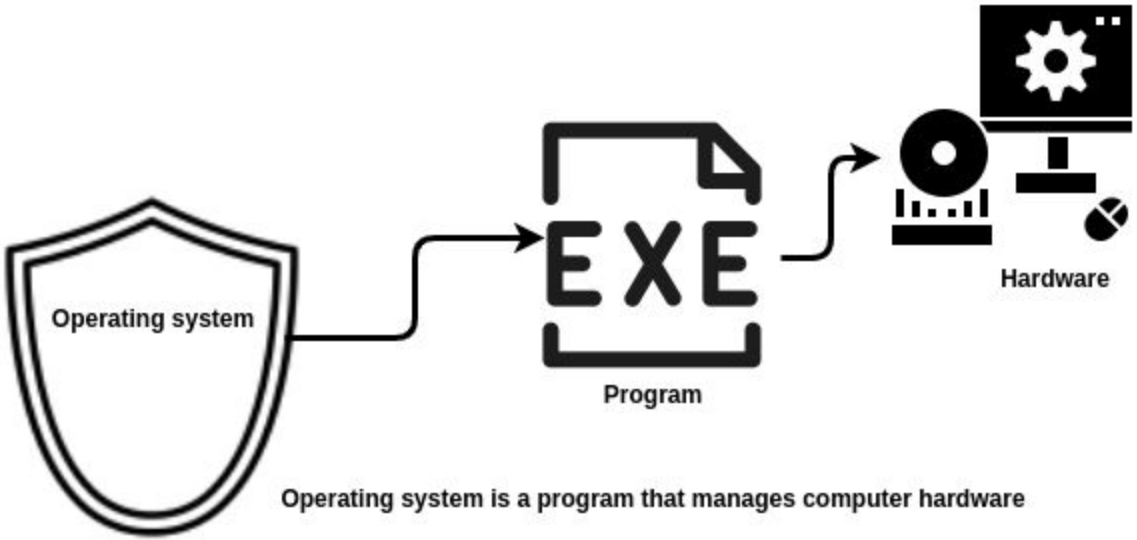
Module: Operating System Concepts
Session: CPU Scheduling

By
Donkada Mohana Vamsi
C-DAC Hyderabad

Operating system



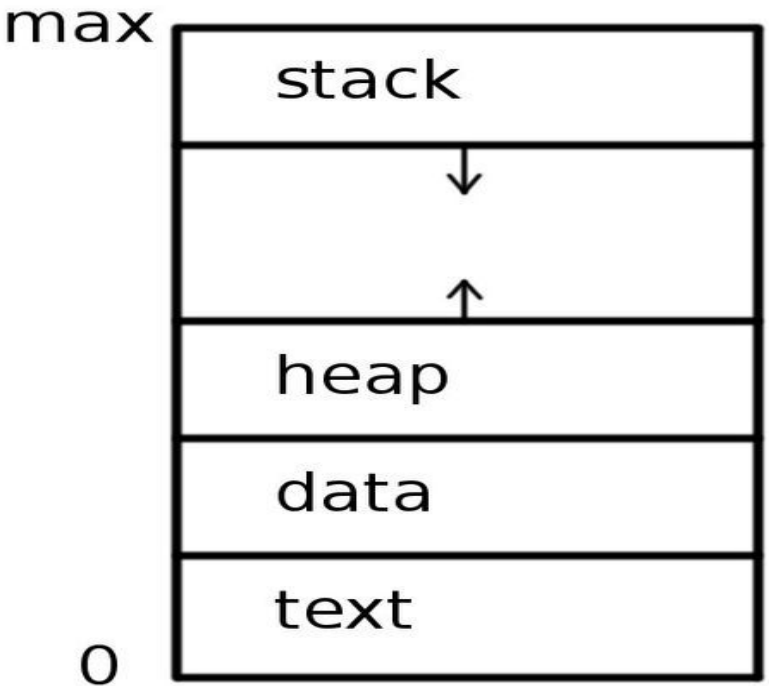
Definition of an Operating System

- Operating System (OS) is a **resource allocator**
  - Manages all resources
  - Ensures an efficient and fair resource use

Definition of a Process

- A Process is a program under execution
- A process includes:
  - program counter
  - Stack
  - data section

