DSC 102 Winter 2020 Midterm Exam

Full Name :

Student ID :

INSTRUCTIONS

1. You have 80 minutes to complete this exam.
2. Mark your answers **on the bubble sheet** by filling the correct options using a dark pen or pencil. Please ensure your markings are **dark, crisp, and not smudged**.
3. You can have up to 2 letter/A4-sized sheets of notes, formulae, etc. Apart from this, the exam is closed book/notes/electronics/peers.
4. Please wait until being told to start reading and working on the exam.
5. When you finish or when the time is over, **return both** this question booklet and your bubble sheet.

Part 1. [6 x 1 = 6pts] For each statement below, indicate if it is True (pick option A) or False (pick option B).

1. An SQL query is a string.

2. In the task parallelism paradigm discussed in class, if no worker has any idle times in the schedule, we will likely get linear speedup for the workload.

3. Serverless setups in the cloud can substantially reduce resource wastage compared to classical IaaS.

4. In the scaleup plot (weak scaling), the dataset size is fixed as the factor is varied.

5. Data processing programs need to go through the OS System Call API to read text files but can typically bypass that API if they want to read binary files.

6. All threads of a multi-threaded process share the same address space.

Part 2. [12 x 2 = 24pts] Answer the following questions. Only one option must be picked–pick the best one.

7. Which of the following is typically *not* considered one of the "3 Vs of Big Data"?

   (A) Variety (B) Vitality (C) Velocity (D) Volume (E) Both B & C

8. Suppose you are given 4 models for a prediction task: M1, M2, M3, and M4 with respective prediction errors of 12%, 8%, 15%, and 5% (lower is better) and respective monetary costs for building and deployment of 10K, 40K, 20K, and 90K dollars (lower is better). Which model is *not* Pareto-optimal when prediction errors and monetary costs are both important?

   (A) M1 (B) M2 (C) M3 (D) M4 (E) None of these

9. Which of the following is *not* a typical file format to store structured data?

   (A) CSV (B) JSON (C) TSV (D) JPEG (E) Both A & B

10. Which of the following properties of data processing programs is sometimes exploited to help reduce runtimes?

    (A) Spatial locality of reference (B) Temporal locality of reference

    (C) Parallelism in computations (D) All of A, B, & C (E) None of these

11. What is the OS term for a virtual slot of DRAM that holds data read in from disk?

    (A) Page frame (B) File frame (C) Folder frame (D) Disk frame (E) Register

12. Which of the following is typically considered SaaS in cloud jargon?

    (A) EC2 (B) S3 (C) EBS (D) Lambda (E) SageMaker

13. Which of the following storage devices in the memory hierarchy typically has a dichotomy for random vs sequential access latency?

    (A) CPU Caches (B) DRAM (C) Magnetic Hard Disk

    (D) Flash SSD (E) None of these

14. Suppose you spin out an EC2 cluster and read a whole dataset from S3 to each node's DRAM. What form of parallelism is this typically called?

    (A) Shared Nothing (B) Shared Disk (C) Shared Memory

    (D) Both B & C (E) None of these

15. Which of the following tools is custom-designed for tensor dataflow graphs?

    (A) RDBMSs (B) Python Pandas (C) PyTorch

    (D) Both A & B (E) All of A, B, & C

16. Among these popular two-dimensional structured data models, which one has no in-built notion of ordering among the rows?

    (A) Matrix (B) Relation (C) DataFrame (D) Both A & B (E) Both B & C

17. Which component of the access latency for reading a disk block from a magnetic hard disk is primarily affected by the RPM of the disk?

    (A) Rotation delay (B) Seek time (C) Transfer time

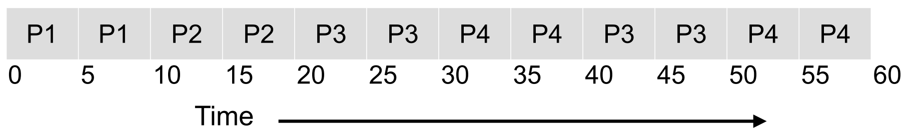
    (D) Both A & B (E) None of these

18. How many positive integers are there with exactly 2 digits in decimal representation that also have the exact same 2 digits in hexadecimal representation?

    (A) 0 (B) 9 (C) 10 (D) 15 (E) 16

Part 3. [18pts] Consider the following Gantt Chart for concurrent execution of 4 processes on a processor.

The processes P1, P2, P3, and P4 arrive at times 0, 5, 10, and 15, respectively. They have lengths 10, 10, 20, and 20, respectively (in time units).



Answer the following questions. Only one option must be picked–pick the best one.

19. **[3pts]** Which process sees the largest response time?

    (A) P1 (B) P2 (C) P3 (D) P4 (E) All 4 see equal response times

20. **[4pts]** What is the average response time (in time units)?

