CS60002: Distributed Systems
Department of Computer Science & Engineering
Indian Institute of Technology, Kharagpur
Spring 2021

Test - 2, April 12, 2021

Time: 1 Hour 30 Minutes

Instructions (Read carefully before starting)

Write on plain paper (white). Final answer must all be handwritten, nothing should be typed.
At the end of the examination, scan all the pages in order in a **single** pdf. All the pages should be clearly readable in the pdf.
Name the pdf file in the following format: <your roll no>_<first name>.pdf. For example 16CS30001_Arijit.pdf
Mail the pdf file to pranav.23890@gmail.com. (Same as the address where you emailed your first test)
The pdf file must reach the TA within 5 minutes of the end of the examination. So make sure you leave enough time for scanning, mailing etc., and check that you are mailing to the correct address. Anything reaching the TA beyond the 5 minutes will incur (exponentially increasing with delay) penalty. The amount of penalty will be completely decided by me, you cannot change it, so do not submit late.

Answer Q. 1, 3, 4, 5, and any one of Q2(a) or Q2(b)

1. Consider the 2-phase commit protocol. Clearly explain how the coordinator recovers if it has failed (i) before writing the commit record in the log but after receiving AGREED messages from all other processes, and (ii) after writing the commit record in the log, sending the COMMIT messages, and getting back some (but not all) acknowledgements from the other processes. (6)

2. Answer **any one** of Q2(a) or 2(b):

   (a) Consider a cluster of servers running a scientific simulation. The main simulation task first spawns several subtasks that start to run on the different servers. Each subtask typically runs for 15-16 hours independently, and then communicates some intermediate results back to the main task. The main task then again does some computation and then distributes the subtasks on the other servers. This process continues for around 4-5 days before the simulation finishes. There is no communication between the subtasks directly. The time taken by the main task to process the intermediate results from the subtasks and distribute the subtasks is very small, of the order of 5-

10 minutes. The history of operation of the cluster show that the failure rate of a server in the cluster is less than once in one year. If you want to checkpoint this simulation application, what type of checkpointing and recovery scheme will you use and with what parameter values (approximate order is fine)? Briefly justify your answer, do not write an essay. (6)

(b) Modify Koo-Toueg's protocol so that it checkpoints a strongly consistent state instead of a consistent state. Show any new data structure that you may want to add and then describe (i) the set of processes to which an initiator process X will send a checkpoint request, and (ii) the action taken by a process Y when it receives a checkpoint request (Use the ids X and Y for the initiator and receiver). Assume that I know other existing data structures used by the algorithm (by their name in the paper given), you can just refer to them without describing them. (6)

3. Give two reasons why the central master node of a cluster does not become a performance bottleneck in GFS (Write 3-4 sentences max. Do not draw any pictures). (3)

4. (a) In Raft, show an example where all processes are currently at term 4 but there is no entry for term 3 in any process's log. You should show the final log of all the processes (with the above constraint), and clearly list the steps (starting from an empty log for all processes at the beginning) by which the final log shown is reached. (5)

   (b) Show an example of how Jajodia-Muechler's dynamic voting algorithm allows an update to happen in at least one partition even when there is a network partition and no partition has majority (among total number of processes). Your example must not be taken from the paper given. For simplicity, start with 7 processes in one partition and do 5 updates on the data item before any network partition. Then proceed with the example from there. Assume that there is no distinguished site. (5)

5. A distributed directory service stores objects. Each object has a type, a set of attributes, and values of those attributes. For example, an employee's information may be kept in an object of type *user*, with attributes like *first_name, last_name, id, phone* etc. A directory service can store objects of different types. A schema specifies the different types of objects allowed in the directory service, and for each type, the attributes allowed on an object of that type. All objects in the directory service must always be consistent with the schema, and are checked against the schema on creation and modification to make sure they conform to the schema (or the creation/modification is rejected). The directory service can be searched with different filters on the attributes to find any information regarding the company. Objects can also be added to and deleted from the directory. The schema can also be modified dynamically. A schema change can add/delete a type or add/delete attributes to a type, but assume that deletes succeed only if it will not make any existing objects inconsistent with the schema after the deletion (do not worry how that will be checked, just that this is what deletes satisfy).

Consider a large company having offices worldwide in 50 cities spread over all continents, communicating over the internet. All information regarding the company is kept in a replicated distributed directory service using different types of objects. Employees of different cities primarily access information related to the company office in that city (read and update), but may need to access information of other cities once in a while (read and update). The global headquarter situated in one particular city X needs to access information of all cities regularly (read and update). Updates to the information stored happen regularly, though reads are much much more frequent than writes. Schema is changed only by a small number of designated administrators worldwide, that too rarely.

You have to design this replicated directory system for the company. First list the design parameters you will consider, then list the goals that you will try to achieve for the system, then list the choices you make for each parameter with a brief justification. **Your answer must follow this order**, do not just write anything. You do not have to consider any security aspects in your answer. If you need to make any other assumptions, state it/them clearly at the beginning. (15)