

# CS 544 (CS 639-4): Intro to Big Data Systems

## Spring 2023 Midterm Practice

Q1. Who are you allowed to discuss the quiz questions with while taking it?

- (A) TAs
- (B) classmates who are taking the quiz at the same time
- (C) friends who have already taken the quiz
- (D) nobody

Q2. What does the JVM execute?

- (A) bytecode
- (B) machine code
- (C) Java source code
- (D) Python source code

Q3. You buy a bigger hard drive for your computer to store more movies. This is an example of "scaling \_\_\_\_\_".

- (A) down
- (B) out
- (C) up
- (D) left
- (E) right

Q4. 4 Mbps is equivalent to what?

- (A) 512 KB/s
- (B) 1 MB/s
- (C) 0.5 GB/s
- (D) 32 KB/s
- (E) 4 MB/s

Q5. If "ls -la" shows the following, which one is NOT a directory?

```
total 76
drwxrwxr-x  3 trh trh  4096 Jan 27 19:23 .
drwxr-x--- 13 trh trh 65536 Jan 27 19:23 ..
-rw-rw-r--  1 trh trh      0 Jan 27 19:23 A
drwxrwxr-x  2 trh trh  4096 Jan 27 19:23 B.out
```

- (A) .
- (B) B.out
- (C) ..
- (D) A

Q6. Which command can use to show you the first few lines of a file?

- (A) echo
- (B) first
- (C) top
- (D) htop
- (E) head
- (F) ps
- (G) lsof

Q7. How can you count the number of files containing "txt"?

- (A) find txt --count
- (B) grep-find txt | lc
- (C) find txt | wc
- (D) cat txt | wc
- (E) find . | grep txt | wc

Q8. You have an LRU cache of size 3. How many hits will you get for the following access pattern?

1, 2, 3, 1, 4, 1, 1, 1

- (A) 3 (B) 7 (C) 5 (D) 0 (E) 8 (F) 6 (G) 4 (H) 2 (I) 1

Q9. You have an FIFO cache of size 3. How many hits will you get for the following access pattern?

1, 2, 3, 1, 4, 1, 1, 1

- (A) 0
- (B) 5
- (C) 8
- (D) 6
- (E) 1
- (F) 3
- (G) 7
- (H) 4
- (I) 2

Q10. Which of the following is capable of holding the largest number?

- (A) Python int
- (B) int64
- (C) int32

Q11. When running `./check-ebay.sh`, you get this error:

```
bash: ./check-ebay.sh: Permission denied
```

What should you try first?

- (A) update the PATH variable
- (B) add a shebang
- (C) use chmod

Q12. The shape of tensor A is (4, 256) and the dtype is float64. Approximately how many bytes does A consume?

- (A)  $4 \times 256$
- (B)  $4 \times 256 \times 64 / 8$
- (C)  $4 \times 256 \times 64$
- (D)  $4 \times 256 \times 64 \times 8$
- (E)  $4 \times 256 + 64$

Q13. You run the following:

```
y = f(x)
y.backward()
```

Now `x.grad` is 8. If you're trying to MINIMIZE `y`, what should you do?

- (A) increase `x`
- (B) decrease `x`

Q14. During the first few epochs of optimization, your loss increases, before becoming Inf. What should you do?

- (A) use a smaller learning rate
- (B) use a bigger learning rate

Q15. Each thread has its own \_\_\_\_\_. Choose the most complete answer that is true.

- (A) instruction pointer and heap
- (B) instruction pointer, stack, and heap
- (C) heap and stack
- (D) instruction pointer
- (E) instruction pointer and stack

Q16. Network packets are often dropped due to congestion. What protocol helps with reliability by resending packets that appear to have been dropped?

- (A) IP
- (B) UDP
- (C) MAC
- (D) TCP
- (E) HTTP

Q17. A number is being represented with 3 bytes. What type might it be?

- (A) PyTorch `int32`
- (B) Python `float`
- (C) PyTorch `int64`
- (D) protobuf `int64`

Q18. After a very long Docker build, you realize you need to make a small change and then build again. Will making the change near the beginning or end of the Dockerfile make the second build FASTER?

- (A) end
- (B) beginning

Q19. You started a container on your VM with the "-p 127.0.0.1:9000:8000". Then you will create an SSH tunnel from your laptop to the VM with "-L localhost:9000:localhost:????". What should ???? be so that you can communicate between a browser on your laptop with a process in the container?

