CSE 538 – Fall 2015 Final
4 Problems – 80 points total

Your Name:

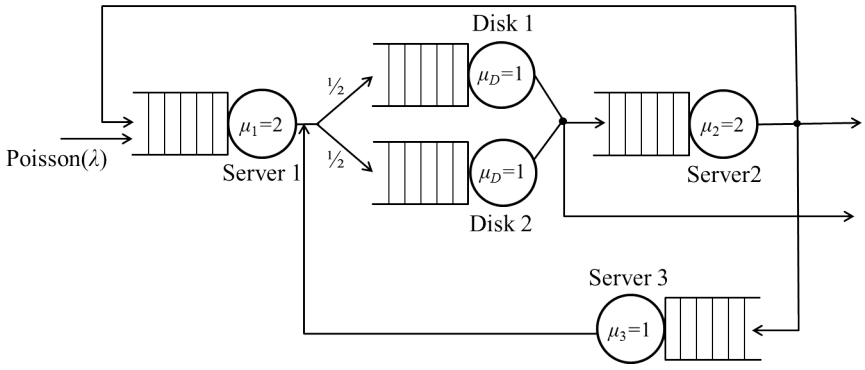

Figure 1: Transactions processing system.

Problem 1 [15 points] Consider the transactions processing system of Fig. 1, which consists of three servers, servers 1 to 3, and a Disk sub-system consisting of two disks, Disks 1 and 2. Transactions arriving at the Disk sub-system are load-balanced equally between the two disks. Transactions enter the system at server 1, but the path they take through the system depends on the transaction type. There are four (4) types of transactions, a to d, that the system handles, each with the following path through the system:

1. T_a: Server 1–Disk–Server 2–Out
2. T_b: Server 1–Disk–Server 2–Server 1-Disk–Out
3. T_c: Server 1–Disk–Server 2–Server 3-Disk–Out
4. T_d: Server 1–Disk–Server 2–Server 3–Disk–Server 2–Out

Transactions arrive to the system according to a Poisson process of rate λ. A new transaction is of type k with probability $p_k = \frac{1}{4}, k = a, b, c, d$. Server i has a service rate Exp$(\mu_i), i = 1, 2, 3$, as shown on the figure, *i.e.,* $\mu_1 = \mu_2 = 2, \mu_3 = 1$ and the two Disks in the Disk sub-system each have a service rate Exp(μ_D), where $\mu_D = 1$.

1. **[5 points]** What is the maximum total transactions arrival rate λ^* below which the system is stable?

   We first need to compute the total arrival rate at each server and at the Disk sub-system. Denote as $\lambda_i, i = 1, 2, 3$, the arrival rate at server i, and as λ_D the arrival rate at the Disk sub-system. Using the paths that each type of transactions takes through the system to determine the number of visits they make to each component, we have

$$\begin{aligned}
\lambda_1 &= \lambda_a + 2\lambda_b + \lambda_c + \lambda_d = \frac{5}{4}\lambda \\
\lambda_D &= \lambda_a + 2(\lambda_b + \lambda_c + \lambda_d) = \frac{7}{4}\lambda \\
\lambda_2 &= \lambda_a + \lambda_b + \lambda_c + 2\lambda_d = \frac{5}{4}\lambda \\
\lambda_3 &= \lambda_c + \lambda_d = \frac{1}{2}\lambda
\end{aligned}$$

   where we have used the fact that $\lambda_a = \lambda_b = \lambda_c = \lambda_d = \frac{\lambda}{4}$.

   Stability then requires

$$\frac{5}{4}\lambda < 2;\ \frac{7}{4}\lambda < 2;\ \frac{5}{4}\lambda < 2;\ \frac{1}{2}\lambda < 1 \tag{1}$$

$$\Rightarrow \lambda < \lambda^* = \frac{8}{7} \tag{2}$$

2. **[5 points]** What is the average response time through the system $E[T_d]$ for transactions of type d, assuming that the total transactions arrival rate is $\lambda = 1$?

We first need to compute the average response time at each one of the three servers and the two Disks of the Disk sub-system. Because they all behave as independent M/M/1 queues, we have:

$$\begin{aligned} E[T_1] &= \frac{1}{2-\frac{5}{4}\lambda} = \frac{4}{8-5\lambda} = \frac{4}{3} \\ E[T_D] &= \frac{1}{1-\frac{7}{8}\lambda} = \frac{8}{8-7\lambda} = 8 \\ E[T_2] &= \frac{1}{2-\frac{5}{4}\lambda} = \frac{4}{8-5\lambda} = \frac{4}{3} \\ E[T_3] &= \frac{1}{1-\frac{1}{2}\lambda} = \frac{2}{2-\lambda} = 2 \end{aligned}$$

The response time of transactions of type d is, therefore, equal to

$$E[T_d] = E[T_1] + E[T_D] + E[T_2] + E[T_3] + E[T_D] + E[T_2] = \frac{4}{3} + 8 + \frac{4}{3} + 2 + 8 + \frac{4}{3} = 22$$

3. **[5 points (3+2)]**

(i) **[3 points]** What is the probability that Server 2 is busy serving a transaction of type d?

