CSE 538 – Fall 2016 Final
3 Problems – 75 points total

Your Name:

Problem 1 [30 points] Consider a single unit rate server system with an infinite storage capacity, where jobs arrive according to a Poisson process of rate λ. Jobs sizes follow an exponential distribution with parameter μ. Assume $\frac{\lambda}{\mu} < 1$.

1. **[10 points]** Assume that the system is configured so that jobs of size $< \Delta$ are queued in FCFS order *ahead* of all the jobs of size $\geq \Delta$ that are also queued in FCFS order. In other words, small jobs (of size less than Δ) are served first ahead of jobs of size greater than or equal to Δ. In both categories jobs are served in order of their arrival, and once in service jobs cannot be preempted.

   Provide and expression, function of λ, μ, and Δ for the system's overall average response time $E[T]$.

   This is essentially a non-preemptive priority system, where small jobs, *i.e.,* jobs of size less than Δ, have priority. In order to use the response time expressions for a non-preemptive priority queue, we first need to compute the loads, ρ_S and ρ_B, for small and big jobs, respectively. Note that because the job size distribution is exponential, the expected excess service time is $E[S_e] = \frac{1}{\mu}$. Defining $p_S = P\{S < \Delta\}$ and $P\{S \geq \Delta\}$, we have

$$p_S = \int_0^{\Delta} \mu e^{-\mu x} dx = 1 - e^{-\mu\Delta} \quad \text{and} \quad p_B = \int_{\Delta}^{\infty} \mu e^{-\mu x} dx = e^{-\mu\Delta}$$

   so that we can write

$$\begin{aligned}
\rho &= \frac{\lambda}{\mu} \\
\rho_S &= \lambda p_S \frac{\int_0^{\Delta} x\mu e^{-\mu x} dx}{p_S} = \lambda \left[-\Delta e^{-\mu\Delta} + \frac{1}{\mu} - \frac{e^{-\mu\Delta}}{\mu} \right] = \rho \left(1 - e^{-\mu\Delta}\right) - \lambda\Delta e^{-\mu\Delta} \\
\rho_B &= \lambda p_B \frac{\int_{\Delta}^{\infty} x\mu e^{-\mu x} dx}{p_B} = \lambda \left[\Delta e^{-\mu\Delta} + \frac{e^{-\mu\Delta}}{\mu} \right] = (\lambda\Delta + \rho)\, e^{-\mu\Delta}
\end{aligned}$$

   From the results of Section 31.2 in the book, we then have

$$\begin{aligned}
E[T_S] &= \frac{1}{\mu}\left(1 + \frac{\rho}{1 - \rho_S}\right) \\
E[T_B] &= \frac{1}{\mu}\left(1 + \frac{\rho}{(1 - \rho_S)(1 - \rho)}\right)
\end{aligned}$$

   This in turn gives the following expression for $E[T]$:

$$E[T] = p_S E[T_S] + p_B E[T_B]$$

2. **[5 points]** Assuming $\lambda = 0.5, \mu = 1$, what is the value Δ^* that minimizes the system response time $E[T]$ and how much smaller is $E[T]$ than if all jobs were served in FCFS order, irrespective of their size?
   Note: Either derive an explicit expression for Δ^* and solve for its value when $\lambda = 0.5, \mu = 1$, or write a small numerical search procedure to compute its values (the latter may be faster).

   From the above expression for $E[T]$ and writing a simple numerical procedure evaluating $E[T]$ for increasing values of Δ, we find that $\Delta^* \approx 1.28$ minimizes $E[T]$ when $\lambda = 0.5, \mu = 1$. It gives, $E[T^*] = 1.782$ versus $E[T_{\text{M/M/1}}] = 2$.

3. **[10 points]** Consider now a configuration where we split the single unit rate server into two servers. Jobs of size less than Δ are assigned to the first server, while jobs of size greater than or equal to Δ are sent to the second server. The unit capacity of the original server is split between the two servers so that both have the same load that, therefore, remains equal to $\rho = \frac{\lambda}{\mu}$.

   Provide and expression, function of λ, μ, and Δ for the system's overall average response time $E[\widehat{T}]$ in this two-server configuration.

   The two queues now behave as independent M/G/1 queues. In order to derive the response time of each queue, we therefore need to characterize the first and second moment of the corresponding service time distribution.

