CS425 Fall 2020 – Homework 2
(a.k.a. “Once upon a time in Distributed Hollywood”)

Out: Sep 22, 2020. Due: Oct 6, 2020 (Start of Lecture. 11 am US Central time.)

Topics: Key-value Stores, Time and Ordering (Lectures 9-13)

Instructions:

1. Attempt any 8 out of the 10 problems in this homework (regardless of how many credits you’re taking the course for). If you attempt more, we will grade only the first 8 solutions that appear in your homework (and ignore the rest). Choose wisely!

2. Please hand in solutions that are typed (you may use your favorite word processor. We will not accept handwritten solutions. Figures and equations (if any) may be drawn by hand (and scanned).

3. All students (On-campus and Online/Coursera) – Please submit PDF only! Please submit on Gradescope. [https://www.gradescope.com/]

4. Please start each problem on a fresh page, and type your name at the top of each page.

5. Homworks will be due at the beginning of class on the day of the deadline. No extensions. For DRES students only: once the solutions are posted (typically a few hours after the HW is due), subsequent submissions will get a zero. All non-DRES students must submit by the deadline time+date.

6. Each problem has the same grade value as the others (10 points each).

7. Unless otherwise specified, the only resources you can avail of in your HWs are the provided course materials (slides, textbooks, etc.), and communication with instructor/TA via discussion forum and e-mail.

8. You can discuss lecture concepts and the questions on Piazza and with your friends, but you cannot discuss solutions or ideas on Piazza.

9. For Fall 2020 semester, we are making the following exceptions:
   a. You can discuss up to 4 (out of 8) problems with a group of at most 2 other students who are in this semester’s CS425 class. This is to encourage remote “group work” among students. (We encourage setting up Zoom sessions to discuss.)
   b. At least 4 (out of 8) problems must be on your own.
Prologue: You have just been made the technical head in a production company that is producing a new Hollywood movie. The movie is sure to be a blockbuster, with a lot of well-known actors and actresses hired to star in it. Amazingly many of them know distributed systems! You run into them every day on the set. Here is what ensues.

All characters and their actions used in this homework are meant to make the homework fun! Any resemblance of their actions or opinions to real events, or places, is purely coincidental. Any stories involving real actors or actresses are fictional.

Problems:

1. One of the producers, Leo Bloom, likes Bloom filters. In order to make more money, he decides to make the film a flop. His mind at ease, he uses his spare time to create a Bloom filter uses \( m = 64 \) bits, and 3 hash functions \( h_1, h_2, h_3 \), and where \( h_i(x) = (x^{2i} + x - i) \mod m \) (if you already did the problem with \( h_i(x) = (i^2 x + x - i) \), that is fine). His program then starts inserting continuous integers starting from 3, 6, 9, (multiples of 3).... and so on. Before inserting each integer, his program checks if it is already in the Bloom filter (i.e., is a false positive) – if it is not, then the integer is inserted; if it is a false positive, the program terminates. What integer does the program terminate on? (Give the integer that is the false positive, not the last-inserted integer).

2. The co-director, C. Nolan, likes to deal with time travel, which means he asks a lot of “What if?” questions. He asks you several questions about Cassandra – for each of these, give one advantage and one disadvantage:
   a. What if there were no Bloom filter (but the rest of the system works the same way)?
   b. What if Cassandra did not use Memtable at all, and instead wrote directly into the “latest” SSTable? (but the rest of the system works the same way)
   c. What if SSTables were mutable, i.e., when a Memtable were flushed, instead of creating a new SSTable, it would go and modify the
corresponding values in already-present SSTables? (but the rest of the system works the same way)

3. (For this question you can search resources on the Web.) One of the actors, Mahershala, is consistently a good actor and consistently wins awards. It’s no surprise that he is very interested when you tell him about consistency models. He asks you about the differences between linearizability, sequential consistency, and causal consistency (for key-value stores with get/put operations on keys).
   a. Can you say briefly, and clearly what the differences are between the three?
   b. Give an example (using 2-3 clients writing and reading objects), where, for a particular read, using one of the 3 models above gives a completely different return value. While you can search the Web to clarify differences between the 3 models, you cannot borrow an example from the Web.