Process in Memory



States of a process

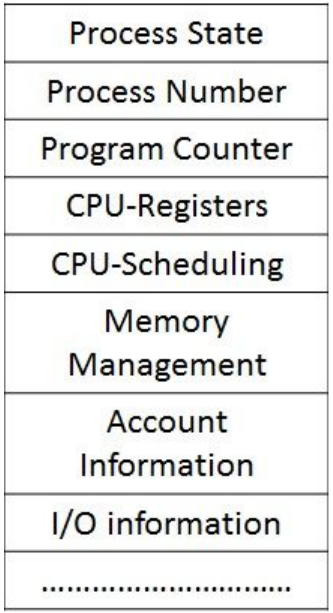
- Various states of a process are:
  - **new**: In this state, the process is about to be created but not yet created
  - **ready**: After the creation of a process, the process enters the ready state i.e. the process is loaded into the main memory
  - **running**: The process is chosen by CPU for execution and the instructions within the process are executed
  - **waiting**: The process is waiting for an I/O operation to happen or needs input from user or needs access to a critical region
  - **terminated**: The process has finished it's execution

Process Control Block (PCB)

A PCB is a data structure which stores all the information related to a process such as

- Process state
- Process number
- Program counter
- CPU registers
- CPU scheduling
- Memory-management

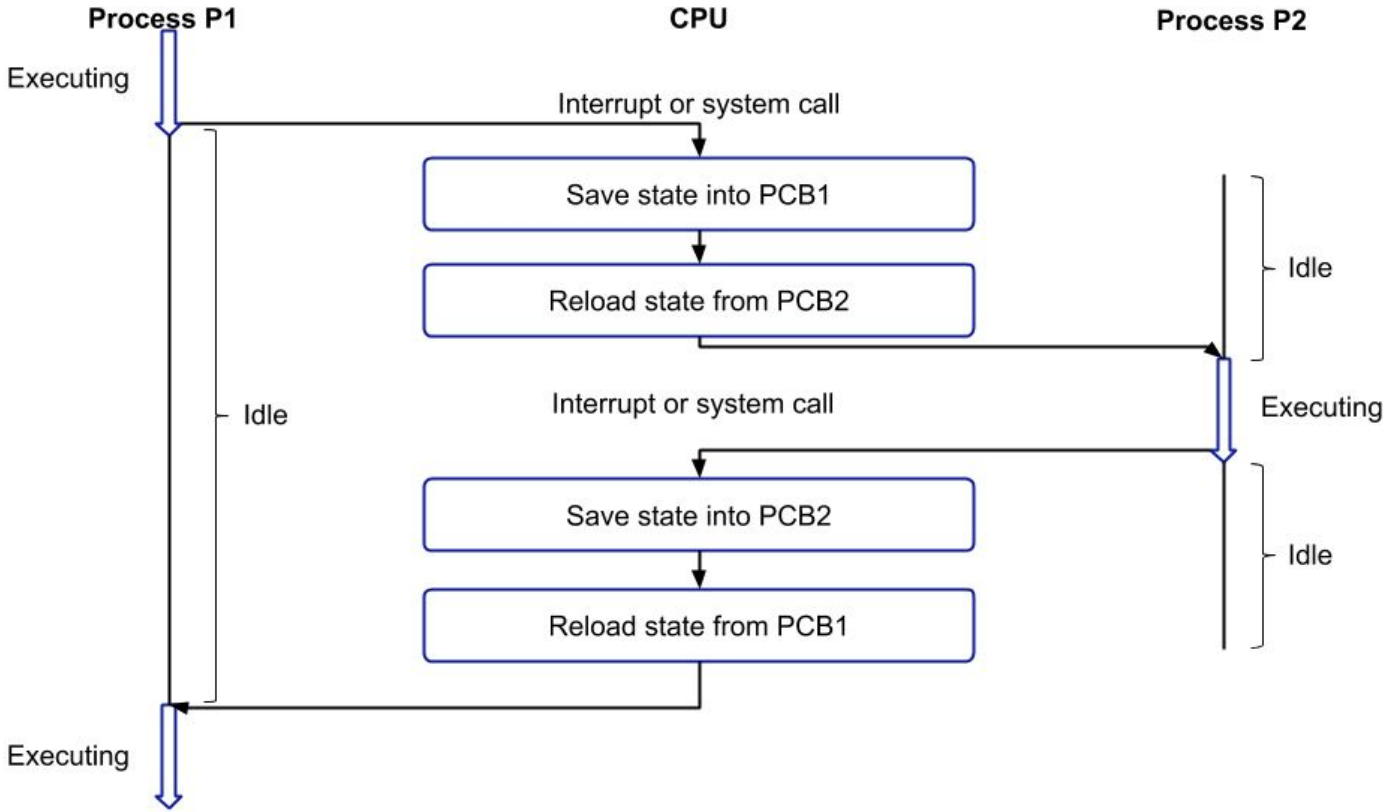
Process Control Block (PCB)



