DSC 102 Winter 2020 Midterm Exam
Answers

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.

   (A) True

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.

   (A) True

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

   (A) True

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

   (B) False

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.

   (B) False

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

   (A) True

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

   ANSWER: (B)

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

   ANSWER: (C). M3 is dominated by M1 on both axes.

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

   ANSWER: (D)

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

    ANSWER: (D)

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

    ANSWER: (A)

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

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

    ANSWER: (E)

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

    ANSWER: (C)

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

    ANSWER: (B)

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

    ANSWER: (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

    ANSWER: (B)

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

    ANSWER: (A)

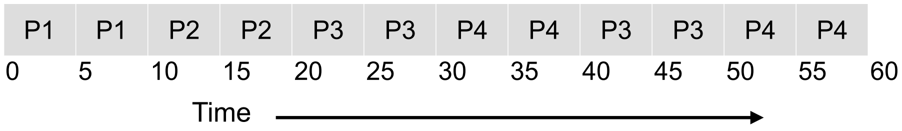
18. How many positive integers are there with exactly 2 digits in decimal representation that also have the exact same 2 digits in hexadecimal representation?

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

    ANSWER: (A). Such a number cannot exist! :) Here is a simple proof by contradiction. Suppose the number has xy as its 2 decimal digits, which means $x \neq 0$. Its value in decimal is $10x + y$ but its value in hexadecimal with the same 2 digits is $16x + y$. These cannot be equal unless $x = 0$.

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

    ANSWER: (D). P4's response time of $30 - 15 = 15$ is the largest.

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

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

    ANSWER: (B). Average response time = $((0-0)+(10-5)+(20-10)+(30-15))/4 = 7.5$.

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

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

    ANSWER: (E). Average turnaround time = $((10-0)+(20-5)+(50-10)+(60-15))/4 = 27.5$.

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

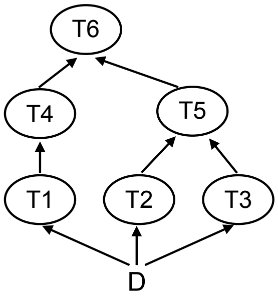
    ANSWER: (D). P4 again; its ratio of $45/20 = 2.25$ is the largest.

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

ANSWER: (D). The schedule has preemption, which immediately rules out FIFO and SJF. SCTF cannot preempt P3 to run P4. So, only Round Robin is left. With a quantum of 10, Round Robin could actually yield the given schedule.

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

    ANSWER: (B). T1, T2, and T3 can run in parallel.

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

    ANSWER: (D). Longest path from D to end is T1 → T4 → T6.

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

    ANSWER: (C). Total time of all tasks is 70; speedup is 70/35 = 2x.

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

    ANSWER: (B). T1, T4, T6 on W1; T2 and T5 on W2; T3 on W3. Idle time is largest on W3: $35 - 10 = 25$. Note that in the task parallel setup we discussed, there is no task preemption or movement across workers.

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

    ANSWER: (E). With only 2 workers, one best possible schedule places T3 after T2 and before T5 on W2, while W1 runs T1 and T4. Note T6 cannot start until both T4 and T5 finish. So, new overall lowest completion time possible is 35 (due to W2) + 5 (for T6).

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:	b4	b3	b2	b1	b0

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

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

    ANSWER: (A). Given 5 bits per weight, we have $2^5 = 32$ possible unique values. The observant will note that this set has separate "+0" and a "-0"! This is one of the quirks of floating points. :) While distinguishing +0 and -0 is mathematically "nonsensical," they have computationally distinct properties.

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

ANSWER: (D). $(-1)^0 \times 2^{1-2} \times (1 + 0.25) = 0.625$.

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

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

ANSWER: (B). $((100 \cdot 1000) + (1000 \cdot 20)) \cdot 5$ bits = 75 KB.

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

ANSWER: (C). The largest number of ASCII characters in a float5 weight is 7, e.g., as given by the following bit sequence: 10001, which is the real decimal -0.3125. We have a total of $(100 \cdot 1000) + (1000 \cdot 20) = 120$K weights. Since it is a CSV, we add 1B to each weight for the comma. So, the largest file size is $120K \times 8B = 960KB$.

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;

ANSWER: (B). Matrix-multiplication is basically a join-aggregate query. The output inherits row numbers of A and column numbers of B (SELECT clause). Mapping columns of A to rows of B becomes the join condition (FROM and WHERE clauses). Summing up on the middle dimension for each cell of the output leads to the GROUP BY SUM.