the local processor
<code>void spin_lock_irqsave(spinlock_t *lock, unsigned long flags)</code>	Like spin lock irq, but also saves the current interrupt state in flags
<code>void spin_lock_bh(spinlock_t *lock)</code>	Like spin lock, but also disables the execution of all bottom halves
<code>void spin_unlock(spinlock_t *lock)</code>	Releases given lock
<code>void spin_unlock_irq(spinlock_t *lock)</code>	Releases given lock and enables local interrupts
<code>void spin_unlock_irqrestore(spinlock_t *lock, unsigned long flags)</code>	Releases given lock and restores local interrupts to given previous state
<code>void spin_unlock_bh(spinlock_t *lock)</code>	Releases given lock and enables bottom halves
<code>void spin_lock_init(spinlock_t *lock)</code>	Initializes given spinlock
<code>int spin_trylock(spinlock_t *lock)</code>	Tries to acquire specified lock; returns nonzero if lock is currently held and zero otherwise
<code>int spin_is_locked(spinlock_t *lock)</code>	Returns nonzero if lock is currently held and zero otherwise

Table 6.4 Linux Spinlocks

Semaphores

- User level:
 - Linux provides a semaphore interface corresponding to that in UNIX SVR4
- Internally:
 - implemented as functions within the kernel for its own use
 - cannot be accessed directly by the user program via system calls more efficient than user-visible semaphores
- Three types of kernel semaphores:
 - binary semaphores
 - counting semaphores
 - reader-writer semaphores
 - allow multiple concurrent readers, but only a single writer



Traditional Semaphores	
<code>void sema_init(struct semaphore *sem, int count)</code>	Initializes the dynamically created semaphore to the given count
<code>void init MUTEX(struct semaphore *sem)</code>	Initializes the dynamically created semaphore with a count of 1 (initially unlocked)
<code>void init MUTEX LOCKED(struct semaphore *sem)</code>	Initializes the dynamically created semaphore with a count of 0 (initially locked)
<code>void down(struct semaphore *sem)</code>	Attempts to acquire the given semaphore, entering uninterruptible sleep if semaphore is unavailable
<code>int down interruptible(struct semaphore *sem)</code>	Attempts to acquire the given semaphore, entering interruptible sleep if semaphore is unavailable; returns <code>-EINTR</code> value if a signal other than the result of an up operation is received
<code>int down trylock(struct semaphore *sem)</code>	Attempts to acquire the given semaphore, and returns a nonzero value if semaphore is unavailable
<code>void up(struct semaphore *sem)</code>	Releases the given semaphore
Reader-Writer Semaphores	
<code>void init rwsem(struct rw_semaphore, *rwsem)</code>	Initializes the dynamically created semaphore with a count of 1
<code>void down read(struct rw_semaphore, *rwsem)</code>	Down operation for readers
<code>void up read(struct rw_semaphore, *rwsem)</code>	Up operation for readers
<code>void down write(struct rw_semaphore, *rwsem)</code>	Down operation for writers
<code>void up write(struct rw_semaphore, *rwsem)</code>	Up operation for writers

Table 6.5

Linux

Semaphores

Barriers

- The reorderings of memory accesses may be done to optimize the use of the instruction pipeline either by the compiler or the processor
- Used to enforce the order in which instructions are executed

Table 6.6 Linux Memory Barrier Operations

<code>rmb()</code>	Prevents loads from being reordered across the barrier
<code>wmb()</code>	Prevents stores from being reordered across the barrier
<code>mb()</code>	Prevents loads and stores from being reordered across the barrier
<code>Barrier()</code>	Prevents the compiler from reordering loads or stores across the barrier
<code>smp_rmb()</code>	On SMP, provides a <code>rmb()</code> and on UP provides a <code>barrier()</code>
<code>smp_wmb()</code>	On SMP, provides a <code>wmb()</code> and on UP provides a <code>barrier()</code>
<code>smp_mb()</code>	On SMP, provides a <code>mb()</code> and on UP provides a <code>barrier()</code>

SMP = symmetric multiprocessor
UP = uniprocessor

6.10 Windows 7 Concurrency Mechanisms

- Windows provides synchronization among threads as part of the object architecture

