RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF ENGINEERING & TECHNOLOGY

BCS-501 Operating System

Lecturer-08

Manisha Verma
Assistant Professor
Computer Science & Engineering

Scheduling Algorithms

- Scheduling Criteria
- Optimization Criteria
- FCFS
- Shortest-Job-First Scheduling
- Priority Scheduling
- Round Robin (RR)

Scheduling Algorithms

Selects from among the processes in memory that are ready to execute, and allocates the CPU to one of them.

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.*

Dispatcher module gives control of the CPU to the process selected by the short-term scheduler; this involves:

- switching context
- switching to user mode
- jumping to the proper location in the user program to restart that program

Dispatch latency – time it takes for the dispatcher to stop one process and start another running.

Scheduling Criteria

- CPU utilization – keep the CPU as busy as possible
- Throughput – # of processes that complete their execution per time unit
- Turnaround time – amount of time to execute a particular process
- Waiting time – amount of time a process has been waiting in the ready queue
- Response time – amount of time it takes from when a request was submitted until the first response is produced, **not** output (for time-sharing environment)

Optimization Criteria

- Max CPU utilization
- Max throughput
- Min turnaround time
- Min waiting time
- Min response time

First-Come, First-Served (FCFS) Scheduling

Example: urst Time	Process	B
	P_1	24
	P_2	3
	P_3	3

Suppose that the processes arrive in the order: P_1 , P_2 , P_3
The Gantt Chart for the schedule is:

Waiting time for P_1 = 0; P_2 = 24; P_3 = 27
Average waiting time: (0 + 24 + 27)/3 = 17

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; P_2 = 0, P_3 = 3
Average waiting time: (6 + 0 + 3)/3 = 3
Much better than previous case.
Convoy effect short process behind long process

P_2	P_3	P_1

Shortest-Job-First (SJR) Scheduling

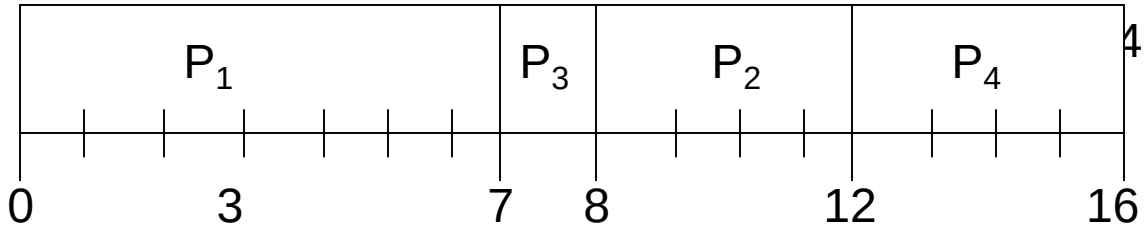
Associate with each process the length of its next CPU burst. Use these lengths to schedule the process with the shortest time.

Two schemes:
nonpreemptive – once CPU given to the process it cannot be preempted until completes its CPU burst.

Preemptive – if a new process arrives with CPU burst length less than remaining time of current executing process.
 preempt. This scheme is know as the Shortest-Remaining-Time-First (SRTF).
SJF is optimal – gives minimum average waiting time for a given set of processes.

Process	Arrival Time	Burst Time
P_1	0.0	7
P_2	2.0	4
P_3	4.0	1
P_4	5.0	4

SJF (non-preemptive)

