ISYE 7201: Production & Service Systems
Spring 2019
Instructor: Spyros Reveliotis
Final Exam (Take Home)
Release Date: April 19, 2019
Due Date: April 28, 2019

While taking this exam, you are expected to observe the Georgia Tech Honor Code. In particular, no collaboration or other interaction among yourselves is allowed while taking the exam.

You can send me your responses as a pdf file attached to an email. This pdf file can be a scan of a hand-written document, but, please, write your answers very clearly and thoroughly. Also, report any external sources (other than your textbook) that your referred to while preparing the solutions.

Finally, the homeworks and past exams that are posted at the course website can be provide useful complementary material (and experience) to the in-class lectures while working on the exam problems.

Problem 1 (20 points): Consider a small call center that provides two types of services, A and B. All calls for each type of service are answered by a single employee. Customers calls occur according to a Poisson process with rate $\lambda = 20\ hr^{-1}$, and each arriving call is directed to service A or B with corresponding probabilities 0.6 and 0.4. A customer who received service of type A might be directed to service B at the end of that first transaction with probability 0.2; otherwise she completes her call. Similarly, customers who received type-B service are routed to service A with probability 0.1, and they complete their calls otherwise. Also, a customer who has received both types of service just completes her call upon completion of the second service. On, the other hand, the employee who supports the call for each of the two service types, answers these calls on a First-Come-First-Serve basis, according to the order with which these calls appeared on his console. Finally, calls concerning type A service are exponentially distributed with a mean duration of 4 minutes, and calls concerning type B service are exponentially distributed with a mean duration of 5 minutes.

Answer the following questions:

i. Model the operation of this call center as a queueing network. Try to be as detailed as possible regarding the features of your model.

ii. What is the average wait for each service?

iii. What is the average number of calls waiting for each type of service?

iv. What is the utilization of each of the two employees?

v. What is the expected duration of a call that receives (a) only type A service, (b) only type B service, (c) both types of services?

Problem 2 (20 points): A local plant possesses two power generators that are desired to be operational all the time. However, each of these generators experiences outages according to an exponential distribution with mean time to failure equal to 2 weeks. When one of these generators breaks down, there is a probability p=0.7 that the damage can be repaired locally. The plant avails of two technicians that can take care of such a local repair, and the corresponding repair time is exponentially distributed with mean value $\tau_l = 1$ week. If the damage cannot be fixed locally, the generator must be sent out for service, to a contracted external expert, and the corresponding service times for this expert are exponentially distributed with mean value $\tau_e = 3$ weeks. Furthermore, there is a probability $q = 0.2$ that the local technicians might not be able to address effectively the experienced damage, in which case the generator eventually must be sent out for external service. A generator that is sent out for external service is always restored successfully back to its operational mode.

Answer the following questions:

i. Model the operation described above as a closed queueing network.

ii. Use mean value analysis in order to determine the percentage of time that (a) both generators are operational, or (b) at least one generator is operational, in steady-state.

Problem 3 (20 points): Consider a non-preemptive priority queue with two customer classes operating in steady-state, and suppose that service times for both classes are exponentially distributed with the same rate μ. Show that the average wait time in the queue across all customers has the form of the average wait time of an M/M/1 queue. Characterize the parameters of this M/M/1 queue.

Problem 4 (20 points): In class we defined the notion of the "busy period" for a stable M/G/1 queue as a maximal time interval T_B during which the queue is non-empty, and we showed that the expected duration of such a busy period is equal to $\tau/(1-\lambda\tau)$, where τ is the mean service time and λ is the rate of the arrival process. In this problem you are called to establish the following:

i. Show that the number of customers that are served during a busy period, N, is geometrically distributed, and define the parameter of this distribution.

ii. Use the results in item #1 above in order to compute $E[N]$ and $Var[N]$. Also, show that your results for this part are in agreement with the aforementioned result that was derived in class for $E[T_B]$.

iii. Use the results in items #1 and #2 above in order to compute $Var[T_B]$.

iv. Briefly describe how the above results could be used in order to support an alternative approach for the mean value analysis of a two-class priority queue with preemptive priorities, similar to that discussed in class.