We simply apply Little's Law to Server 2. As per the derivation of question 1, the arrival rate of transactions of type d to Server 2 is $2\lambda_d$, and each visit lasts on average $\frac{1}{2}$. Hence, the expected number of transactions of type d at Server 2, and therefore, the probability that Server 2 is busy serving a transaction of type d, is equal to

$$E[N_{S_2}^{(d)}] = P\{\text{Server 2 is busy serving a type } d \text{ transactions}\} = 2\lambda_d \cdot \frac{1}{2} = \lambda_d = \frac{\lambda}{4}$$

(ii) **[2 points]** What is the probability that there are exactly one transaction of type c and one transaction of type d at Server 3?

Because the system is a classed network of queues, we know it has a product form. We are looking at the probability that the state of Server 3 consists of one transaction of type c and one transaction of type d. Applying Theorem 18.1 from the book and the expression derived from it in Section 18.6.2, we get

$$\begin{aligned} P\{\text{one transaction of type } c \text{ and one transaction of type } d \text{ at Server 3}\} &= \binom{2}{1} \frac{\lambda_c \lambda_d}{1^2} \cdot (1-\rho_3) \\ &= 2 \cdot \frac{\lambda^2}{16}\left(1-\frac{\lambda}{2}\right) \\ &= \frac{\lambda^2(2-\lambda)}{16} \end{aligned}$$

Problem 2 [30 points] Consider an M/G/1 system with vacations, *i.e.,* arrivals are Poisson with rate λ, and service times follow a general distribution with p.d.f. $f_S(t), t \geq 0$, and whenever the server becomes idle it goes on "vacation" for a period of duration V, where V follows a general distribution with p.d.f. $f_V(t), t \geq 0$. The one difference with the vacation system we have studied is that the server goes on vacation only once. When it returns from vacation, it either starts serving jobs that are present, if any, or waits until the first arrival to start work again, *i.e.,* it does not go on another vacation if it returns from vacation to an empty system.

1. **[15 points]** Under these assumptions, derive an expression for $E[B_{V_1}]$, the average duration of a busy period, *i.e.,* the server is busy serving.
Hint 1: Express the probability f_0 of 0 arrival in a vacation as a function of the LST $\widetilde{V}(s)$ of $f_V(t)$.
Hint 2: Use this result to first obtain an expression for the duration of an average idle period $E[I]$, and then use this to find $E[B_{V_1}]$ by leveraging the relationship between the system utilization $\rho, E[I]$, and $E[B_{V_1}]$.

   The probability f_0 of 0 arrivals during a vacation is given by

$$f_0 = \int_0^\infty \frac{(\lambda t)^0}{0!} e^{-\lambda t} f_V(t) dt = \widetilde{V}(\lambda)$$

   where $\widetilde{V}(s)$ is the LST of $f_V(t)$.

   The average duration $E[I]$ of an idle period is then of the form

$$E[I] = E[V](1 - f_0) + \left(E[V] + \frac{1}{\lambda}\right) f_0 = E[V] + \frac{\widetilde{V}(\lambda)}{\lambda}$$

   We also know that

$$\begin{aligned} \rho &= \frac{E[B_{V_1}]}{E[B_{V_1}] + E[I]} \\ \Rightarrow E[B_{V_1}](1-\rho) &= \rho \left(E[V] + \frac{\widetilde{V}(\lambda)}{\lambda}\right) \\ E[B_{V_1}] &= \frac{\rho}{1-\rho} \left(E[V] + \frac{\widetilde{V}(\lambda)}{\lambda}\right) \end{aligned}$$

2. **[15 points]** Derive an expression for the average waiting time $E[T_Q^{V_1}]$ of an arrival to the system.
Hint: Follow the same derivation as that on p. 397 of the book, but accounting for the fact that the "unfinished work at the server" now has two components; one when the server is busy, as before, as well as one when the server is not busy. In this latter case, the "unfinished work" is the excess time of the vacation **if** the server is still on vacation, and 0 otherwise. You also need to make sure you correctly "weigh" the different components of the unfinished work, *i.e.,* by their respective probabilities.

