Lecture 10:

The one without much prepared content/

Low Latency Scheduling

Announcements

- Presentations/posters on May 4.

  \rightarrow Format:

  \rightarrow Bring laptop

  \rightarrow Presentation/Poster about

  \rightarrow What you were doing

  \rightarrow What you found

  \rightarrow What was hard

  \rightarrow What worked well

  \rightarrow Lessons

Ideally, communicate essence in no more than 3 minutes.
On HWs

- Originally intended to have 3
  - Clearly failed

Plan: Treat them as if everyone got a 100

On to today's topic

- Low latency scheduling

Fundamental tradeoff

\[ \text{SLO} \leftarrow \text{High percentile} \rightarrow \text{Latency} \rightarrow \text{Utilization} \]

Why?
Concurrency = 10 requests at a time

Choices

1. Queue Up Requests
2. Drop Requests

1. How many requests to queue?
   - Process max = 1 ms
     - 99.7% jobs finish in 1 ms
   - Process max is unknown
     - 99.7% jobs finish in 25 ms
N servers: how to distribute load?

When to add servers:

Q: How much to queue?

- Email
- Gaming
- Chat
- Trading

RTO per

Avg process runtime
- # of processors
- Capacity
- E[arrival rate]
- Request distributor

Token use
N-servers how to partition tasks

1. Single queue

   → Work stealing

   → Overhead

Admission control

| Q | > max

- Does it work
- How to decide max
- When does it work