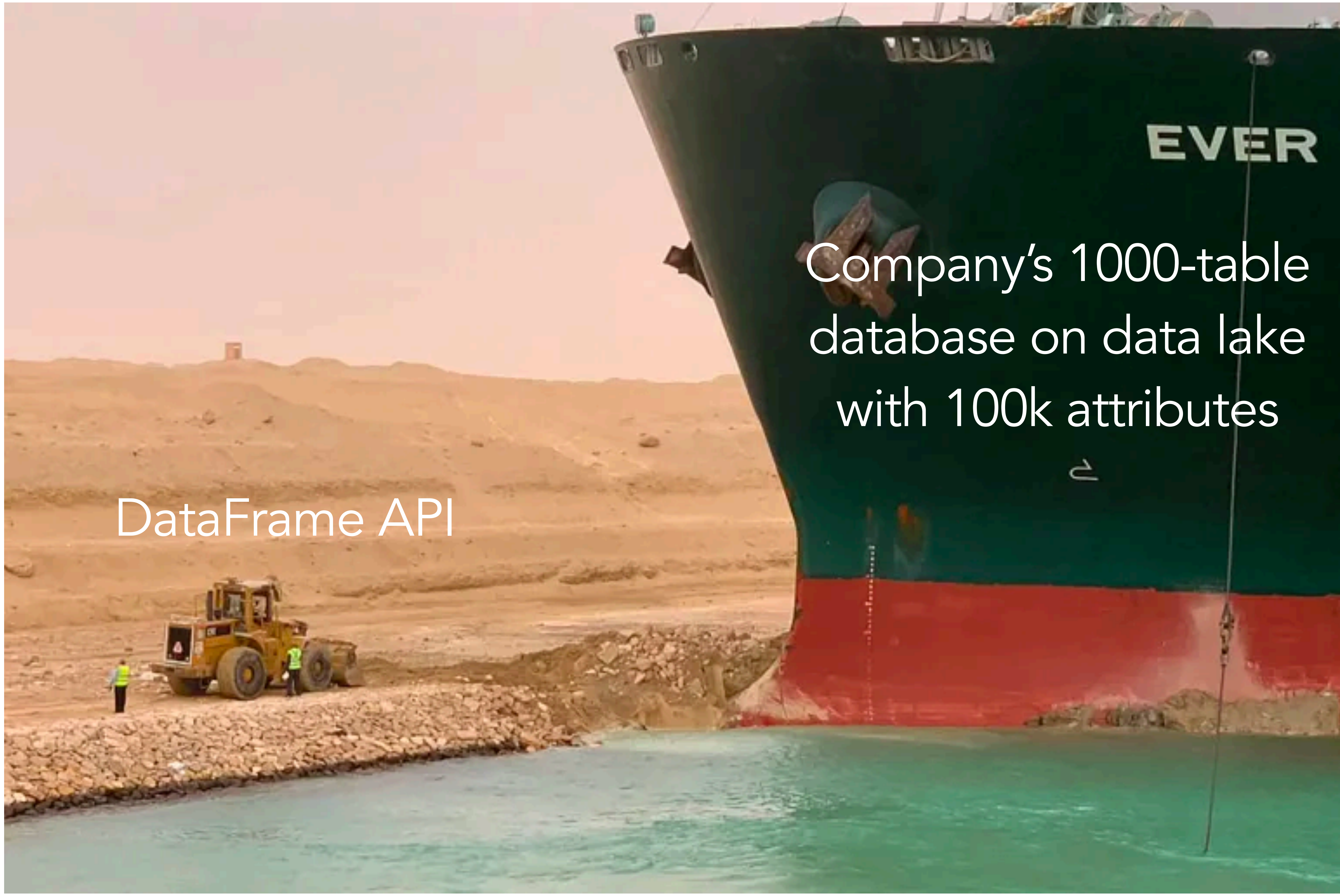
DSC 204a Scalable Data Systems

- Haojian Jin


Company's 1000-table database on data lake with 100k attributes
DataFrame API
EVER

Meme idea credit: https://datasystemsfun.tumblr.com/

Today’s activities

1. Process.
2. PIA
3. Filesystems and Data Files

Where are we in the class?