Problem 5 (20 points): Consider a workstation with 2 servers and a total buffering capacity (including the buffering capacity of the servers) of 4 parts. Processing times at the two servers are exponentially distributed with corresponding rates μ_1 and μ_2 parts per hour, and with $\mu_1 > \mu_2$. Parts arrive according to a Poisson process with rate λ parts per hour, and those parts that find the station full just leave. On the other hand, parts that enter the station are joining a single queue and they are processed on a first-come, first-serve basis. Furthermore, the entire operation follows a non-idling policy for the system servers, and an arriving part that finds both servers idle is directed to the fastest server.

Prove or disprove the following statements:

i. Under the operational regime that is described in the previous paragraph, the faster server will experience a higher utilization than the slower one.

ii. The adopted routing policy maximizes the throughput of this workstation over all routing policies that will route an arriving part that finds both servers idle, to the fastest station with some probability $p \in [0, 1]$.

ISyE 7201 - SPRING 2019

FINAL EXAM SOLUTIONS

Problem 1:

(i) A difficulty with the modeling of the considered call center as a queueing network according to the corresponding theory presented in class, is that the customers that depart from each station follow a routing prob distribution that depends on their origin. Hence, for instance, customers departing from station A will just leave the call center if they came to this station from station B; if, on the other hand, a departing customer from that station joined the station as an external arrival, then this customer leaves the call center with prob. 0.2, and joins station B with the remaining probability. And a similar situation holds for the departures from station B.

From a more technical standpoint, we can capture the above effect by differentiating the customers at each station into two "classes": Class 1 collects all the external arrivals, while class 2 collects all those customers who came from the other workstation. Customers from both classes are served on a FCFS basis and with the same service type distribution. The only difference, as mentioned above, is what happens to the customers from each class when they leave the station. On the basis of the above we can compute the total arrival rates, λ_A and λ_B, for each station, as follows:

$$\lambda_A = 0.6\lambda + \lambda_B \times 0.1 \times \frac{0.4\lambda}{\lambda_B} = 0.64\lambda$$

$$\lambda_B = 0.4\lambda + \lambda_A \times 0.2 \times \frac{0.6\lambda}{\lambda_A} = 0.52\lambda$$

Next, we answer parts (ii) - (v) starting with (iv), in order to check also the stability of each station. We have:

$$u_A = \lambda_A \cdot t_A = 0.64 \times \frac{20}{60} \times 4 \simeq 0.853 < 1$$

$$u_B = \lambda_B \cdot t_B = 0.52 \times \frac{20}{60} \times 5 \simeq 0.867 < 1$$

for (ii) and (iii) we have:

$$W_{q_A} = \frac{u_A}{1-u_A} \cdot t_A = \frac{0.853}{1-0.853} \cdot 4 = 23.21 \text{ min}$$

$$W_{q_B} = \frac{u_B}{1-u_B} \cdot t_B = \frac{0.867}{1-0.867} \cdot 5 = 32.59 \text{ min}$$

$$L_{q_A} = \lambda_A \cdot W_{q_A} = 0.64 \times \frac{20}{60} \times 23.21 \simeq 4.95$$

$$L_{q_B} = \lambda_B \cdot W_{q_B} = 0.52 \times \frac{20}{60} \times 32.59 \simeq 5.65$$

Finally, for (v) we have:

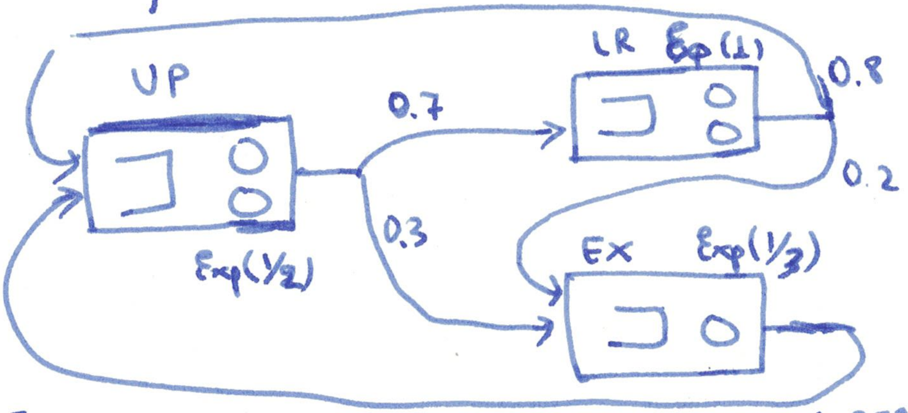
$$W_A = W_{q_A} + t_A = 23.21 + 4 = 27.21 \text{ min}$$

$$W_B = W_{q_B} + t_B = 32.59 + 5 = 37.59 \text{ min}$$

$$W_{A+B} = W_A + W_B = 64.8 \text{ min}$$

Problem 2:

(i) The requested CQN is as follows:



- The "customers" circulating in this CQN are the two generators.
- Workstation UP: A generator is in this workstation when it is functional. Since there are only two generators going through the network, any generator located at this workstation will be in service, and therefore each service time models the operational time of a generator till its next failure.
- Workstations LR and ER: These two workstations model respectively the experience of a generator that was sent for local or external repair. The queueing dynamics of each of these two workstations are the standard dynamics of a workstation with an infinite buffer and the corresponding number of servers. Furthermore, since $N=2$, we can also see that workstation LR will involve no queueing (i.e., any generator arriving at this workstation will be picked up immediately for processing.

Finally, the parameters of the proc.-time distributions for each workstation are the corresponding proc. rates in (weeks^{-1}).

(ii) The requested numbers is essentially the marginal distribution of the number of generators at workstation UP, in steady state. This distribution can be computed using the recursion presented in class for MVA of CQNs with general proc. rates $\mu_i(n)$ for each workstation i. Also, in the application of the above recursion to the considered problem we can apply some "shortcuts" that result from the particular structure of the considered CQN, and the focus of the entire computation to the requested distribution.

We start by computing some values for the relative arrival rates U_x, $x \in \{UP, LR, ER\}$. From the specification of the considered CQN, we have:

$$\begin{cases} U_{UP} = 0.8\, U_{LR} + U_{ER} \\ U_{LR} = 0.7\, U_{UP} \end{cases}$$

Then, setting $U_{UP} = 1$, we get $\begin{cases} U_{LR} = 0.7 \\ U_{ER} = 1 - 0.8 \times 0.7 = 0.44 \end{cases}$

Next we execute the recursion for a single circulating customer ($N=1$), ignoring the computation of the average number of customers at each station since these numbers are not needed for getting to our final objective.

for the average cycle time at each station, $W_x(1)$, we have:

$$W_x(1) = \sum_{n=1}^{1} \frac{n}{\mu_x(n)} P[X_x = n-1; \emptyset]$$

Hence,

$$W_{UP}(1) = \frac{1}{1/2} \cdot 1 = 2; \; W_{LR}(1) = \frac{1}{1} \cdot 1 = 1; \; W_{ER}(1) = \frac{1}{1/3} \cdot 1 = 3$$

Notice that the above numbers are just the mean proc. times at each station, and they could have been obtained more directly by noticing that when there is only one customer in the entire network, the mean cycle time per visit at any station is just the mean proc. time since there is no queueing in this case.

Next, we get $TH(1) = \frac{1}{\sum_x v_x W_x(1)} = \frac{1}{1 \cdot 2 + 0.7 \cdot 1 + 0.44 \cdot 3} \simeq 0.25$

Finally

$$P[X_{UP} = 1; 1] = \frac{v_{UP} \cdot TH(1)}{\mu_{UP}(1)} \cdot P[X_{UP} = 0; 0] = \frac{1 \times 0.25}{1/2} \cdot 1 = 0.5$$

$$P[X_{UP} = 0; 1] = 1 - P[X_{UP} = 1; 1] = 1 - 0.5 = 0.5$$

Similarly

$$P[X_{LR} = 1; 1] = \frac{0.7 \times 0.25}{1} \cdot 1 = 0.175; \; P[X_{LR} = 0; 1] = 0.825$$

$$P[X_{ER} = 1; 1] = \frac{0.44 \times 0.25}{1/3} 1 = 0.33; \; P[X_{ER} = 0; 1] = 0.67$$

6

For the second iteration, that allows two customers in the system, we get:

$$W_{UP(2)} = \sum_{n=1}^{2} \frac{n}{\mu_{UP(n)}} P[X_{UP} = n-1; 1] = 2$$

$$W_{LR}(2) = 1$$

$$W_{FR(2)} = \frac{1}{1/3} \cdot 0.67 + \frac{2}{1/3} \cdot 0.33 = 3.99$$

The first two numbers above were derived immediately from the fact that even with two circulating customers in the network, workstations UP and LR involve no queueing. On the other hand, $W_{FR}(2)$ is obtained by applying the corresponding formula (so, we need $P[X_{FR}; 1]$).

