CSE 538 – Spring 2014 Final Exam

6 questions – 110 points total – 100 points max

Harder questions are indicated with a *

Your Name:

1. **[30 points]*** Consider a single server with an infinite queueing facility, *i.e.,* there is no limit to the number of jobs that can be waiting for service. There are two types of jobs submitted to the server, high and low-priority. High priority jobs arrive according to a Poisson process of rate λ_H. Conversely, low priority jobs arrive according to a Poisson process of rate λ_L. Both types of jobs have service times that follow an exponential distribution with mean value $\frac{1}{\mu}$.

   A high priority job that arrives and finds the server busy is always allowed to queue. Conversely, a low priority job that arrives and finds the server busy is queued provided there are fewer than N jobs in the system (1 in service and $N-1$ in the queue). Once queued, low and high-priority jobs are indistinguishable, and queued jobs are served in FCFS order. We assume that both λ_H and λ_L are such that the system is stable.

   Q: Derive an expression for the expected number of *low-priority* jobs in the system.

   Hint 1: Identify a CTMC for the system and solve for the probability distribution π_i that there are i jobs in the system, from which you can derive the average number of jobs in the system $E[N]$.
   Hint 2: What is the average time in the system of either the high or low priority packets, and can you then use Little's Law to to obtain the average number of packets in each class? Note that there are two ways of doing it; a hard way and an easy way.

   The system can be represented through a CTMC as shown in Fig. 1, for which the balance equations are

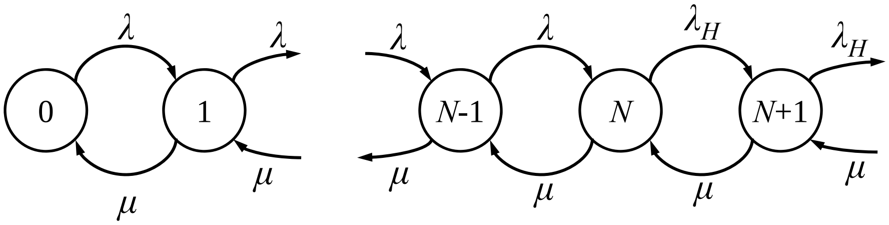

Figure 1: CTMC representation of threshold system.

as follows

$$\begin{aligned}\pi_i &= \rho^i \pi_0, 0 \le i \le N \\ \pi_{N+i} &= \rho_H^i \rho^N \pi_0, i \ge 0\,,\end{aligned}$$

where $\rho = \frac{\lambda_H + \lambda_L}{\mu}$ and $\rho_H = \frac{\lambda_H}{\mu}$. Using the normalization condition then yields the following expression for π_0

$$\pi_0 \left[\sum_{i=0}^{N-1} \rho^i + \rho^N \sum_{j=0}^{\infty} \rho_H^J\right] = 1 \quad \Rightarrow \quad \pi_0 = \left[\frac{1-\rho^N}{1-\rho} + \frac{\rho^N}{1-\rho_H}\right]^{-1}$$

Using the above expressions, we can derive the expected number of jobs in the system $E[N]$.

$$\begin{aligned}E[N] &= \pi_0 \left[\sum_{i=0}^{N-1} i\rho^i + \rho^N \sum_{j=0}^{\infty} (N+j)\rho_H^j\right] = \pi_0 \left[\rho \sum_{i=1}^{N-1} i\rho^{i-1} + N\rho^N \sum_{j=0}^{\infty} \rho_H^j + \rho^N \rho_H \sum_{j=1}^{\infty} j\rho_H^{j-1}\right] \\ &= \pi_0 \left[\frac{\rho\left[(N-1)\rho^N - N\rho^{N-1} + 1\right]}{(1-\rho)^2} + \frac{\rho^N}{1-\rho_H} + \frac{\rho_H \rho^N}{(1-\rho_H)^2}\right],\end{aligned}$$