Foundations of Data Systems

- Digital representation of Data → Computer Organization → Memory hierarchy → Process → **Storage**

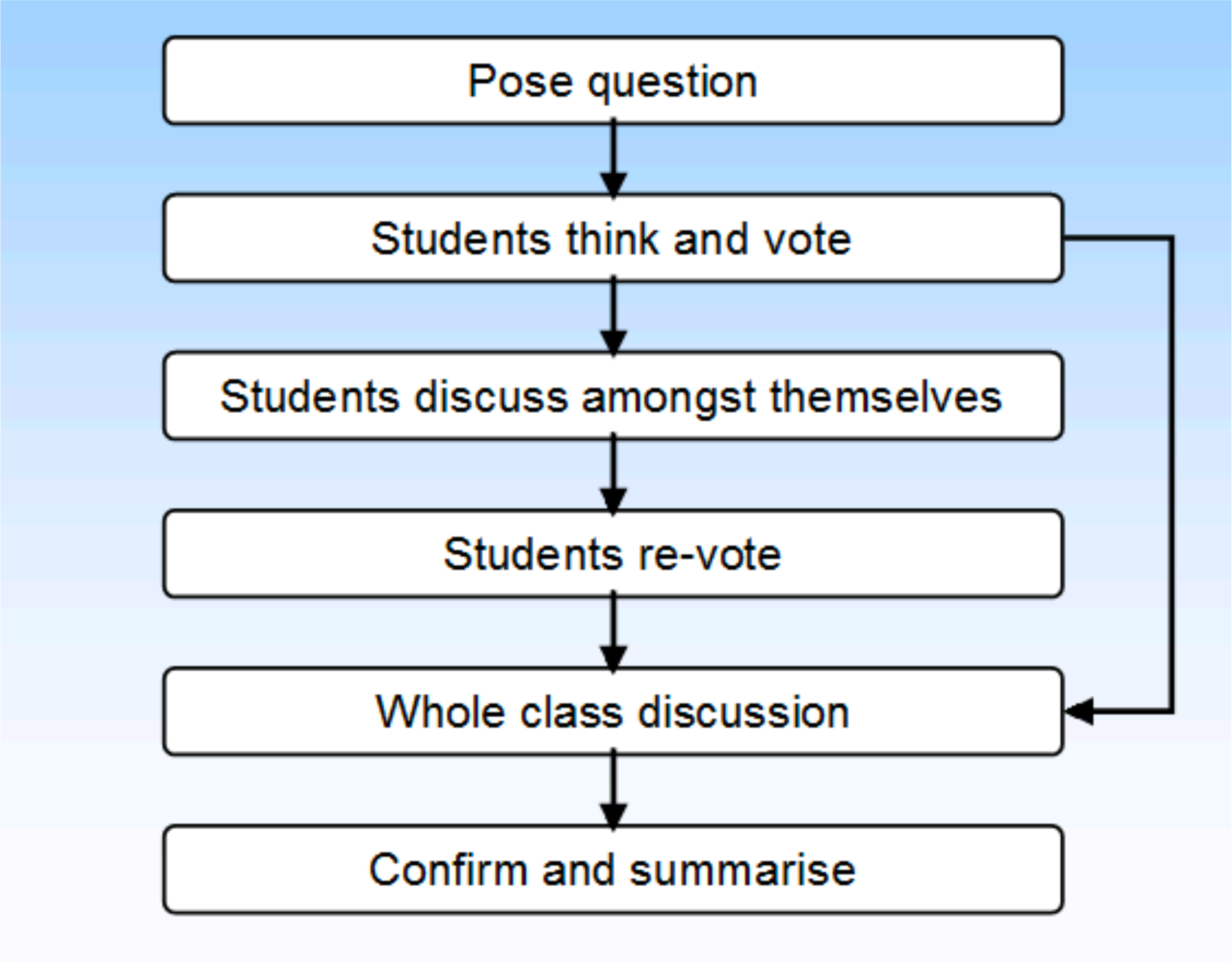
Scaling Distributed Systems

- Cloud → Distributed storage → Partition and replication → HDFS

Data Processing and Programming model

- Data encoding evolution → Batch processing (MapReduce) → Stream processing (Spark)

Peer instruction activity



Pose question

Students think and vote

Students discuss amongst themselves

Students re-vote

Whole class discussion

Confirm and summarise

Peer Instruction Activity

bit.ly/dsc204aApr14

Peer Instruction Activity (About 1min per 1pt)

Q1. **[2pts]** Which of these levels of the memory hierarchy typically has the lowest latency to read data from?

