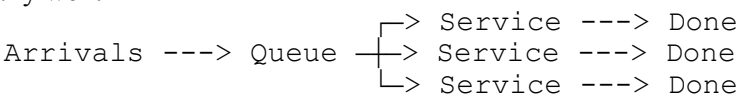
Queuing Theory II

Assignment 8 is on page 7. Assignment 8A is on page 10.

More complex queues:

Multiple Server Single Stage Queue

-- meaning that we have one line that leads to several servers, each of whom can serve any customer equally well.



The formulas use M for the number of servers or "channels".
M = 3 in the above diagram.

λ is still the average arrivals per unit of time
μ is still the average number of customers a server can handle per unit time.
μ is 1divided by the average service time.

ρ changes its definition. $\rho = \lambda/(M\mu)$.

This way ρ still represents the service utilization factor, the proportion of time on average that each server is busy. As with the simple queue, ρ has to be less than 1 for a steady state to exist.

We again assume that arrivals and customer service completion have the Poisson distribution. This is equivalent to saying that inter-arrival and service length times have the exponential distribution.

The probability that at any given time there are no customers waiting or being served:

$$Prob(0\ in\ system) = \frac{1}{\sum_{i=0}^{M-1} \frac{(M\rho)^i}{i!} + \frac{(M\rho)^M}{M!(1-\rho)}}$$

The probability that at any given time there are n customers in the system.

$$If\ n \le M\ then\ Prob(n\ in\ system) = Prob(0)\frac{(M\rho)^n}{n!}$$

$$If\ n \ge M\ then\ Prob(n\ in\ system) = Prob(0)\frac{M^M\rho^n}{M!}$$

To get the length of the queue and the average wait, it's easiest to start with L_q:

L_q, the mean number of customers waiting in queue to be served:

$$L_q = Prob(0)\frac{M^M \rho^{M+1}}{M!(1-\rho)^2}$$

The rest of the quantities can then be calculated. The following formulas are true for any queuing model.

Mean time customers spend in queue: $W_q = L_q / \lambda$

Mean time customers spend in the system: $W = W_q + 1/\mu$

Mean number of customers in system: $L = L_q + \lambda/\mu$

A spreadsheet for this kind of system is in the Queuing Theory Cookbook (use the link on the syllabus).

An example

For an example of a multiple-channel model, let us revisit the pharmacy from Queuing Theory I. How much faster will it be if we add a second pharmacist? Here is how to analyze it:

Two pharmacists are simultaneously filling orders. M=2.
10 orders come in per hour. $\lambda = 10$.
Filling the orders takes an average of 4 minutes each, so each pharmacist can handle 15 orders per hour on average. $\mu = 15$.

$\rho = \lambda/(M\times\mu) =$
$\rho = 10/(2\times15) = 1/3$ Each server is busy, on average, one-third of the time.

Assume that orders and fills are Poisson (i.e. service times are exponential). This assumption lets us use the formulas.

The hardest quantity to calculate is the probability of 0, because it requires expanding that expression with the summation symbol in the denominator.

$$\frac{1}{\sum_{i=0}^{M-1}\frac{(M\rho)^i}{i!}+\frac{(M\rho)^M}{M!(1-\rho)}}$$

Here, M = 2. Substitute 0 for *i* in the formula to the right of the summation sign. Then substitute 1 for *i*. The summation expression thus turns into two terms to be added together. ρ is 1/3. Here is what you get:

$$\text{Probability (0 in system)} = \frac{1}{1 + 2\times1/3 + (2\times1/3)^2/(2\times2/3)} = 1/2$$

$$\text{Probability (1 in system)} = 1/2 \times (2 \times 1/3)^1/1 = 1/3$$

$$\text{Probability (2 in system)} = 1/2 \times (2 \times 1/3)^2/2 = 1/9$$

This system is empty (no one being served or waiting) half of the time.

