

# Scheduling

1

## 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?

2

## 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

3

## 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

4

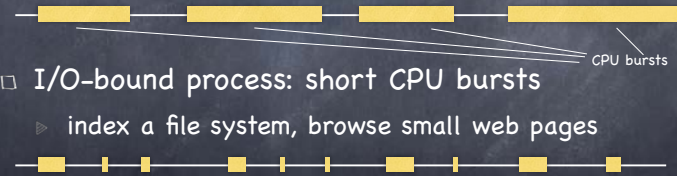
# Schedulers in the OS

- ③ CPU scheduler selects next process to run from the ready queue
- ③ Disk scheduler selects next read/write operation
- ③ Network scheduler selects next packet to send or process
- ③ Page Replacement scheduler selects page to evict

5

# Why scheduling is challenging

- ③ 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



6

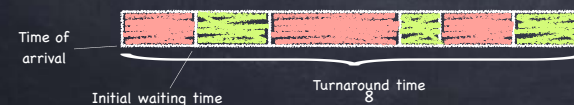
# Terminology and Metrics

- ③ Job/Task
  - ❑ A user request: e.g., mouse click, web request, shell command...
- ③ Turnaround time
  - ❑ Time elapsed between a job's arrival and its completion
- ③ Throughput
  - ❑ Number of tasks completed per unit of time

7

# More Metrics

- ③ Response time
  - ❑ Time between job's arrival and its first response
- ③ Initial waiting time
  - ❑ Time between job's arrival and first time job runs
- ③ 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

8

# Other Concerns

## ④ Fairness

- ❑ Equitable division of resources

## ④ Starvation

- ❑ Lack of progress by some job

## ④ Overhead

- ❑ Time wasted switching between jobs

## ④ Predictability

- ❑ Low variance in response time for repeated requests