4 M / M / 1 Queuing Systems

We discuss now continuous-time queuing systems with the usual approach: consider a discrete-time queuing system and let the frame size $\Delta \to 0$.
First, let us explain the notation:
<u>**Notation**</u> A queuing system is denoted by **A / S / k / C / P**, where
$-$ **A** denotes the distribution of **interarrival times**;
$-$ **S** denotes the distribution of **service times**;
$-$ **k** denotes the number of **servers**;
$-$ **C** denotes the **capacity**;
$-$ **P (or K)** denotes the size of the **source population**.
Usually, the default values for the last two are $C = P = \infty$ and they are dropped from the notation. When the Exponential distribution is considered for A or S, then it is denoted by M, because it is *<u>m</u>emoryless* and the resulting process is *<u>M</u>arkov*. You may see other notations, like G for "general" (any distribution), D for "deterministic" (fixed interarrival time), etc.

Definition 4.1. *An* ***M / M / 1 queuing process*** *is a continuous-time Markov queuing process with the following characteristics:*

- *one server;*
- *unlimited capacity;*
- *Exponential interarrival times with arrival rate λ_A;*
- *Exponential service times with service rate λ_S;*
- *service times and interarrival times are independent.*

Remark 4.2. Let us recall that Exponential interarrival times imply a Poisson process of arrivals with parameter λ_A. This is a very popular model for telephone calls and many other types of arriving jobs.

We study M / M / 1 systems by starting with a B1SQS and letting its frame size Δ go to zero. We want to derive the steady-state distribution and other quantities of interest that measure the system's performance.
Recall that

$$\begin{aligned} p_A &= \lambda_A \Delta, \\ p_S &= \lambda_S \Delta \end{aligned}$$

and as Δ gets small, Δ^2 becomes practically negligible. Then the transition probabilities are

$$\begin{aligned}
p_{00} &= 1 - p_A \;=\; 1 - \lambda_A \Delta \\
p_{01} &= p_A \;=\; \lambda_A \Delta \\
p_{i,i-1} &= (1 - p_A) p_S \;=\; (1 - \lambda_A \Delta) \lambda_S \Delta \;\approx\; \lambda_S \Delta \\
p_{i,i} &= (1 - p_A)(1 - p_S) + p_A p_S \;\approx\; 1 - \lambda_A \Delta - \lambda_S \Delta \\
p_{i,i+1} &= p_A (1 - p_S) \;\approx\; \lambda_A \Delta,
\end{aligned}$$

for $i = 1, 2, \ldots$. The transition probability matrix becomes

$$P \;\approx\; \begin{bmatrix}
1 - \lambda_A \Delta & \lambda_A \Delta & 0 & \ldots & 0 & \ldots \\
\lambda_S \Delta & 1 - \lambda_A \Delta - \lambda_S \Delta & \lambda_A \Delta & \ldots & 0 & \ldots \\
0 & \lambda_S \Delta & 1 - \lambda_A \Delta - \lambda_S \Delta & \ldots & 0 & \ldots \\
0 & 0 & \lambda_S \Delta & \ldots & 0 & \ldots \\
\vdots & \vdots & \vdots & & \ddots &
\end{bmatrix}. \qquad (4.1)$$

Let us find the steady-state distribution from

$$\begin{cases} \pi P &= \pi \\ \sum_{i=0}^{\infty} \pi_i &= 1, \end{cases}$$

a system of infinitely many equations with infinitely many unknowns.

The first equation is

$$\begin{aligned}
[\pi_0 \;\; \pi_1 \;\; \pi_2 \;\; \ldots] \cdot \begin{bmatrix} 1 - \lambda_A \Delta \\ \lambda_S \Delta \\ 0 \\ \vdots \end{bmatrix} &= \pi_0, \;\; \text{i.e.} \\
(1 - \lambda_A \Delta) \pi_0 + \lambda_S \Delta \pi_1 &= \pi_0, \;\; \text{i.e.} \\
-\lambda_A \Delta \pi_0 + \lambda_S \Delta \pi_1 &= 0, \;\; \text{i.e.} \\
\lambda_S \pi_1 &= \lambda_A \pi_0.
\end{aligned}$$