Suppose two chairs are provided at the pharmacy. How much of the time will someone have to stand?

$$\text{Probability (0, 1, or 2 in system)} = P(0) + P(1) + P(2) = 0.9444..$$

$$\text{Probability (3+ in system)} = 1 - (P(0) + P(1) + P(2)) = 0.0555..$$

If there are two chairs, someone will be standing only if there are three or more people in the system, which will happen about 0.056, or 5.6% of the time.

$$L_q = (1/2) \times \frac{2^2 \times (1/3)^{2+1}}{2 \times (2/3)^2} = 1/12 \quad \text{Average length of queue}$$

$$L = 1/12 + 2/3 = 3/4 \quad \text{Average number of customers waiting or being served}$$

$$W_q = (1/12)/10 = 1/120 \quad \text{Average time spent in queue}$$

$$W = 1/120 + 1/15 = 0.075 \quad \text{Average time spent in system}$$

An economic analysis of this two-server queue:

Aides who get order spend W = 0.075 hours average per order.
There are $\lambda = 10$ orders per hour.
Total aide-hours waiting or being served per hour = $\lambda \times W = 10 \times 0.075 = 0.75$.
Aide pay rate is \$6 /hr.
Aide time cost per hour = $0.75 \times \$6 = \$4.50$
Adding a second pharmacist reduces the aide time cost per hour from \$12 to \$4.50. The difference is \$7.50. Therefore, it would pay to hire the second pharmacist if and only if the pharmacist's wage is less than \$7.50 per hour.

More examples:

Two Lines to Two Servers Compared With One Line to Two Servers

This is the McDonalds versus Burger-King comparison. McDonalds restaurants typically have multiple lines, one for each working cash register. Customers come in and get in whichever line they wish. Burger King restaurants generally have one line. Customers at the front of the line go to whichever server is open. (In our area, KFC does it like McDonalds. Wendy's does it like Burger King.)

Let us suppose that one person arrives per minute, and that the service time averages 30 seconds.

We can model Burger King using the methods of this section.

$\lambda = 1$ per minute
$\mu = 2$ per minute
W calculates to 0.53333... minutes, which is 32 seconds average to wait and get served.

Modelling McDonalds is more complex, because we have to model how people behave after they get in line. The worst behavior, from the customer's point of view, is to never switch lines, even if the other server is free. If we further assume that the customer picks a line at random at the time entry to the restaurant, ignoring how long each line is, then we can model each of the two lines individually.

Each line gets half the arrivals, so $\lambda = 0.5$ per minute.
$\mu = 2$ per minute, as above
Using the simple one-server model from last time, W calculates to 0.6666..., which is 40 seconds to wait and get served..

So, the McDonalds method takes longer.

Except that I am being unrealistic by having people pick a line at random and not allowing them to switch lines. If people do switch lines, which we know they do, then the average time to get served in a McDonalds becomes the same as the average time in a Burger King. There is one important difference, though: Burger King's method enforces first-in first-out priorities, while McDonalds' method does not. The result is that, even though the average time is the same in both places, the variation in waiting time is greater at McDonalds than at Burger King.

Two Servers Compared With One Server Who Is Twice as Fast

Imagine that you have a choice between buying two relatively slow copying machines or one machine that is twice as fast as each of the slow ones. Total capacity of throughput is the same with either choice, so you might think that the choices would be equally good.

I make up some numbers for illustrative purposes:

Arrivals: 3 jobs every 5 minutes on the average, so $\lambda = 0.6$ jobs/minute.

Two slow machines:
Service time: 2 minutes average, so $\mu = 0.5$ jobs/minute.
$\rho = 0.6$.
$W = 3.13$ minutes average in system per job.
Each minute, you average
3.13 person-minutes in system/job × 0.6 jobs/minute =
1.878 person-minutes in system/minute.
This equals 1.878 person-hours copying per hour of the day of people standing at the copying machine or waiting for a turn at a machine.

One fast machine.
Service time: 1 minute average, so $\mu = 1$ job/minute.
$\rho = 0.6$.
$W = 1/0.4 = 2.5$ minutes in system average per job.
2.5 person-minutes/job × 0.6 jobs/minute =
1.5 person-minutes in system per minute of the day =
1.5 person-hours at the copier per hour of the day.

The fast machine saves the workers time. They spend 0.378 person-hours per hour of the day less in the copying room.

If the single fast machine is less expensive to own and operate, that should be your choice.

If the fast machine is more expensive than two slow ones, you will choose it anyway if and only if its cost exceeds the cost of two slow machines by more than your gain from freeing 0.378 person-hours per hour.

