CPU Scheduling

Gursharan Singh Tatla

professorgstatla@gmail.com

CPU Scheduling

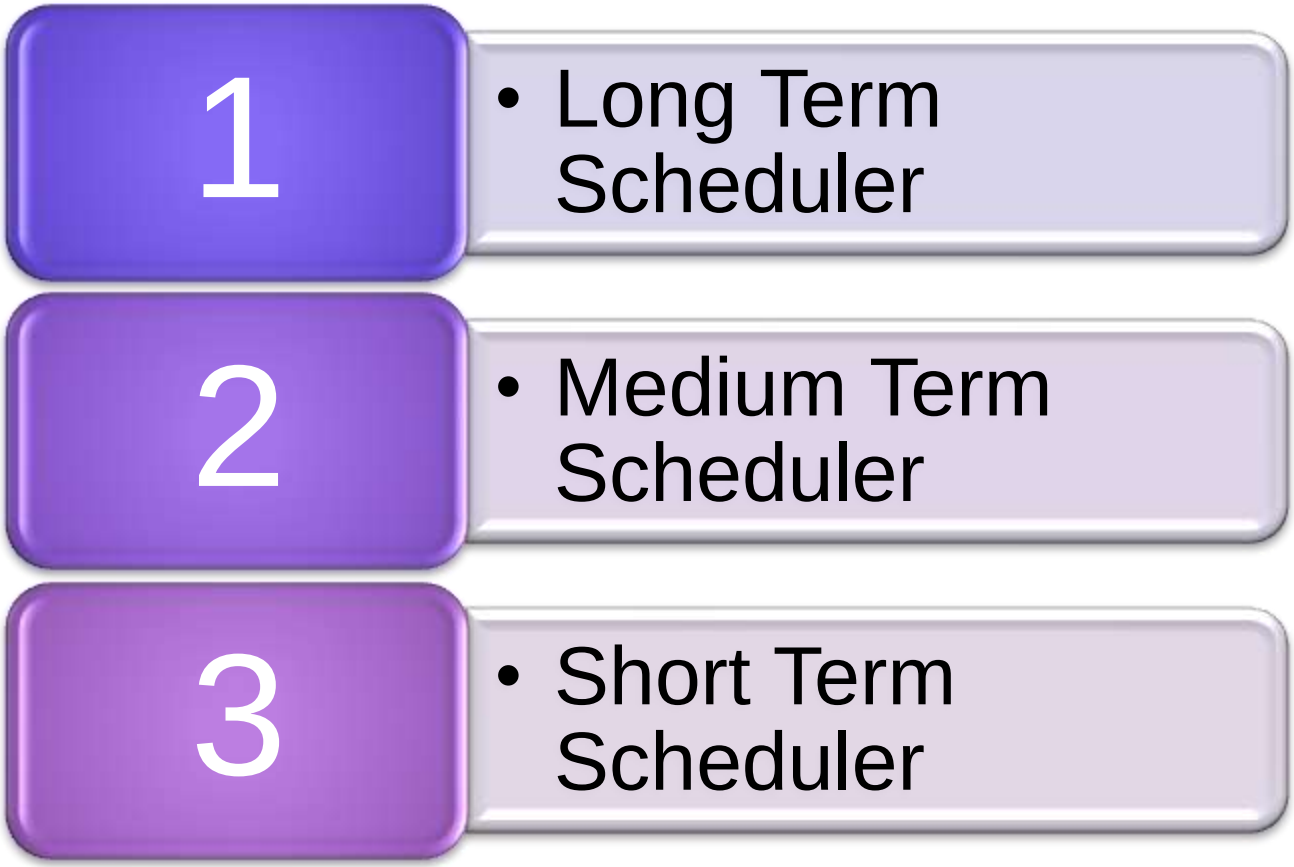
- **Scheduling** refers to selecting a process, from many ready processes, that is to be next executed on CPU.
- In multiprogramming environment, multiple processes are kept in main memory.
- When one process has to wait for I/O completion, operating system takes the CPU from that process and assigns it to another process.
- In this way, CPU is never idle and has some process to work on.

