Operating System

Prof. Vibhuti Patel, Assistant Professor
Parul Institute of Computer Application

CHAPTER-2

Scheduling

What is Scheduling?

- Scheduling is a method used to distribute valuable computing resources, usually processor time, bandwidth and memory, to various processes, threads, data streams, and applications that require them.
- The decision is made to balance the load on the system and to ensure a uniform distribution of resources and to give some priority according to the prescribed rules.
- Scheduling is also known as process scheduling.

Process Scheduling

- **Process scheduling** is the activity of the process manager, which removes the ongoing process from the CPU and selects another process based on a specific strategy.

- **Process scheduling** is the function of the operating system to schedule various state processes such as scheduling, running, and waiting.

- **Process scheduling** is very important in **multiprogramming and multi-tasking operating system**, where multiple processes run simultaneously.

- Scheduling ensures maximum utilization of the central processing unit (CPU) because a process always runs at a certain time.

Process Scheduling Queues

- The OS maintains all PCBs in Process Scheduling Queues.
- When the state of a process changes, its PCB is removed from the current queue and transferred to its new state.

- The operating system performs the following complex process scheduling queues:

1. **Job Queue:** This queue keeps all the processes in the system.
2. **Ready Queue:** This queue keeps a set of all processes residing in main memory, ready and waiting to execute. A new process is always put in this queue.
3. **Device Queue:** The processes which are blocked due to unavailability of an I/O device constitute this queue.

Process Scheduling Queues

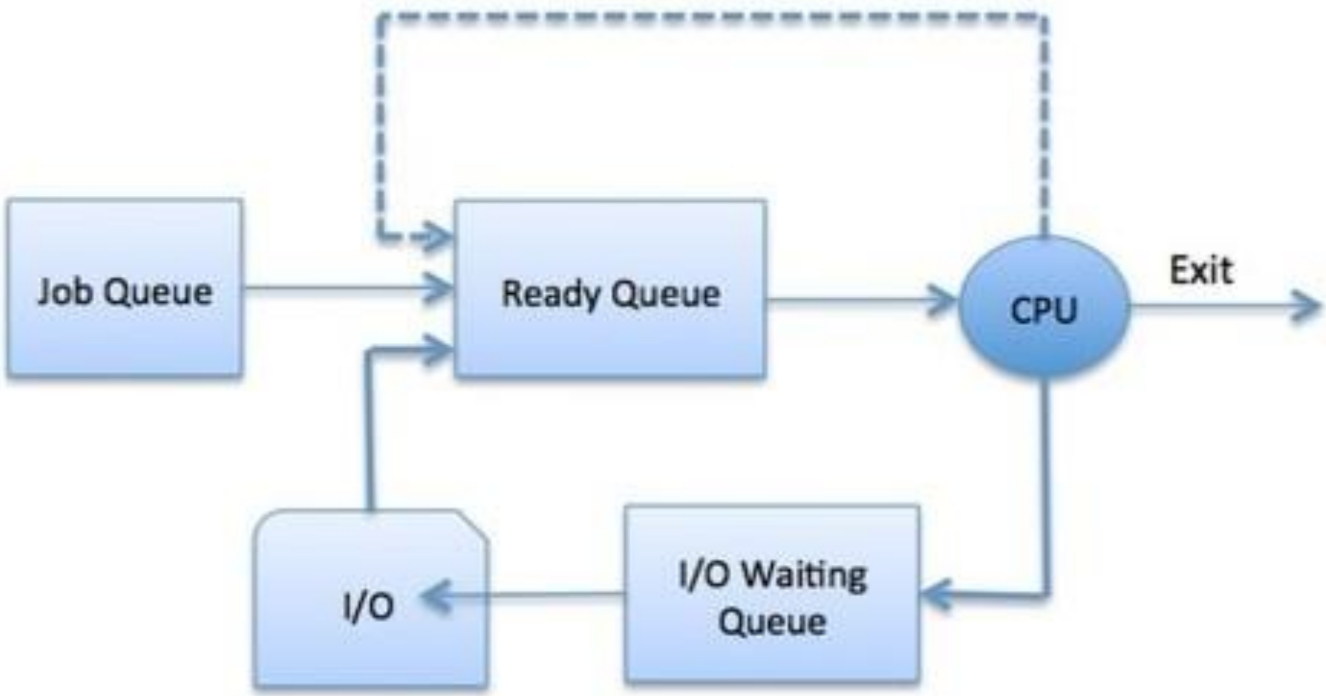

Image source : Book

Schedulers

Schedulers are specialized system software that manage process scheduling in a variety of ways. Their main task is to select the jobs to be performed in the system and to decide which process to run.