This analysis assumes that there is no demand effect from having a faster machine. We did not change λ. What if people do more copying when it's faster? You could use a different λ for the fast machine to allow for demand increasing as the time cost of making a copy decreases.

How many seats do you want in your cafeteria?

In this example, 60 people come per hour. They spend an average of 15 minutes sitting and eating, so mu = 4 per hour.

If, as shown here, you have 20 tables, there will be no one waiting to sit 88% of the time. The other 12% of the time, someone will be standing with a tray, waiting for an open seat.

M servers in parallel

Lambda	60
Mu	4
M	20
Rho	0.75
L	15.48129
Lq	0.481288
W	0.258021
Wq	0.008021

i or n	(RhoM)^I/I!	Prob(n)	Prob(<=n)
0	1	2.93E-07	2.93443E-07
1	15	4.4E-06	4.69508E-06
2	112.5	3.3E-05	3.77074E-05
3	562.5	0.000165	0.000202769
4	2109.375	0.000619	0.00082175
5	6328.125	0.001857	0.002678692
6	15820.31	0.004642	0.007321048
7	33900.67	0.009948	0.017268953
8	63563.76	0.018652	0.035921276
9	105939.6	0.031087	0.06700848
10	158909.4	0.046631	0.113639287
11	216694.6	0.063587	0.17722675
12	270868.3	0.079484	0.25671108
13	312540.3	0.091713	0.348423767
14	334864.6	0.098264	0.446687361
15	334864.6	0.098264	0.544950955
16	313935.6	0.092122	0.637073075
17	277002	0.081284	0.718357298
18	230835	0.067737	0.78609415
19	182238.2	0.053476	0.839570613
20	0	0.040107	0.879677959

Assignment 8

You have two physicians seeing patients in the ER. Patients arrive at a rate of 2 per hour. The average service time per patient is 20 minutes. (Questions 1-5 are copied from assignment 7A, so you can see what difference it makes to add a second server.)

1. To proceed with your analysis, what do you assume about the distributions of arrivals and service times?

2. What is the average number of patients in the ER (waiting or being served?)

3. What is the average length of time that a patient spends from the time they enter the ER to the time they leave?

4. The waiting area is separate from the examining/treatment room. How many chairs should there be in the waiting area to reduce the probability that someone will have no chair to less than 0.01? (No fractions of chairs, please.)

5. Suppose the hospital has announced, as part of a CQI policy, that it will discount each patient's bill by $6.00 per hour that the patient waits in the ER waiting area. How much will this discount cost the hospital per hour on the average? (Note that, unlike the example on the preceding page, patients get paid only for waiting time, not for service time. Also note that patients get paid proportionally for fractions of hours spent waiting. For example, if a patient waits just one minute before being seen, he or she gets $0.10.)

6. Suppose a physician costs $30 per hour. If the $6.00 per hour waiting penalty is in effect, does it pay to add the second physician? You must compare the situation when you have one physician with the situation when you have two.

7. Regardless of the answer to question 6, which gives you shorter waits overall, one ER with two physicians or two separate ER's with one physician in each one, each serving half as many patients on the average? Based on this comparison alone, is it better to have centralized or decentralized ER facilities in a city?

Multiple stages in series compared with single stage
also compared with queues in series

In a model with "stages," the process of being served involves individual steps. The server has to finish all the steps with a customer before the next customer can start. Placing a telephone call involves multiple stages. So does filling out a bunch of forms while a clerk helps you.

For stages models, if the individual stages have the Poisson or exponential distribution, the service time has the Erlang distribution, also called the Gamma distribution. This is an extension of the exponential distribution.

For simplicity, assume that all stages take the same amount of time on average. (Breaking up a long process into equal stages gives you better performance than unequal stages would.)

The parameter k represents the number of stages. If k = 1 the Erlang distribution reduces to the exponential. If k is very large, the service times gets close to being the same for all customers.

As with the Poisson and exponential distributions, the use of the Erlang distribution is based on the assumption that service times are independent of how busy you've been and how many are waiting in line.