$$TH(2) = \frac{2}{1\times2 + 0.7\times1 + 0.44\times3.99} = 0.45$$

Finally,

$$\Rightarrow P[X_{UP} = 2; N=2] = \frac{U_{UP} \cdot TH(2)}{\mu_{UP}(2)} P[X_{UP} = 1; N=1] =$$

$$= \frac{1\times0.45}{2/2} \cdot 0.5 = 0.225$$

$$P[X_{UP} = 1; N = 2] = \frac{U_{UP} \cdot TH(2)}{\mu_{UP}(1)} P[X_{UP} = 0; N=1] =$$

$$= \frac{1\times0.45}{1/2} \cdot 0.5 = 0.45$$

$$\rightarrow P[X_{UP} \geq 1; N=2] = 0.225 + 0.45 = 0.675$$

Problem 3:

In the subsequent discussion, we assume that class 1 has non-preemptive priority over class 2; the workstation is stable. We also adopt the following notation:

- Class arrival rates: λ_1, λ_2
- Proc. rate: μ; mean proc. time $\tau = 1/\mu$ (1)
- Class traffic intensities (and corresp. server utilizations):

  $\rho_1 = \lambda_1 \tau$; $\rho_2 = \lambda_2 \tau$. (2)
- Also, from stability condition: $\rho \equiv \rho_1 + \rho_2 < 1$. (3)

Then, using the corresponding formulae that we derived in class, we have:

$$R^{(n)} \equiv \text{Exp. Remaining Proc. Time} = \frac{\lambda_1+\lambda_2}{2}\left[\sigma^2+\tau^2\right] =$$

$$= \frac{\lambda_1+\lambda_2}{2}\left[\frac{\sigma^2}{\tau^2}+1\right]\tau^2 = \frac{\lambda_1\tau+\lambda_2\tau}{2}\left[\frac{\sigma^2}{\tau^2}+1\right]\tau \quad (4)$$

Since proc. times are exp. distributed, $\sigma^2/\tau^2 = 1$ (5).

From (2), (4) and (5):

$$R^{(n)} = \rho\tau \quad (6)$$

Also, $W_1^{(n)} = \frac{R^{(n)}}{1-\rho_1} = \frac{\rho\tau}{1-\rho}$; $W_2^{(n)} = \frac{R^{(n)}}{(1-\rho_1)(1-\rho_1-\rho_2)} = \frac{\rho\tau}{(1-\rho_1)(1-\rho)}$

and

$$W^{(n)} = \frac{\lambda_1}{\lambda_1+\lambda_2}W_1^{(n)} + \frac{\lambda_2}{\lambda_1+\lambda_2}W_2^{(n)} = \frac{1}{\lambda_1+\lambda_2}\frac{\rho}{1-\rho_1}\left[\rho_1+\frac{\rho_2}{1-\rho}\right]$$

$$= \frac{1}{\lambda_1+\lambda_2}\frac{\rho}{1-\rho_1}\frac{\rho_1-\rho_1\rho+\rho_2}{1-\rho} = \frac{1}{\lambda_1+\lambda_2}\frac{\rho}{\cancel{1-\rho_1}}\frac{\rho\cancel{(1-\rho_1)}}{1-\rho} = \frac{1}{\cancel{\lambda_1+\lambda_2}}\frac{\rho}{1-\rho}\cancel{(\lambda_1+\lambda_2)}\tau$$

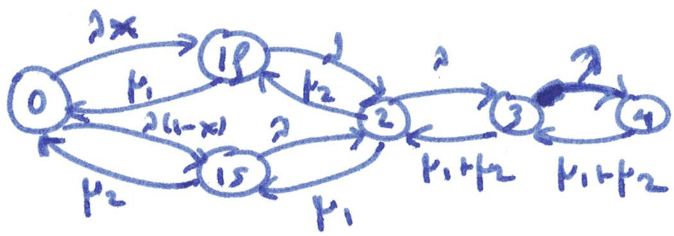
$$= \frac{\rho}{1-\rho}\tau = W_q(M/M/1;\ \lambda=\lambda_1+\lambda_2;\ \mu)$$

