Lab #5

Professor Lydia Y. Chen

02121/62121 - Modeling and Scaling of Generative AI Systems

November 2025

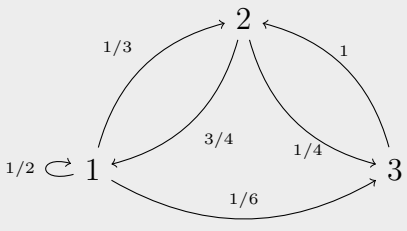
Exercise 0: Walking to the limit **(0 Points)**

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 $n \to \infty$.

Solution 0.

1. 

   A Markov chain is irreducible if it is possible to reach any state from any other state. From state 1, we can access every state; from state 2 we can access state 1 so every state is accessible as well; from state 3 we can access state 2 so it holds. P is irreducible.

   A state has period d if returns to that state can only occur at multiples of d. If d=1, the state is aperiodic. Looking at the graph, we can move between all states in varying steps, hence this chain does not have a fixed period for returns. Therefore, the Markov chain is aperiodic

2. The process is started in state 1, the initialization vector is $\pi_1 = (1, 0, 0)$. We compute the distribution of step 3 with $\pi_3 = \pi_1 \cdot P \cdot P = (1/2, 1/3, 1/6)$ so the probability of reaching state 3 after two steps is 1/6. (Note since we only care about state 3 we could have computed $\pi_{33} = p_{11} * p_{13} + p_{12} * p_{23} + p_{13} * p_{33} = \frac{1}{12} + \frac{1}{12} + 0 = 1/6$)

3. Since P is irreducible and aperiodic, it admits a stationary distribution that is its limit when $n \to \infty$. Thus we have: $\exists \pi^\infty$ s.t. $\pi^\infty \cdot P = \pi^\infty$ and $\|\pi^\infty\|_1 = 1$. Solving this simple system we obtain $\pi^\infty = (1/2, 1/3, 1/6)$. Hence we obtain:

$$\lim_{n \to \infty} \boldsymbol{P}^n = \begin{pmatrix} 1/2 & 1/3 & 1/6 \\ 1/2 & 1/3 & 1/6 \\ 1/2 & 1/3 & 1/6 \end{pmatrix}$$

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.

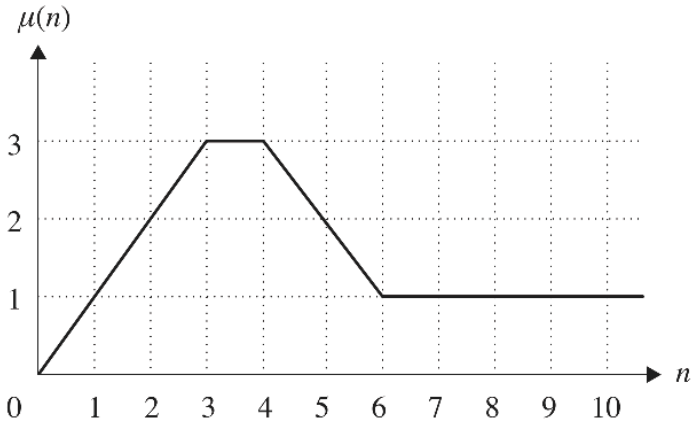


Figure 1: Figure for Exercise 1

1. Solve for the mean response time 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

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?

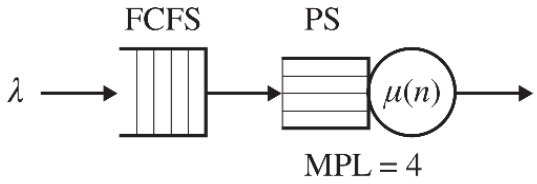
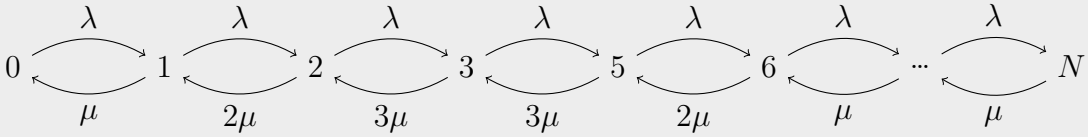


Figure 2: Processor-Sharing with limited multiprogramming level, MPL = 4.

Solution 1.