Here are the formulas for the Erlang queuing model:

Server utilization factor (what proportion of the time the server -- at least one stage -- is busy) = $\rho = \lambda/\mu$

Probability of 0 in system: $\text{Prob}(0) = 1 - \rho$

Probability of n in system: (General formula too messy)

Average number in system: $L = \rho + \frac{(k+1)\rho^2}{2k(1-\rho)}$

Average number in queue: $L_q = L - \rho$

Average time in system: $W = \frac{L}{\lambda}$

Average wait in queue: $W_q = \frac{L_q}{\lambda} = W - \frac{1}{\mu}$

Now to a worked-out example:

An admissions clerk fills out 4 forms that take 2 minutes each. (We use the simplifying assumption that all the forms take the same amount of time.) Six patients arrive per hour.

k = 4 because there are 4 forms.
$\lambda = 6$ -- arrivals per hour
$\mu = 7.5$

μ is the average number of patients served per hour when the clerk is busy. If there are 4 forms to fill out and they average 2 minutes each then the average patient takes $4 \times 2 = 8$ minutes. There are 60 minutes in an hour, so the clerk can handle 60/8 patients per hour, which equals 7.5.

$\rho = 6/7.5 = 0.8$

So the clerk is idle 1 - 0.8 = 0.2 or 20% of the time.

	4 stages		One stage
L	2.8	L	4
Lq	2	Lq	3.2
W	0.466667	W	0.666667
Wq	0.333333	Wq	0.533333

Queues in series:
Each stage takes 2 minutes average so $\mu = 30$.
$\lambda = 6$ for all stages.
L = 0.25, W = 0.041667 for each stage.
For all 4 stages added up, L = 1, W = 0.166667.

Now, it's your turn to work through some examples:

Assignment 8A
Queuing Theory Applications

For each of the following primary care centers ("doc-in-a-box"), calculate the average number of patients in the system, the average length of time spent in the system, and the average length of time spent waiting. Unless stated otherwise, use the usual assumptions about the distribution of the arrivals and service times, and about queue priorities.

Use the Queuing Theory Cookbook, which is linked from the syllabus. Match each of the six models with the appropriate page in the Cookbook.

0. What are those assumptions?

1. Patients arrive at a doc-in-the-box at the average rate of 2 per hour. The doc spends an average of 15 minutes with each patient. (Hint: You can use your spreadsheet from Assignment 7A, with one change of parameter.)

2. Patients arrive at Doc-and-Nurse-in-a-Box at the average rate of 2 per hour. A nurse spends an average of 7.5 minutes with each patient, then the doc spends an average of 7.5 minutes with the patient while the nurse starts with the next patient. (This is not Erlang. This is two queues in series, one after the other. You can again start with your spreadsheet from Assignment 7A.) Compare this clinic with 1. Is the expected total time less?

3. Patients arrive at Compute-O-Doc at the average rate of 2 per hour. Compute-O-Doc deals with each patient in exactly 15 minutes. Compare this answer with number 1's. Does taking the randomness out of the service times make the average wait less?

4. Patients arrive at Doc-in-a-Matchbox at the rate of 2 per hour. The doc spends an average of 15 minutes with each patient. If an arriving patient finds that there are 2 patients in the waiting room, the patient leaves and doesn't come back. Also calculate the probability that an arriving patient will leave. Comparing this clinic with number 1, is the average total time less, for those who do get in?

5. Patients arrive at Tender-Loving-Care-Docs-in-a-Box at the rate of 2 per hour. There are 2 docs working. Patients don't care which doc they see. Patients spend an average of 30 minutes with the doc. Compare this answer with number 1's. Is it faster to have one fast server or two slow ones?

6. Patients arrive at Urgie-Care-in-a-Box at the rate of 2 per hour. There is one doc who spends an average of 15 minutes with each patient. On arrival, patients are judged to be either urgent or non-urgent. If the doc is serving a non-urgent patient and an urgent patient comes in, the doc interrupts the non-urgent visit and takes care of the urgent patient, returning to the non-urgent patient only when no urgent patients are waiting. Assume that, on the average, one patient per hour is urgent and one is not urgent. Calculate answers for both urgent and non-urgent patients. Compared with clinic number 1, which does not triage patients, is the expected wait less in clinic 6 if you are an urgent patient? Is the expected wait less in clinic 6 if you are a non-urgent patient?