ISYE 7201: Production & Service Systems
Spring 2016
Instructor: Spyros Reveliotis
Final Exam
April 29, 2016

Name: SOLUTIONS

Problem 1 (30 points): Consider a (state-dependent) $M/G/1$ queue where arrivals occur with rate λ, and an arriving customer that finds the system idle has an exponential processing time distribution with rate μ_1 while every other customer has an exponential service time distribution with rate μ. For this queue establish the following results:

i. (10 pts) If this queue is stable, the steady-state probability of being empty is equal to $\pi_0 = (1-\rho)/(1-\rho+\rho_1)$, where $\rho \equiv \lambda/\mu$ and $\rho_1 \equiv \lambda/\mu_1$.

ii. (10 pts) The required stability condition is $\rho < 1.0$.

iii. (10 pts) This queue is actually Markovian, i.e., it can be modeled by a CTMC. Provide the corresponding CTMC, and also use this CTMC to interpret the result of part (ii) above.

(i) The distribution of the proc. times of this queue can be perceived as a mixture of two exponentials. In particular, letting T denote the corresponding r.v., we have:

$$T \sim \begin{cases} \exp(\mu_1) & \text{w.p. } \pi_0 \\ \exp(\mu) & \text{w.p. } 1-\pi_0 \end{cases}$$

where π_0 denotes the steady-state prob. that the queue is empty. Then, $\pi_0 = 1-\hat{\rho}$, where $\hat{\rho}$ is the server utilization, i.e., $\hat{\rho} = \lambda E[T] = \lambda\left[\pi_0 \frac{1}{\mu_1} + (1-\pi_0)\frac{1}{\mu}\right]$. Hence,

$$\pi_0 = 1 - \lambda\left[\pi_0 \frac{1}{\mu_1} + (1-\pi_0)\frac{1}{\mu}\right] \Rightarrow \pi_0\left(1 + \frac{\lambda}{\mu} - \frac{\lambda}{\mu_1}\right) = 1 - \frac{\lambda}{\mu} \Rightarrow$$

$$\Rightarrow \pi_0 = \frac{1-\rho}{1-\rho+\rho_1}$$

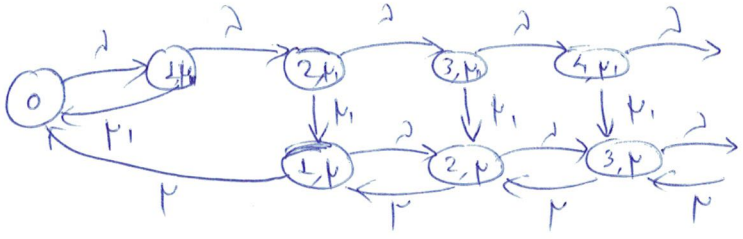
(ii) From the above: $\hat{\rho} = 1-\pi_0 = 1 - \frac{1-\rho}{1-\rho+\rho_1} = \frac{\rho_1}{1-\rho+\rho_1}$.

Then, since, by definition, $\rho_1 > 0$, for a well-defined $\hat{\rho}$ we need $1-\rho+\rho_1 > 0$. Also, for stability: $\hat{\rho} < 1 \Leftrightarrow$

$$\Leftrightarrow \frac{\rho_1}{1-\rho+\rho_1} < 1 \Leftrightarrow \rho_1 < 1-\rho+\rho_1 \quad (\text{since } 1-\rho+\rho_1 > 0) \Leftrightarrow$$

$$\Leftrightarrow 0 < 1-\rho \Leftrightarrow \rho < 1.$$

(iii) It is easy to check that a CTMC modeling the dynamics of the considered queue is as follows:



The component 'μ_1' or 'μ' in the state definition of the above Markov chain denotes the server proc. rate at that state.

Then, from the structure of the above MC, it is clear that, for stability, we essentially need the chain that is defined by the sequence of the states (n, μ) to be stable. But the stability condition for this part is exactly that $\lambda/\mu = \rho < 1$.

As we added during the exam, the routing of calls leaving node A is as follows:
- node B w.p. 0.6
- node C w.p. 0.2
- node A w.p. 0.1

Problem 2 (30 points): Consider a call center that sells tickets for a local baseball team. Upon dialing, a customer first connects to an interactive voice response (IVR) system (node A). This system provides two choices: (1) purchase single-game tickets, or (ii) purchase multi game packages. Also, customers who do not select any option effectively are returned to the beginning of the IVR process, to start over. On the other hand, based on an effective selection, the customer is transferred to either the single-game sales representatives (node B) or to the multi-game sale representatives (node C). After selecting tickets, customers are transferred to another set of representatives (node D) to handle credit card payments. Furthermore, customers routed to node C might change their mind during the transaction process and be re-routed to node B with probability 0.1, while at nodes A, B and C, there is also a probability of abandonment of 0.1.