   Following the same steps as the derivation on p. 397 of the book, we have

$$E[T_Q^{V_1}] = \rho \cdot E[T_Q^{V_1}] + \rho \cdot E[S_e] + (1-\rho)\frac{E[V]}{E[I]} \cdot E[V_e] \tag{3}$$

   The last term in the above expression is due to the fact that an arrival finds the server idle with probability $(1-\rho)$, and given that it find the server idle, the time average probability that it arrives during a vacation interval is $\frac{E[V]}{E[I]}$, where $E[I]$ is as derived in the previous question. In this latter case, the arrival waits for the excess of a vacation, *i.e.,* $E[V_e]$ on average.

Eq. (3) gives

$$\begin{aligned} E[T_Q^{V_1}] &= \frac{\rho}{1-\rho} \cdot \frac{E[S^2]}{2E[S]} + \frac{1-\rho}{1-\rho} \cdot \frac{E[V]}{E[V] + \frac{\widetilde{V}(\lambda)}{\lambda}} \cdot \frac{E[V^2]}{2E[V]} \\ &= \frac{\lambda E[S^2]}{2(1-\rho)} + \frac{E[V^2]}{2\left(E[V] + \frac{\widetilde{V}(\lambda)}{\lambda}\right)} \end{aligned}$$

Problem 3 [25 points] Consider the two servers system of Fig. 2, where the first server is configured to operate according to a FCFS policy, while the second server operates according to a PS policy. Both servers have the same unit service rate. We will be exploring different job assignment policies to the two servers.

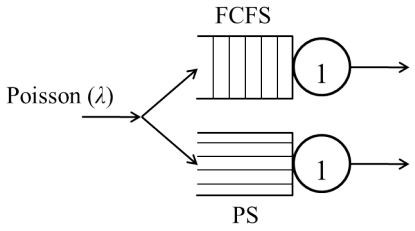

Figure 2: FCFS and PS servers system.

1. **[5 points]** Assume that incoming jobs have a general service time distribution with p.d.f. $f_S(t), t \geq 0$ and first and second moments $E[S] = 1$ and $E[S^2] = 10$, respectively. Our first policy assigns incoming jobs to the FCFS server with probability p and to the PS server with probability $(1-p)$. obtain an expression, function of λ, for the overall average response time $E[T_p]$ of this assignment policy. Explicitly identify constraints that p and λ must satisfy for an assignment to be feasible.

   Both servers see jobs with the original service time distribution but with job arrival rates of $p\lambda$ and $(1-p)\lambda$ for the FCFS and PS servers, respectively. We can, therefore, apply the P-K formula to the FCFS server system, and use the fact that an M/G/1/PS queue behaves just like an M/M/1/FCFS queue to obtain

$$\begin{aligned} E[T_p^{\text{FCFS}}] &= \frac{10\lambda p}{2(1-\lambda p)} + 1 = \frac{10\lambda p + 2(1-\lambda p)}{2(1-\lambda p)} = \frac{1+4\lambda p}{1-\lambda p} \\ E[T_p^{\text{PS}}] &= \frac{1}{1-(1-p)\lambda} \end{aligned}$$

   with the constraints $\lambda < \frac{1}{p}$ and $\lambda < \frac{1}{1-p}$, *i.e.,* $\lambda < \min\left(\frac{1}{p}, \frac{1}{1-p}\right)$ to ensure that both servers are stable. The overall system response time is, therefore, of the form

$$E[T_p] = pE[T_p^{\text{FCFS}}] + (1-p)E[T_p^{\text{PS}}] = \frac{p(1+4\lambda p)}{1-\lambda p} + \frac{1-p}{1-\lambda(1-p)}$$

2. **[5 points]** Assuming that $\lambda = 1$, identify an expression for the value p^* of p that minimizes $E[T_p]$, and use it to numerically compute the value of p^*.

   When $\lambda = 1$, the above expressions simplifies to

$$E[T_p] = \frac{p(1+4p)}{1-p} + \frac{1-p}{p}$$

   Differentiating this expression with respect to p gives

$$\frac{\partial E[T_p]}{\partial p} = \frac{1+8p-4p^2}{(1-p)^2} - \frac{1}{p^2}$$

   The value of p that minimizes $E[T_p]$ can then be found to be $p \approx 0.35$, with $E[T_{p^*}] \approx 3.15$, as shown in Fig. 3.

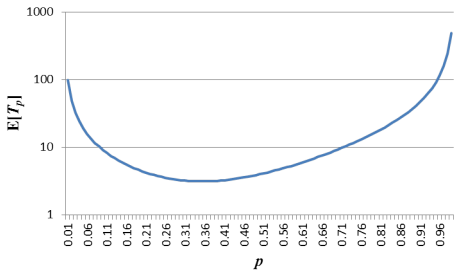

Figure 3: FCFS and PS servers system.

