CS425 Fall 2018 – Homework 2

(a.k.a. "Hollywood Land")

Out: Sep 26, 2018. Due: Oct 9, 2018 (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 **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. **MCSDS (online/Coursera) students** – Please submit Word doc, docx, or pdf only! Please submit on Coursera.
4. **On-campus students**: Please submit PDF only! Please submit on Gradescope.
5. Please **start each problem on a fresh page**, and **type your name at the top of each page**.
6. Homeworks 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.**
7. Each problem has the same grade value as the others (10 points each).
8. 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.
9. 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 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.

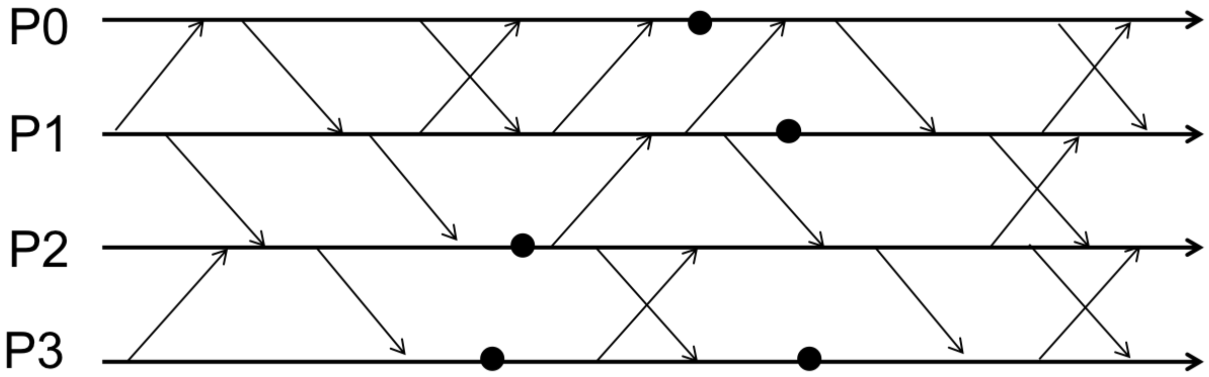
Problems:

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=32 bits, and 4 hash functions h1, h2, h3, and h4, where hi(x) = $(i*x+x^{(i-1)})$ mod m. His program then starts inserting continuous integers starting from 2018, 2019, 2020, …. 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. One of the actors, named Mr. Orlando uses his spare time to design a new Bloom filter-based data structure that uses the same memory but lowers false positive rate. A (regular) Bloom filter's false positive rate is given as $\left(1-e^{-\frac{kn}{m}}\right)^k$ where *k* is the number of hash functions, *n* is the size of the input set and *m* is the size of the Bloom filter in bits. Instead of using a single Bloom filter B with 1024 bits and 4 hash functions, Mr. Orlando's Bloom filter uses L=2^r Bloom filters B1, B2, … BL, each with 1024 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 B1 through BL. When inserting an item it is inserted into ALL of B1 through BL. What are all the values of L (or *r*) for which Orlando's Bloom filter gives a better false positive rate than the original Bloom filter B. Answer this for two cases: (1) when there are typically 10 elements inserted into the datastructure, (2) when there are typically 80 elements inserted into the datastructure. (We recommend though, that you solve the problem with the variables *k, n, m, L,* and then apply these values. But solving with only these two values of 10, 80, would be ok as well.)
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 set up a Hadoop cluster, and Hermione, a magician actress, is in charge of it. To prevent Voldemort from attacking, Hermione wants the Hadoop cluster disconnected from the outside world, that is, no connection to the internet! For synchronization the cluster use the RM server as the central time keeper. All servers run Cristian's algorithm using the RM server as the primary. What problem does this approach suffer from? How would Hermione fix this problem (without requiring a connection to the outside internet)?
5. One of the actors who plays a superhero, is called Chris something-something. You don't quite know his last name---it could be Hemsworth, Evans, Pratt, or something-something. To find out, you hack into Chris' machine, but then get distracted by looking into the logs of the Cristian's algorithm. You find that the round-trip time for one round of synchronization messages is 3.99 ms. You would like to find the error in the run, and so you measure some minimum delays. On the server side, you find that there is a delay of at least 333.3 microseconds for a packet to get from an application to the network interface (NIC) and a delay of 0.33 ms for the opposite path (NIC to application buffer). On the client side, the app to NIC delay is 0.01 ms and the NIC to app delay could not be measured. What is the error, given the data just presented?
6. The lead actress, named Jennifer L. jokingly tells you that her last name initial L is for Lamport. Consequently you chat up with her 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.



7. A consultant on your movie, Radia Perlman (Look her up! She's a famous Computer Scientist!) prefers you solve the problem using vector timestamps. Repeat the previous question by marking vector timestamps on the timeline instead of Lamport timestamps.
8. One of the actresses playing a superhero, named Gal, wants to modify the Lamport algorithm as follows. Each process keeps a local FIFO counter (just like in the FIFO ordering). Let the FIFO counter for an event *e* be *F(e)*. In the Lamport algorithm, instead of incrementing timestamps by +1 (on an instruction/send or in the equation for a receive), Gal would like you increment it instead by (*F(e)*+*1*). Is Gal's algorithm correct? If so, give a formal prove. If no, show a clear counterexample.
9. One of the characters, Thanos, is unhappy, which is not good news, especially since he seems to be wearing some rings on his fingers nowadays around the set and has been saying "Snap!" every once in a while. Anyway, you need 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 always 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 (N large enough where not specified). For each of the following cases, what should the minimum quorum size be in order to satisfy this requirement?
    a. N=50, K=1
    b. K=2 (any N)
    c. K=N/2
    d. K=N/4
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.