Regarding the service provided at each station, the IVR system can handle an arbitrarily large number of calls with the call times at this node being exponentially distributed with an average duration of 0.5min. Calls routed to nodes B and C are answered one at a time with respective (instantaneous) rates $\mu_B = 30\ hr^{-1}$ and $\mu_C = 10\ hr^{-1}$. Station D is supported by two operators and the corresponding service rate is $\mu_D = 15\ hr^{-1}$. The call arrival rate to the center is $r = 30\ hr^{-1}$.

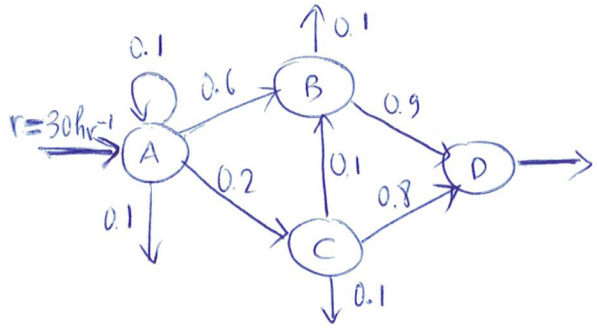
For the above system, apply the results of the open Jackson network in order to compute:

i. (10 pts) the average number of customers in the system;

ii. (10 pts) the average time a customer spends in the system;

iii. (10 pts) the expected number of visits of received each call to each of the nodes of this center.

With the addition provided at the top of this page, the dynamics of this call center can be modelled by the following open queueing network:



The traffic equations for this network are:

$$\begin{cases} \lambda_A = 0.1\lambda_A + r \\ \lambda_B = 0.6\lambda_A + 0.1\lambda_C \\ \lambda_C = 0.2\lambda_A \\ \lambda_D = 0.9\lambda_B + 0.8\lambda_C \end{cases} \Rightarrow \begin{cases} \lambda_A = r/0.9 \\ \lambda_C = 0.2r/0.9 \\ \lambda_B = 0.6r/0.9 + 0.1\times 0.2r/0.9 = 0.62r/0.9 \\ \lambda_D = 0.9\times 0.62r/0.9 + 0.8\times 0.2r/0.9 = 0.718r/0.9 \end{cases}$$

Next, we check the stability conditions for the different nodes of this network when operated with the above arrival rates. For stable nodes, we also compute the expected number of calls in them.

A) From the problem description, it follows that node A is an $M/M/\infty$ queue, and therefore stable for any arrival rate λ_A. Furthermore, from the theory of the $M/M/\infty$ queue, the average number of calls at this station is $L_A = \rho_A = \lambda_A \cdot \tau_A = \frac{30\,hr^{-1}}{0.9} \times \frac{0.5}{60}\,hr \simeq 0.278$

B) Node B is behaving as an $M/M/1$ queue in steady-state. So, for stability, we need $\rho_B = \lambda_B/\mu_B < 1$. We have $\rho_B = \frac{0.62\times 30hr^{-1}}{0.9} \times \frac{1}{30hr^{-1}} = 0.689 < 1$, and from the theory of $M/M/1$ queues, the average number of calls at this node is $L_B = \frac{\rho_B}{1-\rho_B} \simeq \frac{0.689}{1-0.689} \simeq 2.215$

C) Node C corresponds to another $M/M/1$ queue with $\rho_C = \lambda_C/\mu_C = \frac{0.2\times 30\,hr^{-1}}{0.9} \times \frac{1}{10\,hr^{-1}} = 0.667 < 1$ and $L_C = \frac{\rho_C}{1-\rho_C} = \frac{0.667}{1-0.667} \simeq 2.003$

D) Finally, node D is an $M/M/2$ queue and for stability we need $\rho_D = \frac{\lambda_D}{2\mu_D} < 1$. We have $\rho_D = \frac{0.718\times 30hr^{-1}}{0.9} \times \frac{1}{2\times 15hr^{-1}} \simeq 0.798 < 1$