The most important methods of synchronization are:

- executive dispatcher objects
- user mode critical sections
- slim reader-writer locks
- condition variables
- lock-free operations

Wait Functions

Allow a thread to block its own execution

Do not return until the specified criteria have been met

The type of wait function determines the set of criteria used

Dispatcher Objects

- Each dispatcher object instance can be in either a signalled or unsignalled state
 - A thread issues a wait request to the Windows Executive, using the handle of the synchronization object
 - When an object enters the signalled state, the Windows Executive releases one or all of the thread objects waiting on that dispatcher object



Object Type	Definition	Set to Signaled State When	Effect on Waiting Threads
Notification event	An announcement that a system event has occurred	Thread sets the event	All released
Synchronization event	An announcement that a system event has occurred.	Thread sets the event	One thread released
Mutex	A mechanism that provides mutual exclusion capabilities; equivalent to a binary semaphore	Owning thread or other thread releases the mutex	One thread released
Semaphore	A counter that regulates the number of threads that can use a resource	Semaphore count drops to zero	All released
Waitable timer	A counter that records the passage of time	Set time arrives or time interval expires	All released
File	An instance of an opened file or I/O device	I/O operation completes	All released
Process	A program invocation, including the address space and resources required to run the program	Last thread terminates	All released
Thread	An executable entity within a process	Thread terminates	All released

Table 6.7

Windows Synchronization Objects

Note: Shaded rows correspond to objects that exist for the sole purpose of synchronization.

Critical Sections

- Provide a synchronization mechanism similar to that provided *mutex objects*, except that critical sections can be used only by the threads of a single process
 - a much faster, more efficient mechanism for mutual-exclusion synchronization.
- If the system is a multiprocessor, the code will attempt to acquire a spin-lock
 - if the spinlock cannot be acquired within a reasonable number of iterations, a dispatcher object is used to block the thread so that the kernel can dispatch another thread onto the processor

Slim Read-Writer Locks

- Windows Vista added a user mode reader-writer
- The reader-writer lock enters the kernel to block only after attempting to use a spin-lock like critical sections
- It is *slim* in the sense that it normally only requires allocation of a single pointer-sized piece of memory



Condition Variables

- Windows also has condition variables
- The process must declare and initialize a `CONDITION_VARIABLE`
- Used with either critical sections or SRW locks
- Used as follows:
 1. acquire exclusive lock
 2. while (predicate()==FALSE) SleepConditionVariable()
 3. perform the protected operation
 4. release the lock

Lock-free Synchronization

- Windows also relies heavily on interlocked operations for synchronization
 - interlocked operations use hardware facilities to guarantee that memory locations can be read, modified, and written in a single atomic operation
 - *InterlockedCompareExchange*(*Destination, Exchange, Comparand);

“Lock-free” Synchronization Primitives

- synchronization without taking a software lock
- a thread can never be switched away from a processor while still holding a lock

6.11 Android Interprocess Communication

- Android does not use Linux features for IPC but adds to the kernel a new capability known as Binder
 - The binder provides a lightweight remote procedure call (RPC) capability that is efficient in terms of both memory and processing requirements
 - Also used to mediate all interaction between two processes
- The RPC mechanism works between two processes on the same system but running on different virtual machines
- The method used for communicating with the Binder is the *ioctl* system call
 - the *ioctl* call is a general-purpose system call for device-specific I/O operations
 - used to access device drivers
 - include as parameters the command to be performed and the appropriate arguments