3. **[5 points]** Next, we consider a policy that distinguishes between small jobs and large jobs. Jobs still arrive according to a Poisson process of rate λ (assume that λ is small enough to ensure stability) but now come in two sizes, namely, $S_1 = 1$ and $S_2 = 99$, with 99% of the jobs being small and 1% of the jobs being large. Small and large jobs are sent to a different server.
Should small jobs be sent to the FCFS server or the PS server to minimize their average response time? Rigorously justify your answer. No points will be awarded without quantitative reasoning behind it.

   If small jobs are sent to the FCFS server, their average response time is given by the P-K formula as

$$E[T^{\text{FCFS}}_{\text{small}}] = \frac{0.99\lambda}{2(1-0.99\lambda)} + 1 = \frac{2-0.99\lambda}{2(1-0.99\lambda)}$$

   where we have used the fact that $E[S_1] = 1, E[S_1^2] = 1$, and $\rho_1 = \lambda_1 E[S_1] = 0.99\lambda$.
   In contrast, if the small job are sent to the PS server, their average response time is given by

$$E[T^{\text{PS}}_{\text{small}}] = \frac{1}{1-0.99\lambda}$$

   This gives

$$E[T^{\text{FCFS}}_{\text{small}}] - E[T^{\text{PS}}_{\text{small}}] = \frac{2-0.99\lambda}{2(1-0.99\lambda)} - \frac{1}{1-0.99\lambda} = \frac{-0.99\lambda}{2(1-0.99\lambda)} < 0$$

   Hence, The small jobs are always better off using the FCFS server.

4. **[5 points]** Assume the job size distribution of the previous question, and that small jobs are sent to the FCFS server and large jobs to the PS server. Is the average response time of small jobs under this configuration always smaller than the average response time they would have experienced if both job types had been sent to a single FCFS server that is twice as fast? Assume again that λ is small enough to ensure stability.

   Consider first the case where all jobs are sent to an FCFS server that is twice as fast. The first and second moments of the service time are as follows:

$$\begin{aligned} E[S] &= 0.99 \cdot \frac{1}{2} + 0.01 \cdot 49.5 = 0.99 \\ E[S^2] &= 0.99 \cdot \frac{1}{4} + 0.01 \cdot 49.5^2 = 24.75 \end{aligned}$$

   which also implies that $\rho_{2\times} = 0.99\lambda$.
   Using these expressions in the P-K formula gives us the following average response time for small jobs

$$E[T^{2\times\text{FCFS}}_{\text{small}}] = \frac{24.75\lambda}{2(1-0.99\lambda)} + \frac{1}{2} = \frac{1+23.76\lambda}{2(1-0.99\lambda)} \tag{4}$$

From the previous question, we know that the average response time of small jobs sent to a dedicated, but slower, FCFS server is

$$E[T^{\text{FCFS}}_{\text{small}}] = \frac{2 - 0.99\lambda}{2(1 - 0.99\lambda)} \tag{5}$$

Subtracting Eq. (5) from Eq. (4) gives

$$E[T^{2\times\text{FCFS}}_{\text{small}}] - E[T^{\text{FCFS}}_{\text{small}}] = \frac{1 + 23.76\lambda}{2(1 - 0.99\lambda)} - \frac{2 - 0.99\lambda}{2(1 - 0.99\lambda)} \tag{6}$$

Using $\lambda \approx 0$ in the above equation gives a negative difference ($\approx 0.5 - 1 = -0.5$) in average response times, *i.e.,* a single shared but faster server is better. This is intuitive since when λ is very small, there is never any queueing so that the service time is the only contributor to the response time, and a faster server yields a shorter service time. However, as λ increases, the penalty imposed by having to wait behind large jobs quickly takes over (the cross-over happens at $\lambda = \frac{1}{24.75}$), and we get a positive difference, *i.e.,* a slower server dedicated to the small jobs yields a lower average response time for those jobs.

5. **[5 points]** Repeat the previous question but now comparing sending small jobs to a dedicated FCFS server to a configuration where all jobs, small and large, share a single PS server that is again twice as fast. How does the average response time of small jobs in this latter scenario compare to when they have a dedicated FCFS server?

