0301304 FUNDAMENTAL OF OPERATIONG SYSTEM

UNIT	MODULES	WEIGHTAGE
1	INTRODUCATION TO OPERATING SYSTEM	20 %
2	PROCESS MANAGEMENT	20 %
3	PROCESS COMMUNICATION AND SYNCHRONIZATION	20 %
4	MEMORY MANAGEMENT	20 %
5	FILE MANAGEMENT, DISK MANAGEMENT, SECURITY AND PROTECTION	20 %

UNIT -3 Process Communication & Synchronization

- Introduction to Process
- Concurrent Processes
 - Process Communication
- Semaphores
- Solution of Classic Synchronization Problem using Semaphores
 - Solution of Dining Philosophers Problem

UNIT -3 Process Communication & Synchronization

Deadlocks

- Introduction
- Defining Deadlocks
- Conditions for Deadlocks
- Dealing with deadlock
- Thread
 - Process and Thread
 - Multi-Tasking vs. Multi-Threading
 - Thread Control Block
 - Usage of Multi Thread
 - Types of Thread

Introduction to Process

- A process is basically a program in execution. The execution of a process must progress in a sequential fashion.
- To put it in simple terms, we write our computer programs in a text file and when we execute this program, it becomes a process which performs all the tasks mentioned in the program.

Concurrent Processes

- Concurrent processing is a computing model in which multiple processors execute instructions simultaneously for better performance.
- Concurrent means something that happens at the same time as something else. Tasks are broken down into subtasks that are then assigned to separate processors to perform simultaneously, instead of sequentially as they would have to be carried out by a single processor.
- Concurrent processing is sometimes said to be synonymous with parallel processing.

- The processes are interacting or communicating by :
 - Shared Variable
 - Message Passing
- Shared Variable
- There is a shared variable through which they communicate, that is, they are not aware of the existence of each other but coordinate with each other in the execution.
- In this way, there is an indirect communication through shared memory.

Message Passing

- There may be the case that the processes need to share data not required for data access synchronization or control synchronization but for reading purpose.
- In this case there is no need to maintain a shared data as it incurs the cost of accessing.
- The processes can also communicate through messge and be explicitly aware of the existence of each other.
- This type of communication known as message passing.

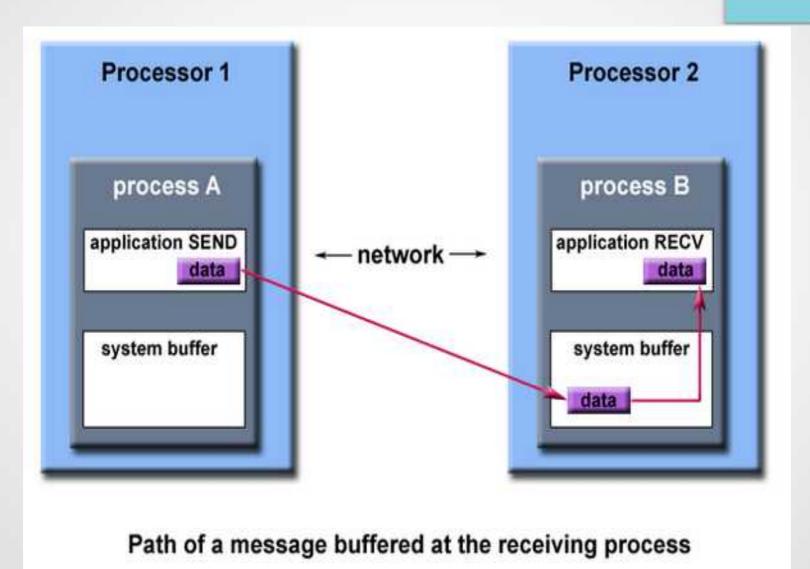
Message Passing

- It is used when processes explicity know each other and exchange message through system calls.
- One system call is used for sending the message and another for receiving it.
- The **message has a fixed format** consisting of a message and the name of its sender or receiver process.
- The process wishing to communicate a message with another process copies the message in its message structure with the specific name of the receiver.

Message Passing

- Similarly, when the receiver receives the message, it copies the messge into its local variable and starts executing.
- Therefore, there is no requirement to update the message concurrently, there is no need to maintain a shared variable.
- So Message passing system is more appropriate.

Message Passing Communication



- Message Passing Synchronization
 - The synchronization is also needed in a message passing system.
 - When a sender sends the message, it is not necessary that the receiver is ready to recieve it.
 - In this case, the sender will be blocked and the messge will be copied to a buffer.
 - It is activated only when the intended receiver will execute its system call for receiving the message.