Also, from the formulae for the M/M/c queue with $c = 2$, we get $L_D = \frac{\lambda_D}{\mu_D} + \frac{(\lambda_D/\mu_D)^2 \rho_D}{2!(1-\rho_D)^2} \times P_0$

where $P_0 = \left[\frac{(\lambda_D/\mu_D)^2}{2!(1-\rho_D)} + \sum_{n=0}^{1} \frac{(\lambda_D/\mu_D)^n}{n!} \right]^{-1} =$

$= \left[\frac{(\lambda_D/\mu_D)^2}{2(1-\rho_D)} + 1 + \frac{\lambda_D}{\mu_D} \right]^{-1} =$

$= \left[\frac{(0.718 \times 30\ hr^{-1} / 15\ hr^{-1})^2}{2(1-0.798)} + 1 + \frac{0.718 \times 30\ hr^{-1}}{15\ hr^{-1}} \right]^{-1} =$

$= \left[\frac{1.436^2}{0.404} + 1 + 1.436 \right]^{-1} \simeq 0.1326$

Then $L_D = 1.436 + \frac{1.436^2 \times 0.798}{2(1-0.798)^2} \times 0.1326 \simeq 4.110$

Finally, the expected number of calls in the entire system is:

$L = L_A + L_B + L_C + L_D \simeq 0.278 + 2.215 + 2.003 + 4.110 =$

$= 8.606$

(ii) Let W denote the average time spent by a customer in this system. Then from Little's law: $W = L/r =$

$= 8.606 / 30\ hr^{-1} \simeq 0.287\ hrs = 17.212\ min.$

(iii) The most straightforward manner to answer this question is by noticing that the only possible recirculation at the network occurs locally at node A, while the flow in the rest of the network has no cycles. Then, the local transitional dynamics of node A can be modeled as a sequence of Bernoulli trials with "success" probabilities $p = 1 - 0.1 = 0.9$. Hence, the expected number of trials until success is $V_A = 1/p = 1/0.9 \simeq 1.111$.

For the remaining nodes, B, C and D, the expected number of visits to these nodes, V_B, V_C and V_D, are equal to the corresponding probabilities P_B, P_C and P_D that a customer who leaves node A will visit the corresponding node. Then,

$$V_B = P_B = \frac{0.6}{0.9} + \frac{0.2}{0.9} \times 0.1 = 0.62/0.9 \simeq 0.689$$

$$V_C = P_C = 0.2/0.9 \simeq 0.222$$

The division by 0.9 in the above calculations accounts for the fact that a call that gets service at node A leaves this node only with probability 0.9.

Finally, $V_D = P_D = P_B \times 0.9 + P_C \times 0.8 = \frac{0.62}{0.9} 0.9 + \frac{0.2}{0.9} 0.8 = \frac{0.718}{0.9} \simeq 0.798$

Remark: Notice that the above values for V_i, $i \in \{A, B, C, D\}$ are equal to λ_i / r. This division could have been another line of reasoning for obtaining the values for V_i: Since the effective (external) arrival rate is r, and every node i gets a total arrival rate λ_i, then each node must be visited λ_i / r times.

Problem 3 (20 points): Consider a stable $M/G/1$ queue with two customer classes, A and B, and with respective arrival and processing rates λ_i and μ_i, $i = A, B$. Furthermore, assume that this system experiences a cost of c_i dollars per time unit spent in the system by a customer of class i.

Show that if these two customer classes are to be served according to a preemptive-resume scheme, then, in order to minimize the accrued cost, class A must be given priority over class B if and only if $c_A\mu_A > c_B\mu_B$ (with the case of the equality of these two products being resolved arbitrarily – this is an application of the, so called, $c\mu$-rule of stochastic scheduling theory). What is the intuitive interpretation of this result?

There might be a glitch with the above problem; in particular, the sought result might not hold in the generality that is implied in the problem statement. More specifically, next we shall show that the claimed result applies for the two-class M/M/1 queue with preemptive-resume priorities. It also holds for the two-class M/G/1 queue with non-preemptive priorities. But I have not been able to show it for the two-class M/G/1 queue with preemptive priorities, even for the simpler case of the M/D/1 queue.

In fact, after writing the above, I also managed to come up with a "counter-example" of a two-class D/D/1 queue where the proposed $c\mu$-rule is suboptimal. I provide this counter-example in the following pages, as well, with the understanding that this example does not resolve completely the considered issue (since the presented queue has deterministic and not Poisson arrivals) but it might shed some light on what are the complications that are introduced by preemption.