This is called the *first balance equation*. From here, we get

$$\pi_1 \;=\; \frac{\lambda_A}{\lambda_S}\pi_0 \;=\; r\pi_0. \tag{4.2}$$

The second equation is

$$\begin{aligned}
[\pi_0 \;\; \pi_1 \;\; \pi_2 \;\; \ldots] \cdot \begin{bmatrix} \lambda_A\Delta \\ 1-\lambda_A\Delta-\lambda_S\Delta \\ \lambda_S\Delta \\ 0 \\ \vdots \end{bmatrix} &= \pi_1, \;\text{ i.e.} \\
\lambda_A\Delta\pi_0 + (1-\lambda_A\Delta-\lambda_S\Delta)\pi_1 + \lambda_S\Delta\pi_2 &= \pi_1, \;\text{ i.e.} \\
\lambda_A\Delta\pi_0 - \lambda_A\Delta\pi_1 - \lambda_S\Delta\pi_1 + \lambda_S\Delta\pi_2 &= 0, \;\text{ i.e. (since } \lambda_A\pi_0 = \lambda_S\pi_1) \\
-\lambda_A\Delta\pi_1 + \lambda_S\Delta\pi_2 &= 0, \;\text{ i.e.} \\
\lambda_S\pi_2 &= \lambda_A\pi_1.
\end{aligned}$$

Thus, we obtained the *second balance equation*, from which

$$\pi_2 \;=\; \frac{\lambda_A}{\lambda_S}\pi_1 \;=\; r\pi_1 \;=\; r^2\pi_0. \tag{4.3}$$

This trend of balance equations will continue exactly the same way, because every next column of matrix P is just the same as the previous column, only shifted down by 1 position. Thus, the general balance equation looks like

$$\lambda_S\pi_i \;=\; \lambda_A\pi_{i-1},$$

or

$$\pi_i \;=\; r\pi_{i-1}.$$

Combining it with the previous equations, we have

$$\pi_i \;=\; r\pi_{i-1} \;=\; r^2\pi_{i-2} \;=\; \ldots \;=\; r^i\pi_0, \; i = 1, 2, \ldots. \tag{4.4}$$

Finally, in the *normalizing equation* $\sum_{i=0}^{\infty} \pi_i = 1$, we get

$$\sum_{i=0}^{\infty} \pi_i \;=\; \sum_{i=0}^{\infty} r^i \pi_0 \;=\; 1. \tag{4.5}$$

Now, the Geometric series $\pi_0 \sum_{i=0}^{\infty} r^i$ is convergent if its ratio $r < 1$, in which case the series is equal to $\frac{\pi_0}{1-r}$. So, assuming the utilization r is less than 1, the last equation becomes

$$\begin{aligned} \frac{\pi_0}{1-r} &= 1, \text{ i.e.} \\ \pi_0 &= 1-r. \end{aligned}$$

Then the steady-state distribution of this queuing process is

$$\pi_i \;=\; r^i(1-r),\ i = 0, 1, \ldots \tag{4.6}$$

So the pdf of $X(t)$ (the total number of jobs in the system at time t) is

$$X(t) \begin{pmatrix} i \\ (1-r)r^i \end{pmatrix}_{i=0,1,\ldots}, \tag{4.7}$$

a $Geo(p)$ distribution with parameter $p = 1 - r$. Then, we know

$$\begin{aligned} E(X) &= \frac{q}{p} = \frac{r}{1-r}, \\ V(X) &= \frac{q}{p^2} = \frac{r}{(1-r)^2}. \end{aligned} \tag{4.8}$$

Evaluation of the system's performance

Now we can analyze the main parameters and distributions that characterize the queuing system, directly from the distribution (4.7).

Utilization

We know $r = \dfrac{\lambda_A}{\lambda_S}$. Now we also have $r = 1 - \pi_0$. What does that mean?

$$\pi_0 \;=\; P(X = 0) \;=\; P(\text{there are } \textit{no} \text{ jobs in the system}) \;=\; P(\text{the system is } \textit{idle})$$

and then

$$r \;=\; P(X > 0) \;=\; 1 - \pi_0 \;=\; 1 - P(\text{the system is idle}) \;=\; P(\text{the system is busy}). \tag{4.9}$$

So, we can say that r is the proportion of time when the system is put to work or *utilized*, hence the name **utilization**.

Obviously, the system is functional only if $r < 1$ (we used this for the convergence of the Geometric series). If $r \geq 1$, the system gets *overloaded.* Arrivals are too frequent compared to the service rate and the system cannot manage the incoming flow of jobs. The number of jobs will accumulate (unless it has a limited capacity) until the system will no longer function.