Message Passing - Synchronization

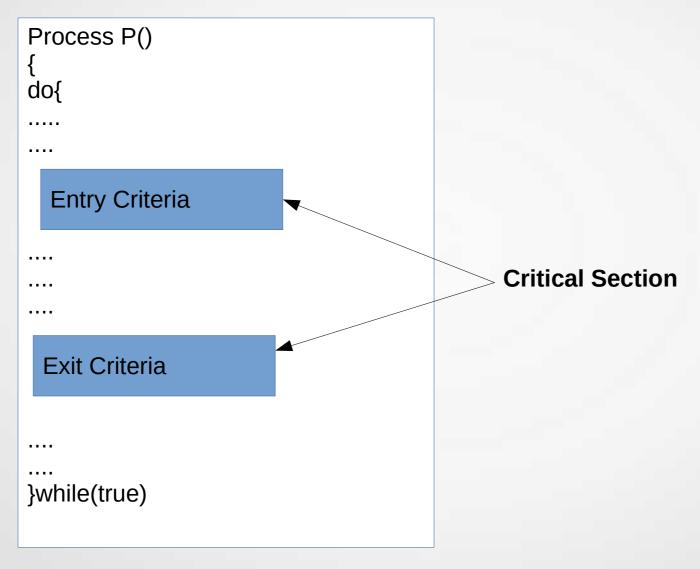
- Similarly, when a process is ready to receive a message, it is not necessary that the sender be ready to send it.
- In this case, the receiver will be blocked and activated only when the intended sender will send the message to it.
- Thus there should be synchronization between the sender and the receiver process.

Shared Variable

 It can be realized in both the cases of communication and synchronization that a shared variable is necessary to have a proper synchronized execution of the process.

- When a process is given access to a non shareable resource such as printer, another process cannot be given the access at the same time; otherwise, output of two processes will be mixed up and no process will get the desired result.
- This requirement is know Mutual Exclusion.
- Mutual Exclusion ensures that the processes do not access or update a resourcce concurrently.

- The issues identified in the execution of concurrent processes lead to the design of a protocol that enforces mutual exclusion and does not cause race condition problem.
- The protocol demands that the code of a process where it modifies the data or accesses the resource be protected and not be allowed to be executed by concurrent processes simultaneously.
- There fore the first thing in the protocol is to identify a section of the code that must be protected. This section is known as a critical section.



- To protect the CS, there should be **some conditional criteria** that a process must pass before entering the critical section.
- Only one process at a time is allowed to enter its critical section.
- If another process requests at this time, then the access will be denied so that mutal exclusion is maintained.
- The criteria for entering the CS are known as entry Criteria.
- The protocol should also define that when one process has finished its execution, the other waiting processes must be informed so that one of them can enter the critical secition.

 The solution for CS problems should satisfy the following characteristics of protocol:

Mutual Exclusion

 The protocol should not allow more than one process at a time to update a global data.

- Progress

• The implementation of this protocol requires that the **track** of all the processes wishing to enter their critical sections be kept.if there are some processes in the queue waiting and, currently, no process is executing inside the critical section, then only the wating processes must be granted for execution.

 The solution for CS problems should satisfy the following characteristics of protocol:

Bounded Wait

 A process in its CS remains there only for a finite time because, generally the execution does not take much time, so the waiting processes in the queue need not wait for a long time.

No Deadlock

 If no process is executing inside the CS and there are processes in the queue wiaiting for the execution but are not allocated the CS yet, then there may be a deadlock in the sytem. There fore the protocol must ensure that all the waiting processes are allotted the critical section.

- The operation that can not be overlapped or interleaved with the execution of any other operations are known as individual or atomic operation
- The semaphore is use to protect any resource such as global shared memory that needs to be accessed and updated by many processes simultaneously.
- Semaphore acts as a guard or lock on the resource.
- Whenever a process needs to access the resource, it first needs to take permission from the semaphore.
- If the resource is free, that is, if no other process is accessing or updating it, the process will be allowed, otherwise permission is denied.
- In case of denial, the requesting process needs to wait until semaphore permits it, that is, when the resource becomes free.

- The semaphore is implememented as an integer variable, say as S, and can be initialized with an positive integer value.
- The semaphore is accessed by only two operations known as wait and signal operations, denoted by
 - Wait ---> P
 - Signal ---> V
- When ever a process tries to enter the critical section, it needs to perform wait operation.
- The wait is an entry criterion according to the designed protocol.

```
Operation Wait
P(S)
   while (S <= 0);
   S=S-1;
Operation Signal
V(S)
   S = S + 1;
```

```
do{
wait (Semaphore)
    Critical
    Section
signal (semaphore)
```

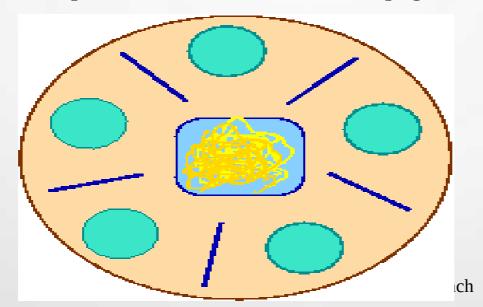
- Whenever a process tries to enter the critical section, it needs to perform wait operation. The wait is an entry criterion according to the designed protocol.
- If the CS is free or no other process is using it, then it is allowed, otherwise denied. The count of semaphore is decremented when a process accesses the available critical section; hence, the count of semaphore tells us the availability of the critical section.
- Initially, the count of semaphore is 1. If it is accessed by a process, then the count is decremented and becomes zero. Now, if another process tries to access the critical section, then it is not allowed to enter unless the semaphore value becomes greater than zero.
- When a process exits the critical section, it performs the signal operation, which is an exit criterion.

