SEGR 550 Distributed Computing

Final Exam, Fall 2011

(100 points total)

1) This is a take-home examination. You must send your solutions in a PDF or text file to zhuy@seattleu.edu by the deadline. Late submissions will not be accepted.

2) You are forbidden from consulting any other person by any means of communication.

3) You are allowed to consult the text book, the research papers in the reading list, and the class notes.

4) The final exam has zero-tolerance policy on collaboration and plagiarism (no copy-and-paste from any source). The score for all the involved parties result in zero for collaboration and plagiarism.

5) Keep your solutions and answers short. Limit your solution or answer to THREE SENTENCES or fewer per problem!

6) Cleary specify the problem and sub-problems for each solution in your solution file.
**Part I: Multiple choices (12 points).**

1. The Domain Name Service (DNS):
   (a) finds the IP address corresponding to a host name.
   (b) finds the MAC (e.g., ethernet) address corresponding to an IP address.
   (c) resolves the route to the destination host.
   (d) finds the IP address corresponding to a MAC (e.g., ethernet) address.

2. Event a has a Lamport timestamp of 4. Event b has a Lamport timestamp of 8. What can we tell about events a and b?
   (a) Events a and b are causally related.
   (b) Events a and b are concurrent.
   (c) Event a happened before event b.
   (d) If events a and b are causally related, then event a happened before event b.

3. Which event is concurrent with the vector timestamp (2, 4, 6)?
   (a) (3, 5, 7)
   (b) (1, 3, 5)
   (c) (1, 4, 6)
   (d) (1, 4, 7)

4. A client has a time of 5:05 and a server has a time of 5:25. Using the Berkeley algorithm, the client's clock will be set to:
   (a) 5:15
   (b) 5:20
   (c) 5:25
   (d) 5:30

5. A bully election algorithm:
   (a) picks the first process to respond to an election request.
   (b) relies on majority vote to pick the winning process.
   (c) assigns the role of coordinator to the process holding the token at the time of election.
   (d) picks the process with the largest ID.

6. The portmapper under Sun (ONC) RPC:
   (a) allocates ports to server processes that host RPC functions.
   (b) allocates ports to clients that use RPC.
   (c) maps external port numbers to internal ones to allow Internet access to RPC services.
   (d) allows clients that use RPC to look up the port number of a server that hosts RPC functions.

**Part II: True or False (6 points).**

1. A file is replicated on 10 servers. The following combinations of (read quorum, write quorum) are permitted by the voting algorithm: (1, 10), (2, 9), (3, 7).
2. Update operations on replicated data store can be done using a pull-based or push-based protocol. The push-based approach is more efficient if read-to-update ratio is low; i.e., many updates between reads.

3. Lamport’s happen-before logical clock guarantees that if the event $E_1$ happened before the event $E_2$, then the timestamp of event $E_1$ is less than the timestamp of event $E_2$, but not vice versa.

Part III: Short answers (82 points).

1. Explain the purpose of an IDL in RPC. (4 points)

2. Election Algorithms. (8 points)
   a. Explain the Bully algorithm. (4 points)
   b. Can the Bully algorithm cope with simultaneous elections? Justify your answer (4 points).

3. Consistency & Replications (8 points)
   a. Describe the primary-backup-replication protocol with respect to client writes. (4 points)
   b. Imagine a quorum-based protocol is being used to implement replicated writes to a set of servers. Each time a write is successful, a version number is incremented. This is used by clients to work out which are the most recent copies of files stored at the servers. Consider a system with 10 servers, a read quorum of 6 and a write quorum of 5. Describe how this could lead to a write-write conflict and how this problem can be fixed by adjusting the quorums. (4 points)

4. In this question we are concerned with a diary that is replicated across datastores. There are two users, Helen and Don. They always interact with the closest replica of the diary datastore. While in Auckland, Helen reads her diary and notes down her appointments. While Helen is traveling from Auckland to Melbourne, Don updates Helen’s dairy by changing the time of her Melbourne appointment. Helen subsequently misses her appointment because the Melbourne replica doesn’t reflect Don’s changes. What client-centric consistency model best describes this behavior. Justify your answer. (5 points)

5. Consider the behavior of two machines in a distributed system. Both have clocks that are supposed to tick 1000 times per millisecond. One of them actually does, but the other ticks only 990 times per millisecond. If UTC updates come in once a minute, what is the maximum clock skew that will occur? (5 points)

6. Consider the following three concurrently-executing processes, assume $x$, $y$, $z$ are (distributed) shared variables and initially all 0: (12 points)

   
<table>
<thead>
<tr>
<th>Process P1</th>
<th>Process P2</th>
<th>Process P3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x := 1;$</td>
<td>$y := 2;$</td>
<td>$z := 3;$</td>
</tr>
<tr>
<td>print($y,z$);</td>
<td>print($x,z$);</td>
<td>print($x,y$);</td>
</tr>
</tbody>
</table>

   (a) What is “Causal Consistency” model? (4 points)
(b) For the above example, is it possible to produce: 000000 under causal consistency? (4 points)
(c) Assuming sequentially consistent data store in the above example, identify all possible execution sequences that would produce: 231312, where 23 means that y and z were 2 and 3, respectively, when P1 executes its print; 13 means that x and z were 1 and 3, respectively, when P2 executes its print; and 12 means that x and z were 1 and 2, respectively, when P3 executes its print. You only need to give the total number of the possible execution sequences and justify your answer. (4 points)

7. Is the following data store sequentially consistent? Explain your answer. (4 points)

\[
\begin{array}{c|c|c}
A & W(x)a & W(x)b \\
B & R(x)a & R(x)b \\
C & R(x)b & R(x)a \\
\end{array}
\]

8. In a group of 10 processes using the Ricart and Agrawala’s algorithm for Mutual Exclusion, all the 10 processes simultaneously (in physical time) generate a request to enter the same Critical Section. What is the worst-case number of messages that the algorithm might transmit? (Calculate this number after everyone has exited their critical section.) You can assume that each process generates only the one request for the Critical Section. (4 points)

9. In class we discussed how varying the waiting timeout (in heartbeat and ping/ack protocols) trades off between failure detection time and false positive rate. What parameter in these failure detector protocols would you vary in order to trade off between bandwidth and detection time? Explain briefly why. (4 points)

10. GFS: the Google File System. (8 points)
   a) Explain how a client write operation is performed. (4 points)
   b) How does GFS maintain a consistent view of chunk stores (consistent data across replicas)?
   c) Explain how GFS copes with failures such as Master failures, chunkserver failures and disk failures. (4 points)

11. MapReduce (6 points)
Given a large set of web documents, write the pseudocode for map and reduce functions in order to produce inverted indices.

12. Facebook’s photo storage: Haystack (7 points)
   a) Haystack needs to handle many requests for the long tail photos. Using conventional storage systems, each single photo retrieval incurs multiple disk accesses. In order to overcome this issue, how does Haystack do to reduce the number of disk accesses for retrieval of a photo? (3 points)
   b) How does a storage machine do fast recovery to recover its in-memory mapping/structure after failures? (4 points)

13. Amazon’s Dynamo (7 points)
   a) What consistency protocol is used? Describe the protocol. (3 points)
b) If an AMAZON service uses Dynamo, and the reads:writes ratio for its workload is 10:1, how would you tune this protocol's parameter to optimize the performance. (2 points)

c) What efficient data structure is used for replicas (the nodes responsible for the same range of keys) to synchronize? Briefly describe it. (2 points)