There are three types of schedulers:

1. Long-Term Scheduler
2. Medium-Term Scheduler
3. Short-Term Scheduler

Schedulers

1. Long-Term Scheduler:

- It is also known as job scheduler.
- A long-term scheduler determines which programs are allowed on system for processing.
- It selects the process to be placed in the ready queue.
- The long-term scheduler basically specifies the priority in which processes must be placed in main memory.
- Processes of the long-term scheduler are kept in the ready state because in this state the process is ready to execute waiting for calls of execution from the CPU which takes time and this is called as long term scheduler.

Schedulers

2. Mid-Term Scheduler:

- Medium-term scheduling is a part of swapping.
- It removes processes from memory.
- It reduces the level of multiprogramming.
- The medium-term scheduler is responsible for maintaining the swapped out-processes.
- A running process will be stopped upon I / O request. A suspended process cannot make any progress towards completion.

Schedulers

2. Mid-Term Scheduler:

- In this case, process can be temporarily swapped from main memory (or move) to secondary storage (disk) and make that memory available to other processes. These processes are called swapping.
- Task of moving from main memory to secondary memory is called swapping out.
- Task of moving back a swapped out process from secondary memory to main memory is called as swapping in.
- Process swapping is performed to ensure the optimal utilization of main memory.

Schedulers

3. Short-Term Scheduler:

- It is also called as CPU scheduler.
- It is a change of ready state to running state of process.
- The CPU scheduler selects one of the processes that is ready to execute and assigns CPU to one of them.
- It sets the priority in which processes is in ready queue are allocates the central processing unit (CPU) time for their execution.
- Short-term schedulers, also known as dispatchers, determine which process to pursue next. Short-term schedulers are faster than long-term schedulers.

Schedulers

3. Short-Term Scheduler:

•An operating system uses two types of scheduling processes:

Preemptive process:

- In preemptive scheduling policy
  - a low priority process has to be suspend its execution
  - if high priority process is waiting in the same queue for its execution.

Non - Preemptive process:

- In non - preemptive scheduling policy, processes are executed in first come first serve basis,
  - which means the next process is executed only when currently running process finishes its execution.

Context Switching

- Context Switching involves storing a context or state of a process so that it can be reloaded when needed and the execution can be resumed as before.
- It is a feature of a multi-tasking operating system and allows a single CPU to be shared by multiple processes.
- Allowing different processes to run simultaneously.
- Context switch is the mechanism for storing and restoring the state or context of a CPU in Process Control Block so that a process execution can be restarted from the same point at a later time.

