

Midterm Review

DS 5110: Big Data Systems (Spring 2023)

Yue Cheng



UNIVERSITY
of
VIRGINIA

Midterm exam

- Wednesday, March 1, 11 am – 5 pm
 - Open book, open notes
- Covering three topics from Lec 2 to Lec 4
 - CPU job scheduling
 - MapReduce
 - Spark

Midterm exam

- The exam sheet will be available on Canvas (under “Assignment”) at 11am
- You may work directly on the PDF document
 - Or, you may print it and write on printed papers, make sure you scan it to PDF with **visible resolution**
 - If you choose to scan using a smartphone camera, make sure it **covers everything clearly** – unrecognizable photos will not be graded
- Submission closes at 5pm
 - If you choose to scan, make sure your printer & scanner are handy

CPU job scheduling

turnaround time.

Arrival time → completion

• FIFO

- How it works?
- FIFO's inherent issues (why we need SJF)?

• SJF

- How it works?
- Any limitations (why we need STCF)?

• STCF (preemptive SJF)

SRTF → Shortest Remaining Time First.

- How it works? How it solves SJF's limitations?

• RR (Round Robin)

non-preemptive

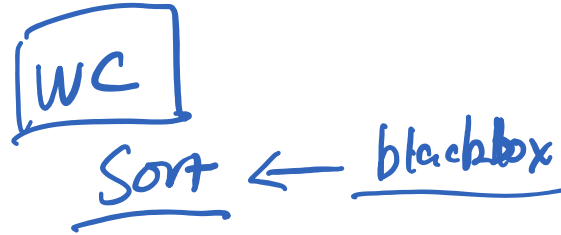
- How it works?

responsiveness

CPU scheduling worksheet

MapReduce

- How MapReduce works



- The performance characteristics of different phases of a MapReduce job (TeraSort)

- Fault tolerance in MapReduce
 - Backup tasks

stragglers

idempotence

- TeraSort evaluation discussion

Spark + RDDs.



- Transformations and actions

Lazy eval.

- • .persist() sched hint

reliable / save-to-disk
HDFS.

- Not an action nor a transformation – tell which RDDs should materialize (memorize)

- • PageRank example

- How iterative PR algorithm works
- How .persist() helps during a PageRank job

Question types

- Multi-choice questions (37.5%) CPU sched.
- Problem solving (62.5%) { MR
Spark.

Good Luck!