   Assuming that short and large jobs share a PS server that is twice as fast, and recalling that the average response time of a job of size x in an M/G/1/PS queue is equal to $\frac{x}{1-\rho}$ (see Theo. 30.4), we find that the average response time of small jobs in a shared PS server that is twice as fast is of the form

$$E[T^{2\times\text{PS}}_{\text{small}}] = \frac{1}{2(1 - 0.99 \cdot \lambda)} \tag{7}$$

   where we have used the fact that $E[S^{2\times}_{\text{small}}] = \frac{1}{2}$ and $\rho_{2\times} = \frac{\lambda}{2}(0.99 \cdot 1 + 99 \cdot 0.01) = 0.99\lambda$.
   Subtracting Eq. (5) from Eq. (7) gives

$$E[T^{2\times\text{PS}}_{\text{small}}] - E[T^{\text{FCFS}}_{\text{small}}] = \frac{1}{2(1 - 0.99 \cdot \lambda)} - \frac{2 - 0.99\lambda}{2(1 - 0.99\lambda)} = \frac{0.99\lambda - 1}{2(1 - 0.99\lambda)}$$

   The above difference can be seen to be negative for all values of λ that ensure stability, *i.e.,* $\lambda < \frac{1}{0.99}$. In other words, a shared but faster PS server offers a lower average response time to short jobs than a dedicated but slower server. This is because the PS scheduling policy largely mitigates the impact of having small jobs coexist with large jobs in the same system.

Problem 4 [10 points] Consider a single server system where arrivals are according to a Poisson process of rate λ, and where service times follow a two-phase Coxian distribution with a first phase of average duration $E[S_1] = \frac{1}{\mu}$ and a second phase of average duration $E[S_2] = \frac{10}{\mu}$. Jobs transition from the first phase to the second phase with probability $p = 0.5$. The server uses a processor sharing (PS) service discipline. Assume $\lambda E[S] < 1$.

1. **[5 points]** What are, as a function of μ and λ, the average response times of jobs that only require the first phase of the Coxian service distribution, and of jobs that require both phases of the Coxian distribution?

   Recall again that the average response time of a job of size x in a PS system is $\frac{x}{1-\rho}$, where $\rho = \lambda E[S]$. We note that $E[S] = \frac{1}{\mu} + 0.5 \cdot \frac{10}{\mu} = \frac{6}{\mu}$ so that $\rho = \frac{6\lambda}{\mu}$. This means that the average response times of jobs requiring either only the first or both phases of the Coxian service time distribution are of the form

   $$\begin{aligned} E[T_1] &= \frac{E[S_1]}{1-\rho} = \frac{\frac{1}{\mu}}{1-\frac{6\lambda}{\mu}} = \frac{1}{\mu - 6\lambda} \\ E[T_2] &= \frac{E[S_1 + S_2]}{1-\rho} = \frac{\frac{11}{\mu}}{1-\frac{6\lambda}{\mu}} = \frac{11}{\mu - 6\lambda} \end{aligned}$$

2. **[5 points]** How much faster should the processor be in order to halve the average response time of jobs that only require the first phase of the Coxian distribution? By how much would such a faster processor improve the overall average response time over all job types? Rigorously justify your answer. No points will be awarded without quantitative reasoning behind it.

   Let $\alpha > 1$ denote the processor speed-up factor. We have

   $$\begin{aligned} E[S_1^{(\alpha)}] &= \frac{E[S_1]}{\alpha} = \frac{1}{\alpha\mu} \\ \rho^{(\alpha)} &= \frac{\rho}{\alpha} = \frac{6\lambda}{\alpha\mu} \end{aligned}$$

   and therefore

   $$E[T_1^{(\alpha)}] = \frac{E[S_1^{(\alpha)}]}{1-\rho^{(\alpha)}} = \frac{1}{\alpha\mu - 6\lambda}$$

   We want

   $$\begin{aligned} E[T_1^{(\alpha)}] &= \frac{E[T_1]}{2} \\ \Rightarrow \frac{1}{\alpha\mu - 6\lambda} &= \frac{1}{2(\mu - 6\lambda)} \\ \Rightarrow \alpha &= 2 - \frac{6\lambda}{\mu} = 2 - \rho \end{aligned}$$

   The overall system response time is of the form

   $$E[T] = \frac{E[S]}{1-\rho} \quad \text{and} \quad E[T^{(\alpha)}] = \frac{\frac{E[S]}{\alpha}}{1-\frac{\rho}{\alpha}} = \frac{\frac{E[S]}{2-\rho}}{1-\frac{\rho}{2-\rho}} = \frac{E[S]}{2-\rho-\rho} = \frac{E[S]}{2(1-\rho)}$$

   so that when increasing the server's rate by α, we also improve the overall system response time by

   $$\frac{E[T]}{E[T^{(\alpha)}]} = 2$$

   as expected.