   From the results of the first question, we know that

$$\begin{aligned} E[S_S] &= -\Delta e^{-\mu\Delta} + \frac{1}{\mu} - \frac{e^{-\mu\Delta}}{\mu} \\ E[S_B] &= \Delta e^{-\mu\Delta} + \frac{e^{-\mu\Delta}}{\mu} \end{aligned}$$

   Similarly, we readily obtain that

$$\begin{aligned} E[S_S^2] &= -\Delta^2 e^{-\mu\Delta} - \frac{2\Delta}{\mu} e^{-\mu\Delta} - \frac{2}{\mu^2} e^{-\mu\Delta} + \frac{2}{\mu^2} \\ E[S_B^2] &= \Delta^2 e^{-\mu\Delta} + \frac{2\Delta}{\mu} e^{-\mu\Delta} + \frac{2}{\mu^2} e^{-\mu\Delta} \end{aligned}$$

   This implies

$$\begin{aligned} E[\widehat{T}_S] &= E[S_S] + \frac{\rho}{1-\rho} \cdot \frac{E[S_S^2]}{2E[S_S]} \\ E[\widehat{T}_B] &= E[S_B] + \frac{\rho}{1-\rho} \cdot \frac{E[S_B^2]}{2E[S_B]} \end{aligned}$$

   where we have used the fact that both queues see the same load ρ.

   The final expression for $E[\widehat{T}]$ can then be obtained as follows

$$E[\widehat{T}] = p_S E[\widehat{T}_S] + p_B E[\widehat{T}_B]$$

   where p_S and p_B are as derived in question 1.

4. **[5 points]** Assuming as in question 2 that $\lambda = 0.5, \mu = 1$, what is the value $\widehat{\Delta}^*$ that minimizes the system response time $E[\widehat{T}]$, and how does it compare to the value obtained in question 2 for the configuration where small jobs were served ahead of big jobs? Can you explain the difference?
   Note: Again, either derive an explicit expression for $\widehat{\Delta}^*$ and solve for its value when $\lambda = 0.5, \mu = 1$, or write a small numerical search procedure to compute its values.

   From the above expression for $E[\widehat{T}]$ and writing a simple numerical procedure evaluating $E[\widehat{T}]$ for increasing values of Δ, we find that $\widehat{\Delta}^* \approx 0.73$ minimizes $E[\widehat{T}]$ when $\lambda = 0.5, \mu = 1$. It gives, $E[\widehat{T}^*] = 1.962$ versus $E[T] = 1.782$ when using the configuration where small jobs are served ahead of big jobs. The reason for the difference is that while small jobs are still protected from bigger jobs, they are now served by a slower server so that individual service times are increased. Conversely, while big jobs are now "protected" from small jobs, they also see a longer service time due to the slower server. This combines for a lesser benefit than when a single server is used that gives priority to smaller jobs.

Problem 2 [30 points] Consider an M/G/1 system with a unit rate server where jobs arrive according to a Poisson process of rate λ, but where when the system empties out, service resumes only after k jobs have arrived (as opposed to as soon as the first job arrives). Once service resumes, it proceeds as in a regular M/G/1 system until the system is again empty. Assume also that $\lambda \cdot E[S] < 1$. Under those assumptions, show that

1. **[5 points - 1, 2, 2 points]** In steady-state, we have

$$\begin{aligned} P\{\text{system is non-empty and serving}\} &= \rho \\ P\{\text{system is non-empty and waiting}\} &= \frac{(k-1)(1-\rho)}{k} \\ P\{\text{system is empty}\} &= \frac{1-\rho}{k} \end{aligned}$$

The probability that the system is non-empty and serving is the same thing as the probability that the server is busy, or alternatively the expected number of jobs in the server. From Little's Law, we immediately get that

$$P\{\text{system is non-empty and serving}\} = \rho$$

Conversely, this implies that the fraction of time that the server is not busy is still $(1-\rho)$. Those periods of time consist of k inter-arrival times, each of average duration $\frac{1}{\lambda}$. The first (out of k) inter-arrival time corresponds to the period of time during which the system is empty, which gives

$$P\{\text{system is empty}\} = \frac{1-\rho}{k}$$

and similarly, the remaining $(k-1)$ inter-arrival times correspond to the period time when the system is non-empty and waiting, so that

$$P\{\text{system is non-empty and waiting}\} = \frac{(k-1)(1-\rho)}{k}$$

2. **[5 points]** The average length of a "busy" period is given by

$$E[B_{(k)}] = \frac{\rho + k - 1}{\lambda(1-\rho)}$$

where a busy period is defined as a period of time during which the system is not-empty.

