6 Scheduling

Motivation and definition of terms

Criterià

Process user perspective

Performance criteria

- minimizes the... true execution time.
- minimize (max) execution time.
- minimizes the... allowed.
- minimizes the... allows.
- minimizes the... allows.

Predictability criteria

- maximizes predictability.
- minimizes the... allowed.
- minimizes the... allows.
- minimizes the... allows.
- minimizes the... allows.

Time scales of scheduling

- Preemptive scheduling
- Non-preemptive scheduling

First come, first served (FCFS)

- Optimize for swift initial responses.
- Prefers short tasks but all tasks will be handled.
- Suffer starvation.
- Very short tasks can suffer starvation.
- Low overhead.
- Low packet latency.

Round Robin (RR)

- Optimizes for small task responses.
- High overhead.
- High packet latency.
- Low performance.

Shortest job first

- Optimize for high priority (short task) requests.
- High overhead.
- High packet latency.
- High performance.

Feedback with 2^i pre-emption

- Use new high priority task arrivals.
- Optimize for the initial response time.
- Prefers short tasks and longer tasks can suffer starvation.
- Low overhead.
- High packet latency.

Performance scheduling

- Optimize for the initial response time.
- Prefers short tasks and longer tasks can suffer starvation.
- Low overhead.
- High packet latency.

Feedback with 2^i pre-emption intervals - sequential

- Optimize for the initial response time.
- Prefers short tasks and longer tasks can suffer starvation.
- Low overhead.
- High packet latency.

Feedback with 2^i pre-emption intervals - overlapping

- Optimize for the initial response time.
- Prefers short tasks and longer tasks can suffer starvation.
- Low overhead.
- High packet latency.

References for this chapter

- Stallings2001 Prentice Hall, 2001

Purpose of scheduling

Two scenarios for scheduling algorithms

1. Ordering resource assignments (CPU time, network access, ...)
   - Linear, online application of scheduling algorithm.
2. Predicting system behaviour under anticipated loads
   - Simulated, off-line application of scheduling algorithms.

Facts and figures:

- A complete time to control the ability of the system on required resource needs...
-Arrange time to permit adherence with requirements for load balancing...

Performance scheduling

- First come, first served (FCFS)
- Shortest possible maximal turnaround time!
- Optimize for good average performance with minimal task-switches.