Scheduler

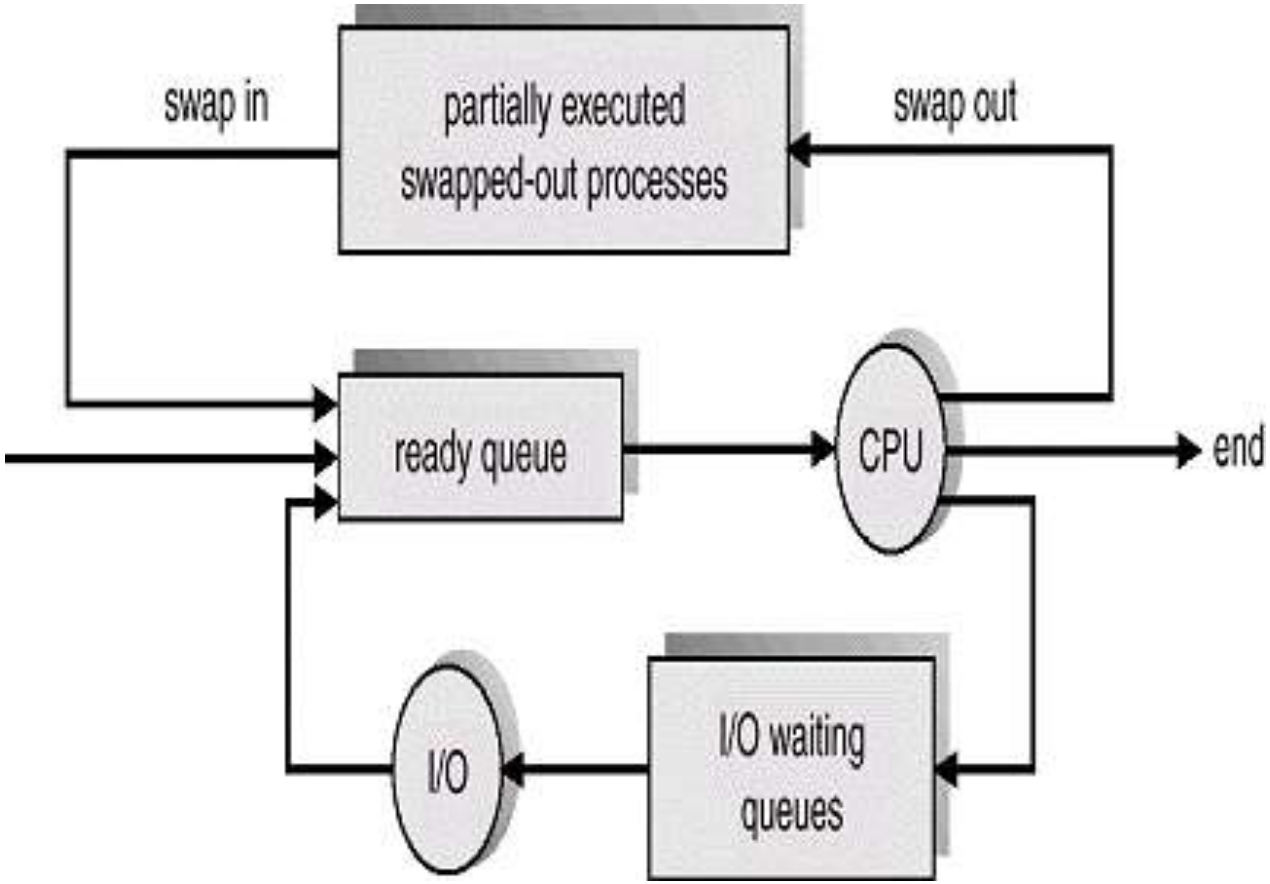
- **Scheduler** is an operating system module that selects the next job or process to be assigned to CPU.
- Thus, scheduler selects one of the many processes in memory that are ready to execute and allocates CPU to it.