Context Switching

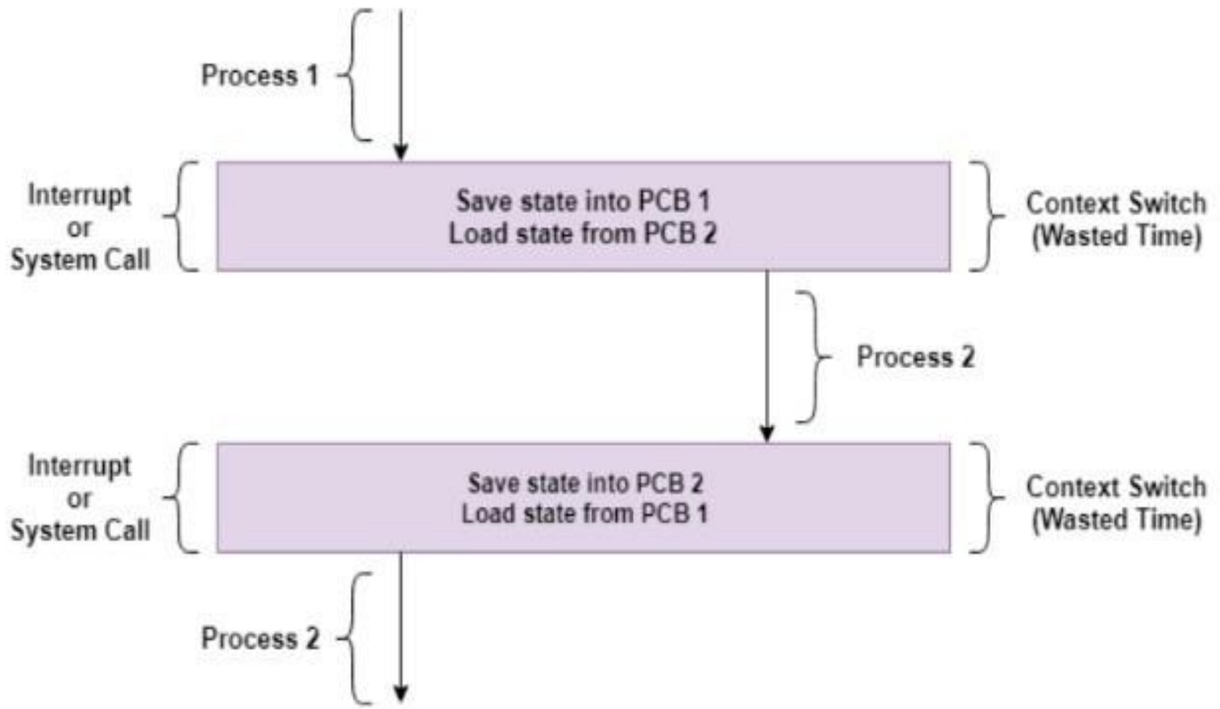

Image source : Google

Context Switching

- In the diagram above, initially Process 1 is running.
- Process 1 is switched out and Process 2 is switched in due to interruption or system call.
- Context switching involves saving the state of Process 1 into PCB1 and loading the state of process 2 from PCB2.
- After some time again a context switch occurs and Process 2 is switched out and Process 1 is switched in again.
- This involves saving the state of Process 2 into PCB2 and loads the state of process 1 from PCB1.

Scheduling Algorithms

- The process scheduler determines the various processes assigned to the CPU based on a specific scheduling algorithm.

- Algorithms are either non-preemptive or preemptive.

- **Non-preemptive** algorithm is designed so that once a process enters a running state, it cannot be preempted until it has completed its allotted time, whereas the **preemptive** scheduling is based on priority where a scheduler may preempt a low priority running process anytime when a high priority process reach into a ready state.

First Come First Serve (FCFS)

- Jobs are executed on first come, first serve basis.
- Easy to understand & implement.
- Its implementation depends on FIFO queue.
- Poor in performance as average wait time is high.
- Burst time = finish time

First Come First Serve (FCFS)

Consider the set of processes that arrived at time 0.

Process	Burst Time
P1	24
P2	3
P3	3

If the CPU scheduling policy is FCFS, calculate the average waiting time.

First Come First Serve (FCFS)

Gantt Chart:

P1	P2	P3

0 24 27 30

Process	Burst Time	Waiting Time
P1	24	0
P2	27	24
P3	30	27

First Come First Serve (FCFS)

- Total Waiting Time = 51ms
- Average Waiting Time =51/3ms =17ms

First Come First Serve (FCFS)

Consider the set of 5 processes whose arrival time and burst time are given below-

Process	Arrival Time	Burst Time
P1	0	4
P2	1	3
P3	2	1
P4	3	2
P5	4	5

If the CPU scheduling policy is FCFS, calculate the average waiting time and average turn around time.

First Come First Serve (FCFS)

Gantt Chart:

P1	P2	P3	P4	P5

0 4 7 8 10 15

Process	Arrival Time	Burst Time	Waiting Time	Turnaround Time
P1	0	4	0	4
P2	1	3	3	6
P3	2	1	5	6
P4	3	2	5	7
P5	4	5	6	11

First Come First Serve (FCFS)

- Total Waiting Time = 19ms
- Average Waiting Time = 19/5ms =3.8ms
- Total Turnaround Time = 34ms
- Average Turnaround Time = 34/5ms =6.8ms

Shortest Job First (SJF)

- This is also called as shortest job first or SJF
- The best approach to minimize waiting time.
- Easy to implement in Batch systems where required CPU time is known in advance.
- Impossible to run in interactive systems that do not required CPU time.
- The processer need to know in advance how long process will take.

Shortest Job First (SJF)

Consider the set of 5 processes whose arrival time and burst time are given below-

Process	Arrival Time	Burst Time
P1	0	4
P2	1	3
P3	2	1
P4	3	2
P5	4	6

If the CPU scheduling policy is SJF(Non Preemptive), calculate the average waiting time and average turn around time.

Shortest Job First (SJF)

Gantt Chart:

P1	P3	P4	P2	P5

0 4 5 7 10 16

Process	Arrival Time	Burst Time	Waiting Time	Turnaround Time
P1	0	4	0	4
P2	1	3	6	9
P3	2	1	2	3
P4	3	2	2	4
P5	4	6	6	12

Parul® University

DIGITAL LEARNING CONTENT

Shortest Job First (SJF)

- Total Waiting Time = 16ms
- Average Waiting Time = 16/5ms =3.2ms
- Total Turnaround Time = 32ms
- Average Turnaround Time = 32/5ms =6.4ms

Priority Scheduling

- Priority is assigned for each process.
- The highest priority process is implemented first.
- Processes with same priority are executed in FCFS manner.
- Priority can be determined based on memory needs; time needs or any other resource need.