1. Let us first draw a Markov Chain of the problem



The proposed $M/M/1/PS$ system behaves like an $M/M/1/FCFS$ system. Thus the time-reversibility equations can be defined as.

$$\begin{aligned}
\lambda\pi_0 &= \pi_1 \\
\lambda\pi_1 &= 2\pi_2 \\
\lambda\pi_2 &= 3\pi_3 \\
\lambda\pi_3 &= 3\pi_4 \\
\lambda\pi_4 &= 2\pi_5 \\
\lambda\pi_i &= \pi_{i+1} \qquad \forall i \geq 5 \\
\sum_{i=0}^{\infty} \pi_i &= 1 \\
\Rightarrow \pi_0(\lambda + \frac{1}{2}\lambda^2 + \frac{1}{6}\lambda^3 + \frac{1}{18}\lambda^4 + \frac{1}{36}\sum_{i=5}^{\infty}\lambda^i) &= 1
\end{aligned}$$

Leverage geometric sum to remove summation

$$\pi_0(1+\lambda+\frac{1}{2}\lambda^2+\frac{1}{6}\lambda^3+\frac{1}{18}\lambda^4+\frac{\lambda^5}{(1-\lambda)36}) = 1$$
$$\Rightarrow \pi_0 = (1+\lambda+\frac{1}{2}\lambda^2+\frac{1}{6}\lambda^3+\frac{1}{18}\lambda^4+\frac{\lambda^5}{(1-\lambda)36})^{-1}$$

Plugging in $\lambda = 0.9$, gives us $\pi_0 \approx 0.3807$. Then finding $\mathbf{E}[N] = \sum_{i=0}^{\infty} i\pi_i$

$$\begin{aligned}\mathbf{E}[N] &= \sum_{i=0}^{\infty} i\pi_i \\ &= \pi_0(t\lambda+\frac{2}{2}\lambda^2+\frac{3}{6}\lambda^3+\frac{4}{18}\lambda^4+\frac{1}{36}\sum_{i=5}^{\infty} i\lambda^i) \\ &= \pi_0(\lambda+\frac{2}{2}\lambda^2+\frac{3}{6}\lambda^3+\frac{4}{18}\lambda^4+\frac{(5-4\lambda)\lambda^5}{36(1-\lambda)^2}) \approx 1.72\end{aligned}$$

Finally, leveraging Little's Law $\mathbf{E}[T] = \frac{\mathbf{E}[N]}{\lambda}$

$$\mathbf{E}[T] = \frac{\pi_0(\lambda+\frac{2}{2}\lambda^2+\frac{3}{6}\lambda^3+\frac{4}{18}\lambda^4+\frac{(5-4\lambda)\lambda^5}{36(1-\lambda)^2})}{\lambda} \approx 1.9104$$

2. An $M/M/3$ system can characterize the system's design. **N.B.** that the parallelism is 'only' three, as $\mu(n)$ plateaus after 3 before it starts to decrease.

   Bianca's hybrid FCFS/PS architecture essentially works as follows: up to four jobs are being served at a time: if there are at most three jobs, the total service rate for the i jobs being served is i and therefore each job being served at a rate of 1, if there are 4 jobs, the total service rate is 3 and each job is being served at a rate of 3/4. When there are five or more jobs in the system, all but the first four jobs wait in a FCFS queue. However, note that since service times are exponential, for the purposes of calculating mean response time, it doesn't matter how many jobs are being served at any given time. All that we need to know is that when there are $i \leq 3$ jobs in the system, a departure occurs with rate i, and when there are $i \geq 4$ jobs in the system, a departure occurs with rate 3. Hence, this architecture behaves exactly like an M/M/3 queue. In fact, it would behave in exactly the same way if the PS portion had a MPL of 3.

$$
\begin{aligned}
\mathbf{E}[T] &= \frac{\mathbf{E}[T_q]}{\lambda} + \mathbf{E}[T_s] \\
&= \frac{1}{\lambda} \cdot P_q \cdot \frac{\rho}{1-\rho} + \frac{1}{\mu} \\
P_q &= P(N > 3)
\end{aligned}
$$

Using equation Erlang-C formula, to calculate queueing probability

$$
\begin{aligned}
&= \frac{(k\rho)^k}{k!(1-\rho} \cdot \Big[\frac{(k\rho)^k}{k!(1-\rho)} + \sum_{i=0}^{k-1} \frac{(k\rho)^i}{i!}\Big]^{-1} \\
&= \frac{(3\rho)^3}{3!(1-\rho)} \cdot \Big[\frac{(3\rho)^3}{3!(1-\rho)} + \sum_{i=0}^{2} \frac{(3\rho)^i}{i!}\Big]^{-1} \\
&= \frac{(3\rho)^3}{3!(1-\rho)} \cdot \Big[1 + 3\rho + \frac{(3\rho)^2}{2} + \frac{(3\rho)^3}{3!(1-\rho)}\Big]^{-1} \\
&= \frac{(3\rho)^3}{3!(1-\rho) \cdot \Big[1 + 3\rho + \frac{(3\rho)^2}{2} + \frac{(3\rho)^3}{3!(1-\rho)}\Big]} \\
&= \frac{(3\rho)^3/6}{-9\rho^3 + \frac{3\rho^2}{2} + 2\rho + 1}
\end{aligned}
$$