Problem 5

(i) This statement is wrong. For a proof of this result see parts (i) and (ii) of Problem 4 in the 2nd Midterm Exam of Spring 2018. (Appended at the end of this writeup)

(ii) This statement is true. Its validity can be proven as follows:

Let $x \in [0,1]$ denote the optimal probability for routing to the fastest station when both servers are idle. Then we get the following CTMC for the dynamics of the resulting queue:



The balance equations for this CTMC are as follows:

0: $\lambda p_0 = \mu_1 p_{1f} + \mu_2 p_{1s}$

1f: $\lambda x p_0 + \mu_2 p_2 = (\lambda + \mu_1) p_{1f}$

1s: $\lambda(1-x) p_0 + \mu_1 p_2 = (\lambda + \mu_2) p_{1s}$

2: $\lambda (p_{1f} + p_{1s}) + (\mu_1 + \mu_2) p_3 = (\mu_1 + \mu_2 + \lambda) p_2$

3: $\lambda p_2 + (\mu_1 + \mu_2) p_4 = (\lambda + \mu_1 + \mu_2) p_3$

4: $\lambda p_3 = (\mu_1 + \mu_2) p_4$

We also have: $TH = \lambda_{eff} = \lambda(1 - p_4)$.

So, TH is maximized when p_4 is minimized.

Next, we shall express P_4 as a function of x, and we shall show that P_4 is minimized when $x = 1$.

From balance Eq. (4), we have: $P_3 = \frac{\mu_1+\mu_2}{\lambda} P_4$

" " " (3), " " : $P_2 = \frac{1}{\lambda}\left[\frac{(\lambda+\mu_1+\mu_2)(\mu_1+\mu_2)}{\lambda} - (\mu_1+\mu_2)\right] P_4$
$= \left(\frac{\mu_1+\mu_2}{\lambda}\right)^2 P_4$

" " " (2), " " : $P_{1f} + P_{1s} = \frac{\mu_1+\mu_2+\lambda}{\lambda} P_2 - \frac{\mu_1+\mu_2}{\lambda} P_3 =$
$= \left(\frac{\mu_1+\mu_2}{\lambda}\right)^3 P_4$

from (0), (1f), (1s), we have:

$$\begin{cases} x(\mu_1 P_{1f} + \mu_2 P_{1s}) + \mu_2 P_2 = (\lambda+\mu_1) P_{1f} \\ (1-x)(\mu_1 P_{1f} + \mu_2 P_{1s}) + \mu_1 P_2 = (\lambda+\mu_2) P_{1s} \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} (\lambda+\mu_1-x\mu_1) P_{1f} - x\mu_2 P_{1s} = \mu_2 P_2 \\ -(1-x)\mu_1 \; P_{1f} + [\lambda+\mu_2-(1-x)\mu_2] P_{1s} = \mu_1 P_2 \end{cases} \Rightarrow \text{from Crammer's rule}$$

$$\Rightarrow \frac{P_{1f}}{P_{1s}} = \frac{\left.\begin{vmatrix} \mu_2 & -x\mu_2 \\ \mu_1 & \lambda+x\mu_2 \end{vmatrix}\right\}\Delta_1}{\left.\begin{vmatrix} \lambda+(1-x)\mu_1 & \mu_2 \\ -(1-x)\mu_1 & \mu_1 \end{vmatrix}\right\}\Delta_2} \Rightarrow \frac{P_{1f}}{P_{1f}+P_{1s}} = \frac{\Delta_1}{\Delta_1+\Delta_2} \Rightarrow$$

$$\begin{cases} P_{1f} = \frac{\Delta_1}{\Delta_1+\Delta_2}(P_{1f}+P_{1s}) = \frac{\mu_2\lambda + x\mu_2^2 + x\mu_1\mu_2}{x(\mu_2^2-\mu_1^2)+(\mu_1+\mu_2)(\lambda+\mu_1)} \cdot \left(\frac{\mu_1+\mu_2}{\lambda}\right)^3 P_4 \\ P_{1s} = \frac{\Delta_2}{\Delta_1+\Delta_2}(P_{1f}+P_{1s}) = \frac{\mu_1\lambda + (1-x)\mu_1^2 + (1-x)\mu_1\mu_2}{x(\mu_2^2-\mu_1^2)+(\mu_1+\mu_2)(\lambda+\mu_1)} \cdot \left(\frac{\mu_1+\mu_2}{\lambda}\right)^3 \cdot P_4 \end{cases}$$