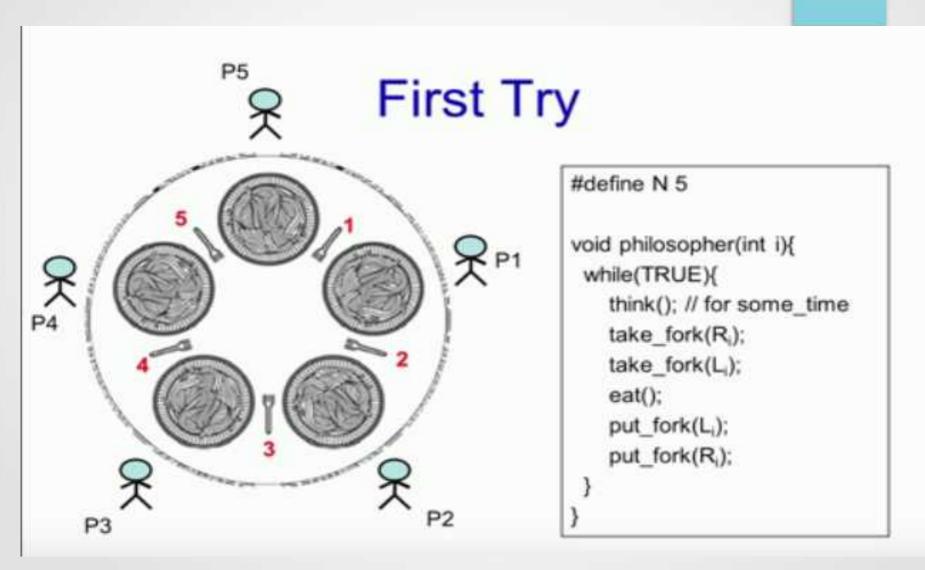
- The semaphore whose value P(S) is either zero or one is known as binary semaphore.
- There is one problem in the implementation of such a semaphore. When a process does not get access to the critical section, it loops continually waiting for it.
- This does not produce any result but consumes CPU cycles, thereby wasting the processor time. This busy waiting is a problem in a multi-programming system where processor time is shared among the processes.
- This type of semaphore is known as a spinlock, since the process spins while waiting for the lock.

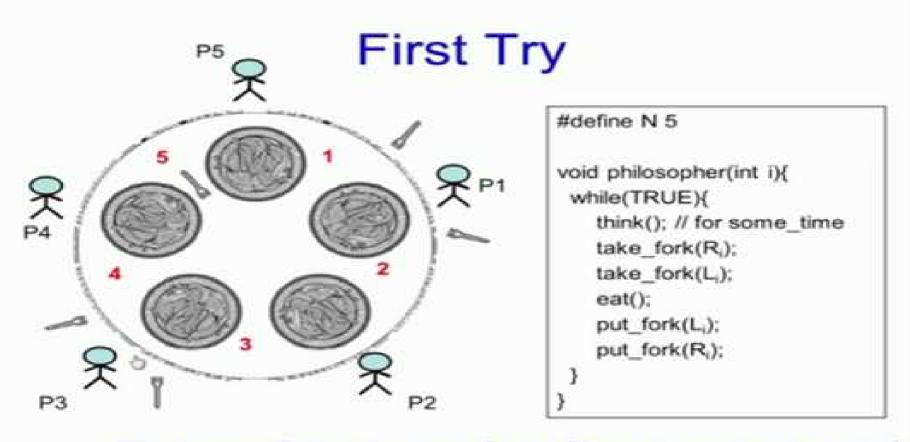
- The semaphore whose value is either zero or one is known as binary semaphore.
- The type of semaphore that takes a value greater than one is known as **counting semaphore**.
- In a binary semaphore, the CS locked by process may be unlocked by any other process is called **mutex semaphore.**