A. Flash SSD
B. Magnetic hard disk
C. DRAM
D. All have similar latency
E. None of the above

Peer Instruction Activity (About 1min per 1pt)

Q2. **[2pts]** Which of these levels of the memory hierarchy typically has the lowest latency to write data from?

A. Flash SSD
B. Magnetic hard disk
C. DRAM
D. All have similar latency
E. None of the above

Peer Instruction Activity (About 1min per 1pt)

Q3. **[6pts]** Suppose you have a 750 x 250 matrix of 32-bit integers in Python memory. Its entries are known to be non-negative and < 10. You write it to disk as a CSV file with one row per line.

What is the rough maximum ratio of the size of your file on disk versus the size of the matrix as an object in DRAM?

The answer is one of the following. If you pick the right answer, you need not explain. If you pick the wrong answer and seek partial credits, please explain your answer succinctly and clearly.

A. 0.25　　B. 0.5　　C. 0.75
D. 1.0　　E. 1.33　　F. 1.5
G. 1.67　　H. 2.0

Peer Instruction Activity (About 1min per 1pt)

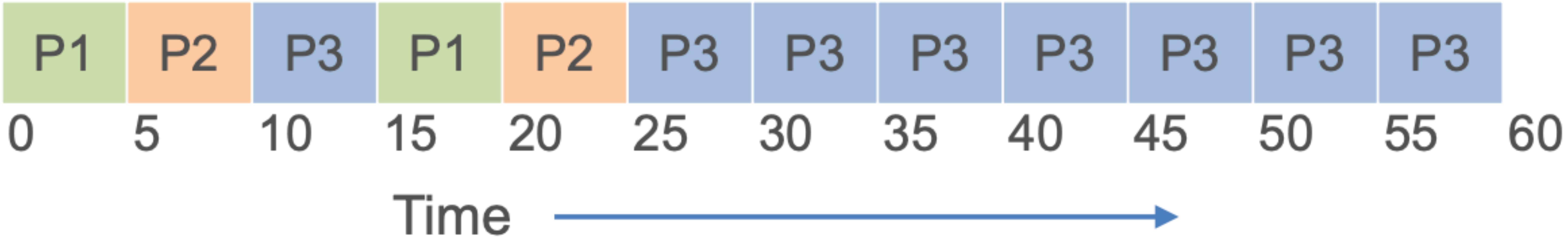
Q4. **[3x6pts]** Here is a Gantt Chart for 3 processes of the given lengths that all arrive at time 0.

A) What is the rough *average response time*?

B) What is the rough *average turnaround time*?

C) Which scheduling policy/policies discussed in class (FIFO, SJF, SCTF, RR) may produce this given schedule? Explain clearly.

P1, P2, and P3 are of lengths 10, 10, and 40 units, resp.



Peer Instruction Activity (About 1min per 1pt)

Q5. **[3x2pts]** For each of the following answer True of False:

A) An OS typically has mechanisms to wrest control of hardware back from a user process.

B) SCTF is the fairest scheduling policy we discussed in class.

C) CPU caches are usually cheaper per MB than Flash SSD.

Peer Instruction Activity (About 1min per 1pt)

Q5.**(Advanced; Optional) Q9)** Suppose you are given that *n* processes of the same lengths (*k* units each) all arrive at roughly the same time. Answer these questions.

A. **[8pts]** Which scheduling policy/policies discussed in class (FIFO, SJF, SCTF, RR) will give the *lowest average turnaround time*? What is that lowest value?

B. **[8pts]** Which scheduling policy/policies discussed in class (FIFO, SJF, SCTF, RR) will give the *lowest average response time*? What is that lowest value?

C. **[4pts]** For what *k* will all 4 policies yield the same metrics?

Today

Process management: virtualization & Concurrency
Filesystems and Data Files
Main memory management
IO & Unix Pipes