~~Then~~:

Also, from 0 :

$$P_0 = \frac{\mu_1}{\lambda} P_f + \frac{\mu_2}{\lambda} P_{1s} = \frac{\mu_1 \mu_2}{\lambda} \frac{\lambda + x\mu_2 + x\mu_1 + \lambda + (1-x)\mu_1 + (1-x)\mu_2}{x(\mu_2^2 - \mu_1^2) + (\mu_1+\mu_2)(\lambda+\mu_1)} \left(\frac{\mu_1+\mu_2}{\lambda}\right)^3$$

$$= \frac{\mu_1 \mu_2}{\lambda} \underbrace{\left(\frac{2\lambda + \mu_1 + \mu_2}{x(\mu_2^2-\mu_1^2) + (\mu_1+\mu_2)(\lambda+\mu_1)}\right)}_{f(x) > 0 \text{ since } P_0 \text{ is well defined}} \left(\frac{\mu_1+\mu_2}{\lambda}\right)^3 P_4$$

Finally, we have:

$$P_4 \left[f(x) \frac{\mu_1\mu_2}{\lambda} \left(\frac{\mu_1+\mu_2}{\lambda}\right)^3 + \left(\frac{\mu_1+\mu_2}{\lambda}\right)^3 + \left(\frac{\mu_1+\mu_2}{\lambda}\right)^2 + \frac{\mu_1+\mu_2}{\lambda} + 1 \right] = 1$$

Hence P_4 is minimized when $f(x)$ is maximized.

But since $\mu_2 < \mu_1$, $f(x) \equiv \dfrac{2\lambda + \mu_1 + \mu_2}{x(\mu_2^2-\mu_1^2) + (\mu_1+\mu_2)(\lambda+\mu_1)}$ is

maximized when $x = 1$. —

Problem 4:

Part (i) of this problem is wrong!

My line of reasoning when I was designing this problem, was as follows:

(i) By an argument similar to that used when we computed the steady-state distribution of ~~an~~ M/G/1 queue, we can establish that the probability that a departing customer leaves the system empty during steady-state operation, is $1-\rho$ where $\rho = \lambda\tau$.

(ii) Then, we could consider each departure as a Bernoulli trial with "success" probability that the system gets empty upon that departure, and have the result that was suggested in that part of the problem.

But this reasoning is erroneous because when we study a busy period for this queue, we start with an arrival that find the queue empty and we are looking for the next departure that will leave the system empty; so, we need to consider the transient dynamics of the system that will generate this busy period.

I realized this discrepancy when I was starting to prepare the solution for this problem. It turns out that the derivation of distributional results of the type that are considered in this problem is a **quite** nontrivial task. Some results along these lines can be found in your textbook (Gross et al.) in pages 97–103 for the M/M/1 queue and in pages 239–240 for the M/G/1 queue.

Also, some further developments can be found in Ross's book on Stochastic Processes, in pages 73–78.

This material characterizes the probability distribution

$B(t,n) = \text{Prob}\{\text{busy period length} \leq t \text{ AND}$
$\text{number of customers served} = n\}$

It is shown that

$$\frac{d\,B(t,n)}{dt} = e^{-\lambda t}\frac{(\lambda t)^{n-1}}{n!}\,dG_n(t)$$

where $G_n(t)$ is the n-fold convolution of the service time distribution $G(t)$.

The above result further implies that

$$B(n) = \text{Prob}\{\text{number of customers served} = n\} =$$
$$= \int_0^\infty \frac{dB(t,n)}{dt}\,dt = \int_0^\infty e^{-\lambda t}\frac{(\lambda t)^{n-1}}{n!}\,dG_n(t)$$

and this is the correct result for part (i) of this problem.

To give you some more context and perspective about this problem, I started thinking about all this development in an effort to develop the alternative approach to the MVA of the 2-class queue with preemptive priorities that is outlined in part (iv) of the problem. If we can get the first and the second moments of the busy periods for class 1 (higher priority) customers, then we should be able to develop a MVA for class 2 customers thinking of the preemptions that are caused by class 1 as non-destructive outages of the type that we modeled in the next-to-last lecture. There is a certain caveat in the application of the above reasoning in the considered case that arises from the fact that a class-2 arrival might find the system empty of class-2 jobs, and the server is still busy due to a class-1 busy period; but the necessary adjustments can be worked out. In fact, I worked out the complete derivation in the last two days, getting the second moment for the busy period from the corresponding results in Gross' book that I mentioned above. The computation of this moment is an interesting exercise in itself and demonstrates the power of transforms (Laplace-Stieltjes, etc.) and generating functions in these computations. I can discuss the complete derivation of the aforementioned results with any of you who might be interested; just let me know.