Context Switch

- Whenever CPU switches from one process to the other process, there is a need to save the state of the old process and load the state of the new process
- This mechanism is called context switching
- This saved state would be used whenever the execution of the process is resumed
- Context switching is a very time consuming process and a overhead to the CPU as CPU does no useful task during the context switching

Context Switching



Process Scheduling Queues

- Job queue – list of all processes that are in the secondary storage and wait for main memory allocation
- Ready queue – set of all processes residing in main memory and are awaiting to be scheduled on a CPU
- Device queues – Processes which are waiting for any particular device are put in that particular device queue

I/O Bound vs CPU Bound processes

- I/O Bound Process: A process which performs more I/O operations and spends less time in CPU
- CPU Bound Process: A process which spends more time in CPU and performs less I/O operations

Operating system main goal

- Effective CPU utilization

The above goal can be achieved through - multiprogrammed operating system.

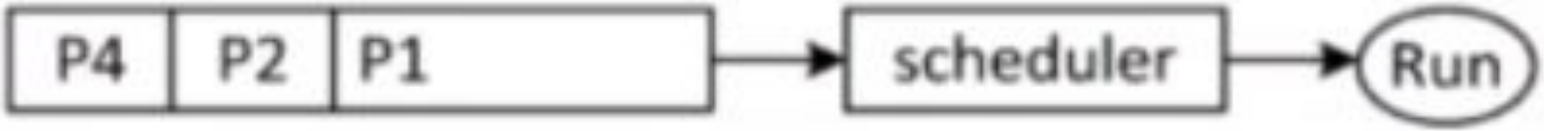
The multiprogrammed operating systems bring the necessity of switching CPU among multiple process to get the better productivity

The concept of sharing cpu across multiple process can be effectively explained with the concept of

CPU-Scheduling algorithms.

CPU Scheduling

When the jobs are ready in ready queue then process to be scheduled to execute on CPU.



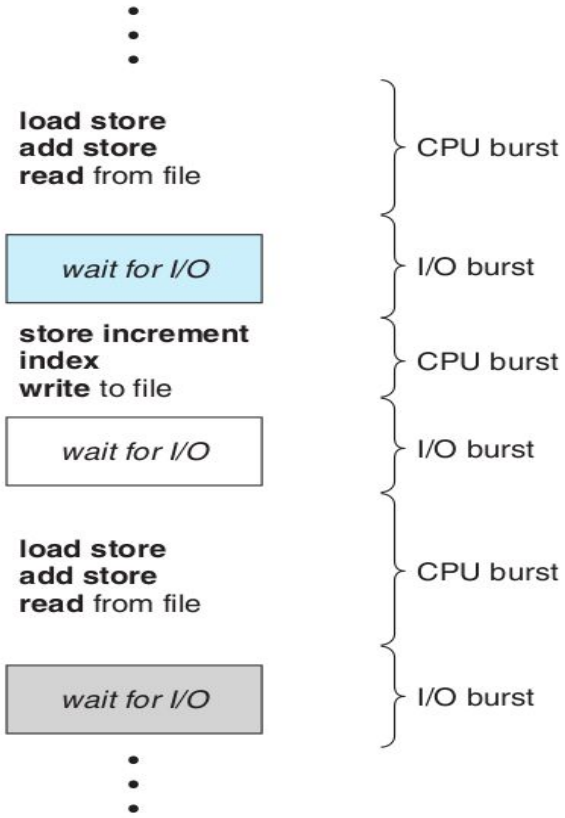
CPU Scheduling - Terminology

CPU Burst - The amount of time period for which a particular process utilizes the CPU is called as CPU Burst

Normally during process execution , it need go for some I/O operations (i.e Reading data from a file / Writing data to a particular file)

The amount of time a particular process spend for I/O operations is called as **I/O Burst**

Alternative sequence of CPU and I/O bursts.



