Midterm Review

DS 5110/CS 5501: Big Data Systems

Spring 2024

Yue Cheng

UNIVERSITY of VIRGINIA

Midterm exam

- Wednesday, February 28, 3:00 pm – 4:45 pm
  - Open book, open notes

- Covering four topics from Lec 2 to Lec 5
  - CPU job scheduling policies
  - Caching policies
  - MapReduce + HDFS
  - Spark

Midterm exam

- The exam sheet will be available on `gradescope` at 3:00 pm (you will receive entry code after the class)

- You should 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 5:00 pm
  - If you choose to scan, make sure your printer & scanner are handy

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?

CPU scheduling worksheet

Caching policy

- LRU (least recently used)

- FIFO (first-in, first-out)

MapReduce + HDFS

- How MapReduce works
- The performance characteristics of different phases of a MapReduce job (TeraSort)
- Fault tolerance
  - Replication for HDFS
  - Backup tasks for MapReduce

Spark

- Motivation

- Transformations and actions
  - Narrow vs. wide transformation

- PageRank example
  - How iterative PR algorithm works
  - Optimizations on baseline PageRank
    - Co-partitioning for communication-efficient join
    - Apply `.persist(StorageLevel.DISK_ONLY)` for fault tolerance

Question types

- Multi-choice questions (40%)

- True or false questions (25%)

- Problem solving (35%)

Good Luck!

Quizzes Q&A

Quiz 1: What's the hit ratio for Problem 3 in the Caching Policy worksheet?

Quiz 2 (Lec 3 – Slide 42):
Q1: What's the job completion time with 1 worker?
A: 65

Q2: What's the job completion time with 3 workers?
A: 35

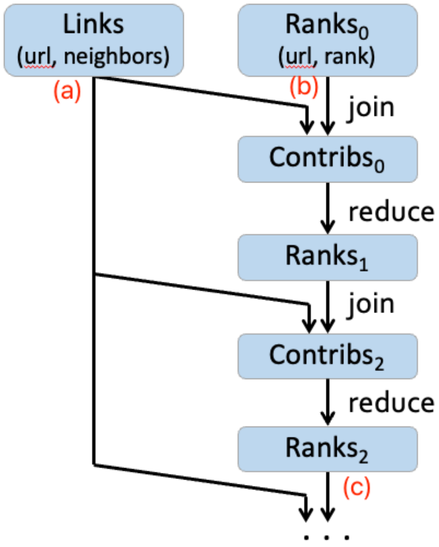
Q3: What's the speedup?
A: 65/35

Quiz 3 (Lec 4b - Stragglers):
Q1: Would backup tasks cause correctness issue in MapReduce jobs? Why or why not?
A: No. Because the output of a MapReduce task is the same no matter how many times you run that task.

Q2: What property of MapReduce the backup tasks exploit?
Idempotence

Quiz 4 (Lec 5b – Slide 26):
Q1: Which RDD should one apply .persist() to?
A: Ranks.

Q2: Where might we have placed .persist() for better fault tolerance?
A: (c)


Links
(url, neighbors)
Ranks$_0$
(url, rank)
(a)
(b)
join
Contribs$_0$
reduce
Ranks$_1$
join
Contribs$_2$
reduce
Ranks$_2$
(c)
. . .