CPU Scheduling

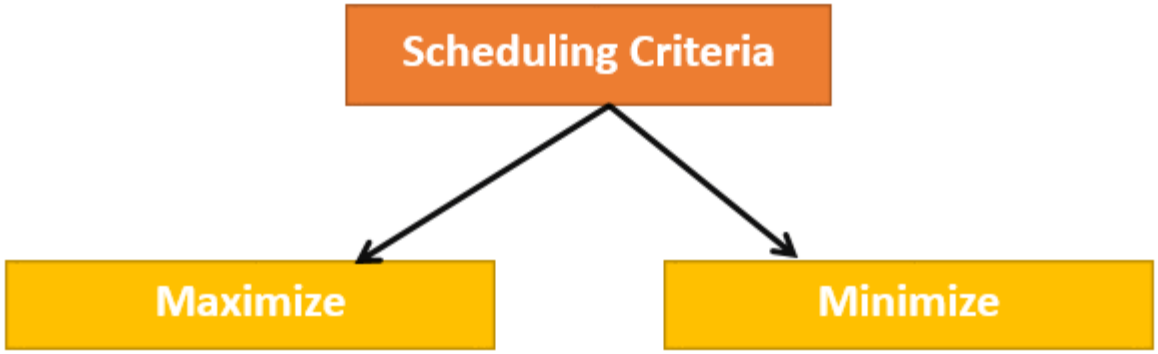
Basic Concepts

- Maximum CPU utilization obtained with multiprogramming
- CPU utilization – keep the CPU as busy as possible

Important Terminologies:

- Arrival time: It is the time instance at which new process arrives into the system
- Finishing time: It is the time instance at which process gets terminated in the system
- Burst time: It is the predicted time required for the process to complete its job
- Waiting time: It is the some of discrete time instances during which process under goes waiting state
- Turnaround time: Time taken by a process from its arrival to completion.
- Throughput – Number of processes completed per time unit

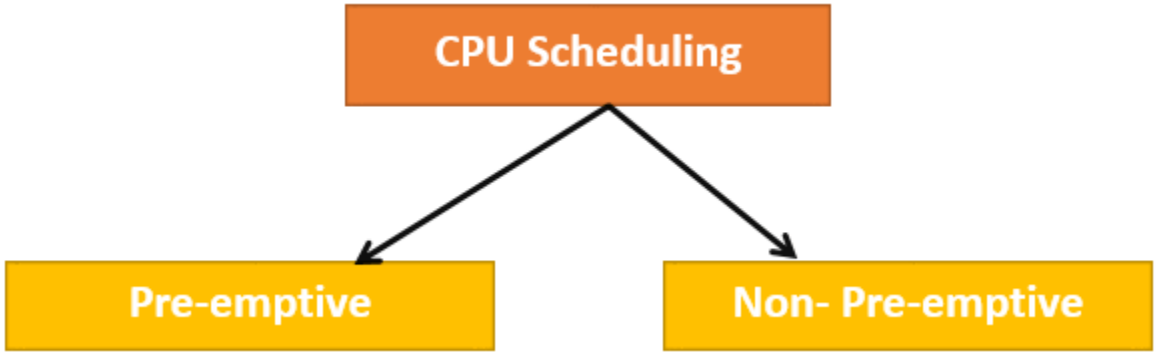
Scheduling Criteria

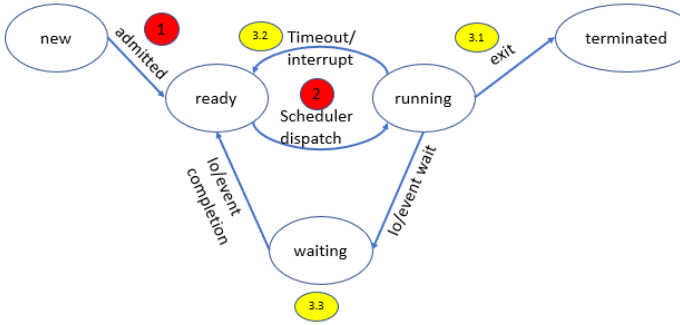


Maximize:
CPU Utilization
Throughput

Minimize:
Turnaround time
Waiting time

Types of CPU Scheduling Methods





Preemptive Scheduling

- In Preemptive Scheduling, the tasks are mostly assigned with their priorities. Sometimes it is important to run a task with a higher priority before another lower priority task, even if the lower priority task is still running. The lower priority task holds for some time and resumes when the higher priority task finishes its execution.
- Switch the process from running to ready state due to time quantum finished.

