CS425 Fall 2019 – Homework 2

(a.k.a. "Once upon a time in Distributed Hollywood")

Out: Sep 25, 2019. Due: Oct 10, 2019 (Start of Lecture. 2 pm 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. 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.**
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. 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.

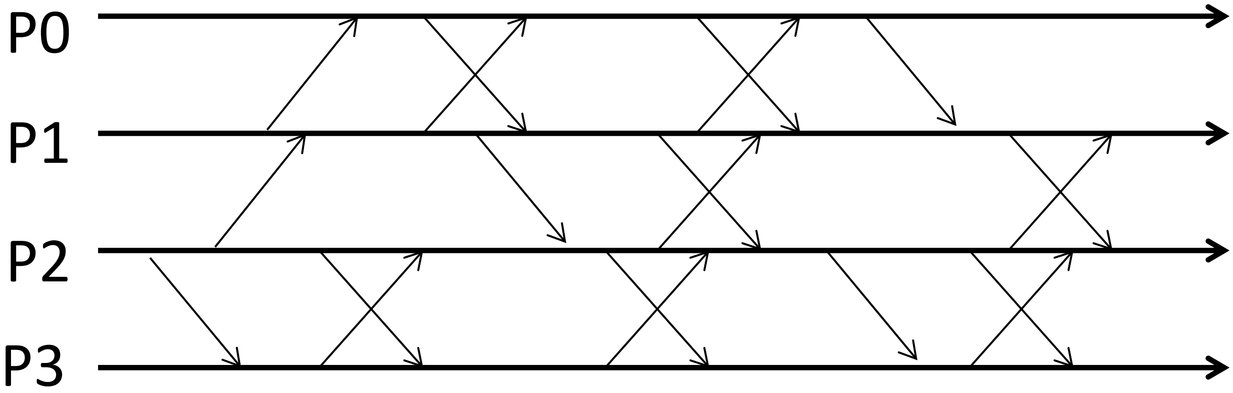
All characters and their actions used in this homework are meant to make the homework fun! Any resemblance of their actions or opinions to real events, or places, is purely coincidental. Any stories involving real actors or actresses are fictional.

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=32 bits, and 4 hash functions h1, h2, h3, and h4, where hi(x) = (x +x^i) mod m. His program then starts inserting continuous integers starting from 2016, 2017, 2018, …. 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. An actor named Orlando uses his spare time to design a new Bloom filter-based data structure. 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. He says that instead of using a single Bloom filter B with 2048 bits and 3 hash functions, his new datastructure uses 2 Bloom filters B1 and B2, each with 1024 bits, and each using 3 hash functions (different from each other, and different from the above 3 hash functions). When checking for an item, it returns true only if the item is present in both B1 and B2. When inserting an item it is inserted into both B1 and B2. Which of the above two approaches—original using only B vs. Orlando's Bloom filter using B1 and B2—gives better false positive rates? Answer this for two cases: (1) when there are typically 4 elements inserted into the datastructure, (2) when there are typically 50 elements inserted into the datastructure.
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 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. To run the video processing services (it's a 3D movie after all!), you set up a Hadoop cluster, and leave one of the characters in the movie in charge of it. The

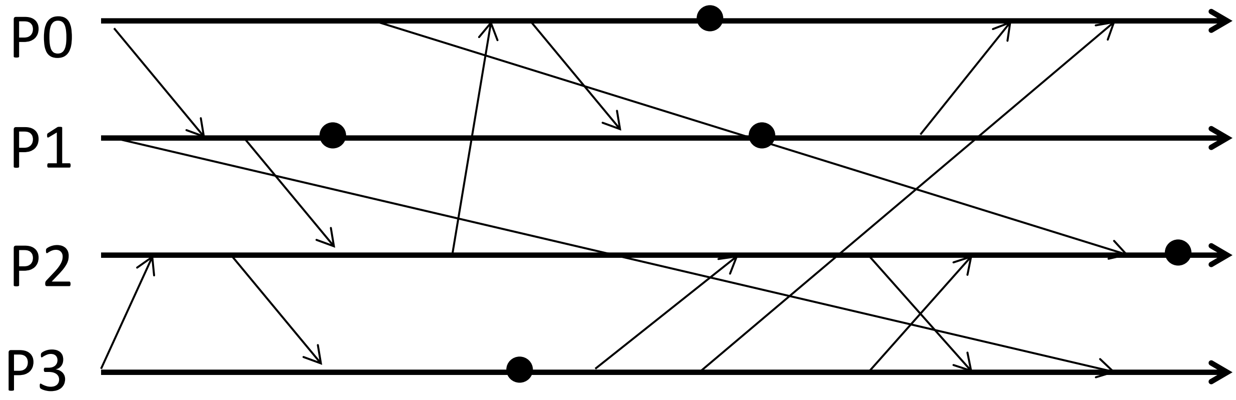
character in charge is named Agent Smith. For reasons of security, Agent Smith wants the Hadoop cluster disconnected from the outside world, that is, no connection to the internet! For synchronization the cluster use the RM server as the central time keeper. All servers run Cristian's algorithm using the RM server as the primary. What problem does this approach suffer from? How would Agent Smith fix this problem (without requiring a connection to the outside internet)?

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 0.66 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 66.6 microseconds for a packet to get from an application to the network interface and a delay of 0.33 ms for the opposite path (network interface to application buffer). On the server side, he only knows that the time to get from the network interface to the application buffer is 20 microseconds, but before he can measure the reverse path, he is called out to do his shot. What is the error, given the data just presented?
6. The lead actress, named Harley Quin jokingly tells you she is related to 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 (if any) represent instructions executed at the corresponding process.



7. It is wrap-up time for the movie! Unfortunately for you, The Joker arrives on set, sees you with Harley Quin, smiles weirdly, and challenges you to mark vector timestamps on the timeline in the previous question #6 (or else!). Can you do it and escape the clones, and get back home safe? The dots represent instructions executed at the corresponding process.

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. Would you mark Lamport timestamps for her? The dots (if any) represent instructions executed at the corresponding process. You can type out the sequence of timestamps for the events at each process (be careful you don't miss any!).



9. At the premiere, you are tasked with babysitting Angelina's kids. You decide to hire someone named Brad, who has had experience with babysitting. So when he has put the kids to sleep, he decides to solve the previous question #8 with vector timestamps. Can you help him?
10. The movie is a hit! The breakout star of the movie, Lois Lane, was so happy with your work that she has asked the production company give you one last puzzle to solve before you can be paid the millions of $ you are owed. The puzzle concerns a modified version of the Lamport timestamp marking algorithm. Instead of incrementing timestamps by +1 each time (as in the original algorithm), you the new algorithm increments it by +K where K is a positive integer.
    a. If K=6 ~~(as in Question #9),~~ would these new timestamps still preserve (obey) causality? Describe why or why not.
    b. If instead, K is selected to be a *random* positive integer (randomly selected for each event), would these new timestamps still preserve (obey) causality? Describe why or why not.
    c. What if K were a negative integer. Describe why or why not.