Dining Philosophers Problem

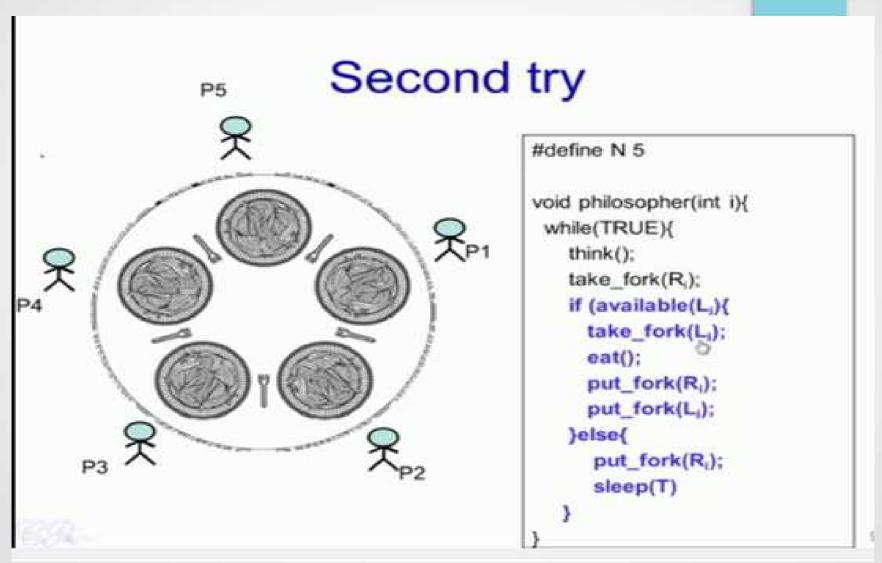
- The Dining Philosophers problems is a classic synchronization problem.
- There is a dining room containing a circular table with five chairs. At each chair is a plate, and between each plate is a single chopstick. In the middle of the table is a bowl of spaghetti. Near the room are five philosophers who spend most of their time thinking, but who occasionally get hungry and need to eat so they can think some more.
- In order to eat, a philosopher must sit at the table, pick up the two chopsticks to the left and right of a plate, then serve and eat the spaghetti on the plate.





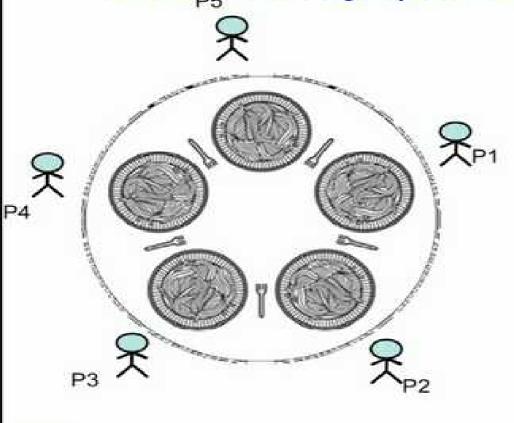


What happens if only philosophers P1 and P3 are always given the priority? P4, P5, and P2 starves... so scheme needs to be fair



Second try P5 Imagine, All philosophers start at the same time Run simultaneously And think for the same time This could lead to philosophers taking fork and putting it down continuously, a deadlock. while(TRUE){ think(); take_fork(R_i); if (available(Li){ take_fork(Li); eat(); put_fork(R_i); put_fork(Li); }else{ put_fork(R_i); sleep(T)

Second try (a better solution)



```
#define N 5
void philosopher(int i){
 while(TRUE){
    think():
    take_fork(R<sub>i</sub>);
    if (available(L<sub>i</sub>){
       take_fork(Li);
       eat();
       put_fork(L<sub>i</sub>);
       put_fork(Ri);
    }else{
       put_fork(R<sub>i</sub>);
       sleep(random_time);
```

Solution with Mutex

Solution using Mutex

- Protect critical sections with a mutex
- Prevents deadlock
- But has performance issues
 - Only one philosopher can eat at a time

```
#define N 5

void philosopher(int i){
  while(TRUE){
    think(); // for some_time
    lock(mutex);
    take_fork(R<sub>i</sub>);
    take_fork(L<sub>i</sub>);
    eat();
    put_fork(L<sub>i</sub>);
    put_fork(R<sub>i</sub>);
    unlock(mutex);
}
```



Solution with Semaphores

Solution with Semaphores

Uses N semaphores (s[1], s[2],, s[N]) all initialized to 0, and a mutex Philosopher has 3 states: HUNGRY, EATING, THINKING A philosopher can only move to EATING state if heither neighbor is eating

```
void philosopher(int i){
    while(TRUE){
        think();
        take_forks(i);
        eat();
        put_forks();
    }
}
```

```
void take_forks(int i){
    lock(mutex);
    state[i] = HUNGRY;
    test(i);
    unlock(mutex);
    down(s[i]);
}
```

