The purpose of this problem set is to give you some practice with concepts related to recovery, parallel query processing, two-phase commit, and other papers we read during the second half of the course.

1 Serializability

1. [6 points]: Below, we provide 3 interleavings of several concurrent transactions (consisting of READ and WRITE statements on objects).

Your job is to indicate whether, for each of the interleavings:

- The interleaving has an equivalent serial ordering. If so, indicate what the serial ordering is.
- Whether the interleaving would be permitted/could arise under strict two-phase locking (where write locks are held until end of transaction).
- Whether the interleaving would be permitted/could arise under rigorous two-phase locking (where read and write locks are held until end of transaction).
- Whether the ordering would be valid using snapshot isolation. If not, indicate which transaction will be aborted.

Assume that snapshot isolation is implemented in the same way as optimistic concurrency control in the paper by Kung and Robinson, except that read sets are not tracked (and read-write / write-read conflicts are ignored.) You can assume that serial validation is used the write phases of optimistic concurrency control happen during the COMMIT statement (and no sooner.)

Assume in all cases that written values can depend on previously read values. (The interleavings are shown on the next page.)
Interleaving 1:
1 T1: READ A
2 T2: READ B
3 T1: WRITE A
4 T2: WRITE B
5 T1: READ B
6 T1: COMMIT
7 T2: COMMIT

Interleaving 2:
1 T1: READ A
2 T2: WRITE B
3 T2: COMMIT
4 T1: WRITE B
5 T1: COMMIT

Interleaving 3:
1 T1: READ B
2 T2: READ A
3 T3: READ A
4 T2: WRITE B
5 T2: COMMIT
6 T3: WRITE A
7 T3: COMMIT
8 T1: WRITE A
9 T1: COMMIT
2  Recovery

2. [1 points]: If your DBMS always STOLE pages (as discussed in class and in the paper by Franklin), how would that affect the design of the database recovery manager?

3. [4 points]: Suppose you are told that the following transactions are run concurrently on a database system that has just been restarted and is fully recovered, and is running Rigorous 2PL.

\[
\begin{array}{cccc}
T1 & T2 & T3 & T4 \\
\-- & -- & -- & -- \\
RX & -- & -- & -- \\
RY & -- & -- & -- \\
RW & -- & -- & -- \\
RZ & -- & -- & -- \\
WX & -- & -- & -- \\
RV & -- & -- & -- \\
WY & -- & -- & -- \\
\text{ABORT} & -- & -- & -- \\
WV & -- & -- & -- \\
WW & -- & -- & -- \\
\end{array}
\]

Suppose the system crashes after executing WW in T3. Assume that each object (e.g., X, Y, etc.) occupies exactly one page of memory.

A. Show all of the records that should be in the log at the time of the crash (given your serial order), assuming that there have been no checkpoints and that you are using an ARIES-style logging and recovery protocol. Your records should include all of the relevant fields described in Section 4.1 of the ARIES paper. Also show the status of the transaction table (as described in Section 4.3 of the ARIES paper) after the analysis phase of recovery has run.

B. Suppose you have 2 pages of memory, and are using a STEAL/NO-FORCE buffer management policy as in ARIES. Given the interleaving you showed above, for each of the 5 pages used in these transactions, show one possible assignment of LSN values for those pages as they are on disk before recovery runs. You should use the value “?” if the LSN is unchanged from the prior state of the page before these transactions began. Finally, indicate which pages will be modified during the UNDO pass, and which will be modified during the REDO pass.
3 Parallel Query Processing

The standard way to execute a query in a shared-nothing parallel database is to horizontally partition all tables across all nodes in the database. Under such a setting, distributed joins can be computed by repartitioning tables over the join attributes.

Suppose these tables are partitioned across a 4 node distributed database, where each node has 1 TB of disk storage and 10 GB of RAM. Also, assume that:

- The disk can read at 100 MB/sec (for the purposes of this problem, you may ignore differences between sequential and random I/O),
- The network on each node can transmit data at 100 MB/sec, regardless of the number or rate at which other nodes are simultaneously transmitting,
- A computer cannot send over the network and read or write from its disk at the same time,
- CPU costs are negligible relative to disk or network I/O time.

For each of the following queries, indicate what partitioning strategy you would recommend. This is the only query run by the system, and your job is to choose the best up-front partitioning for this particular query / setting. Choose between replicating each table on all nodes, or hash, range, or round-robin partitioning it. For hash partitioning, specify the attribute you would use; for range partitioning specify the ranges you would use.

4. [3 points]:

SELECT COUNT(*)
FROM R JOIN S
ON R.a = S.b // S.b is a primary key
WHERE S.c > 100 // selectivity of .1

Here, S.b is a primary key, R is 2 TBs and several billion records, and S is 100 MB and several million records. Indicate the best partitioning / replication strategy (and partitioning attributes) for R and S, as well as a brief explanation.

5. [3 points]:

SELECT COUNT(*)
FROM R JOIN S
ON R.a = S.b // S.b is a primary key
JOIN T on R.c = T.d // T.d is a primary key
WHERE T.e = 1 // T.e is uniformly distribute in [1..100]
AND S.c > 100

Here, S.b and T.d are primary keys, R is 2 TBs and several billion records, and S is 1 TB and several hundred million records, and T is 1 GB and several million records. Indicate the best partitioning / replication strategy (and partitioning attributes) for R, S, and T as well as a brief explanation.
4 Dynamo

You are given a Dynamo instance with nodes A–L, as shown below.

6. [2 points]: Suppose this graph has no virtual nodes, a sloppy quorum, and uses an N/R/W configuration of 4/2/3. If a write for key x arrives and node K is off-line, which nodes could participate in the write?

7. [2 points]:
Now suppose you are using the same Dynamo instance, with an N/R/W configuration of 4/1/4. If a read for key y arrives and no nodes in the system have ever been off-line, which nodes could participate in the read?