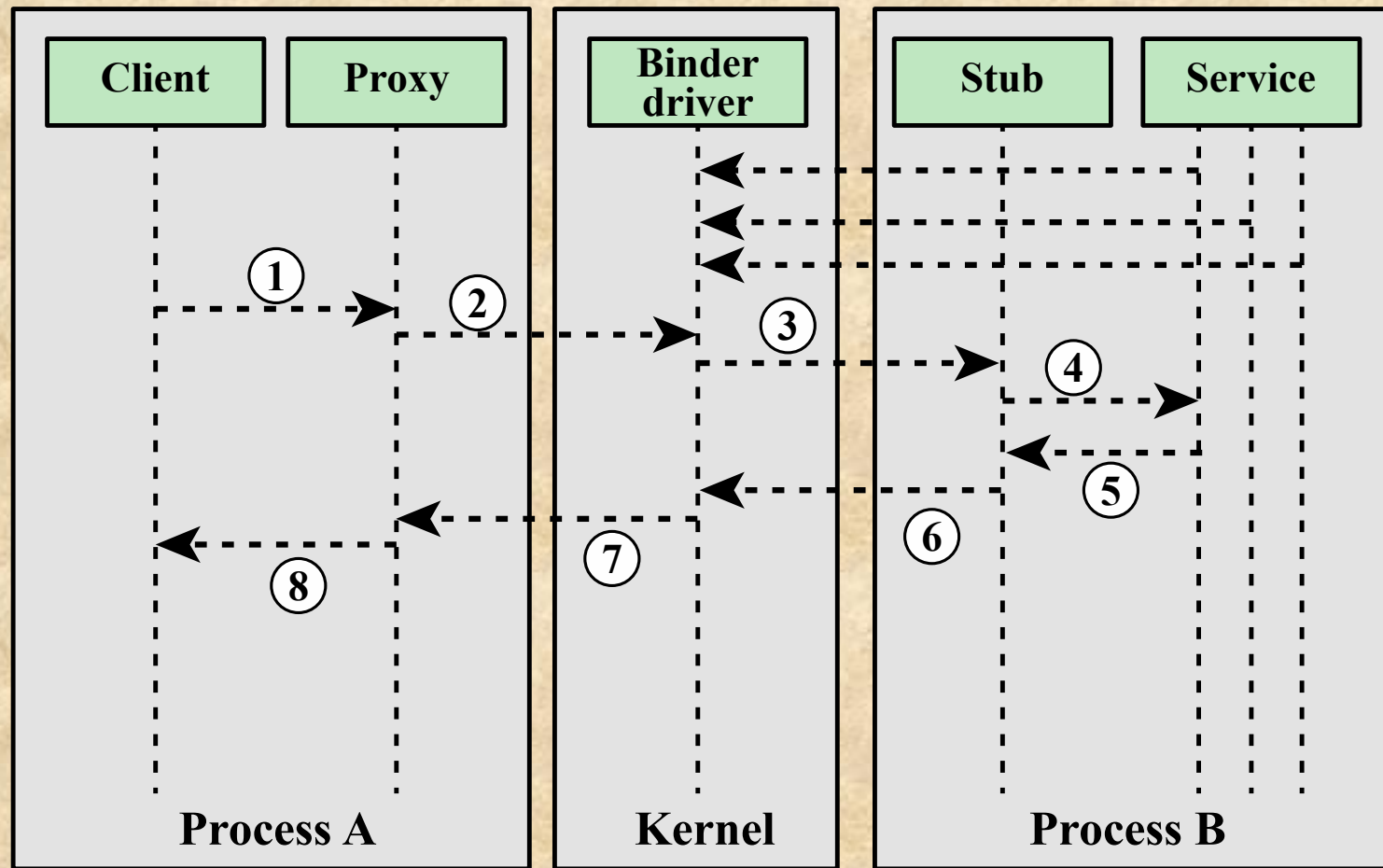


Figure 6.16 Binder Operation

Summary

- Principles of deadlock
 - Reusable/consumable resources
 - Resource allocation graphs
 - Conditions for deadlock
- Deadlock prevention
 - Mutual exclusion
 - Hold and wait
 - No preemption
 - Circular wait
- Deadlock avoidance
 - Process initiation denial
 - Resource allocation denial
- Deadlock detection
 - Deadlock detection algorithm
 - Recovery
- UNIX concurrency mechanisms
 - Pipes
 - Messages
 - Shared memory
 - Semaphores
 - Signals
- Linux kernel concurrency mechanisms
 - Atomic operations
 - Spinlocks
 - Semaphores
 - Barriers
- Windows 7 concurrency mechanisms
- Android interprocess communication