One way to understand $c\mu$-rule, especially in the context of the subsequent discussion, is by noticing that $c_A\mu_A \geq c_B\mu_B \iff c_A\tau_B \geq c_B\tau_A$.

In the case of the two-class M/M/1 queue with preemptive-resume priorities we can use the formulae for the average number of customers for the two classes that are provided in pg. 157 of your textbook (these formulae can also be derived by the corresponding results that were presented in class). To establish the claimed result for this case, we need to show that:

$$c_A L_A^{(1)} + c_B L_B^{(2)} < c_B L_B^{(1)} + c_A L_A^{(2)} \Rightarrow$$

$$\Rightarrow c_A \frac{\rho_A}{1-\rho_A} + c_B \frac{\rho_B - \rho_A\rho_B + \rho_A\rho_B(\mu_B/\mu_A)}{(1-\rho_A)(1-\rho_A-\rho_B)} < c_B \frac{\rho_B}{1-\rho_B} + c_A \frac{\rho_A - \rho_A\rho_B + \rho_A\rho_B(\mu_A/\mu_B)}{(1-\rho_B)(1-\rho_A-\rho_B)}$$

$$\Rightarrow c_B \left[\frac{\rho_B - \rho_A\rho_B + \rho_A\rho_B(\mu_B/\mu_A)}{(1-\rho_A)(1-\rho_A-\rho_B)} - \frac{\rho_B}{1-\rho_B} \right] < c_A \left[\frac{\rho_A - \rho_A\rho_B + \rho_A\rho_B(\mu_A/\mu_B)}{(1-\rho_B)(1-\rho_A-\rho_B)} - \frac{\rho_A}{1-\rho_A} \right]$$

$$\Rightarrow c_B \frac{[\rho_B - \rho_A\rho_B + \rho_A\rho_B(\mu_B/\mu_A)](1-\rho_B) - \rho_B(1-\rho_A)(1-\rho_A-\rho_B)}{(1-\rho_A)(1-\rho_B)(1-\rho_A-\rho_B)} <$$

$$c_A \frac{[\rho_A - \rho_A\rho_B + \rho_A\rho_B(\mu_A/\mu_B)](1-\rho_A) - \rho_A(1-\rho_B)(1-\rho_A-\rho_B)}{(1-\rho_A)(1-\rho_B)(1-\rho_A-\rho_B)}$$

But we also have:

$$[\rho_B - \rho_A\rho_B + \rho_A\rho_B(\mu_B/\mu_A)](1-\rho_B) - \rho_B(1-\rho_A)(1-\rho_A-\rho_B) =$$

$$= \rho_B - \rho_B^2 - \rho_A\rho_B + \rho_A\rho_B^2 + \rho_A\rho_B(\mu_B/\mu_A) - \rho_A\rho_B^2(\mu_B/\mu_A) - \rho_B + 2\rho_A\rho_B - \rho_B\rho_A + \rho_B^2 + \rho_A\rho_B^2$$

$$= \rho_A\rho_B\left(\mu_B/\mu_A - \rho_B\mu_B/\mu_A + 1 - \rho_A\right) = \rho_A\rho_B \frac{\mu_B - \lambda_B + \mu_A - \lambda_A}{\mu_A}$$

Then, by symmetry, the above inequality becomes:

$$\frac{c_B}{\mu_A}\rho_A\rho_B(\mu_B - \lambda_B + \mu_A - \lambda_A) < \frac{c_A}{\mu_B}\rho_A\rho_B(\mu_B - \lambda_B + \mu_A - \lambda_A) \Rightarrow$$

$$\Rightarrow c_B\mu_B < c_A\mu_A.$$

Also, for the case of the two-class M/G/1 queue with non-preemptive priority we have:

$$L^{(1)} = W_q^{(1)} \lambda_1 + \rho_1 \quad ; \quad L^{(2)} = W_q^{(2)} \lambda_2 + \rho_2$$

and the resulting total cost rate is:

$$C = c_1 W_q^{(1)} \lambda_1 + c_1 \rho_1 + c_2 W_q^{(2)} \lambda_2 + c_2 \rho_2$$