FROM 2nd MIDTERM EXAM

OF SPRING 2018

Problem 4 (30 points): Consider a queueing station with 2 servers and a total buffering capacity (including the buffering capacity of the servers) of K units. Processing times at the two servers are exponentially distributed with corresponding rates μ_1 and μ_2 parts per hour, and with $\mu_1 > \mu_2$. Parts arrive according to a Poisson process with rate λ parts per hour, and those parts that find the station full just leave. On the other hand, parts that enter the station are joining a single queue and they are processed on a first-come, first-serve basis. Furthermore, the entire operation follows a nonidling policy for the system servers, and an arriving part that finds both servers idle is directed to the fastest server.

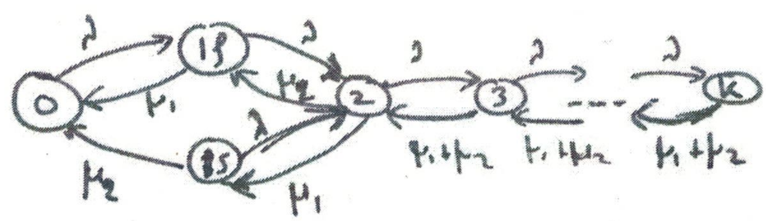
i. (10 pts) Model the operation of the above station as a CTMC. Under what conditions will this MC possess a steady-state?

ii. (10 pts) Is it true that under the considered operational scheme, the fastest server will experience a higher utilization than the slowest one?

iii. (10 pts) How can we modify the above scheme in order to attain equal server utilization while maintaining the nonidling policy for this station? Will such a modification always be possible?

Please, provide a complete analysis for all the above questions.

(i) The corresponding CTMC is as follows:



where '1f' denotes the state with just one ~~customer~~ part in the station at the faster server, and '1s' denotes the state with just one ~~customer~~ part in the station, at the slower server.
The remaining states are defined by the corresponding number of parts in the station.

Since the embedded DTMC of this CTMC is irreducible and finite-state, the CTMC is ergodic.

(ii) Letting $\pi(\cdot)$ denote the stationary distribution for the CTMC of part (i), we can express the utilization of the two servers as follows:

Faster: $\pi(1f) + \pi(2) + \dots + \pi(K) = u_f$

Slower: $\pi(1s) + \pi(2) + \dots + \pi(K) = u_s$

Hence, to compare u_f and u_s, it suffices to compare $\pi(1f)$ and $\pi(1s)$.

From the balance equations at nodes 0, 1f and 1s, we get:

$$\left\{\begin{array}{l} \lambda\pi(0) = \mu_1\pi(1f) + \mu_2\pi(1s) \\ (\lambda+\mu_1)\pi(1f) = \lambda\pi(0) + \mu_2\pi(2) \\ (\lambda+\mu_2)\pi(1s) = \mu_1\pi(2) \end{array}\right\} \Rightarrow (\lambda+\mu_1)\pi(1f) = \mu_1\pi(1f) + \mu_2\pi(1s) + \mu_2\frac{\lambda+\mu_2}{\mu_1}\pi(1s)$$

$$\Rightarrow \pi(1f) = \left[\frac{\mu_2}{\lambda} + \frac{\mu_2}{\mu_1} + \frac{\mu_2}{\lambda}\frac{\mu_2}{\mu_1}\right]\pi(1s) \qquad (1)$$

From (1) we can see that if $\mu_2 \geq \lambda$, then, clearly, $\pi(1f) > \pi(1s)$ and therefore $u_f > u_s$.

But it is also possible to get values for (λ, μ_1, μ_2) s.t. $\pi(1f) < \pi(1s)$; e.g., set $\mu_2/\lambda \leq 1/3$ and $\mu_2/\mu_1 \leq 1/3$. Then,

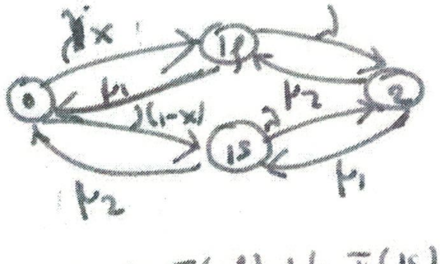
$\pi(1f) = [1/3 + 1/3 + (1/3)^2]\,\pi(1s) = 7/9\,\pi(1s)$.