CPU Scheduling - Terminology

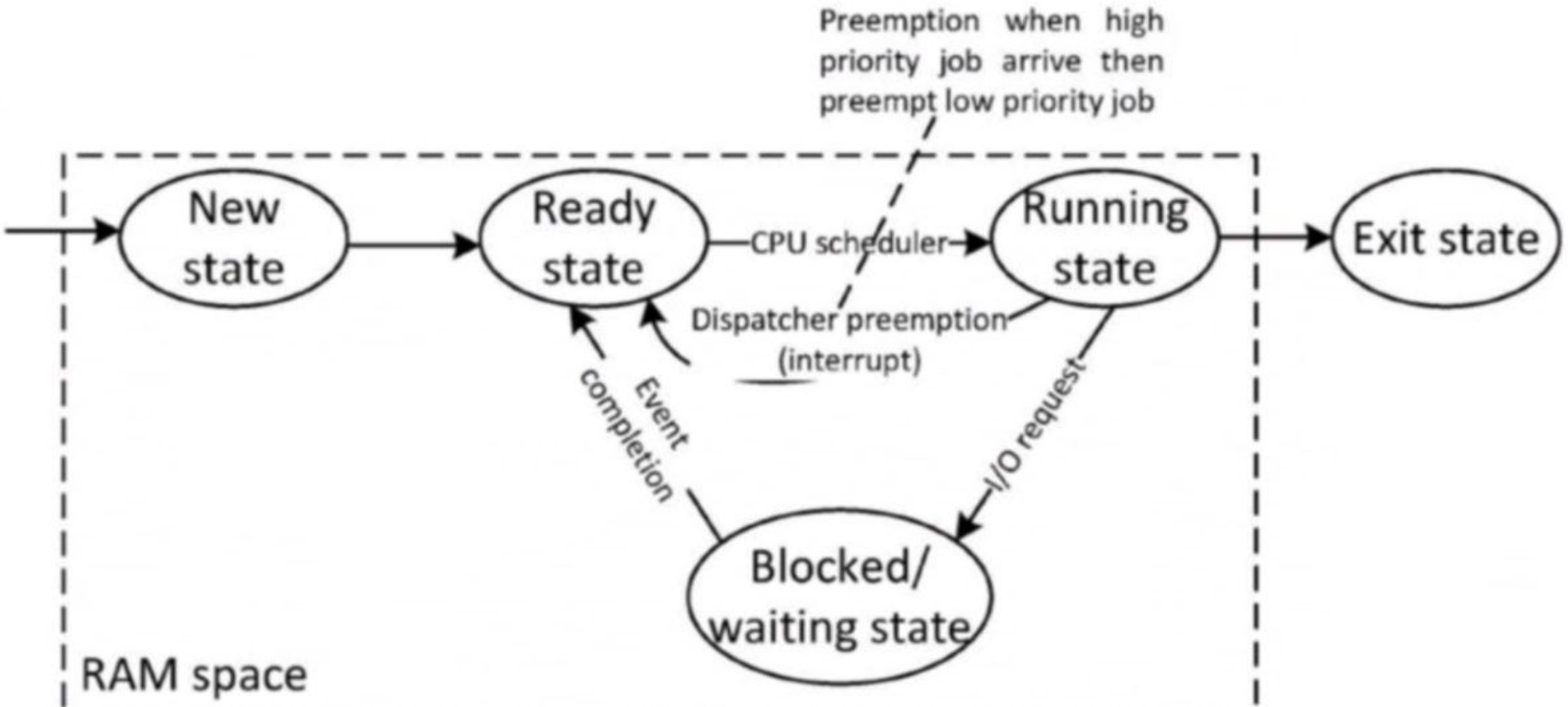
CPU Scheduling - The CPU scheduling can be defined as a task of choosing a process which is waiting in the ready queue and assigning that process to the CPU

- CPU scheduling decisions may take place when a process:

1. Switches from running to waiting state
2. Switches from running to ready state
3. Switches from waiting to ready
4. Terminates

- Scheduling under 1 and 4 is *nonpreemptive*
- All other scheduling is *preemptive*

CPU Scheduling



Non Preemptive transitions

Running ---- Waiting (Due to I/O request of running process / Self invocation of wait [i.e for termination of child process first])

Terminate - A process terminating itself after fulfilling its execution is also comes under non preemptive

Process leaving the cpu voluntarily

Preemptive transitions

Running - Ready (i.e when high priority job comes then preempt the low priority job)

Waiting - Ready (i.e the completion of I/O request)

Scheduler vs Dispatcher

Scheduler

Schedulers are system software that handle the removal of the running process from the CPU and the selection of another process. It selects a process out of several processes that need to be executed. [**Selection of the Process**]

Dispatcher .[**Transfer of the process state**]

Dispatcher starts functioning once the scheduler's functions are over. It takes the process to the desired state or queue. When the short term scheduler selects the process from the ready queue, the dispatcher performs the task of allocating the selected process to the CPU. When the running process goes to the waiting state than CPU is allocated to some other process. The switching of CPU from one process to another is called **context switching**

Scheduling Criteria

Different CPU-scheduling algorithms have different properties .

