CS425 Fall 2017 – Homework 2
(a.k.a. “Hollywood Land”)

Out: Sep 26, 2017. Due: Oct 10, 2017 (Start of Lecture. 2 pm US Central time.)

Topics: Key-value Stores, Time and Ordering (Lectures 9-12)

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 hardcopy 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. Online students can email solutions, and MCS-DS students must upload on Coursera. If you’re not online/MCS-DS, and are traveling, please make other arrangements – we don’t accept email submissions.
3. Please start each problem on a fresh sheet (not just page), and type your name at the top of each sheet. Staple all your sheets together.
4. Homeworxs 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.
5. Each problem has the same grade value as the others (10 points each).
6. 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.
7. You can discuss lecture concepts and the questions on Piazza and with your friends, but you cannot discuss solutions or ideas. All work must be 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 either know Computer Science. You run into them every day on the set. Here is what ensues.

All characters used in this homework are meant to make the homework fun. Any resemblance to persons, places, or events, living or dead, past, present, or future, is purely coincidental.
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 4 hash functions \( h_1, h_2, h_3, \) and \( h_4 \), where \( h_i(x) = (i * x^i - 2x) \mod m \). His program then starts inserting continuous integers starting from 2017, 2018, 2019, ..., 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. (For this question you can search resources on the Web.) One of the actors, named Mr. Orlando uses his spare time to design a new Bloom filter-based data structure. He says that instead of using a single Bloom filter \( B \) with 1024 bits and 4 hash functions, his new data structure uses 4 Bloom filters \( B_1, B_2, B_3, B_4 \), each with 256 bits, and each using 1 hash function (different from each other, and different from the above 4 hash functions). When checking for an item, it returns true only if the item is present in all of \( B_1 \) through \( B_4 \). When inserting an item it is inserted into all of \( B_1 \) through \( B_4 \). Which of the above two approaches—original using only \( B \) vs. Orlando’s Bloom filter—gives better false positive rates? Answer this for two cases: (1) when there are typically 5 elements inserted into the data structure, (2) when there are typically 100 elements inserted into the data structure. You can use the Web as a resource to find false positive rates for Bloom filters (but solve the problem yourself!).

3. (For this question you can search resources on the Web.) One of the actresses, named Meryl, is consistently a good actress and consistently wins awards. It’s no surprise that she is very interested when you tell her about consistency models. She 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 at least 2 clients writing and reading objects), where, for a particular read, using each 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. To run the video processing services (it’s a 3D movie after all!), you hire Neytiri. She gets very interested in the NTP algorithm. She tells you first that the NTP algorithm assumes zero clock drift (true! look at the algorithm slides again!). She wants to repurpose NTP for the case where there is zero clock skew initially but non-zero clock drift. The goal is to measure clock drift, defined as \( \frac{(\text{Parent’s clock speed})}{(\text{Child’s clock speed})} - 1 \). (Thus when drift is zero, both clocks have equal speed.) The approach remains similar: send a message from child to parent, measure its send and receive times \((ts1 \text{ and } tr1)\), then send a message in the reverse path, and measure its send and receive times \((ts2 \text{ and } tr2)\). You can assume that at the moment the first message is sent \((ts1)\), the clock skew is zero. You can also assume \(ts1=0\). To make life easier, you are told that the network is symmetric: Message 1 and Message 2 both take the exact same latency (but you don’t know the latency of course). Can you derive the algorithm, and write the equation(s) to derive the clock drift? If it makes life simpler you can assume the Parent’s clock is faster than the Child’s (though your solution should really work in the general case). (Note: Terms “parent” and child are used differently in this modified NTP algorithm than in the NTP algorithm discussed in class.)

5. 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 1.93 ms. He would like to find the error in the run, and so he measures some minimum delays. On the server side, he finds that there is a delay of at least 0.21 ms for a packet to get from an application to the network interface and a delay of 55.6 microseconds for the opposite path (network interface to application buffer). On the client side, before he can measure the delays, he is called out to do his shot, so you never gets around to calculating the minimum delays. What is the error, given the data just presented?

6. The producers hire a wonderful director named Patty J. One day after watching “Wonder Woman” you’re near the water cooler, talking about how much you liked the movie. Patty overhears you and gets into a conversation with you. About key-value stores. She is so intrigued by your work that she quickly looks up the Web, learns up everything about key-value stores, and starts a conversation with you. Her claim is that in order to detect and avoid conflicts for objects in key-value stores (e.g., for conflicting writes) and to implement causal consistency, it is better to use vector clocks in key-value stores, where each of the clients and each of the servers has an entry in the vector. What are two major issues you see with this? How would you address these issues while still using the vectors?
7. The lead actress, named Jennifer L. jokingly tells you that her last name initial L is for Lamport. Consequently you chat her up and tell her all about Lamport timestamps. She looks at the CS425 website, sees the logo on top, and draws the following timeline for you, and challenges you to mark Lamport timestamps on all events. The dots represent instructions executed at the corresponding process.

8. It is wrap-up time for the movie! Unfortunately for you, one of the villainous actors, Agent Smith, played a character that had multiple clones in the movie. That evening, 4 of these clones surround you and challenge you to mark vector timestamps on the timeline in the previous question #7 (or else Mr. Anderson!). Can you do it and escape the Agents Smith, and get back home safe? The dots represent instructions executed at the corresponding process.

9. During pre- and post-production, you’ve having a hard time getting all the actors to attend the sessions to do the dubbing and voice-overs and lots of inconsistency issues across shoots. To solve this, you decide to implement a quorum approach with fixed-size quorums of size Q. The requirement is that any two arbitrary quorum sets must intersect in at least K actors with each other. 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=100, K=1
   b. K=1 (any N)
   c. K=10 (any N)
   d. K=N/3
   e. K=N/2
   f. K=2N/3
   g. K=N

10. At the movie premiere, you run into another actress from your movie, Madonna. 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. She draws a timeline with processes P0, P1, P2, P3. The only exception is that instead of incrementing timestamps by +1 (as in the original algorithm), she would like you increment an event at Pi by +(i+1) each time. Does this ensure correctness of Lamport timestamps, i.e., does it preserve causality? Say clearly why or why not. Would you mark the Madonna timestamps for her in the following figure? The dots represent instructions executed at the corresponding process.