Homework #1

18-847F: Foundations of Cloud and ML Infrastructure

Prof. Gauri Joshi

Due: Monday, Sept 16, 2019 at 11:59:59pm ET

Instructions

Please remember to show your work for all problems and to write down the names of any students that you collaborate with. The reading material required for the homework is posted on the course website: https://andrew.cmu.edu/course/18-847F/.

Your solutions should be uploaded to Gradescope (https://www.gradescope.com/) in PDF format by the deadline. We will not accept hardcopies. If you choose to hand-write your solutions, please make sure the uploaded copies are legible. Gradescope will ask you to identify which page(s) contain your solutions to which problems, so make sure you leave enough time to finish this before the deadline. We will give you a 30-minute grace period to upload your solutions in case of technical problems.

Problem 1: Complete by Monday Sept 9th, the Presentation skills workshop, 20 pts. The goal of this assignment is to give you an opportunity to practice a new method of slide design that you will use in your full presentation. While you will use a different article for your full presentation, we will practice the type of thinking you will need to do later to tease out any questions you have about slide design and professional research presentations. To prepare for the GCC workshop on Monday, Sept 9, please complete the following tasks:

(a) Watch this 10-minute video distilling a research-backed method of slide design: https://www.youtube.com/watch?v=8FrKLgcxnYA

(b) Read the "Tail at Scale" paper posted on the course website, and imagine that you have to present it in class. What is some critical information that you would want to convey on your slides?

(c) Create 2 slides using the assertion-evidence model of slide design

(d) Submit the slides on Gradescope

(e) Bring your slides to class (we will be using them in the workshop)

Problem 2: (Probability Review, 15 pts)

(a) If a random variable X is distributed uniformly between 2 and 4. What are the mean and variance of X? What is the mean and variance of $Y = X + 2$?

(b) Consider the following joint distribution between X and Y. What is $P(X = T|Y = b)$?

$P(X,Y)$		Y		
		a	b	c
X	T	0.2	0.1	0.2
	F	0.05	0.15	0.3

Problem 3: (Markov Chains, 15 pts) Google's PageRank algorithm ranks a webpage according to the number of links pointing to that page from other webpages. If a webpage is highly ranked then it will result in the webpages that it points to getting a high rank as well. The (simplified) PageRank algorithm uses a Markov chain model to rank pages, where each state corresponds to a webpage. The transition probability to go from state I to state J is proportional to the number of links pointing from webpage I to webpage J. The rank of page I is the steady-state probability π_I of state I of the Markov chain.

Consider an example with three webpages A, B, and C with the following links:

- B has one link to A and one link to C.
- C has one link to B
- A has one link to B and one link to itself

Draw the Markov chain used by the PageRank algorithm and compute the ranks π_A, π_B and π_C of the three webpages. Since these are steady-state probabilities, $\pi_A + \pi_B + \pi_C = 1$.

Problem 4: (Scheduling in Parallel Computing, 15 pts) Consider the three queueing systems shown in Figure 1 below. Job arrivals are Poisson with rate λ and service times are exponential with rate μ.

(a) Write the expressions for their expected response times (waiting time plus service time) in terms of λ, μ and k. We denote the system load by $\rho = \lambda/k\mu$.

(b) Compare the expected response times when the load on the system $\rho = \lambda/k\mu \to 0$ and identify the fastest and slowest system(s).

(c) Compare the expected response times when the load on the system $\rho = \lambda/k\mu \to 1$ and identify the fastest and slowest system(s).

Problem 5: (Grid Computing, 5 pts)

(a) What are the key distinctions between Cloud and Grid computing?

(b) What are the different methods of estimating the job runtime/length in backfilling?

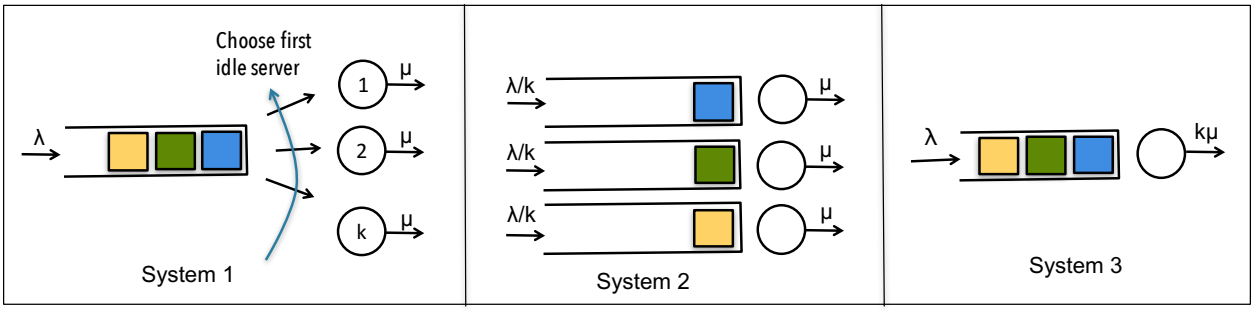

Figure 1: Three Queueing Systems

Give brief answers (1-3 sentences each).

Problem 6: (Tail at Scale, 10 pts)

(a) What are canary requests and how do they reduce unnecessary system load? Give a brief answer (1-3 sentences).

(b) Suppose a server finishes one task in 2 second with probability $p = 0.8$, and 15 seconds with probability $1 - p = 0.2$. What is the expected task execution time? If $m = 10$ tasks are run in parallel, what is the expected task execution time to complete all of them? Please also plot the expected latency versus number of tasks run in parallel ($m = \{1, 2, 3, \ldots, 30\}$).

Problem 7: (Map Reduce, 20 pts) Write a program to count the occurrence of each word in the works of William Shakespeare. You can find the code and the data in **hw1.tar** on Canvas. Specifically,

(a) Implement your own algorithm to count the word frequency

(b) Implement single node MapReduce to count the word frequency

(c) Implement multi-node MapReduce simulated on multiple processes to count the word frequency

Make sure to note the computational time in seconds for each implementation and give comparative analysis. Upload your code (without the data) on **Canvas** and write your comparative analysis here.