Recall from Eq. (27.7) that the expected duration of a period during which the server is busy serving given that it started with an amount of work W is of the form

$$E[B_W] = \frac{E[W]}{1-\rho}$$

In our system the server starts with k jobs, each of average size $E[S]$ so that $E[W] = k \cdot E[S]$. In addition the average duration of a period during which the server is waiting and the system is non-empty (after the first arrival) is equal to $(k-1)$ inter-arrival times, each of average duration $\frac{1}{\lambda}$. This gives

$$\begin{aligned} E[B_{(k)}] &= \frac{k \cdot E[S]}{1-\rho} + \frac{k-1}{\lambda} \\ &= \frac{k\rho}{\lambda(1-\rho)} + \frac{(k-1)(1-\rho)}{\lambda(1-\rho)} \\ &= \frac{\rho + k - 1}{\lambda(1-\rho)} \end{aligned}$$

3. **[15 points]** Assuming that a busy period is divided into busy/waiting and busy/serving periods, show that the average number of jobs in the system during a busy/waiting period is $\frac{k}{2}$ and the average number of jobs in the system during a busy/serving period is

$$E[N_{\text{busy/serving}}] = \frac{E[N_{\text{M/G/1}}]}{\rho} + \frac{k-1}{2}$$

where $E[N_{\text{M/G/1}}]$ is the average number of customers in a regular M/G/1 system.
Hint: Relate the busy/serving period to k independent busy periods of a regular M/G/1 system.

Consider a service discipline that rather than serving jobs in FCFS order, first serves the first job that arrived during the waiting period followed by all the jobs that arrive during the busy period induced by that job. Once all those jobs have been served, the server then moves to serving the second job that arrived during the waiting period and similarly continues serving all the jobs that arrive during its own induced busy period. The process repeats until the k^{th} job that arrived during the waiting period, and the system eventually empties out after all the jobs that arrive during this last induced busy period ends.

Because, this service discipline does not depend on job sizes, it does not affect the number of jobs in the system. In addition, it is easy to see that each one of the induced sub-busy period is identical to the busy period of a regular M/G/1 queue, so that the distribution of the number of packets is the same except for the remaining additional "first" packets that arrived during the busy/waiting period. There are $(k-i)$ such additional first packets in the $i^{\text{th}}, i = 1, 2, \ldots, k$, sub-busy period. This means that

$$E[N|b_i] = E[N_{\text{M/G/1}}|\text{busy}] + k - i$$

where $b_i, i = 1, 2, \ldots, k$, indicates the i^{th} sub-busy period.

In addition, since the duration of all k sub-busy periods have the same distribution (that of the busy period of a regular M/G/1 queue), we have

$$P\{b_i|\text{busy/serving}\} = \frac{1}{k}$$

This allows us to write

$$\begin{aligned}
E[N_{\text{busy/serving}}] &= \sum_{i=1}^{k} E[N|b_i] \cdot P\{b_i|\text{busy/serving}\} \\
&= \sum_{i=1}^{k} \frac{1}{k} \cdot \left(E[N_{\text{M/G/1}}|\text{busy}] + k - i\right) \\
&= \frac{1}{k} \cdot \left(kE[N_{\text{M/G/1}}|\text{busy}] + \frac{k(k-1)}{2}\right) \\
&= \frac{k-1}{2} + E[N_{\text{M/G/1}}|\text{busy}]
\end{aligned}$$

Additionally,

$$\begin{aligned}
E[N_{\text{M/G/1}}] &= E[N_{\text{M/G/1}}|\text{busy}] \cdot \rho \\
\Rightarrow E[N_{\text{M/G/1}}|\text{busy}] &= \frac{E[N_{\text{M/G/1}}]}{\rho}
\end{aligned}$$

This finally allows us to write

$$E[N_{\text{busy/serving}}] = \frac{E[N_{\text{M/G/1}}]}{\rho} + \frac{k-1}{2}$$

Similarly, the expected number of jobs while the system is busy waiting can be written as

$$\begin{aligned} E[N_{\text{busy/waiting}}] &= \sum_{i=1}^{k-1} E[N|\text{busy/waiting with } i] \cdot P\{\text{busy/waiting with } i|\text{busy/waiting}\} \\ &= \frac{1}{k-1}\sum_{i=1}^{k-1} i = \frac{k}{2} \end{aligned}$$

where we have used the fact that all waiting periods with $i, i = 1, 2, \ldots, k-1$, jobs have the same distribution, *i.e.,* an inter-arrival time.