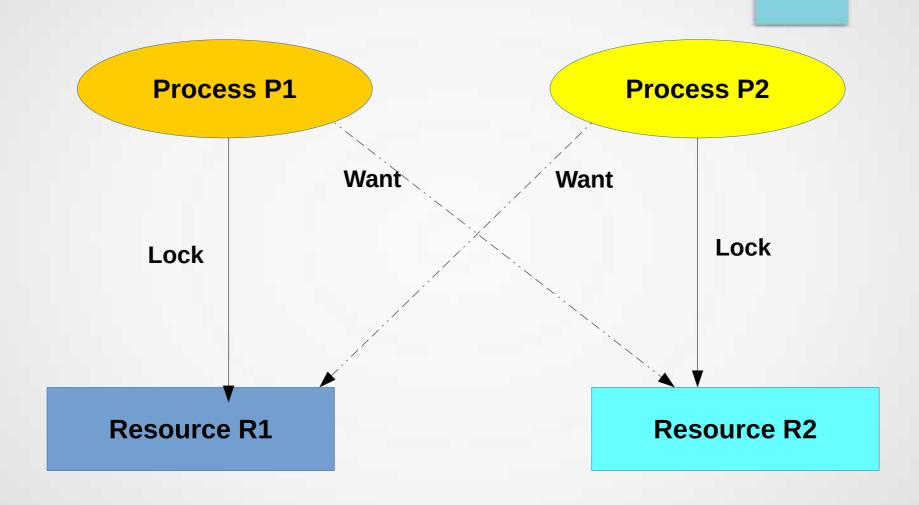
```
void put_forks(int i){
    lock(mutex);
    state[i] = THINKING;
    test(LEFT);
    test(RIGHT)
    unlock(mutex);
}
```

```
void test(int i){
   if (state[i] = HUNGRY && state[LEFT] != EATING && state[RIGHT] != EATING){
        state[i] = EATING;
        up(s[i]);
   }
}
```

Deadlocks

- If there is no control on the competing processes for accessing multiple resources, then it can lead to a severe problem in the system.
- For example, a Process P1 is accessing a Resource R1 and needs another Resource R2 to complete its execution, but at the same time, another Process P2 is holding a Resource R2 and requires R1 to proceed.
- Here, both the processes are waiting for each other to release the resources held by them. This situation is called a deadlock.
- There can be many situations like this causing deadlock problems in the system

Deadlocks



DeadLock – Defining a DeadLock

- The computer system uses several types of resources, such as consumable or non-consumable, and pre-emptable or non-pre-emptable.
- In a multi-programming environment, the problem starts with non-pre-emptable resources. These resources have to be used in a mutually exclusive manner. But the problem is that resources are always limited, compared to the number of processes.
- In this case, when the concurrent processes request the resources, a deadlock occurs.

- In general, the OS follows a protocol to use non-pre-emptive resources.
 The protocol consists of the following three events:
- 1. Request: Before using a resource, a process must request for it. This request is in the form of a system call, and it is passed on to the kernel. This uses a data structure resource table to store the information about all the resources in the system. When- ever the request is received, it checks the status of the resource. If the resource is free, the requesting process will get it. Otherwise, the requesting process will be blocked, and it must wait until the resource is free.
 - **2.Use:** If a process gets the resource, the kernel changes the status of resource to allocated and the state of process to ready, if it has been blocked earlier. The process uses the resource in a mutually exclusive manner.
 - 3.Release: After its use, a process releases the resource again through the use of a system call. The kernel can now allocate this resource to any blocked process or change the status of the resource to free.

 37

Conditions for Deadlock

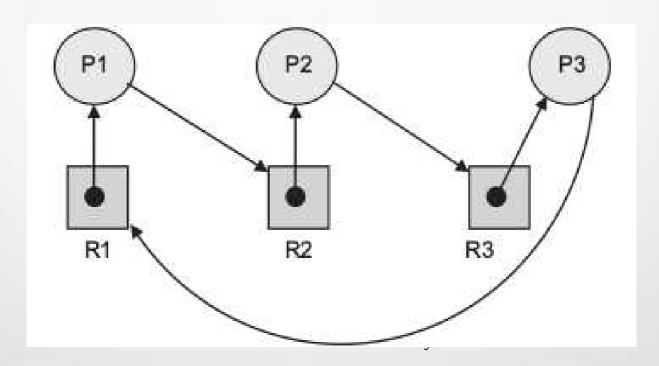
- There are certain conditions that give rise to a dead lock.
 The following are the conditions:
 - Mutual Exclusion
 - Hold and Wait
 - No Pre-emption
 - Circular Wait

Conditions for Deadlock

- It is not necessary that a deadlock will always occur in concurrent process environments. In fact, there are certain conditions that give rise to a deadlock. The following are the conditions:
 - 1 Mutual Exclusion The resources which require only mutually exclusive access, may give rise to a deadlock. These resource types cannot allow multiple processes to access it at the same time. For example, a memory location, if allocated to a process, cannot be allowed to some other process. A printer cannot print the output of two processes at the same time. Therefore, a printer, if allocated to a process, cannot be allocated to another.
 - 2 Hold and Wait When all the processes are holding some resources and waiting for other resources, a deadlock may occur.
 - 3 No Pre-emption The resources in a deadlock situation are of nonpre-emptable nature, as discussed above. If a resource cannot be pre-empted, it may lead to a deadlock.
 UNIT - 3 Process Communication & Synch

Conditions for Deadlock

• These three conditions together give rise to a fourth condition. A chain of processes may be produced, such that every process holds a resource needed by the next process. In this way, a circular chain of processes is built up in the system, due to mutually exclusive and non-pre-emptable nature of resources.



