Lab #4

Professor Lydia Y. Chen

02121/62121 - Modeling and Scaling of Generative AI Systems

November 2025

Exercise 0: Walking to the limit **(0 Points)**

Consider the Markov chain with transition matrix

$$\boldsymbol{P} = \begin{pmatrix} 1/2 & 1/3 & 1/6 \\ 3/4 & 0 & 1/4 \\ 0 & 1 & 0 \end{pmatrix}$$

1. Draw the corresponding Markov chain, and show that this is irreducible and aperiodic.
2. The process is started in state 1; find the probability that it is in state 3 after two steps.
3. Find the matrix which is the limit of $\boldsymbol{P}^n$ as $n \to \infty$.

Exercise 1: Database Throughput Analysis **(0 Points)**

Bianca observes that her database throughput drops when she runs too many transactions concurrently (this is typically due to thrashing). She also observes that if she runs too few transactions concurrently, her database throughput drops as well (this is often due to insufficient parallelism). To capture these effects, Bianca models her time-sharing database system as an M/M/1/PS queue with load-dependent service rate, $\mu(n)$, where n denotes the number of concurrent transactions. The function $\mu(n)$ is shown in Figure 1.

1. Solve for the mean number of job in the system under Bianca's M/M/1/PS system. Assume arrival rate $\lambda = 0.9$. [Hint: Use a Markov chain.]
2. Bianca has a great idea: Rather than allow all transactions into the database as before, she decides to allow at most 4 transactions to run concurrently in the database, where all remaining transactions are held in a FCFS queue. Bianca's new queueing architecture is shown in Figure 2. Compute the mean response time for Bianca's new

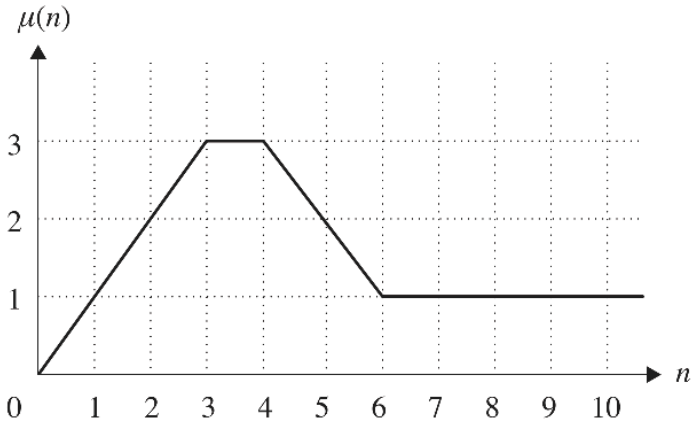

Figure 1: Figure for Exercise 1

architecture, again assuming $\lambda = 0.9$, and Exponentially distributed service times with rates from Figure 2. What is the intuition behind Bianca's hybrid FCFS/PS architecture?

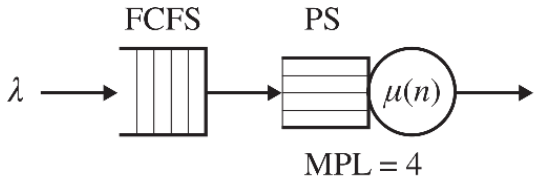

Figure 2: Processor-Sharing with limited multiprogramming level, MPL = 4.

Exercise 2: A Simple Queueing Network **(0 Points)**

Figure 3 shows a simple queueing network. Jobs arrive according to a Poisson process with rate λ. When a job completes service, the job goes back into the queue with probability p and and leaves the system with probability $1-p$. Thus a single job may serve multiple times before leaving the system. Each time the job serves, its service time is a new Exponentially distributed random variable with rate μ.

Your goal is to derive the mean response time for a job in two different ways. A job's response time is the time from when the job first arrives until it finally leaves, including possibly multiple visits to the queue.

1. To start, use the theory of Jackson networks from this chapter to derive an expression for the mean response time in terms of λ, μ and p.

2. Now again derive the mean response time, but this time do it by solving a CTMC that tracks the number of jobs in the system (draw only transitions that *change* the state).

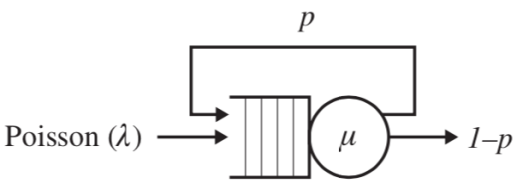

Figure 3: Exercise 2