Plugging in result in the equation above, using $\mu = 1$, $\rho = \frac{\lambda}{3\mu}$, $\lambda = 0.9$

$$
\begin{aligned}
\mathbf{E}[T] &= \frac{\mathbf{E}[T_q]}{\lambda} + \mathbf{E}[T_s] \\
&= \frac{1}{\lambda} \cdot \frac{\rho}{1-\rho} \cdot \frac{(3\rho)^3/6}{-9\rho^3 + \frac{3\rho^2}{2} + 2\rho + 1} + \frac{1}{\mu} \\
&\approx 1.033
\end{aligned}
$$

The intuition behind the system is that we prevent overloading the system (exceeding a parallelism of , resulting in thrashing. By letting the jobs wait in a queue, the DB system once $n > 4$, we ensure that the databases' performance is not degraded by the degree of parallelism.

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.

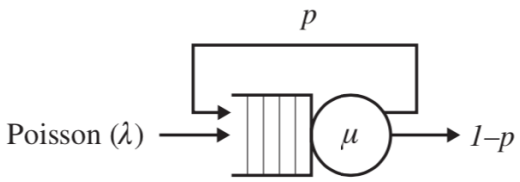


Figure 3: Exercise 2

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).

Solution 2.

1. This is a single queue Jackson network. To use Jackson theory, we need to compute the total arrival rate into the queue. Let $\hat{\lambda}$ denote the total arrival rate into the queue. Since the queue is visited a Geometric number of times with parameter $1p$, we have that:

$$\hat{\lambda} = \frac{\lambda}{1-p}$$

   The queue has service rate μ.

   Although the queue is not an M/M/1, Jackson theory tells us that we can view it as an M/M/1 with respect to the number of jobs.

   Let

$$\rho = \frac{\hat{\lambda}}{\mu} = \frac{\lambda}{\mu(1-p)}$$

   Then

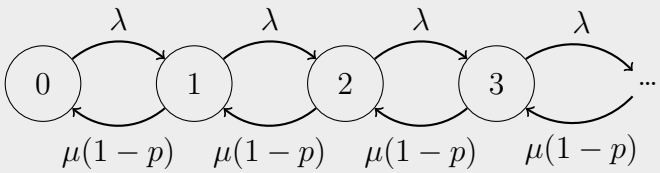
$$\mathbb{E}[N] = \frac{\rho}{1-\rho}$$

   Now, we have to be careful in defining $\mathbb{E}[T]$, via Little's Law, because the *outside* arrival rate to the network is λ, not $\hat{\lambda}$.

So, we have:

$$\begin{aligned}
\mathbb{E}[T] &= \frac{1}{\lambda}\mathbb{E}[N] \\
&= \frac{1}{\lambda}\frac{\rho}{1-\rho} \\
&= \frac{1}{\lambda}\frac{\frac{\lambda}{\mu(1-p)}}{1-\frac{\lambda}{\mu(1-p)}} \\
&= \frac{1}{\lambda}\frac{\frac{\lambda}{\mu(1-p)}}{\frac{\mu(1-p)-\lambda}{\mu(1-p)}} \\
&= \frac{1}{\mu(1-p)}\frac{1}{1-\frac{\lambda}{\mu(1-p)}} \\
&= \frac{1}{\mu(1-p)-\lambda}
\end{aligned}$$

2. We let the state denote the number of jobs in the system. We label those transitions that cause a change in the state. Here is the CTMC:



This is an easy chain to solve.

Let

$$\rho = \frac{\lambda}{\mu(1-p)}$$

$$\pi_i = \mathbb{P}\{N = i\} = \rho^i(1-\rho)$$

$$\mathbb{E}[N] = \frac{\rho}{1-\rho}$$

Now applying Little's law, where the outside arrival rate to the system is λ, we have:

$$\mathbb{E}[T] = \frac{1}{\lambda} \cdot \mathbb{E}[N] = \frac{1}{\mu(1-p)-\lambda}$$