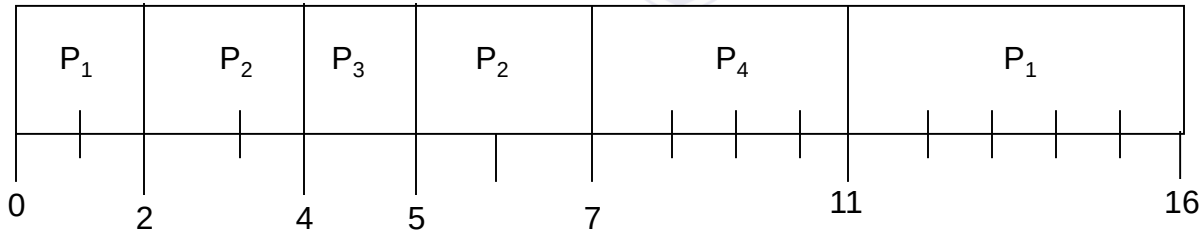


Preemptive SJF

Process	Arrival Time	Burst Time
P_1	0.0	7
P_2	2.0	4
P_3	4.0	1
P_4	5.0	4

SJF (preemptive)

Average waiting time = (9 + 1 + 0 +2)/4 - 3



Priority Scheduling

•A priority number (integer) is associated with each process

The CPU is allocated to the process with the highest priority (smallest integer ≡ highest priority).

Preemptive

Nonpreemptive

•SJF is a priority scheduling where priority is the predicted next CPU burst time.

Problem ≡ Starvation – low priority processes may never execute.

Solution ≡ Aging – as time progresses increase the priority of the process.

Round Robin (RR)

Each process gets a small unit of CPU time (*time quantum*), usually 10-100 milliseconds. After this time has elapsed, the process is preempted and added to the end of the ready queue.

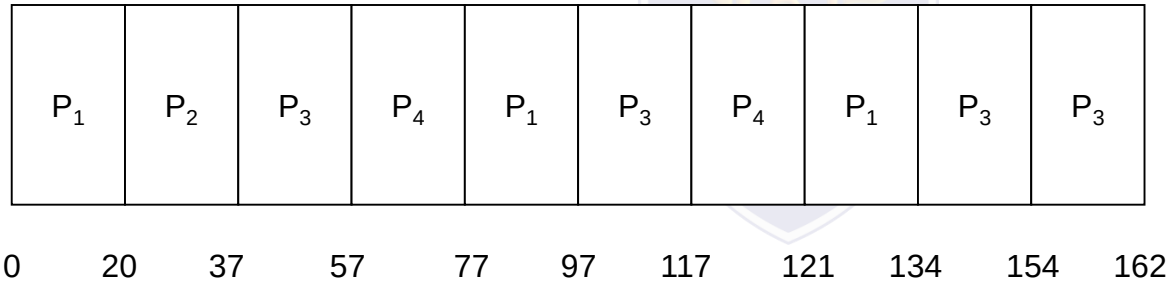
If there are n processes in the ready queue and the time quantum is q, then each process gets $1/n$ of the CPU time in chunks of at most q time units at once. No process waits more than $(n-1)q$ time units.