Waiting time

When a job arrives, it finds the system with X jobs in it. The new job waits in a queue, while those X jobs are being serviced. Thus, its waiting time is the sum of service times of X jobs

$$W \;=\; S_1 + S_2 + \ldots + S_X.$$

Recall that service times are Exponential and this distribution has the *memoryless* property (i.e. $P(S > x + y \mid S > x) = P(S > y)$). So, even if the first job has already started service, its *remaining* service time still has $Exp(\lambda_S)$ distribution, regardless of how long it has already been served or *when* its service time began. Then, the expected waiting time is

$$\begin{aligned} E(W) &= E(S_1 + S_2 + \ldots + S_X) \\ &= E(S \cdot X) \;=\; E(S)E(X) \\ &= \mu_S \cdot \frac{r}{1-r} \;=\; \frac{r}{\lambda_S(1-r)}. \end{aligned} \tag{4.10}$$

Remark 4.3.
1. At the step $E(S \cdot X) = E(S)E(X)$, we actually used the fact that service times are *independent* of the number of jobs in the system at that time.

2. The random variable W, the waiting time, is a rare example of a variable whose distribution is neither discrete nor continuous. Notice that it has a probability *distribution (mass)* function at 0, because

$$P(W=0) \;=\; P(\text{the system is } \textit{idle}) \;=\; 1-r$$

is the probability that the server is idle and available and there is no waiting time for a new job. On the other hand, for all $x>0$, it has a probability *density* function. Given any positive number of jobs $X=n$, the waiting time is the sum of n independent $Exp(\lambda_S)$ times, which is a $Gamma(n, 1/\lambda_S)$ random variable, so continuous. Such a distribution is called *mixed*.

Response time

Response time is the time a job spends in the system, from its arrival to its departure. It consists of waiting time (if any) and service time. So, the expected response time is then

$$\begin{aligned} E(R) &= E(W)+E(S) \\ &= \mu_S\cdot\frac{r}{1-r}+\mu_S \;=\; \mu_S\Big(\frac{r}{1-r}+1\Big) \\ &= \frac{\mu_S}{1-r} \;=\; \frac{1}{\lambda_S(1-r)}. \end{aligned} \tag{4.11}$$

Queue

The length of the queue is the number of waiting jobs

$$X_w \;=\; X-X_s.$$

As we have discussed in Example 3.2. (Lecture 8), the number of jobs being serviced, X_s, at any time is either 0 or 1 (because there is only one server), so it has a Bernoulli distribution with parameter

$$P(\text{the system/server is busy}) \;=\; r$$

and, hence,

$$E(X_s) \;=\; 0\cdot(1-r)+1\cdot r \;=\; r.$$

Then, the expected queue length is

$$\begin{aligned} E(X_w) &= E(X) - E(X_s) \\ &= \frac{r}{1-r} - r \;=\; r\Big(\frac{1}{1-r} - 1\Big) \\ &= \frac{r^2}{1-r}. \end{aligned} \tag{4.12}$$

So, to summarize:

Main performance characteristics of an M / M / 1 queuing system

- Expected number of jobs in the system

$$E(X) \;=\; \frac{r}{1-r},$$

- Expected queue length

$$E(X_w) \;=\; \frac{r^2}{1-r},$$

- Expected number of jobs being serviced

$$E(X_s) \;=\; r,$$

- Expected response time

$$E(R) \;=\; \frac{\mu_S}{1-r} \;=\; \frac{1}{\lambda_S(1-r)},$$

- Expected waiting time

$$E(W) \;=\; \frac{\mu_S r}{1-r} \;=\; \frac{r}{\lambda_S(1-r)},$$

- Expected service time

$$E(S) \;=\; \mu_S,$$

- Utilization

$$\begin{aligned} r &= P(X>0) \;=\; 1-\pi_0 \;=\; P(\text{system is busy}), \\ 1-r &= P(X=0) \;=\; \pi_0 \;=\; P(\text{system is idle}). \end{aligned}$$