The effectiveness of an algorithm also depends on the class of processes.

The important factors (i.e characteristics) helps to judge an algorithm as best normally include

- CPU utilization
- Throughput
- Turnaround time
- Waiting time
- Response time

Scheduling Criteria

CPU utilization - The CPU utilization is effective, when the maximum effort of CPU goes in process execution.

Keeping the CPU busy all the time is the benchmark for CPU utilization.

The CPU utilization normally ranges from 0 to 100 percentage

Scheduling Criteria

Throughput - It is defined as no of processes that completed per unit time.

It means with effective CPU utilization, throughput also gets increased

Arrival time - The time when the process has arrived into read state is call the arrival time of the process

Burst time - The time required for the process to complete its execution is calle burst time

Completion time - The time when the process complete its execution is called the completion time of the process

Scheduling Criteria

Turnaround time - Interval between the submission of the process and its completion is called Turnaround time .

The time difference between the completion time and arrival time is called the turnaround time of the process.

TAT = Completion/Finishing time - Arrival time

Or

Waiting time + Burst time

Scheduling Criteria

Waiting time - The time difference between turnaround time and burst time is called waiting time

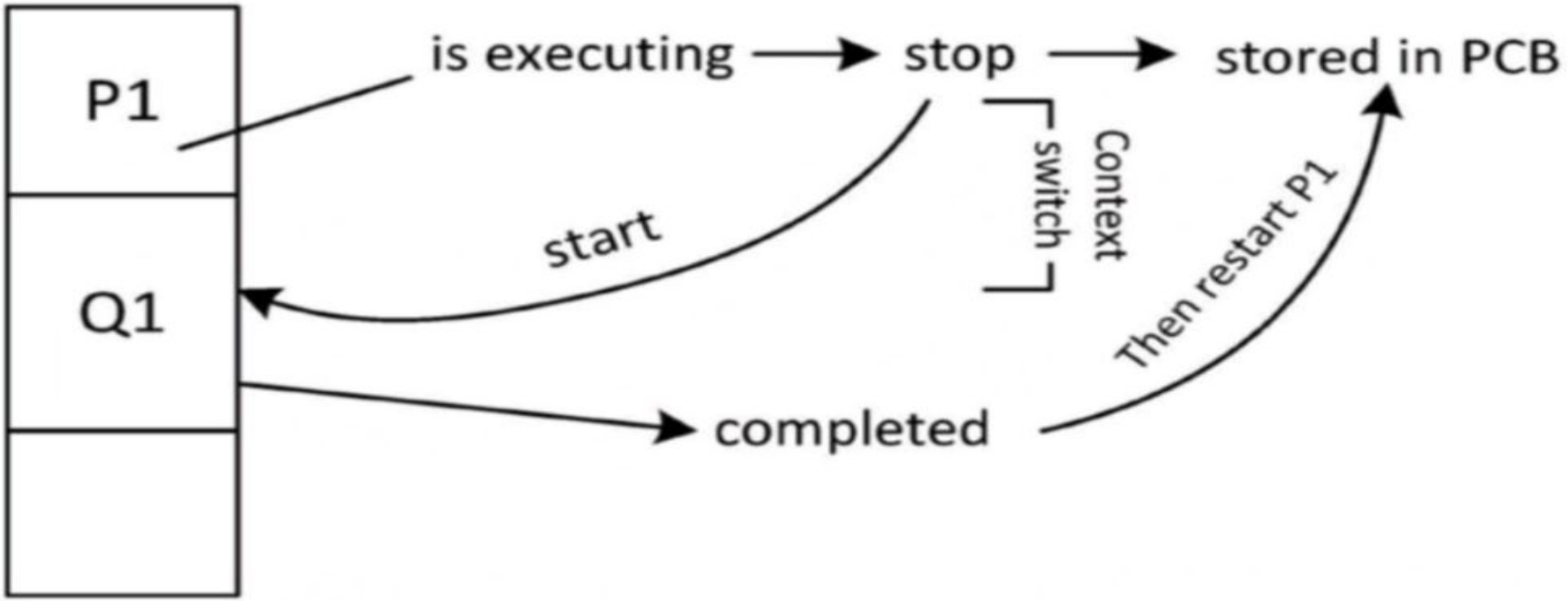
WT = TAT- BT

It is nothing but, The amount of time a job is not executing.

Response time - The time difference between first response and arrival time is called Response time.

Dispatch latency - The time taken by the dispatcher to do the context switch

Scheduling Criteria



Thank you