Performance

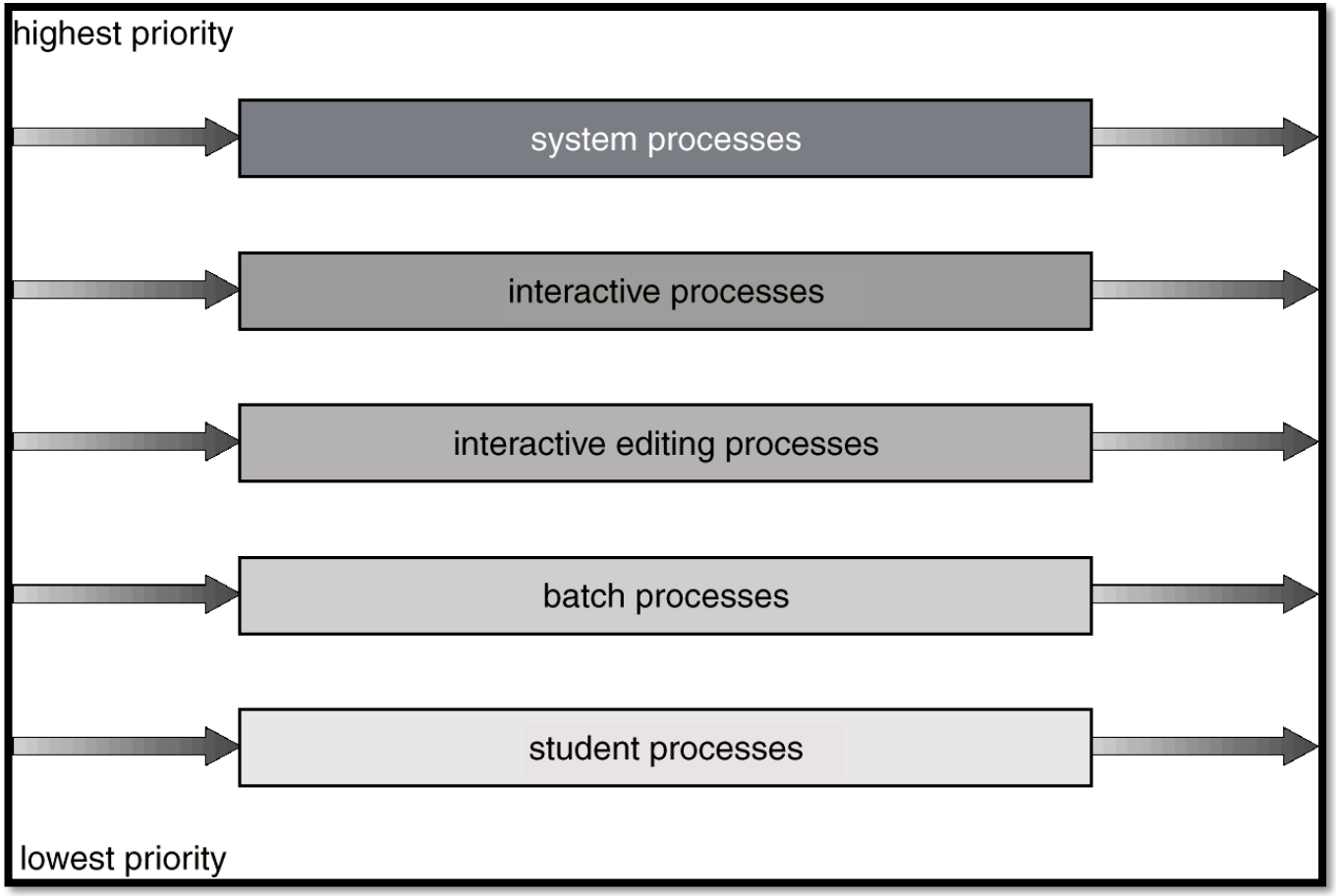
- q large $\Rightarrow$ FIFO
- q small $\Rightarrow$ q must be large with respect to context switch, otherwise overhead is too high.

Process	Burst Time
P_1	53
P_2	17
P_3	68
P_4	24

The Gantt chart is:



