Scheduling

Readings: Sections 9.1 and 9.2

Processor Scheduling

- Aim is to assign processes to be executed by the processor in a way that meets system objectives, such as response time, throughput, and processor efficiency
- Broken down into three separate functions:
Long-Term Scheduler

- Determines which programs are admitted to the system for processing
- Controls the degree of multiprogramming

© 2017 Pearson Education, Inc., Hoboken, NJ. All rights reserved.
Medium-Term Scheduling

- Part of the swapping function

- Swapping-in decisions are based on the need to manage the degree of multiprogramming

  - Considers the memory requirements of the swapped-out processes

Short-Term Scheduling

- Known as the dispatcher

- Executes most frequently

- Makes the fine-grained decision of which process to execute next

- Invoked when an event occurs that may lead to the blocking of the current process or that may provide an opportunity to preempt a currently running process in favor of another

Examples:
- Clock interrupts
- I/O interrupts
- Operating system calls
- Signals (e.g., semaphores)
Evaluation metrics

User-oriented
- Measurable: Turnaround time, Response time, Deadlines
- Non-measurable: Predictability

System-oriented
- Measurable: Throughput, Utilization
- Non-measurable: Fairness, Ensuring priorities, Load balancing

Table 9.3
Characteristics of Various Scheduling Policies

<table>
<thead>
<tr>
<th>FCFS</th>
<th>Round Robin</th>
<th>SPN</th>
<th>SRT</th>
<th>HRRN</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection function</td>
<td>max[w]</td>
<td>constant</td>
<td>min[s]</td>
<td>min[s – c]</td>
<td>max((\frac{w + s}{p})) (see text)</td>
</tr>
<tr>
<td>Decision mode</td>
<td>Non-preemptive</td>
<td>Preemptive (at time quantum)</td>
<td>Non-preemptive</td>
<td>Preemptive (at arrival)</td>
<td>Non-preemptive</td>
</tr>
<tr>
<td>Through-Put</td>
<td>Not emphasized</td>
<td>May be low if quantum is too small</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Response time</td>
<td>May be high, especially if there is a large variance in process execution times</td>
<td>Provides good response time for short processes</td>
<td>Provides good response time for short processes</td>
<td>Provides good response time</td>
<td>Provides good response time</td>
</tr>
<tr>
<td>Overhead</td>
<td>Minimum</td>
<td>Minimum</td>
<td>Can be high</td>
<td>Can be high</td>
<td>Can be high</td>
</tr>
<tr>
<td>Effect on processes</td>
<td>Penalizes short processes, penalizes I/O bound processes</td>
<td>Fair treatment</td>
<td>Penalizes long processes</td>
<td>Penalizes long processes</td>
<td>Good balance</td>
</tr>
<tr>
<td>Starvation</td>
<td>No</td>
<td>No</td>
<td>Possible</td>
<td>Possible</td>
<td>No</td>
</tr>
</tbody>
</table>

(Table can be found on page 405 in textbook)

© 2017 Pearson Education, Inc., Hoboken, NJ. All rights reserved.
First-Come-First-Served (FCFS)

- Simplest scheduling policy
- Also known as first-in-first-out (FIFO) or a strict queuing scheme
- As each process becomes ready, it joins the ready queue
- When the currently running process ceases to execute, the process that has been in the ready queue the longest is selected for running
- Performs much better for long processes than short ones
- Tends to favor processor-bound processes over I/O-bound processes

Round Robin

- Uses preemption based on a clock
- Also known as time slicing because each process is given a slice of time before being preempted
- Principal design issue is the length of the time quantum, or slice, to be used
- Particularly effective in a general-purpose time-sharing system or transaction processing system
- One drawback is its relative treatment of processor-bound and I/O-bound processes
Shortest Process Next (SPN)

- Nonpreemptive policy in which the process with the shortest expected processing time is selected next
- A short process will jump to the head of the queue
- Possibility of starvation for longer processes
- One difficulty is the need to know, or at least estimate, the required processing time of each process
- If the programmer’s estimate is substantially under the actual running time, the system may abort the job

Shortest Remaining Time (SRT)

- Preemptive version of SPN
- Scheduler always chooses the process that has the shortest expected remaining processing time
- Risk of starvation of longer processes
- Should give superior turnaround time performance to SPN because a short job is given immediate preference to a running longer job
Highest Response Ratio Next (HRRN)

- Chooses next process with the greatest ratio
- Attractive because it accounts for the age of the process
- While shorter jobs are favored, aging without service increases the ratio so that a longer process will eventually get past competing shorter jobs

\[
\text{Ratio} = \frac{\text{time spent waiting} + \text{expected service time}}{\text{expected service time}}
\]

© 2017 Pearson Education, Inc., Hoboken, NJ. All rights reserved.