Scheduling

The Problem

- You are the cook at the State Street Diner
  - Customers enter and place orders 24 hours a day
  - Dishes take varying amounts of time to prepare
- What are your goals?
  - Minimize average latency
  - Minimize maximum latency
  - Maximize throughput
- Which strategy achieves your goal?

Context matters!

- What if instead you are:
  - the owner of an expensive container ship, and have cargo across the world
  - the head nurse managing the waiting room of an emergency room
  - a student who has to do homework in various classes, hang out with other students and (occasionally) sleep

Scheduling processes

- OS keeps PCBs, TCBs on different queues
  - Ready processes are on ready queue-OS chooses one to dispatch
  - Processes waiting for I/O are on appropriate device queue
  - Processes waiting on a condition are on an appropriate condition variable queue
- OS regulates PCB migration during life cycle of corresponding process

Schedulers in the OS

- CPU scheduler selects next process to run from the ready queue
- Disk scheduler selects next read/write operation
- Network scheduler selects next packet to send or process
- Page Replacement scheduler selects page to evict

Why scheduling is challenging

- Processes are not created equal!
  - CPU-bound process: long CPU bursts
    - mp3 encoding, compilation, scientific applications

  - I/O-bound process: short CPU bursts
    - index a file system, browse small web pages

  - Balanced
    - playing video, moving windows around

Terminology and Metrics

- Job/Task
  - A user request: e.g., mouse click, web request, shell command...
- Turnaround time
  - Time elapsed between a job's arrival and its completion
- Throughput
  - Number of tasks completed per unit of time

More Metrics

- Response time
  - Time between job's arrival and its first response
- Initial waiting time
  - Time between job's arrival and first time job runs
- Total waiting time
  - Time on the ready queue but not running
    - sum of "red" intervals below

Response time depends on job: we'll assume it equal to the initial waiting time

Other Concerns

- Fairness
  - Equitable division of resources
- Starvation
  - Lack of progress by some job
- Overhead
  - Time wasted switching between jobs
- Predictability
  - Low variance in response time for repeated requests