Operating Systems - Chapter 7 Summary (Detailed)

Scheduling: Introduction

• Scheduling policies decide which job runs on the CPU at a given time.
• Origins in operations management (assembly lines, efficiency).
• Challenge: balance performance and fairness.

Workload Assumptions

• 1. Each job runs for the same amount of time.
• 2. All jobs arrive at the same time.
• 3. Once started, each job runs to completion.
• 4. Jobs only use the CPU (no I/O).
• 5. Runtime of each job is known.
• These assumptions are unrealistic, but useful for building intuition.

Scheduling Metrics

• Metrics let us evaluate how good a scheduling policy is.
• Two main metrics: turnaround time and response time.
Turnaround Time:
– Definition: Tturnaround = Tcompletion – Tarrival.
– Measures total time a job spends in the system from arrival to completion.
– Lower turnaround time means jobs finish sooner, improving throughput.
– Useful metric for batch systems where users care about when jobs complete.
Response Time:
– Definition: Tresponse = Tfirstrun – Tarrival.
– Measures how quickly the system reacts to a new job.
– Critical for interactive systems: users care about system responsiveness.
– Lower response time = faster feedback to the user.
• Trade-off: Improving turnaround may worsen response time, and vice versa.

FIFO (First In, First Out)

• Simple and fair in order of arrival.
• Problem: Convoy effect — long job delays many short jobs.

SJF (Shortest Job First)

• Runs shortest jobs first, minimizing average turnaround time.
• Optimal if all jobs arrive simultaneously.
• Still suffers if long jobs arrive before short ones.

STCF (Shortest Time-to-Completion First)

• Preemptive version of SJF.
• When new job arrives, scheduler picks job with least remaining time.

• Optimal for turnaround when jobs arrive at different times.

Response Time and Round Robin

• STCF is bad for response time (long waits for late jobs).
• Round Robin (RR): each job runs for a time slice, then rotates.
• Great for response time and fairness, but bad for turnaround.
• Time-slice length trade-off: shorter improves responsiveness but increases context-switch cost.

Incorporating I/O

• Jobs often alternate between CPU bursts and I/O waits.
• Scheduler should overlap I/O of one job with CPU of another.
• Treat CPU bursts as separate sub-jobs for scheduling decisions.

Unknown Job Length

• In real systems, OS usually does not know job length.
• Goal: approximate SJF/STCF behavior without perfect knowledge.
• Leads to adaptive schedulers (e.g., multi-level feedback queue, next chapter).

Summary

Schedulers balance performance (turnaround) and fairness/response time. Turnaround time measures job completion speed (important for throughput). Response time measures interactivity (important for user experience). FIFO is simple but suffers from convoy effect. SJF optimizes turnaround but is non-preemptive. STCF preempts for optimal turnaround with varying arrivals. RR improves response time and fairness but worsens turnaround. Schedulers must also handle I/O and job length uncertainty, motivating advanced approaches like MLFQ.