Q1. [10pts] Suppose you deploy a logistic regression model on a custom-built inference accelerator for a low-resource IoT device. The model has 1000 weights, which are all represented using a custom 4-bit float-like real number format as shown below.

<table>
<thead>
<tr>
<th>Bit number:</th>
<th>b3</th>
<th>b2</th>
<th>b1</th>
<th>b0</th>
</tr>
</thead>
<tbody>
<tr>
<td>sign (1 bit)</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>value (2 bits)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>doubling (1 bit)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Answer the following questions. Pick only the single best option for the first three sub-questions. Partial credits are possible for the last sub-question; show all your steps.

1. [2pts] What is the rough total size of the logistic regression model in bytes?

   (A) 400  (B) 100  (C) 500  (D) 1000  (E) None of the rest

   ANSWER: (C). $1000 \times 4 \text{ bits} = 500 \text{ bytes}$.

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

   (A) 24  (B) 8  (C) 32  (D) 16  (E) None of the rest

   ANSWER: (D). Given 4 bits per weight, we have $2^4 = 16$ unique values.

3. [3pts] Suppose one of the model’s weights is 1101 in binary, as in the figure. What is its value in real decimal?

   (A) -1.5  (B) -2.5  (C) -1.0  (D) -0.5  (E) None of the rest

   ANSWER: (B). $(-1)^{1} \times (0.25 + 1) \times 2^{1} = -2.5$.

4. [3pts] Suppose you serialize the weights into a human-readable TSV file with ASCII text (1 byte per character). Roughly, what is the largest possible size of this file in bytes? Ignore trailing zeros but do not ignore a pre-decimal point zero.

   **Hint:** $-0.5$ will be converted to the 4-character string “-0.5”.

   ANSWER: 6000. The largest number of ASCII characters in this weight format is 5, e.g., 1100 representing the decimal $-1.25$. Since it is a TSV, we add 1 byte to each weight for a tab space. So, the largest file size is $1000 \times 6B = 6000B$. 

Q2. [10pts] Consider the following Gantt Chart showing the full concurrent schedule of 4 processes on a 2-core machine. All processes arrived at time 0.

```
C1:  P1  P3  P1  P3  P1
C2:  P2  P4  P2  P4  P4
```

Answer the following questions. Pick only the single best option for the first three sub-questions. Partial credits are possible for the last sub-question; show all your steps.

1. [2pts] What is the average response time?
   (A) 2.5     (B) 0     (C) 3.3     (D) 5     (E) None of the rest

   **ANSWER: (A).** P1 and P2 start at 0; P3 and P4 start at 5. So, average response is (0+0+5+5)/4 = 2.5.

2. [2pts] Which process has the lowest turnaround time?
   (A) P1     (B) P2     (C) P3     (D) P4     (E) All 4 are equal

   **ANSWER: (B).** Since all arrived at the same time, just look for which finished first.

3. [2pts] Which scheduling policy from class is most likely to produce this schedule?
   (A) FIFO     (B) SJF     (C) SCTF     (D) Round Robin     (E) None of the rest

   **ANSWER: (D).** Round Robin with quantum of 5.

4. [4pts] One crude way to think about the fairness of a schedule is to divide each process’s turnaround time by its job length. The lower this ratio, the fairer the deal a process got. Viewed this way, which process got the least fair deal here?

   **ANSWER: P3.** The ratios are as follows: P1: 25/15; P2: 15/10; P3: 20/10; P4: 25/15. So, P3 has the largest ratio.

Q3. [10pts] Consider the following task graph and given task lengths (in time units). The unspecified x and y are non-negative numbers.

