Module 19: Multi-core computing Operating Systems
Lecture 38: Priority and Schedule

The Lecture Contains:

Previous Next

Module 19: Multi-core computing Operating Systems

Lecture 38: Priority and Schedule

Issues With Priority

- Priority definition.
  - How to define a priority?
    - System policies: Charge, job quantum, I/O
    - External or internal definition.
- Pre-emptive vs. non-preemptive
  - When a process arrives, should we remove the running process? (e.g. SRTF + Priority)
- Starvation
  - When high priority jobs come at a frequency that a low priority job is blocked.
  - Solution: add “aging”.

Aging with Priority

- Each time a process is scheduled,
  - Increase the priority of each pending task by 1. (could even be periodic)
- Priority: defined at admission time only.
- In pre-emptive scheduling.
  - When a job is pre-empted, reset the priority.
- Guaranteed “No starvation”.

Previous Next

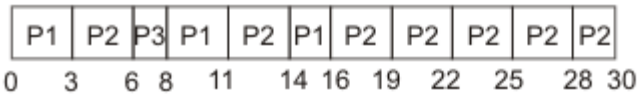
Adding Time Quantum: Round Robin Scheduling

- FCFS + Preemption
  - Each time a process is scheduled, a time quantum is given to this.
- When time quota expires, process is moved to ready queue.
  - Job is entered at the end of the queue and scheduled from the beginning of the queue.
    - Round robin scheduling

RR Scheduling

Process	CPU Burst
P1	8
P2	20
P3	2

Time Quantum: 3 Units



Smaller Quantum: Large number of context switching. High response.

Larger Quantum: Slow response but fewer context switches

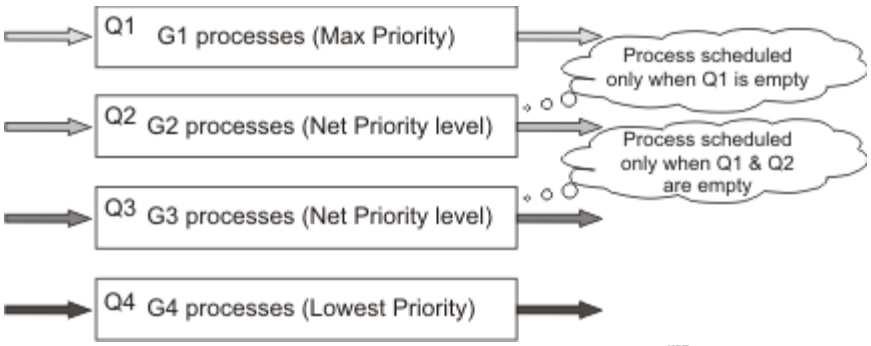
Previous Next

Module 19: Multi-core computing Operating Systems

Lecture 38: Priority and Schedule

Multilevel Queue Scheduling

- Processes may be grouped.
  - System processes, interactive processes, batch processes etc.
- Ready queue may be made one for each group.
- Within a queue the scheduling can be one of the earlier defined scheduling (typically RR is used)
- The queues have associated priorities



Previous Next

Issues With Multilevel Queue

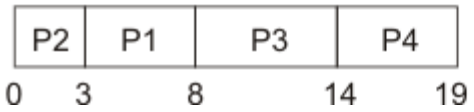
- Processes show a varying character
  - Interactive or non-interactive
  - Response requirement: high to low
- Processes in low level of priority may starve
- Solution: Add movement of processes from one queue to another
  - Multilevel Feedback Queue Scheduling
  - Movement can be due to
    - Change in character of the process
    - Aging

Scheduling in Real time OS

- Real time jobs have an additional criterion
  - Deadline, or time guarantee to execute a process
- Earliest Deadline First (EDF) Scheduling
  - Schedule a task that has earliest deadline

EDF Examples

Process	Arrival Time	CPU Burst	Deadline
P1	0	5	15
P2	0	3	6
P3	2	6	12
P4	4	5	15



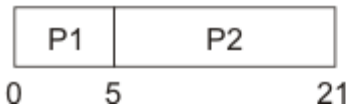
EDF Examples

Process	Arrival Time	CPU Burst	Deadline
P1	0	5	15
P2	0	16	6

EDF may miss a deadline even when a scheduling is possible without missing the deadline

Possible Feasible Schedule

Process	Arrival Time	CPU Burst	Deadline
P1	0	5	15
P2	0	16	6



No process misses deadline

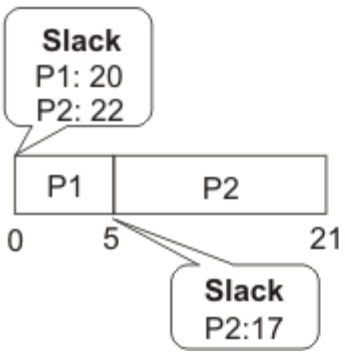
Previous Next

Least Slack First Scheduling

- Slack: Burst + Deadline –Current time
- Least slack first is optimal scheduling for meeting the deadline.
- In real time systems, the task are well behaved
  - Easier to guess the burst accurately.
  - Or, the tasks announce their burst time.

LSF Schedule

Process	Arrival Time	CPU Burst	Deadline
P1	0	5	15
P2	0	16	6



RT Scheduling

- If task declare their CPU burst
  - A task that finishes within the time is a good task.
  - A task that leaves a lot of slack is an issue and can influence the admission policy.
    - Denial of service??
- Solution: “Charge” tasks based on their declared burst.

Previous Next

Characteristics of Real-Time (RT) Systems

- Determinism
- Responsiveness
- User control
- Reliability

Deterministic Response

- External event and timings dictate the request of service.
- OS's response depends on
  - speed at which it can respond to interrupts
  - whether the system has sufficient capacity to handle requests.
- Can we put an upper bound on time for OS response?
  - Factor of the OS design.
- In non-RT this delay averages around 50 to 500 ms,
  - Also is usually non-deterministic.
- In an RT, delay is usually guaranteed to have an upper-bound (usually small: few µsto 1-2ms.

Responsiveness

- The time for servicing the interrupt once it has been acknowledged.
- Comprises:
  - Time to transfer control, (and context switch) and execute the ISR
- Depends upon
  - Interrupt latency of the hardware (usually very small)
  - Priority of interrupts.
    - Classic Problem: Priority inversion.
  - Priority of tasks.
- response time = *f* (responsiveness, determinism)

Previous Next