Dealing With Deadlock

- Every system may be prone to deadlocks in a multi-programming environment, with few dedicated resources. To deal with the deadlock, the following three approaches can be used:
- **Deadlock Prevention:** This method is very idealistic. It is based on the fact that if any of the four necessary conditions is prevented, a deadlock will not occur.
- Although it is very difficult to achieve this kind of prevention, this method is better, in the sense that the conditions for deadlock will not be allowed to arise in the system. Moreover, a principle is established that a discipline is maintained, while the processes request for the resources.

Deadlock Avoidance

- **Deadlock Avoidance:** Deadlock prevention is not always possible. Another method to deal with the deadlock is to avoid it.
- But avoidance is possible when there is complete knowledge in advance, about which resources will be requested and used by the processes.
- At any instant of time, information such as, how much resources are available, how many are allocated to the processes, and what the future requests are, must be known.
- With this information, a process may be delayed, if its future request may produce a deadlock. In this way, the deadlock may be avoided.

Deadlock Detection & Recovery

- **Deadlock Detection and Recovery:** If a system does not employ either prevention or avoidance methods, then a deadlock may occur.
- In this case, it becomes necessary to detect the deadlock. At least, it must be known that a deadlock has occurred in the system.
- The processes, due to which the deadlock has occurred, should also be known.
- We should have a mechanism to detect the deadlock in the system and resolve it, that is, recover from the deadlock situation.

Thread

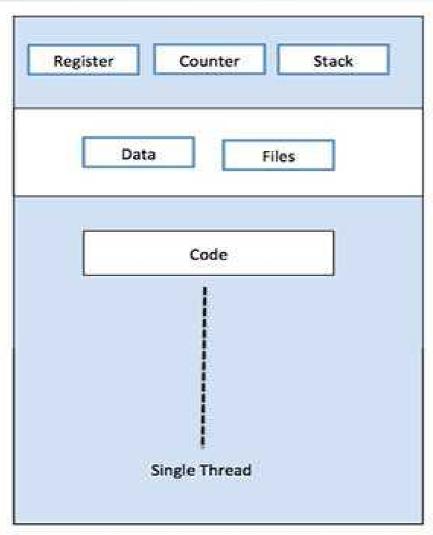
Thread: A thread is a flow of execution through the process code, with its own program counter that keeps track of which instruction to execute next, system registers which hold its current working variables, and a stack which contains the execution history.

• A thread is also called a lightweight process. Threads provide a way to improve application performance through parallelism. Threads represent a software approach to improving performance of operating system by reducing the overhead thread is equivalent to a classical process.

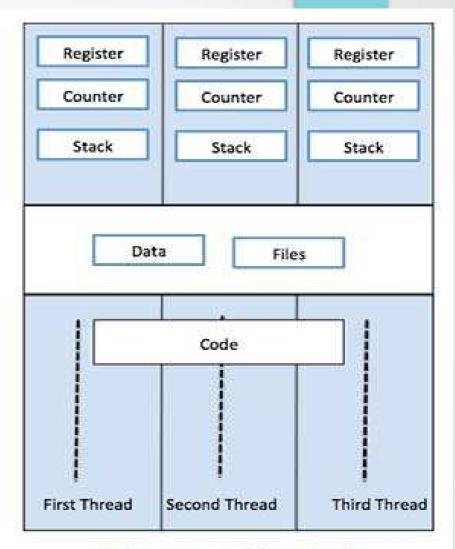
Thread

• Each thread belongs to exactly one process and no thread can exist outside a process. Each thread represents a separate flow of control. Threads have been successfully used in implementing network servers and web server. They also provide a suitable foundation for parallel execution of applications on shared memory multiprocessors.