Priority Scheduling

Consider the set of 6 processes whose arrival time and burst time are given below-

Process	Arrival Time	Burst Time	Priority
P1	0	4	4
P2	1	5	5
P3	2	1	7
P4	3	2	2
P5	4	3	1
P6	5	6	6

If the CPU scheduling policy is priority (non-preemptive), calculate the average waiting time and average turn around time.

Priority Scheduling

Gantt Chart:

P1	P3	P6	P2	P4	P5

0 4 5 11 16 18 21

Process	Arrival Time	Burst Time	Turnaround Time	Waiting Time
P1	0	4	4	0
P2	1	5	15	10
P3	2	1	3	2
P4	3	2	15	13
P5	4	3	17	14
P6	5	6	6	0

Priority Scheduling

- Total Waiting Time = 39ms
- Average Waiting Time = 39/6ms =6.5ms
- Total Turnaround Time = 60ms
- Average Turnaround Time = 60/6ms =10ms

Round Robin (RR)

- It is also called as time slicing because each process is given a slice in advance.
- The Clock interrupt is generated in periodic intervals
- When an interruption occurs, the currently running process is placed in ready queue.
- The next ready job is selected.

Round Robin (RR)

Consider the set of 4 processes whose arrival time and burst time are given below-

Process	Arrival Time	Burst Time
P1	0	8
P2	1	4
P3	2	9
P4	3	5

If the CPU scheduling policy is Round Robin with time quantum = 4 unit, calculate the average waiting time and average turn around time.

Round Robin (RR)

Gantt Chart:

P1	P2	P3	P4	P1	P3	P4	P3

0 4 8 12 16 20 24 25 26

Process	Arrival Time	Burst Time	Turnaround Time	Waiting Time
P1	0	8	20	12
P2	1	4	7	3
P3	2	9	24	15
P4	3	5	22	17

Round Robin (RR)

- Total Waiting Time = 47ms
- Average Waiting Time = 47/4ms =11.75ms
- Total Turnaround Time = 73ms
- Average Turnaround Time = 73/4ms =18.25ms

Multiprocessor Scheduling

In a multiprocessor, the scheduler must decide which process to run and which central processing unit to run.

1.Timesharing: On multiprocessor, the simplest scheduling algorithm for dealing with unrelated processes is to have a single system-wide data structure for ready processes possibly just a list, but more likely a set of lists for the processes at different priorities.

2.Space Sharing: Multiprocessor scheduling can be used when processes are related to one another. Scheduling two or more than two threads at the same time across multiple central processing units is called as space sharing. The big advantage of space sharing is the elimination of multiprogramming which eliminates the context switching overhead.

Multiprocessor Scheduling

3. Gang Scheduling: Gang scheduling has the following three parts:

- Groups of related threads are scheduled as a unit, a gang.
- All members of a gang run simultaneously, on different timeshared central processing unit.
- All the gang members start and finish their time slices together.

Real Time Scheduling

- Real-time system means that the system is subject to real-time, that is, response must be guaranteed within a specified time frame or system should meet the specified time limit. For example: flight control system, real time monitors etc.

Parul® University

DIGITAL LEARNING CONTENT

Types of Real Time Scheduling

1. Hard Real Time System: This type of system can never miss its deadline. Missing the deadline may have disastrous consequences. Tardiness means how late a real time system completes its task with respect to its deadline. Example: Flight controller system.

2. Soft Real Time System: This type of system can miss its deadline occasionally with some acceptably low probability. The usefulness of result produced by a soft real time system decreases gradually with increase in tardiness. Example: Telephone switches.

Reference Model of Real Time Scheduling

Real Time Reference Model: Our reference model is classified by three components:

- A workload model: Specifies the application supported by the system.
- A resource model: Specifies the resources available for the application.
- Algorithms: Specifies how the application system uses resources.

Real Time Scheduling Terms

Terms related to real time system:

Job: The job is a small piece of work that can be allocated to a processor and may or may not require resources.

Task: A set of related jobs that jointly provide certain system functionality.

Release time of a job: This is the time at which job becomes ready for execution.

Execution time of a job: This is the time taken to execute job.

Real Time Scheduling Terms

- Deadline of a job: It is time to complete job. There are two types of deadlines: absolute deadline and relative deadline.

- Response time of a job: This is the length of time from release time of a job to the instant when it finishes.

DIGITAL LEARNING CONTENT

Parul® University

www.paruluniversity.ac.in