4. **[5 points]** Finally, show that the average number of jobs in the system, $E[N_{(k)}]$ is given by

$$E[N_{(k)}] = E[N_{\text{M/G/1}}] + \frac{k-1}{2}$$

where $E[N_{\text{M/G/1}}]$ is again the average number of customers in a regular M/G/1 system.

We simply write

$$\begin{aligned} E[N_{(k)}] &= E[N_{(k)}|\text{busy/waiting}] \cdot P\{\text{busy/waiting}\} + E[N_{(k)}|\text{busy/serving}] \cdot P\{\text{busy/serving}\} \\ &= \frac{k}{2} \cdot \frac{(k-1)(1-\rho)}{k} + \left(\frac{E[N_{\text{M/G/1}}]}{\rho} + \frac{k-1}{2}\right) \cdot \rho \\ &= E[N_{\text{M/G/1}}] + \frac{k-1}{2} \end{aligned}$$

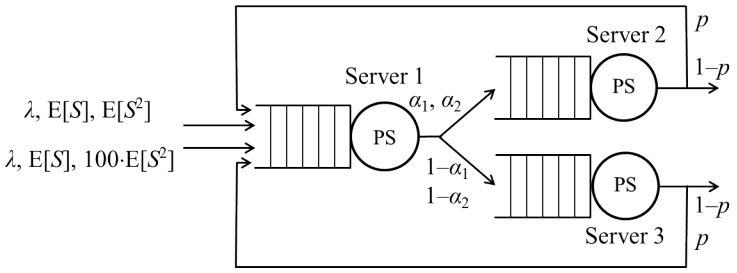

Figure 1: Open network of PS servers with two types of jobs.

Problem 3 [15 points] Consider the open network of Fig. 1 consisting of three unit rate PS servers with infinite storage capacity. Two types of jobs arrive to the network at server 1. Upon leaving server 1, jobs of type $i, i = 1, 2$, can be directed to either server 2 or server 3 with probabilities α_i and $1 - \alpha_i$, respectively, and when leaving either server 2 or 3, jobs of both types depart the system with probability $1 - p$ or return to server 1 with probability p. Both types of jobs arrive according to a Poisson process of the same rate λ, and have general service time distributions. We also know that $E[S_1] = E[S_2] = E[S]$, but that $E[S_1^2] = 100 \cdot E[S_2^2]$.

1. **[5 points - 2, 3 points]** What relationship should λ satisfy to ensure that the system is stable, and what is the arrival rate of type $i, i = 1, 2$, jobs at server 2?

   The total arrival rate λ_{tot} to server 1 is of the form

   $$\lambda_{tot} = (\lambda + \lambda) + p\lambda_{tot} \Rightarrow \lambda_{tot} = \frac{2\lambda}{1-p}$$

   So that stability requires $\frac{2\lambda E[S]}{1-p} < 1$. Conversely, the arrival rate $\lambda_i^{(2)}$ of type $i, i = 1, 2$, jobs at server 2 is equal to

   $$\lambda_i^{(2)} = \frac{\alpha_i \lambda}{1-p}$$

2. **[10 points]** Identify a pair of values (α_1^*, α_2^*) that minimizes the overall system response time $E[T]$. Justify your answer.

   Recall that a network of PS servers behaves as a set of independent M/M/1 queues, independent of the service time distribution except through its mean. Hence, from the point of view of servers 2 and 3, jobs of type 1 and 2 are indistinguishable when it comes to mean response time. This implies that the mean response is minimized when the mean response time at servers 2 and 3 are equal, *e.g.*, $\alpha_1^* = \alpha_2^* = \frac{1}{2}$. Note that there are many other possible combinations that realize the same goal, *e.g.*, $\alpha_1^* = 1, \alpha_2^* = 0$ yields a similar outcome.