Thread



Single Process P with single thread



Single Process P with three threads

Difference: Process-Thread

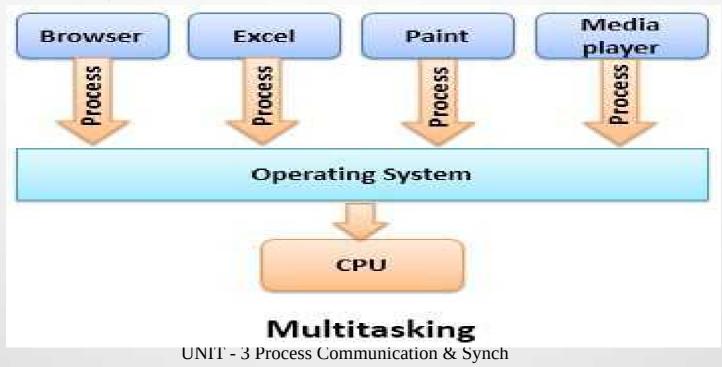
S.N.	Process	Thread
1	Process is heavy weight or resource intensive.	Thread is light weight, taking lesser resources than a process.
2	Process switching needs interaction with operating system.	Thread switching does not need to interact with operating system.
3	In multiple processing environments, each process executes the same code but has its own memory and file resources.	All threads can share same set of open files, child processes.
4	If one process is blocked, then no other process can execute until the first process is unblocked.	While one thread is blocked and waiting, a second thread in the same task can run.
5	Multiple processes without using threads use more resources.	Multiple threaded processes use fewer resources.
6	In multiple processes each process operates independently of the others.	One thread can read, write or change another thread's data.

Multitasking vs Multi threading

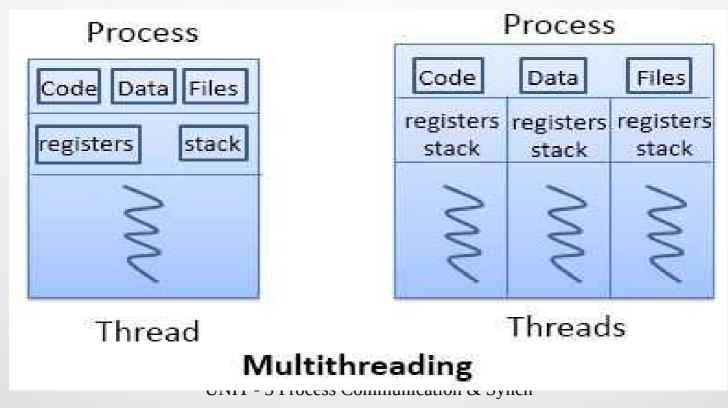
- "The concept of implementing multiple threads to achieve concurrency within a single process is known as multithreading."
- Multi threading is implemented at thread level.
- Multi tasking or Multi user is implemented at process level.

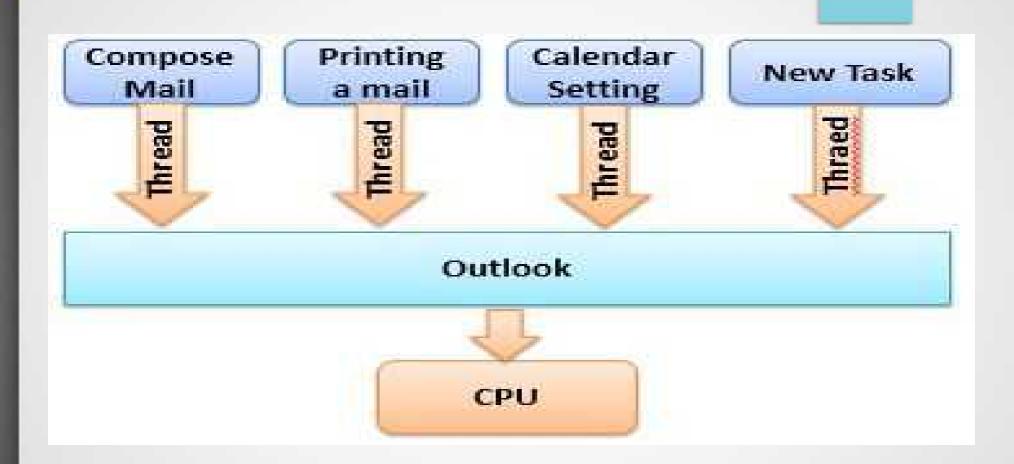
Multitasking vs Multi threading

• Multitasking is when a single CPU performs several tasks (program, process, task, threads) at the same time. To perform multitasking, the CPU switches among theses tasks very frequently so that user can interact with each program simultaneously.



• Multithreading is different from multitasking in a sense that multitasking allows multiple tasks at the same time, whereas, the Multithreading allows multiple threads of a single task (program, process) to be processed by CPU at the same time.





BASIS FOR COMPARISON	MULTITASKING	MULTITHREADING
Basic	Multitasking let CPU to execute multiple tasks at the same time.	Multithreading let CPU to execute multiple threads of a process simultaneously.
Switching	In multitasking CPU switches between programs frequently.	In multithreading CPU switches between the threads frequently.
Memory and Resource	In multitasking system has to allocate separate memory and resources to each program that CPU is executing.	In multithreading system has to allocate memory to a process, multiple threads of that process shares the same memory and resources allocated to the process.