Remark 4.4. Little's Law applies to M / M / 1 queuing systems and their components, the queue and the server. Assuming the system is functional ($r < 1$), all jobs go through the entire system, and thus, each component is subject to the same arrival rate λ_A. Notice that, indeed, we have

$$\begin{aligned}
\lambda_A E(R) &= \lambda_A \cdot \frac{1}{\lambda_S(1-r)} = \frac{r}{1-r} = E(X)\\
\lambda_A E(W) &= \lambda_A \cdot \frac{r}{\lambda_S(1-r)} = \frac{r^2}{1-r} = E(X_w)\\
\lambda_A E(S) &= \lambda_A \cdot \mu_S = \frac{\lambda_A}{\lambda_S} = r = E(X_s).
\end{aligned}$$

Example 4.5. Messages arrive to a communication center at random times according to a Poisson process, with an average of 5 messages per minute. They are transmitted through a single channel in the order they were received. On average, it takes 10 seconds to transmit a message. Compute the main performance characteristics for this center.

Solution. Recall that a Poisson process of arrivals *implies* Exponential interarrival times (and the other way around). Since messages are transmitted (i.e. jobs are being serviced) in the order they arrive, we also have Exponential service times. Thus, conditions of an M / M / 1 queuing system are satisfied.
We have

$$\begin{aligned}
\lambda_A &= 5 \,/\, \text{minute},\\
\mu_S &= 10\,\text{seconds} = \frac{1}{6}\,\text{minutes},\\
\lambda_S &= 6 \,/\, \text{minute},\\
r &= \frac{5}{6} = 0.833 < 1.
\end{aligned}$$

This is also the proportion of time, 83.3%, when the channel is busy and the probability of a nonzero waiting time. Then, we have:
Average number of messages stored in the system at any time

$$E(X) = \frac{r}{1-r} = 5.$$

Out of these, average number of messages waiting to be transmitted

$$E(X_w) \;=\; \frac{r^2}{1-r} \;=\; \frac{25}{6} \;\approx\; 4.17.$$

Average number of messages being transmitted

$$E(X_s) \;=\; r \;=\; \frac{5}{6} \;\approx\; 0.83.$$

When a message arrives to the center, its average waiting time until transmission is

$$E(W) \;=\; \frac{\mu_S r}{1-r} \;=\; \frac{r}{\lambda_S(1-r)} \;=\; \frac{5}{6}\,\text{minutes} \;=\; 50\,\text{seconds}.$$

The total time from arrival until the end of transmission has an average of

$$E(R) \;=\; \frac{\mu_S}{1-r} \;=\; \frac{1}{\lambda_S(1-r)} \;=\; 1\text{ minute} \;=\; 60\,\text{seconds}.$$

■

Notice that the utilization was less than 1, but not by much. Let us try a little bit of forecasting for this system and see what happens when the arrival rate is *slightly* increased, keeping the service rate the same.

Example 4.6. Suppose that next year the customer base of this transmission center is projected to increase by 10%, and thus, its incoming traffic rate, λ_A, increases by 10%, also. How will this affect the center's performance?

Solution. So, with that increase, we now have

$$\begin{aligned} \lambda_A &= 5 + 0.1 \cdot 5 \;=\; 5.5 \;=\; \frac{11}{2}/\ \text{minute}, \\ r &= \frac{11}{2} \cdot \frac{1}{6} \;=\; \frac{11}{12} \;<\; 1. \end{aligned}$$

The new system's performance parameters are

$$\begin{aligned}
E(X) &= \frac{r}{1-r} = 11 \text{ (compared to 5 before)},\\
E(X_w) &= \frac{r^2}{1-r} = 10.08 \text{ (compared to 4.17 before)},\\
E(X_s) &= r = 0.92 \text{ (compared to 0.83 before)},\\
E(W) &= \frac{\mu_S}{1-r} = 110 \text{ seconds (compared to 50 before)},\\
E(R) &= \frac{\mu_S}{1-r} = 120 \text{ seconds (compared to 60 before)}.
\end{aligned}$$

■

Notice that the response time, the waiting time, the average number of stored messages (and hence, the average required amount of memory) more than doubled when the number of customers increased by a mere 10%. The utilization r is still less than 1, but *dangerously close* to 1, when the system gets overloaded. For high values of r, various parameters of the system increase rapidly.

We could forecast the two-year future of the system, assuming a 10% increase of a customer base each year. It appears that during the second year the utilization will exceed 1, making the system unable to function. What solutions are there? Either increase the service rate (by using better equipment, higher internet speed, etc) **or** add more channels (servers) to help handle all the arriving messages, so have a *multiserver* queuing system. The new system will then have more than one channel-server, and it will be able to process more arriving jobs.