where π_0 is a given above. Because of PASTA, we know that a high-priority arrival sees on average $E[N]$ jobs in the system, so that the average time in the system $E[T_H]$ of a high-priority job is of the form

$$E[T_H] = \frac{E[N]+1}{\mu} \,,$$

where $E[N]$ is as we just computed. Note that this does not apply to low-priority jobs, as their arrival process is not Poisson. From Little's Law, we then get

$$E[N_H] = \lambda_H E[T_H] = \rho_H (E[N]+1)$$

Since we also have that $N = N_H + N_L$, this implies

$$E[N_L] = E[N] - E[N_H] = E[N](1-\rho_H) - \rho_H \,.$$

2. **[20 points]** Consider a system fed with jobs that arrive according to a Poisson process of rate $\lambda = 1$. Job sizes are exponentially distributed. The system has two processors, which both process jobs at a rate of $\mu = 1$ jobs/sec, *i.e.,* job processing times are exponentially distributed with mean 1 unit of time. Jobs that find both processors busy wait in a queue that for practical purposes can be assumed to have infinite capacity.

   Q: Assuming you want to replace the two original processors by a single faster processor, what should the processing rate of the new processor be to halve the system's response time, *i.e.,* reduce it by a factor 2?

   The original system is essentially an M/M/2 system. The expected time in the system $E[T]$ for an M/M/2 system is of the form (see Eq. (14.9) in the textbook):

$$E[T] = \frac{1}{\lambda} P_Q \frac{\rho}{1-\rho} + \frac{1}{\mu} \,,$$

   where $\rho = \frac{\lambda}{2\mu}$, and P_Q is as given by Eq. (14.5) in the textbook. Specializing these expressions to the case $\lambda = \mu = 1$, *i.e.,* $\rho = \frac{1}{2}$, gives

$$P_Q = \frac{(2\rho)^2}{2(1-\rho)} \cdot \frac{1}{1 + 2\rho + \frac{(2\rho)^2}{2(1-\rho)}} = \frac{2\rho^2}{1+\rho} = \frac{1}{3}$$

   so that we obtain $E[T] = \frac{4}{3}$.

   If we now consider a single processor system with a processor of rate μ', and the same arrival rate $\lambda = 1$, the system behaves like an M/M/1 queue for which we have

$$E[T'] = \frac{1}{\mu' - 1} \,.$$

   Given that we want $E[T'] = \frac{E[T]}{2} = \frac{2}{3}$, this gives

$$\mu' = 2.5 \,.$$

   In other words, the new single processor needs to be 2.5 times faster than an individual processor to halve the response time of the original system.

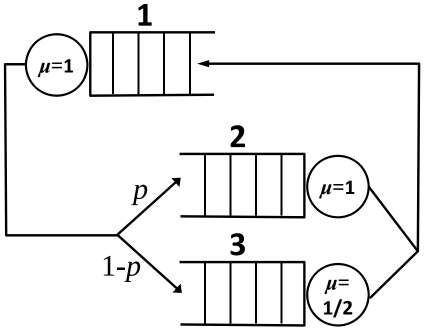

Figure 2: Closed network processing system.

3. **[20 points]** Consider the closed network processing system of Fig. 2, which consists of three servers. Servers 1 and 2 have a unit processing rate, *i.e.,* $\mu_1 = \mu_2 = 1$, while server 3 is twice as slow, *i.e.,* $\mu_3 = \frac{1}{2}$. There are $N = 2$ jobs in the system. All jobs visit server 1, and upon completing their service at server 1 are assigned to either server 2 or server 3 with probabilities p and $1-p$, respectively. All jobs return to server 1 after completing this second service, and the cycle repeats.

   Q: Compute the value of p that minimizes the system response time $E[T]$. Note that the derivation must rely on the *exact* value of $E[T]$ and not bounds.