- (A) 7000
- (B) 8888
- (C) 8000
- (D) 22
- (E) 9000
- (F) 80

Q20. Which access pattern is most problematic for SSDs?

- (A) sequential reads
- (B) random reads
- (C) random writes
- (D) sequential writes

Q21. Which file system is a pseudo file system?

- (A) HDFS
- (B) ext4
- (C) tmpfs
- (D) NFS
- (E) procfs

Q22. A row-oriented layout is generally most useful for what kind of database?

- (A) Data warehouse
- (B) OLTP
- (C) OLAP

Q23. For which kind of database would a row-oriented file layout generally be better?

- (A) Data warehouse
- (B) OLAP
- (C) OLTP

Q24. Which SQL clause is responsible for projection?

- (A) WHERE
- (B) HAVING
- (C) GROUP BY
- (D) SELECT

Q25. There is a 3x replicated HDFS file in a cluster, of size 10 MB. A client copies this to a 2x replicated HDFS file in another cluster. About how many bytes will be read and written to disks overall, assuming no caching?

- (A) 30 MB read and 10 MB written
- (B) 10 MB read and 10 MB written
- (C) 10 MB read and 20 MB written
- (D) 60 MB read and 60 MB written
- (E) 30 MB read and 20 MB written

Q26. Which webhdfs operation makes use of HTTP redirects (status code 307)?

- (A) GETFILESTATUS
- (B) OPEN
- (C) LISTSTATUS
- (D) MKDIRS

Q27. Say a CPU provides 0.7 TFLOPS. That is the same as \_\_\_\_ GFLOPS.

- (A) 0.007
- (B) 7
- (C) 700
- (D) 7000

Q28. A computer can download 16 MB/s. What is this in Mbps?

- (A) 1 Mbps
- (B) 2 Mbps
- (C) 8 Mbps
- (D) 16 Mbps
- (E) 32 Mbps
- (F) 128 Mbps

Q29. About how much memory does A use to store elements?

`A = torch.rand(1024, 1024, dtype=torch.float32)`

- (A) 8 bytes
- (B) 32 bytes
- (C) 32 KB
- (D) 4 MB
- (E) 8 MB
- (F) 32 MB

Q30. Assume any bytecode-level interleaving is possible, and total starts at 0. What is the SMALLEST possible final value for total?

**thread 1 (T1)**

load total  
load 2  
add  
store total

**thread 2 (T2)**

load total  
load 3  
add  
store total

- (A) 0 (B) 1 (C) 2 (D) 3 (E) 4 (F) 5

Q31. What is "42:01:0a:80:00:25" an example of?

- (A) port number
- (B) IP address
- (C) MAC address

Q32. You started a container on your VM with the "-p 127.0.0.1:8000:9000". Then you created an SSH tunnel from your laptop to the VM with "-L localhost:7000:localhost:8000". What port number should you use in the browser on your laptop when trying to communicate with the server inside the container?

- (A) 127
- (B) 127.0.0.1
- (C) 7000
- (D) 8000
- (E) 9000

Q33. A CSV like this is loaded:

x,y  
a,b  
c,d

The table is written to another format such that the values appear in the file in the following order: acbd.

- (A) the new format is column oriented
- (B) the new format is row oriented

Q34. A client is writing 5 MB of data to a 4x replicated HDFS file. About how much data does the client send over the network?

- (A) 4 MB (B) 5 MB (C) 20 MB (D) 9 MB

Q35. For which node type is a format necessary before the first launch?

- (A) Client only
- (B) NameNode only
- (C) DataNode only
- (D) NameNode and DataNode only
- (E) Client, NameNode, DataNode



## **ANSWER KEY**

Q1: B  
Q2: A  
Q3: C  
Q4: A  
Q5: D  
Q6: E  
Q7: E  
Q8: G  
Q9: F  
Q10: A  
Q11: C  
Q12: B  
Q13: B  
Q14: A  
Q15: E  
Q16: D  
Q17: D  
Q18: A  
Q19: E  
Q20: C  
Q21: E  
Q22: B  
Q23: C  
Q24: D  
Q25: C  
Q26: B  
Q27: C  
Q28: F  
Q29: D  
Q30: C  
Q31: C  
Q32: C  
Q33: A  
Q34: B  
Q35: B