5 Multiserver Queuing Systems

We now consider queuing systems with several servers. We assume that each server can perform the same range of services; however, in general, some servers may be faster than others. Thus, the service times for different servers may potentially have different distributions.

When a job arrives, it either finds all servers busy serving jobs, or it finds one or several available servers. In the first case, the job will wait in a queue for its turn, whereas in the second case, it will be routed to one of the idle servers. A mechanism assigning jobs to available servers may be random, or it may be based on some rule.

The number of servers may be finite or infinite. A system with infinitely many servers can afford an unlimited number of concurrent users (e.g. any number of people can watch a TV channel simultaneously), so there is no queue, no waiting time.

As before (the single server case), we start with a discrete-time k-server queuing process (described in terms of Bernoulli trials), verify that the number of jobs in the system at time t is a Markov process, find its transition probability matrix, then get a continuous-time process by letting the frame size $\Delta \to 0$, compute its steady-state distribution π and finally use it to evaluate the system's long-term performance characteristics.

We treat a few common and analytically simple cases in detail. Sure enough, advanced theory goes further, but it is beyond the scope of this course. However, as mentioned previously, more complex and non-Markov queuing systems can be analyzed by Monte Carlo methods.

Remark 5.1. The utilization r no longer has to be less than 1. A system with k servers can handle k times the traffic of a single-server system; therefore, it will function with any $r < k$.

5.1 Bernoulli k-Server Queuing Process

Definition 5.2. *A* ***Bernoulli*** k***-server queuing process (BkSQP)*** *is a discrete-time queuing process with the following characteristics:*

- k *servers;*
- *unlimited capacity;*
- *arrivals occur according to a Binomial process with probability of a new arrival during each frame* p_A*;*
- *during each frame, each busy server completes its job with probability* p_S*, independently of the other servers and independently of the process of arrivals.*

So, all interarrival times and all service times are independent Shifted Geometric random variables (multiplied by the frame length Δ) with parameters p_A and p_S, respectively. Therefore, since Shifted Geometric variables have a memoryless property, again this process is Markov. The novelty is that now several jobs may finish during the same frame.

Suppose that $X_s = n$ jobs are currently getting service. During the next frame, each of them may finish and depart, independently of the other jobs. Then the number of departures, X_d, is the number of successes in n independent Bernoulli trials (with "success" meaning that a job's service is finished), and thus, has $Bino(n, p_S)$ distribution. Let us recall the pdf

$$X_d \begin{pmatrix} l \\ C_n^l p_S^l (1-p_S)^{n-l} \end{pmatrix}_{l=\overline{0,n}}.$$

This will help us compute the transition probability matrix.

Transition probability matrix

Suppose there are i jobs in the k-server system. Then, the number of busy servers, n, is the smaller of the number of jobs i and the total number of servers k,

$$n = \min\{i, k\}.$$

Indeed,
− for $i \le k$, the number of servers is sufficient for the current jobs, all jobs are getting service, and the number of departures X_d during the next frame is $Bino(i, p_S)$;
− for $i > k$, there are more jobs than servers. Then all k servers are busy, and the number of departures X_d during the next frame is $Bino(k, p_S)$.

Again, at most 1 job can arrive during each frame and that happens with probability p_A. Let us compute the transition probabilities

$$p_{ij} \;=\; P\big(X(t+\Delta) = j \mid X(t) = i\big).$$

We have

$$\begin{aligned}
p_{00} &= P\big(\text{ 0 arrivals }\big) \;=\; 1 - p_A,\\
p_{01} &= P\big(\text{ 1 arrival }\big) \;=\; p_A,\\
p_{i,i+1} &= P\big(\text{ 1 arrival } \;\cap\; \text{0 departures }\big) \;=\; p_A(1-p_S)^n,\\
p_{i,i+j} &= 0,\ \forall j > 1,\\
p_{i,i} &= P\big((\text{ 1 arrival } \;\cap\; \text{1 departure }) \cup (\text{ 0 arrivals } \;\cap\; \text{0 departures })\big)\\
&= p_A\, C_n^1\, p_S(1-p_S)^{n-1} + (1-p_A)(1-p_S)^n,\\
p_{i,i-1} &= P\big((\text{ 1 arrival } \;\cap\; \text{2 departures }) \cup (\text{ 0 arrivals } \;\cap\; \text{1 departure })\big)\\
&= p_A\, C_n^2\, p_S^2(1-p_S)^{n-2} + (1-p_A)\, C_n^1\, p_S(1-p_S)^{n-1},\\
p_{i,i-2} &= P\big((\text{ 1 arrival } \;\cap\; \text{3 departures }) \cup (\text{ 0 arrivals } \;\cap\; \text{2 departures })\big)\\
&= p_A\, C_n^3\, p_S^3(1-p_S)^{n-3} + (1-p_A)\, C_n^2\, p_S^2(1-p_S)^{n-2},\\
&\cdots\\
p_{i,i-n} &= P\big(\text{ 0 arrivals } \;\cap\; n \text{ departures }\big) \;=\; (1-p_A)p_S^n,\\
p_{i,i-j} &= 0,\ \forall j > n.
\end{aligned}$$