A more general condition implied by (1) for having $\pi(1f) \leq \pi(1s)$, is: $\frac{\mu_2}{\lambda} + \frac{\mu_2}{\mu_1} + \frac{\mu_2}{\lambda}\frac{\mu_2}{\mu_1} \leq 1 \Rightarrow$

$$\Rightarrow \frac{\mu_2}{\lambda} \leq \frac{\mu_1 - \mu_2}{\mu_1 + \mu_2}. \quad (2)$$

<u>Rem</u>: The above developments demonstrate that high utilization does not mean (necessarily) high productivity. A server might experience high utilization just because it is <u>very slow</u>!

(iii) If $\pi(1s) \geq \pi(1f)$ (or $U_s \geq U_f$) under the considered scheme (i.e., if Eq (2) above is true), then it is not possible to equate the utilization of the two servers. In the opposite case, we can try to achieve $U_f = U_s$ by redirecting some arrivals that find the system empty to the slower server. More specifically, under this new scheme, an arrival finding the system empty will be routed to the faster server with probability $x \in (0,1)$, and to the slower server with probability $1-x$. Next we determine the right value for x, considering, without loss of generality the case where $K=2$. The corresponding CTMC is:



and the new "flow balance" equations are:

$$\left\{\begin{array}{l} 0:\ \lambda\pi(0) = \mu_1\pi(1f) + \mu_2\pi(1s) \\ 1f.\ \lambda x\pi(0) + \mu_2\pi(2) = (\lambda+\mu_1)\pi(1f) \\ 1s:\ \lambda(1-x)\pi_0 + \mu_1\pi(2) = (\lambda+\mu_2)\pi(1s) \end{array}\right\} \Rightarrow \left\{\begin{array}{l} (\lambda+\mu_1 - x\mu_1)\pi(1f) - x\mu_2\pi(1s) = \mu_2\pi \\ -(1-x)\mu_1\pi(1f) + [\lambda+\mu_2 - (1-x)\mu_2]\pi(1s) \\ \qquad = \mu_1\pi \end{array}\right.$$

$\Rightarrow$

$$\Rightarrow \quad \frac{\pi(1f)}{\pi(1s)} = \frac{\begin{vmatrix} \mu_2 & -x\mu_2 \\ \mu_1 & \lambda + x\mu_2 \end{vmatrix}}{\begin{vmatrix} \lambda + (1-x)\mu_1 & \mu_2 \\ -(1-x)\mu_1 & \mu_1 \end{vmatrix}} \qquad (3)$$

Setting $\frac{\pi(1f)}{\pi(1s)} = 1$, Eq. (3) gives:

$$\mu_2\lambda + x\mu_2^2 + x\mu_1\mu_2 = \lambda\mu_1 + (1-x)\mu_1^2 + (1-x)\mu_1\mu_2 \Rightarrow$$

$$\Rightarrow \quad x = \frac{\lambda(\mu_1-\mu_2) + \mu_1(\mu_1+\mu_2)}{(\mu_1+\mu_2)^2} > 0.$$

We also need $x \leq 1 \Rightarrow$

$$\Rightarrow \quad \lambda(\mu_1-\mu_2) + \cancel{\mu_1^2} + \cancel{\mu_1\mu_2} \leq \cancel{\mu_1^2} + \mu_2^2 + \cancel{2}\mu_1\mu_2 \Rightarrow$$

$$\Rightarrow \quad \lambda(\mu_1-\mu_2) \leq \mu_2(\mu_1+\mu_2) \Rightarrow$$

$$\Rightarrow \quad \frac{\mu_2}{\lambda} \geq \frac{\mu_1-\mu_2}{\mu_1+\mu_2}$$

which is consistent with our working assumption [redacted] that Eq. (2) does not hold.

<u>Remark:</u> As some additional "sanity checking", show also that the system of the two linear equations [redacted] of the previous page that we used to obtain the ratio $\pi(1f)/\pi(1s)$, is <u>non-singular</u>.