Non-Preemptive Scheduling

In this type of scheduling method, the CPU has been allocated to a specific process. The process that keeps the CPU busy will release the CPU either by I/O (or event) wait or terminating. Easy to implement as compared to Preemptive Scheduling.

- 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*

Types of scheduling algorithm

- First Come First Served (FCFS)
- Shortest-Job-First (SJF) Scheduling
- Shortest Remaining Time First (SRTF)
- Priority Scheduling
- Round Robin Scheduling
- Highest Response Ratio First (HRRF)
- Multilevel Queue Scheduling
- Multilevel Feedback Queue

First-Come, First-Served (FCFS) Scheduling

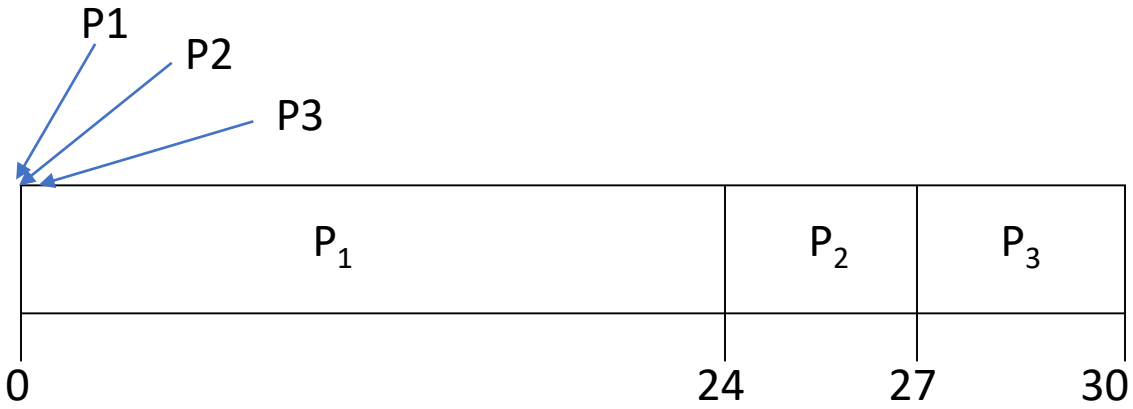
- Example:

Process	Burst Time (ns)
P_1	24
P_2	3
P_3	3

- Suppose that the processes arrive in the order: P_1 , P_2 , P_3

First-Come, First-Served (FCFS) Scheduling

- The Gantt Chart for the schedule is:



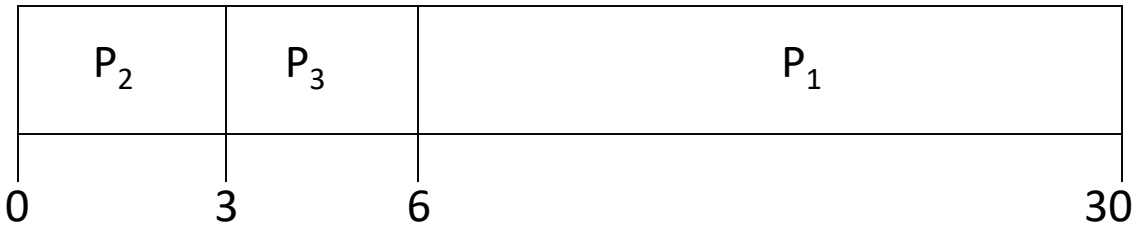
- Starting time of P1 = 0 ns; P2 = 24 ns; P3 = 27 ns.
- Finishing time of P1 = 24 ns; P2 = 27 ns; P3 = 30 ns.
- Waiting time for P_1 = 0 ns; P_2 = 24 ns; P_3 = 27 ns
- Average waiting time: (0 + 24 + 27)/3 = 17 ns
- TT: P1 = 24 ns; P2 = 27 ns; P3 = 30 ns.
- ATT: (24+27+30)/3 = 27 ns

FCFS Scheduling (Cont.)

Suppose that the processes arrive in the order

$$P_2 , P_3 , P_1 .$$

- The Gantt chart for the schedule is:



- Waiting time for $P_1 = 6$ ns ; $P_2 = 0$ ns , $P_3 = 3$ ns
- Average waiting time: (6 + 0 + 3)/3 = 3 ns
- TT for $P_1 = 30$ ns; $P_2 = 3$ ns, $P_3 = 6$ ns
- ATT: (30+3+6)/3=13 ns
- Much better than previous case.

Do following problem

- Example:

Process	Burst Time (ns)
P_1	2
P_2	9
P_3	3
P_4	20
P_5	5

- Suppose that the processes arrive in the order: P_1 , P_3 , P_2 P_4 , P_5
- *Average waiting time: 11ns*
- *ATT = 18.8*