Answer the following questions. Pick only the single best option for the first three sub-questions. Partial credits are possible for the last sub-question; show all your steps.
1. [2pts] What is the degree of parallelism of this workload for task-parallel execution?
(A) 2  (B) 4  (C) 3  (D) 1  (E) None of the rest

**ANSWER:** (C). T6, T4, and T7 can run in parallel.

2. [2pts] Suppose \((x, y) = (10, 20)\). What is the lowest possible completion time of this workload with task-parallelism?
(A) 50  (B) 55  (C) 40  (D) 45  (E) None of the rest

**ANSWER:** (A). Longest path is T2 \(\rightarrow\) T5 \(\rightarrow\) T7.

3. [2pts] Suppose \(x = y = 15\). Suppose the workload is executed in a task-parallel manner for the lowest possible completion time on 3 workers. All workers are on till the last task finishes. What is the total idle time (add across workers)?
(A) 25  (B) 30  (C) 45  (D) 35  (E) None of the rest

**ANSWER:** (B). We assign 3 workers. T1-T3 and T2-T5 can run in parallel on only 2 workers, followed by T4, T6, T7 running in parallel on 3 workers. So, total idle time is T1-T3 or T2-T5 time, which is 30.

4. [4pts] Suppose \((x, y) = (10, 20)\). What is the lowest possible completion time with task-parallelism when given only 2 workers? Show the Gantt Chart clearly.

**ANSWER:** 55. T1-T3 and T2-T5 run in parallel on 2 workers; both run for 30. After those, among T4, T6, T7, place the longest on first worker (20) and the other two on the second \((10+15 = 25)\). So, the time to completion now is \(30 + 25 = 55\).

Q4. \([10 \times 2 = 20 pts]\) Answer the following questions. Only one option must be circled—pick the best one.
1. Which of these structured data models has native support for mixed types in columns?
   (A) Relation  (B) Matrix  (C) DataFrame  (D) Both A and B  (E) Both A and C
   ANSWER: (C)

2. Which of these paradigms of parallelism is most common in data systems?
   (A) Shared-Nothing  (B) Shared-Memory  (C) Shared-Disk
   (D) Shared-CPU  (E) Shared-Everything
   ANSWER: (A)

3. Which of these processors are typically the fastest for deep learning inference?
   (A) Multi-core CPU  (B) FPGA  (C) GPU  (D) TPU  (E) All 4 are similar
   ANSWER: (D)

4. Which of the following is a disadvantage of Parquet when compared to CSVs?
   (A) Column pruning  (B) Immutability  (C) Lossy compression
   (D) Closed source format  (E) Supports only relational data
   ANSWER: (B)

5. Suppose a data-intensive program takes 100min to run on one machine. You use
   a 5-machine cluster to speed it up. Which of the following runtimes represent a
   sub-linear speedup?
   (A) 10min  (B) 20min  (C) 30min  (D) Both A and B  (E) Both B and C
   ANSWER: (C)

6. For an ML classification application, which of these metrics is such that higher val-
   ues are considered better on the Pareto frontier surface?
   (A) Inference Runtime  (B) Training Cost  (C) Accuracy
   (D) Training Throughput  (E) None of the rest
   ANSWER: This question will be waived for all. I meant to include multi-option an-
   swers. Both C and D and correct.
7. Which layer of the public cloud are you programming with for your PAs?
   (A) SaaS  (B) PaaS  (C) IaaS  (D) Serverless  (E) None of the rest
   ANSWER: (C)

8. Which component of a PU is responsible for steering floating point addition?
   (A) ALU  (B) CU  (C) Registers  (D) Cache  (E) None of the rest
   ANSWER: (A)

9. What is the name of the lowest level NVIDIA library that allows a software developer to run a program on a GPU?
   (A) BUDA  (B) CUDA  (C) HUDA  (D) RUDA  (E) None of the rest
   ANSWER: (B)

10. Which of these layers in a typical memory/storage hierarchy is the fastest persistent-by-default layer for data access latency?
    (A) CPU Cache  (B) DRAM  (C) Flash SSD  (D) HDD  (E) Tape
    ANSWER: (C)

Extra Credit Question (Optional). [2pts] You are given a square matrix $A_{n \times n}$ represented as a relation with one tuple per cell in the following schema: $A(row, column, value)$. Write a single SQL query each to compute $\text{trace}(A)$, the sum of its main diagonal cells. No partial credits for this question.

   ANSWER: SELECT SUM(value) FROM A WHERE row = column;