

Processes

A process is an instance of a program running in a computer.

It is close in meaning to task, a term used in some operating systems.

In UNIX and some other operating systems, a process is started when a program is initiated (either by a user entering a shell command or by another program).

Like a task, a process is a running program with which a particular set of data is associated so that the process can be kept track of.

An application that is being shared by multiple users will generally have one process at some stage of execution for each user.

OS can support multiple processes at the same time based on the scheduling policy and its ability to perform concurrency.

Concurrency is the appearance of simultaneous tasks. It means that many tasks/processes run in an operating system however they are not all using the CPU at the same time. The OS will schedule them in such a way that user feels that all of them are running parallelly.

The scheduling algorithm can be vastly pre-emptive & non-pre-emptive. Linux Kernel 2.6 and above are all pre-emptive kernels. Pre-emptive kernel is where a running process be picked & put into wait/ready state so that another process (based on priority or time-slice) can execute.

Time Slicing:

Time slicing is a concept where a process are given certain amount of time to execute and after the time is over, the execution shifts to the next process.

Some of the basic scheduling policies are:

First Come First Serve (FCFS):

Simplest scheduling algorithm that schedules according to arrival times of processes. First come first serve scheduling algorithm states that the process that requests the CPU first is allocated the CPU first.

Shortest Job First (SJF):

Process which have the shortest burst time are scheduled first. If two processes have the same burst time then FCFS is used to break the tie. It is a non-pre-emptive scheduling algorithm.

Longest Job First (LJF):

It is similar to SJF scheduling algorithm. But, in this scheduling algorithm, we give priority to the process having the longest burst time. This is non-pre-emptive in nature i.e., when any process starts executing, can't be interrupted before complete execution.

Shortest Remaining Time First (SRTF):

It is pre-emptive mode of SJF algorithm in which jobs are schedule according to shortest remaining time.

Longest Remaining Time First (LRTF):

It can be pre-emptive or non-preemptive.

If it is preemptive then it can be considered as preemptive mode of LJF algorithm in which we give priority to the process having largest burst time remaining.

If it is non pre-emptive then it is very similar to LJF.

Round Robin Scheduling:

Each process is assigned a fixed time(Time Quantum/Time Slice) in cyclic way. It is designed especially for the time-sharing system. So if a system has round robin with time slicing design for scheduling of processes. It is primarily based on the polling mechanism.

If each time slice is for 20ns, then the total time can be decided based on the :

1. number of processes running
2. order of round robin

Sometimes is a process is critical then it can also occur twice in the robin.

Priority Based scheduling (Non-Pre-emptive):

In this scheduling, processes are scheduled according to their priorities, i.e., highest priority process is scheduled first. If priorities of two processes match, then schedule according to arrival time.

interrupts:

Interrupts are signals sent to the CPU by external devices, normally I/O devices. They tell the CPU to stop its current activities and execute the appropriate part of the operating system.

Interrupts generally have higher priority than normal processes.

Among themselves they can be of same priority or can have different priorities.

Normally when an interrupt occurs the control switches to the ISR (Interrupt Service Routine).

Normally all other interrupts may be ,asked when an ISR runs.

In some cases the other interrupts are sent into a queue when an ISR is running. The other interrupts are executed in queue after the current interrupt finishes execution.

Interrupt latency is the time taken to start the ISR after the interrupt occurs.