Since $c_1 \rho_1 + c_2 \rho_2$ does not depend on the set priorities, we just focus on the remaining part of the above expression. Then, using the results for the waiting times in the two customer classes that were derived for this queue (see also pg. 154 in your textbook), we need to show that:

$$c_A \lambda_A \frac{\sum_{j=1}^{2} \lambda_j E[S_j^2]}{2(1-\rho_A)} + c_B \lambda_B \frac{\sum_{j=1}^{2} \lambda_j E[S_j^2]}{2(1-\rho_A)(1-\rho_A-\rho_B)} <$$

$$< c_B \lambda_B \frac{\sum_{j=1}^{2} \lambda_j E[S_j^2]}{2(1-\rho_B)} + c_A \lambda_A \frac{\sum_{j=1}^{2} \lambda_j E[S_j^2]}{2(1-\rho_B)(1-\rho_A-\rho_B)} \Leftrightarrow$$

$$\Leftrightarrow c_A \lambda_A \left[(1-\rho_B)(1-\rho_A-\rho_B) - (1-\rho_A)\right] < c_B \lambda_B \left[(1-\rho_A)(1-\rho_A-\rho_B) - (1-\rho_B)\right]$$

In this case, $(1-\rho_B)(1-\rho_A-\rho_B) - (1-\rho_A) = (1-\rho_B)(1-\rho_A) - \rho_B(1-\rho_B) -$

$$-(1-\rho_A) = -(1-\rho_A)\rho_B - \rho_B(1-\rho_B) = -\rho_B(2-\rho_A-\rho_B)$$

And, again, using symmetry, we get:

$$-c_A \lambda_A \rho_B (2-\rho_A-\rho_B) < -c_B \lambda_B \rho_A (2-\rho_A-\rho_B) \Rightarrow$$

$$\Rightarrow c_A \frac{\lambda_A \lambda_B}{\mu_B} > c_B \frac{\lambda_B \lambda_A}{\mu_A} \Rightarrow c_A \mu_A > c_B \mu_B$$

A "counterexample" regarding the suboptimality of the cμ-rule for two-class M/G/1 queues operated with preempt-resume priorities

Consider an M/D/1 queue serving two classes A and B with service times $\tau_A = 1$ and $\tau_B = 2$. The corresponding cost rates are $c_A = 2$ and $c_B = 3$. Finally the deterministic inter-arrival times are $t_A = t_B = 3$.

Notice that in this example $c_A \mu_A = 2 \cdot 1 > 3/2 = c_B \mu_B$.

And assuming that there is an one-time-unit lag between the arrivals of type B and type A jobs, under the preempt-resume priority set by the above rule, we shall have the following event sequence for the execution of this queue:

time ∅: Arrival of a B instance; immediate service initiation

time 1: " " " A "; B is preempted and A initiates service

time 2: Completion of A; B resumes service

time 3: Completion of B; initiation of a new cycle with the arrival of a new B instance and initiation of its service.

The cost of the above cycle is $3 \cdot c_B + 1 \cdot c_A = 3 \cdot 3 + 1 \cdot 2 = 11$.

But it can be easily checked that if A does not preempt B, still both jobs will have completed by time 3, but in this case the cycle cost will be: $2 \cdot c_B + 2 \cdot c_A = 2 \cdot 3 + 2 \cdot 2 = 10 < 11$.

The key idea behind the above example is that even if the cμ rule gives higher priority to a certain job class, a preemption of a job of the other class that is quite advanced in its processing cycle might be disadvantageous. On the other hand, this remark is not an issue in the non-preemptive case, or even in the case of preempt-resume but with memoryless processing times.

Problem 4 (20 points): Consider an $M/D/1$ queue with service time equal to b time units. Suppose that the system size is measured when time is a multiple of b, and let X_n denote the system size at time $t = n \cdot b$. Show that the stochastic process $\{X_n, n = 0, 1, 2, \ldots\}$ is a Markov chain, and find its transition matrix. Please, define carefully all the notation used in this characterization.

Let $X_n = k$ for some $k \in \{0, 1, 2, \ldots\}$

Since the sampling interval is equal to the deterministic proc. time of this queue, over the next sampling interval there will be exactly one service completion, provided that $k > 0$. Otherwise, the service completion will be zero. Hence,

$X_{n+1} = [k-1]^+ +$ arrivals during the interval $[nb, (n+1)b]$

Since the arrival process is Poisson, the aforementioned arrivals are determined independently from the history $\{X_\ell, \ell = 0, \ldots, n\}$, and X_n is a DTMC.

The transition probability matrix for this MC can be easily deduced from the above equation and the Poisson distribution that characterizes the arrivals over the considered interval.

Remark: Notice that the sampling times $t = n \cdot b$ are not necessarily departure times for the underlying $M/D/1$ queue. This is due to the randomness in the arrival process.