Big Data & Machine Learning

Meeting 1

Plan

- Course Staff
- What (Why) We Are Going To Look At This Semester
- Course Mechanics
- Technical Material Useful For Next Week

Course Staff

- Office Hours
Vinayak Agarwal
- Master's Student
- Took the course last year
- Final

Others in the class ...

Our focus for this offering:
Scheduling & resource management

What?

Why? Isn't this a course about big data & machine learning?
Both Resource Allocation & Execution Order Matter.

**Execution Order**

1. **First Come First Serve**
   
   200μs 1μs 22s 1μs 1μs

2. **Shortest Job First**
   
   ... 200μs 1μs 22s 1μs 1μs

3. ...

**Resource Allocation**

- What jobs end up queuing for the same resource?
- How long does the job run for
  - Different Hardware
  - Different Data Access Latencies
  - ...

No single "Policy" / Algorithm is good for all.

Why Scheduling?
What has changed/enabled Big Data & Machine Learning
• Easy Access to Compute, Storage & Network Resources (Multi-Tenant Data Centers, Etc.)

• Emergence of New Hardware (GPUs, TPUs, Smart NICs, Etc.)

• Emergence of New Programming Models (Map-Reduce, Graph Processing, Data Flow Graphs, Serverless, Etc.)

Each Imposes New/Different Scheduling Reimts

• Multi-Tenant Data Centers

• Hardware: GPU/TPU

• Programming Models
What Have These Changes Enabled

- New Types Of Applications (with New Requirements)
- More Complex Policies

Bottom Line: Scheduling Gives Us A Unifying Lens To Look At All These Changes
**Class Mechanics**

- 4 Homework: 30%
- Midterm: 10%
- Final Exam: 20%
- Final Project: 25%
- Class Participation: 15%

**Class Participation**

5% Participate in Class Discussions

Q: What if I am attending remotely over Zoom?
Class Participation

10% Look Beyond Assigned Work

Examples

Getting Help
How To Submit

- Post on Campuswire by 5pm
- Indicate if you want to speak in class

How Graded

- You will get feedback on this within the week
- Can submit as often as you want
- Compute grades based on two best submissions

Final Project

See Piazza notes for suggestions.
SEE TOSSED NOTES FOR SUGGESTIONS

SOME MATERIAL FOR NEXT WEEK

- Fairness \& GPS \leftarrow Generalized Processor Sharing

\begin{itemize}
  \item Alice
  \item Bob
  \item ABA AB -
  \item ABB ABB -
  \item ABA A
\end{itemize}

DESIRABLE

- Fairness - Equal shares
- Do not leave
Maximize the minimal allocation

\[ s(C_n) = \frac{A_n}{\delta_n} \rightarrow n \text{-dim}. \]

Extending to multiple machines/sources

[Diagram showing allocation and resources]
ALICE
25 job
1 job/8sec

BOB
25 job
1 job/4.8

\[
\text{DR1} = \frac{R}{2}
\]