Scheduler

- Scheduler is of three types:



Scheduler


swap in
partially executed
swapped-out processes
swap out
ready queue
CPU
end
I/O
I/O waiting
queues

Long Term Scheduler

- Long Term Scheduler selects the processes from secondary storage and loads them into memory for execution.
- It is called “long term” because the time for which the scheduling is valid is long.
- The frequency of execution of a long term scheduler is usually low, as there may be minutes between the creation of new processes in the system.

Long Term Scheduler

- The primary objective of long term scheduler is to control the "**degree of multiprogramming**".
- Degree of multiprogramming refers to the total number of processes present in the memory.
- If the degree of multiprogramming is stable, then the average rate of process creation is equal to the average terminate rate.

Long Term Scheduler

- This scheduler shows the best performance by selecting the good mixture of I/O bound and CPU bound processes.
- I/O bound processes are those that spend most of their time in I/O.
- CPU bound processes are those that spend most of their time in computations.

Medium Term Scheduler

- The medium term scheduler is required at the time when a swapped-out process is to be brought into pool of ready processes.
- A running process may be suspended because of I/O request.
- Such a suspended process is then removed from main memory and stored in secondary memory.

Medium Term Scheduler

- This is done because there is a limit on the number of active processes that can reside in main memory.
- Therefore, a suspended process is swapped-out from main memory.
- At some later time, the process can be swapped-in into the main memory.
- All versions of Windows use swapping.

Short Term Scheduler

- Short term scheduler selects one process from many ready processes that are residing in main memory and allocates CPU to one of them.
- Thus, it handles the scheduling of the processes that are in ready state.
- Short term scheduler is also known as **CPU Scheduler**.

Short Term Scheduler

- As compared to long term scheduler, a short term scheduler has to work very often.
- The frequency of execution of short term scheduler is high.
- It must select a new process for CPU frequently.

Preemptive & Non-Preemptive Scheduling

- A scheduling algorithm can be:

1. Preemptive Scheduling
2. Non-Preemptive Scheduling

Non-Preemptive Scheduling

- A scheduling is **non-preemptive** if, once a process has been given the CPU, the CPU cannot be taken away from the process.
- In other words, in non-preemptive scheduling, once the CPU has been allocated to a process, the process keeps the CPU until it releases the CPU either by terminating or by entering the waiting state.

Preemptive Scheduling

- A scheduling is preemptive if the CPU can be taken away from a process after being allocated.
- In other words, even if the CPU has been allocated to a certain process, it can be snatched from the process any time either due to time constraint or due to priority reason.

Dispatcher

- **Dispatcher** is a program responsible for assigning the CPU to the process, which has been selected by the short term scheduler.
- Dispatching a process involves context switching.

Scheduling Criteria

- The goal of a scheduling algorithm is to identify the process whose selection will result in the best possible system performance.
- The various scheduling criteria for evaluating an algorithm are discussed next.

Scheduling Criteria

- **CPU Utilization:**
  - CPU utilization is the average fraction of time during which the processor is busy.
  - The level of CPU utilization depends on the load on the system.
  - CPU utilization may range from 0 to 100%.

Scheduling Criteria

- **Throughput:**
  - It refers to the number of processes the system can execute in a period of time.
  - For long processes, this rate may be 1 process per hour.
  - For short processes, throughput may be 10 processes per second.
  - Thus, evaluation of throughput depends on the average length of a process.

Scheduling Criteria

- **Turnaround Time:**
  - This is the interval of time between the submission of a process and its completion.
  - Thus, turnaround time is an average period of time it takes a process to execute.
  - Turnaround time includes actual execution time plus time spent waiting for resources and doing I/O.

Scheduling Criteria

- **Waiting Time:**
  - It is the average period of time a process spends waiting.
  - Waiting time can be expressed as

$$W(x) = T(x) - x$$

  - where, $W(x)$ is the waiting time
  - $T(x)$ is the turnaround time
  - x is the actual execution time.

Scheduling Algorithm Optimization Criteria

- The optimization criteria is:
  - Max. CPU Utilization
  - Max. Throughput
  - Min. Turnaround Time
  - Min. Waiting Time
  - Min. Response Time

Thank You

Have a Nice Day