Cheatsheets / **Fundamentals of Operating Systems**

Process Scheduling

Process States

A process typically exists in five states: New, Ready, Running, Blocked, or Finished

Context Switching

Context switching allows CPU cores to alternate between ready and blocked processes to best take advantage of limited computing resources.

Preemption

Preemption occurs when a process is temporarily interrupted by an external scheduler to prioritize a more important task.

Blocked Process

A process is *blocked* when it has to wait for a contested, limited, or slow resource, such as accessing a specific file or waiting for a network request.

The Three Process Schedulers

There are three process schedulers:

- The *long-term scheduler* which admits processes to the Ready queue.
- The *medium-term scheduler* which blocks processes for access to resources.
- The *short-term scheduler* which admits processes from the Ready queue to the CPU to actually be executed.

Long-term Scheduler

The *long-term scheduler* manages the memory of the system and its degree of multitasking by deciding which and how many applications may be loaded into memory.

Medium-term Scheduler

The *medium-term scheduler* is in charge of moving processes out of memory to prioritize others. This can be due to these processes being blocked for resources, lack of activity, low priority, or overly high memory usage.

Short-term Scheduler

The *short-term scheduler* decides which processes in the Ready state to pass onto the CPU. They can also be preemptive, meaning they can forcibly recall processes from the CPU to stop executing if necessary.

First Come, First Served Scheduling

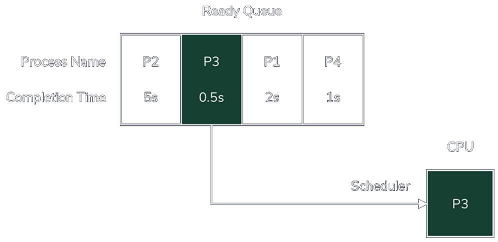
First come, first served is a scheduling algorithm where processes are put into a queue and then executed in the order that they arrive.

Process Starvation

Starvation is the situation in which some processes are never able to be executed due to not being favored by the scheduler. Process starvation can be mitigated by aging tasks such that the priority of a process increases the longer it has been waiting.

Shortest Job First Scheduling

Shortest job first is a scheduling algorithm that prioritizes running the process with the shortest execution time first.



Process Name	P2	P3	P1	P4
Completion Time	5s	0.5s	2s	1s

Multiple-level Queue Scheduling

Multiple-level queue scheduling is a scheduling algorithm that attempts to categorize processes and then place them in multiple queues or levels with different priorities. Tasks are executed by level, such that all of the processes in the topmost level are executed first before moving on to lower levels. If a process is placed in a higher level while a longer one is being processed, the scheduler will move back up to take care of the higher level task first.

Shortest Remaining Time Scheduling

Shortest remaining time is a preemptive scheduling algorithm that prioritizes running the process with the shortest remaining execution time first.

Round Robin Scheduling

Round robin is a scheduling algorithm where a fixed amount of time is chosen and assigned to each process. The scheduler then cycles through all of these processes until they are all completed. Processes that do not finish during their assigned time are rescheduled to allow all other processes an opportunity to run first.

Priority Scheduling

Priority scheduling is a scheduling algorithm that assigns each process a numeric priority and then organizes those processes according to that priority.

Process Scheduler

A *scheduler* is used to organize a computer's limited resources based on some predetermined goal.

Scheduling Throughput

Throughput is the total amount of processes completed per unit of time.

Process Wait Time

Wait time is the amount of time it takes for a process to become ready after being executed.

Process Response Time

Response time is the amount of time it takes a process to finish after becoming ready.