   Hint: Use MVA to first compute $E[T_1^{(2)}]$, $E[T_2^{(2)}]$, and $E[T_3^{(2)}]$ for $N = 2$, and then $E[T]$. The optimal p value can then be computed from this expression.

   The fractions $p_i, i = 1, 2, 3$, of arrivals that are arrivals to server i are given by

$$p_1 = \frac{1}{2}, p_2 = \frac{p}{2}, p_3 = \frac{1-p}{2}\,.$$

   Applying the MVA result for $N = 1$ gives

$$E[T_1^{(1)}] = 1, E[T_2^{(1)}] = 1, E[T_3^{(1)}] = 2, \quad \text{and} \quad \lambda^{(1)} = \frac{1}{\frac{1}{2} + \frac{p}{2} + 1 - p} = \frac{2}{3-p}\,,$$

   Repeating the process for $N = 2$ gives

$$\begin{aligned}
E[T_1^{(2)}] &= 1 + \frac{\frac{1}{2} \cdot \frac{2}{3-p} \cdot 1}{1} = \frac{4-p}{3-p} \\
E[T_2^{(2)}] &= 1 + \frac{\frac{p}{2} \cdot \frac{2}{3-p} \cdot 1}{1} = \frac{3}{3-p} \\
E[T_3^{(2)}] &= 1 + \frac{\frac{1-p}{2} \cdot \frac{2}{3-p} \cdot 2}{\frac{1}{2}} = \frac{10-6p}{3-p}
\end{aligned}$$

   From the above expressions, we obtain

$$\begin{aligned}
E[T^{(2)}] &= E[T_1^{(2)}] + pE[T_2^{(2)}] + (1-p)E[T_3^{(2)}] \\
&= \frac{4-p}{3-p} + \frac{3p}{3-p} + \frac{(1-p)(10-6p)}{3-p} \\
&= \frac{2(7-7p+3p^2)}{3-p}\,.
\end{aligned}$$

Taking the derivative of $E[T^{(2)}]$ with respect to p yields

$$\frac{\partial E[T^{(2)}]}{\partial p} = 0 \quad \Rightarrow \quad -3p^2 + 18p - 14 = 0$$

The above equation gives $p_{\text{opt}} \approx 0.92$, and correspondingly $E[T^{(2)}_{\text{opt}}] \approx 2.98$.

4. **[20 points]** Consider the feedback processing system of Fig. 3, which consists of two servers in tandem. Server 1 has a unit processing rate, *i.e.*, $\mu_1 = 1$, while server 2 is twice as fast, *i.e.*, $\mu_2 = 2$. Jobs arrive according to an external Poisson process of rate λ. After jobs complete their service at server 1, they are either forwarded to server 2 with probability $1 - p$, or go back to server 1 with probability p. Similarly, after jobs complete their service at server 2, they either leave the system altogether with probability $1 - q$, or go back to server 1 with probability q.

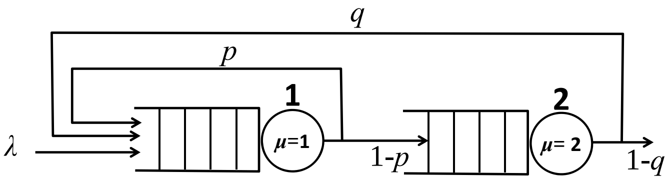

Figure 3: Feedback processing system.

Q: Characterize first **[5 points]** as a function of p and q the maximum feasible arrival rate $\lambda_{\max}$ that keeps the system stable. Assuming $\lambda < \lambda_{\max}$, identify next **[15 points]**, again as a function of p and q, the system response time $E[T]$.

