Midterm Review

DS 5110: Big Data Systems

Spring 2025

Yue Cheng

Midterm exam

- ~~Thursday, February 27, 2:00 pm – 3:15 pm~~

  Tuesday, March 4, 2:00 pm – 3:15 pm
  - Open book, open notes

- Covering four topics from Lec 2 to Lec 8
  - CPU scheduling policies
  - Caching policies
  - Hadoop MapReduce + HDFS
  - Spark RDD

Logistics

- The exam will be remote + synchronous over `gradescope`
- The exam sheet will be available on `gradescope` at 2 pm
- You should work directly on `gradescope`
- Submission closes at 3:25 pm (a grace period of 10 minutes for submission)

CPU job scheduling

- FIFO
  - How it works?
  - FIFO's problems (why we need SJF)?
- SJF
  - How it works?
  - Any limitations (why we need STCF)?
- STCF (preemptive SJF)
  - How it works? How it solves SJF's limitations?
- RR (Round Robin)
  - How it works?

Caching policy

- LRU (least recently used)

- FIFO (first-in, first-out)

MapReduce + HDFS

- How HDFS and MapReduce work
  - The composition of layers (MR atop HDFS)
- The performance characteristics of different phases of a MapReduce job (TeraSort)
- Fault tolerance
  - Storage level: Replication for HDFS
  - Compute level: Backup tasks for MapReduce

Spark

- Motivation
- Transformations and actions
  - Narrow vs. wide transformation
- `.cache()` to pin a computed RDD into memory to avoid recomputation
  - Difference between `.cache()` and `.persist()`

Question types

- Multi-choice questions (~40%)
- True or false questions (~30%)
- Problem solving (~30%)