Multilevel Queue Scheduling



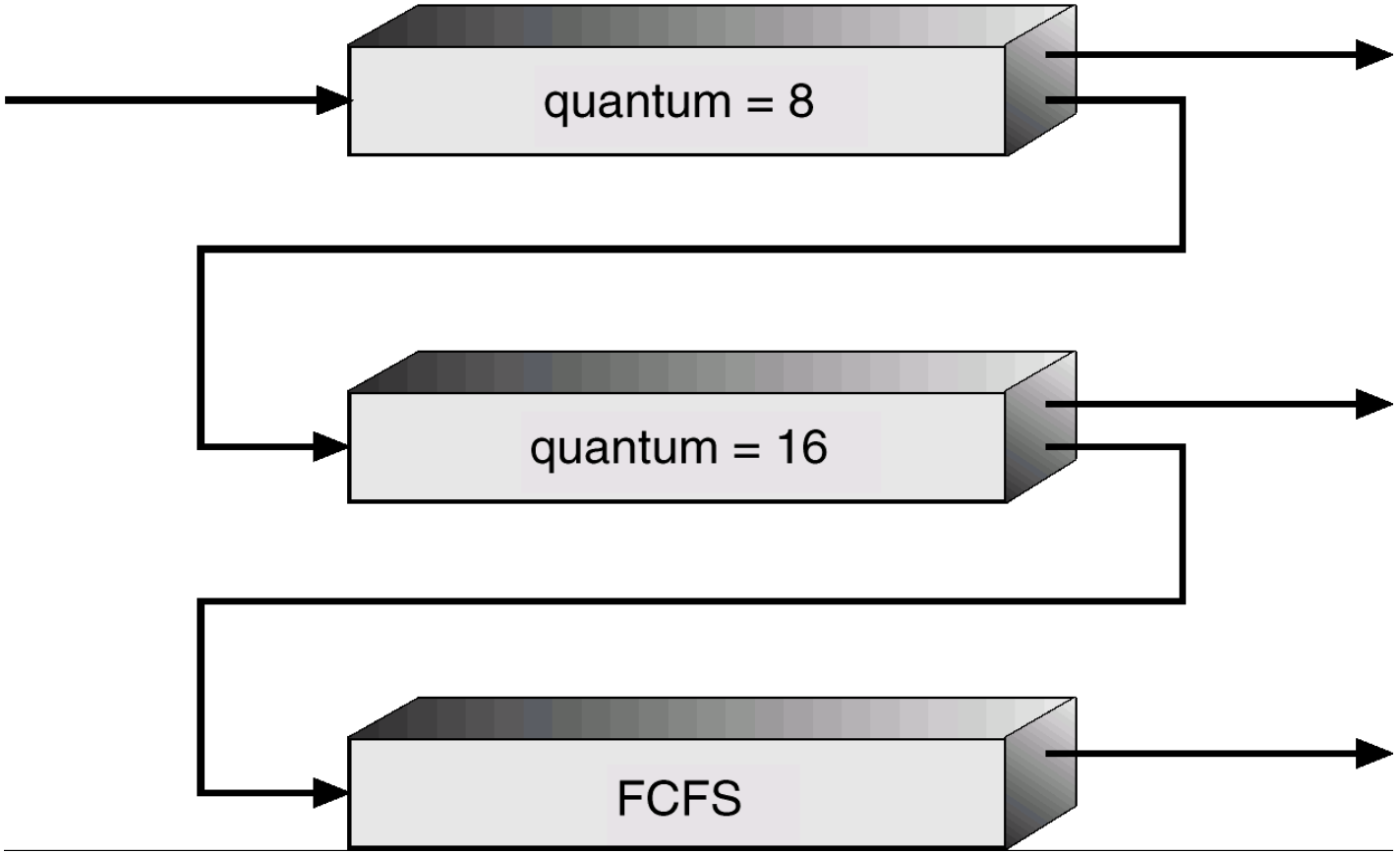
Multilevel Feedback Queue

A process can move between the various queues; aging can be implemented this way.

Multilevel-feedback-queue scheduler defined by the following parameters:

- number of queues
- scheduling algorithms for each queue
- method used to determine when to upgrade a process
- method used to determine when to demote a process
- method used to determine which queue a process will enter when that process needs service.

Multilevel Feedback Queues



MCQ

Dispatch latency is a……..
A. dispatcher to stop one process and start another running
B. dispatcher to stop two process and start first running process
C. dispatcher to stop one process running
D. None

CPU scheduling decisions may take place when a process:
A. Switches from running to waiting state.
B. Switches from running to ready state.
C. Switches from waiting to ready.
D. All of these

Dispatcher module gives control of the CPU to the process selected by the……..

A. short-term scheduler
B. long-term scheduler
C. Both
D. none

CPU utilization contain……

A. Throughput
B. Turnaround time
C. Waiting time
D. All

Preemptive means –

A. process arrives with CPU burst length less than remaining time of executing process.
B. process arrives with CPU burst length greater than remaining time of executing process.
C. Both
D. None