The arrival rates λ_1 and λ_2 to servers 1 and 2 satisfy

$$\begin{aligned}\lambda_1 &= \lambda + p\lambda_1 + q\lambda_2 \Rightarrow (1-p)\lambda_1 = \lambda + q\lambda_2 \\ \lambda_2 &= (1-p)\lambda_1\end{aligned}$$

Combining the first and the second equations above gives

$$\lambda_2 = \frac{\lambda}{1-q} \quad \text{and therefore} \quad \lambda_1 = \frac{\lambda}{(1-p)(1-q)}$$

In order for the system to be stable, we need $\rho_1 = \frac{\lambda_1}{1} < 1$ and $\rho_2 = \frac{\lambda_2}{2} < 1$, which then implies

$$\begin{aligned}\rho_1 < 1 \quad &\Rightarrow \quad \lambda < (1-p)(1-q)\,, \quad \surd \\ \rho_2 < 1 \quad &\Rightarrow \quad \lambda < 2(1-q)\,,\end{aligned}$$

where the first equation represents the tighter constraint.

Since the system is a Jackson network, each server behaves as an independent M/M/1 queue, so that the expected number of jobs in each server is given by

$$\begin{aligned}E[N_1] &= \frac{\rho_1}{1-\rho_1} = \frac{\frac{\lambda}{(1-p)(1-q)}}{1 - \frac{\lambda}{(1-p)(1-q)}} = \frac{\lambda}{(1-q)(1-p) - \lambda} \\ E[N_2] &= \frac{\rho_2}{1-\rho_2} = \frac{\frac{\lambda}{2(1-q)}}{1 - \frac{\lambda}{2(1-q)}} = \frac{\lambda}{2(1-q) - \lambda}\end{aligned}$$

Since $N = N_1 + N_2$, we then have

$$E[N] = \lambda \left[\frac{2(1-q) - \lambda + (1-p)(1-q) - \lambda}{[2(1-q) - \lambda][(1-p)(1-q) - \lambda]} \right],$$

which from Little's Law gives

$$E[T] = \left[\frac{2(1-q) - \lambda + (1-p)(1-q) - \lambda}{[2(1-q) - \lambda][(1-p)(1-q) - \lambda]} \right].$$

5. **[10 points]** Consider an M/M/1 queue with arrival rate λ and service times of average duration $\frac{1}{\mu}$.

   Q: Obtain an expression, function of λ and μ, for the average number of arrivals during a service time.

   Let $\overline{N}$ denote the average number of arrivals during a service time. We have

$$\begin{aligned} \overline{N} &= \int_0^\infty \sum_{i=0}^\infty i \frac{(\lambda t)^i}{i!} e^{-\lambda t} \mu e^{-\mu t} dt = \int_0^\infty \sum_{i=1}^\infty \frac{(\lambda t)^i}{(i-1)!} e^{-\lambda t} \mu e^{-\mu t} dt \\ &= \int_0^\infty \lambda t \left(\sum_{i=1}^\infty \frac{(\lambda t)^{i-1}}{(i-1)!} e^{-\lambda t} \right) \mu e^{-\mu t} dt = \lambda \int_0^\infty t \mu e^{-\mu t} dt = \frac{\lambda}{\mu} = \rho, \end{aligned}$$

   where we have used the fact that $\sum_{i=1}^\infty \frac{(\lambda t)^{i-1}}{(i-1)!} e^{-\lambda t} = \sum_{i=0}^\infty \frac{(\lambda t)^i}{i!} e^{-\lambda t} = 1$.

6. **[10 points]** Consider the same M/M/1 queue as in the previous question, *i.e.,* with and arrival rate λ and exponentially distributed service times of average duration $\frac{1}{\mu}$.

   Q: Obtain an expression for the average duration I of an *idle* period of the queue, *i.e.,* period of time during the server is idle.

   Assume the server becomes idle at time t_0, which is then the start of an idle period. The idle period ends at the time of the next arrival. Because inter-arrival times are exponentially distributed, the average time to the next arrival after t_0 is simply the average inter-arrival time, *i.e.,* $\frac{1}{\lambda}$. Hence, we have $I = \frac{1}{\lambda}$.