Thread Control Block

- Just like the PCB, in multi-threading, there is a thread control block (TCB) to save and restore the context of a thread in case of thread switching.
- The only difference is that the information in the TCB is lesser as compared to the PCB. The fields associated with a TCB are
 - 1. Thread ID (TID): It is a unique identification number of the thread.
 - 2. PC: Indicates the address value at which the next instruction of the thread will be executed by the processor.
 - 3. Registers: CPU registers are used for the execution of a thread. While the thread is in execution, data registers, address registers, control registers, and status registers are used for executing and controlling the process. The information in registers must be saved when there is a change in the state of the threads so that it may resume its execution in its nexturn Process Communication & Synch

TCB

- 4. State: A thread also has a number of states just like a process. For scheduling the threads, the current state of a thread must be known.
- 5. Priority: A priority number may be assigned to a thread as provided to a process.
- 6. Event information: This is the event for which a blocked thread is waiting. If the awaited event is over, the information regarding this event must be stored in this field so that the status of the blocked thread is changed to ready.
- 7. Information related to scheduling: The information related to scheduling of a thread, such as the waiting time and the execution span of the thread the last time it was running, is also stored.

TCB

- 8. Pointer to owner Thread: The thread will be created within a thread and it needs to access the execution environment of its owner/ parent thread. Thus, the TCB contains a pointer to this information.
- 9. TCB pointer: This is a pointer to another TCB that is used to maintain the scheduling list.

UNIT – 3 Usage of Multi-Threading

Low Context Switch Overhead

- Switching the inforamtion to be save and retrieved is less as compared to the process swithcing.

High Computation Speed up

 Due to less information in the TCB and low context switch overhead, there is a tremendous increase in computation speed.

No Need of IPC

 The threads in a multithreaded process do not require system calls to communicate with each other.

UNIT – 3 Usage of Multi-Threading

Decreased Response time

 As a consequence of the increased computation speed, the response time to a user decrease inspite of the execution of many other tasks.

Sharing of Resources

 All the threads share the resources allocated to their owner process.

Efficient Management

 It is easy and economical in terms of memory and speed to create and destroy a thread as compared to a process.

UNIT – 3 When to use Multi Threading instead of Multi Processing

- If the application need to divided into multiple prcesses.
- If process have several tasks inside.
- If application are highly interactive.

UNIT – 3 Types of Threads

- Threads can be of TWO types:
 - User Threads
 - Kernel Threads

UNIT – 3 User Threads

- The user threads are managed by the application in user space.
- By default, an application starts running with a single thread of control, that is as a process managed by the kernel.
- User Threads are created and managed through a Thread library provided by the Os.

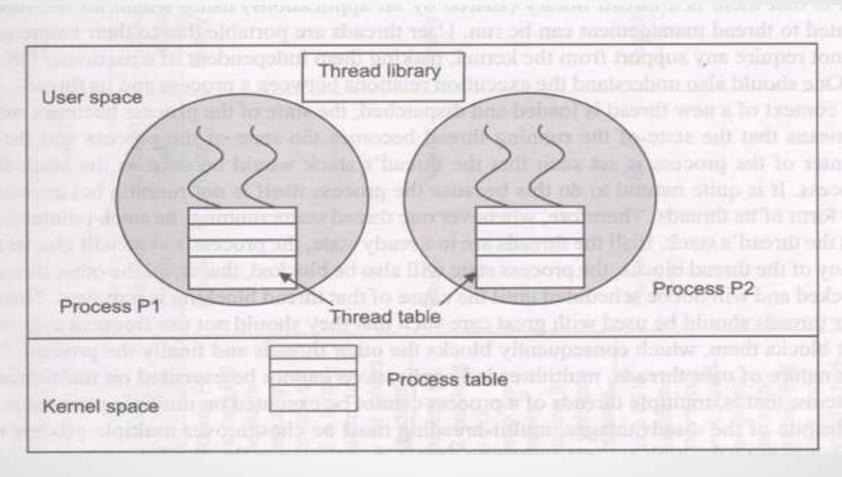
UNIT - 3 User Threads

- Thread Library / Package has following functions:
 - Thread creation and deletion
 - Assigning priorites to the threads
 - Thread Scheduling
 - Thread Synhronization
 - Communication between the threads
 - Saving and restoring contexts of the threads
 - Blocking and resuming the threads

UNIT - 3 User Threads

- A user thread, is managed through the library functions.
- First of all a process spawns a thread by calling a procedure.
- The thread related data structures are then stored in the TCB.
- The process needs to keep track of all the threads within it.
 For that process used Thread table.
- Thread table contains entries as a pointer to the TCB of each thread.

UNIT - 3 User Threads



User Thread

UNIT – 3 Kernel Threads

- The kernel threads are managed by the kernel.
- Kernel threads, implemented in the kernel space, are managed through system call.
- Operation such as thread creation, delection, context saving, and restoring, synchronization and so on, are implemented through system call.
- There is no seperated thread table corresponding to each thread as in user threads.
- All the thread are managed through a single thread table maintained in the kernel space.

UNIT – 3 Kernel Threads