    (A) 5 (B) 7.5 (C) 10 (D) 12.5 (E) 15

21. **[4pts]** What is the average turnaround time (in time units)?

    (A) 15 (B) 20 (C) 22.5 (D) 25 (E) 27.5

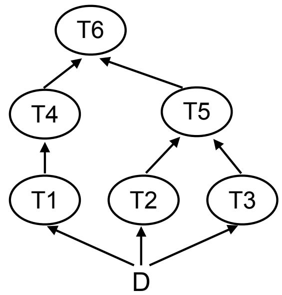
22. **[3pts]** One crude way to think about fairness in scheduling is to divide each process's turnaround time by its length; the lower this ratio, the better the deal a process got in the schedule. Viewed this way, which process got the worst deal in the given schedule?

    (A) P1 (B) P2 (C) P3 (D) P4 (E) All 4 got equally good deals

23. **[4pts]** Which of the following scheduling algorithms discussed in class could plausibly yield the given schedule?

    (A) FIFO (B) SJF (C) SCTF (D) Round Robin (E) None of these

Part 4. [20pts] Consider the following task graph and given task lengths (in time units).



Task	Length
T1	10
T2	15
T3	10
T4	20
T5	10
T6	5

Answer the following questions. Only one option must be picked–pick the best one.

24. **[2pts]** What is the largest degree of parallelism if you were to execute this workload using task parallelism as discussed in class?

    (A) 2 (B) 3 (C) 4 (D) 5 (E) 6

25. **[4pts]** Suppose you are given 3 identical worker nodes and use task parallelism as discussed in class (i.e., preemption or migration of tasks is *not* allowed) . What is the lowest possible completion time of this workload?

    (A) 20 (B) 25 (C) 30 (D) 35 (E) 40

26. **[4pts]** Continuing with the above question's setup, what is the highest possible speedup against running on only one worker node?

    (A) 1x (B) 1.5x (C) 2x (D) 2.5x (E) 3x

27. **[5pts]** Continuing with the above question's setup, in a task-parallel schedule that offers the highest possible speedup, what is the highest possible idle time of one worker among the 3 workers? Assume all workers are on from start to the end of the whole workload.

    (A) 20 (B) 25 (C) 30 (D) 35 (E) 40

28. **[5pts]** Now suppose you are given only 2 identical worker nodes for task parallelism as discussed in class. What is the lowest possible completion time of this workload now?

    (A) 20 (B) 25 (C) 30 (D) 35 (E) 40

Part 5. [12pts] Suppose you are given an artificial neural network (ANN) model consisting of 2 weight matrices of respective dimensions 100×1000 and 1000×20. Also suppose you use the following custom *float5* representation for the model weights.

	sign (1 bit)	exponent (2 bits)		fraction (2 bits)	
	0	0	1	0	1
Bit number:	b_4	b_3	b_2	b_1	b_0

$$(-1)^{sign} \times 2^{exponent-2} \times (1 + b_1 2^{-1} + b_0 2^{-2})$$

Answer the following questions. Only one option must be picked–pick the best one.

29. **[2pts]** What is the largest possible number of unique weight values?

    (A) 32 (B) 10 (C) 24 (D) 256 (E) 64

30. **[4pts]** Suppose one of the ANN weights is "00101" in binary (as in the figure). What is its value in real decimal?

    (A) 0.25 (B) 0.375 (C) 0.5 (D) 0.625 (E) 0.75

31. **[2pts]** What is the rough total size of the ANN model in Kilobytes?

    (A) 50 (B) 75 (C) 100 (D) 120 (E) 150

32. **[4pts]** Suppose you serialize the ANN weights to a human-readable CSV file with ASCII text (1B per character). Roughly, what is the largest possible size of this file in Kilobytes? Ignore trailing zeros but do not ignore a pre-decimal point zero.
    Hint: -0.25 will be converted to the 5-character string "-0.25".

    (A) 600 (B) 840 (C) 960 (D) 1080 (E) 1200

33. (Optional) Extra Credit. [4pts] You are given two matrices $\mathbf{A}_{m\times n}$ and $\mathbf{B}_{n\times p}$ represented as relations with one tuple per cell in the following schemas: *A(row, column, value)* and *B(row, column, value)*. How do you compute the matrix product **AB** in SQL?

(A) SELECT A.column, B.row, SUM(A.value * B.value) FROM A, B WHERE A.row = B.column;

(B) SELECT A.row, B.column, SUM(A.value * B.value) FROM A, B WHERE A.column = B.row GROUP BY A.row, B.column;

(C) SELECT A.row, B.column, SUM(A.value * B.value) FROM A, B WHERE A.column = B.row;

(D) SELECT A.column, B.row, SUM(A.value * B.value) FROM A, B WHERE A.column = B.row GROUP BY A.column, B.row;

(E) SELECT A.row, B.column, SUM(A.value * B.value) FROM A, B WHERE A.row = B.column GROUP BY A.row, B.column;