4. One of the actors, fresh from a hit role in a TV series, is named Slater. Slater likes Cristian’s algorithm, and gets to calibrating it in the above cluster. He finds that the round-trip time for one round of synchronization messages is 1.23 ms. He would like to find the error in the run, and so he measures some minimum delays. On the client side, he finds that there is a delay of at least 23.4 microseconds for a packet to get from an application to the network interface and a delay of 0.05 ms for the opposite path (network interface to application buffer). The minimum latency due to the routers in the network (each way) is 0.3 ms. On the server side, before he can measure the latency between network interface to the application buffer, he is called out to do his shot. What is the error, given the data just presented?

5. Katherine Johnson was a Black computer mathematician and scientist who worked at NASA in building/writing the first NASA computer programs to do mathematical calculations for the geometrical trajectories of the Apollo moon spacecraft, being one of the first to write computer programs involving space and time. The astrophysicist consultant for your current film, a Dr. N. Tyson, is also a big fan of space and time. He is intrigued by the NTP algorithm. He poses the following question – suppose you are dealing with a network which is symmetric (e.g., an enhanced and controlled ATM network) wherein latencies are identical in reverse directions, i.e., $A \rightarrow B$ message latency and $B \rightarrow A$ message latency (for same message) are known to be identical (except of course you don’t know their absolute values). From the equations, for NTP, can you derive the new error bound? Don’t just show the last portion, please work through the entire derivation just like in the slides, from the beginning to the end.
6. The lead actress, named Viola, also likes to play with time as her role is a part-time law professor. Consequently you chat her up and tell her all about Lamport timestamps. She draws the following timeline for you about interactions between her notorious law students, and challenges you to mark Lamport timestamps on all events. The dots (if any) represent instructions executed at the corresponding process.

7. Can you mark vector timestamps on the previous question?

8. At the movie premiere, you run into another actress from your movie, Angelina. She corners you in the after-party and tells you that she has been mapping all the emails among her 4 eldest kids, and would like to find some causality among their communications. But Angelina is quirky. The algorithm she writes increments Lamport timestamps by +4 instead of +1 (in each place there is a +1 in the original algorithm). Furthermore, there is a bug in the algorithm (because it was written by a dude named Brad) wherein process P2 (only) does not take a max on receiving a message but instead just adds +3 to the received message timestamp. The question is - are there any causality violations? That is, is there at least one pair of events $e_1 \rightarrow e_2$ so that the modified timestamps do not obey the rule $\text{TS}(e_1) < \text{TS}(e_2)$? But you are interested in a specific case only – the figure below. Dots represent local instructions. Can you answer the question?
9. You are amazed to learn that the first traffic light was invented by a Black scientist, Garrett Morgan! Inspired, you are trying to build a traffic light system for all the vehicle traffic on the film sets, and this requires you to build a highly fault-tolerant quorum system. You decide to implement a quorum approach with fixed-size quorums of size $Q$. All quorum sets are chosen randomly. The requirement is that any THREE arbitrary quorum sets must intersect in at least $K$ actors, i.e., there must be $K$ actors in common across all three quorum sets. There are $N$ total actors. For each of the following cases, what should the minimum quorum size be in order to satisfy this requirement?
   a. $N=50$, $K=2$
   b. $K=N/2$, Arbitrary $N$
   c. $K=2$, Arbitrary $N$

10. Alas, Thanos makes a comeback. To defeat the villain for good, you need to solve the immediately previous question for arbitrary $K$, arbitrary $N$, and arbitrary number of intersecting members $M$ (i.e., previous question has $M=3$, but you should solve for arbitrary $M$). You may want to verify that your solution leads to the values you came up with in the previous question.