A transition diagram for a 2-server system is shown in Figure 1. The number of concurrent jobs can make transitions from i to $i-2, i-1, i$ and $i+1$.

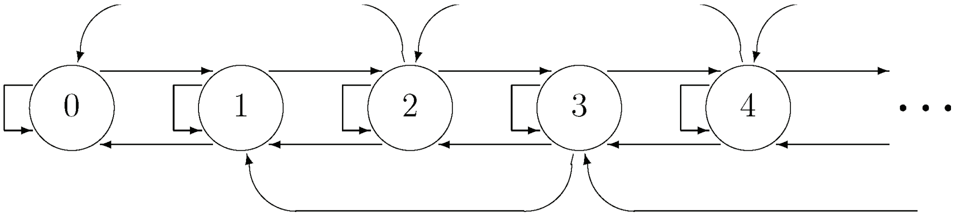

Fig. 1: Transition diagram for a B2SQS

For systems with a limited capacity $C < \infty$, the last probability changes:

$$\begin{aligned} p_{C,C} &= P\big((\text{ 1 arrival } \cap \text{ 1 departure }) \cup (\text{ 0 arrivals } \cap \text{ 0 departures }) \\ &\quad \cup (\text{ 1 arrival } \cap \text{ 0 departures })\big) \\ &= p_A\, C_n^1\, p_S(1-p_S)^{n-1} + (1-p_A)(1-p_S)^n + p_A(1-p_S)^n \\ &= np_Ap_S(1-p_S)^{n-1} + (1-p_S)^n. \end{aligned}$$

Example 5.3. There are two customer service representatives on duty answering customers' calls. When both of them are busy, two more customers may be "on hold", but other callers will receive a busy signal. Customers call at the rate of 1 call every 5 minutes and the average service takes 8 minutes. Assuming a B2SQS with limited capacity and 1-minute frames, find

a) the steady-state distribution of the number of concurrent jobs in the system;

b) the proportion of callers who get a busy signal;

c) the percentage of time each representative is busy, if each of them takes 50% of all calls.

Solution.

a) We have $k = 2$ servers, capacity $C = 4$ and parameters

$$\begin{aligned} \lambda_A &= 1/5 \text{ / minute } = 0.2\text{/ minute,} \\ \lambda_S &= 1/8 \text{ / minute } = 0.125\text{/ minute,} \\ \Delta &= 1 \text{ minute.} \end{aligned}$$

So,

$$\begin{aligned} p_A &= \lambda_A\Delta = 0.2,\ \ 1-p_A = 0.8, \\ p_S &= \lambda_S\Delta = 0.125,\ \ 1-p_S = 0.875. \end{aligned}$$

There are 5 states, $\{0,1,2,3,4\}$. The transition probability matrix is

$$P = \begin{bmatrix} 0.8000 & 0.2000 & 0 & 0 & 0 \\ 0.1000 & 0.7250 & 0.1750 & 0 & 0 \\ 0.0125 & 0.1781 & 0.6562 & 0.1531 & 0 \\ 0 & 0.0125 & 0.1781 & 0.6562 & 0.1531 \\ 0 & 0 & 0.0125 & 0.1781 & 0.8094 \end{bmatrix}.$$

The steady-state distribution (obtained the usual way) is

$$\pi = [\pi_0\ \pi_1\ \pi_2\ \pi_3\ \pi_4] = [0.1527\ \ 0.2753\ \ 0.2407\ \ 0.1837\ \ 0.1476].$$

b) Callers hear a busy signal when the system is full, i.e. $X = C = 4$. So that probability is

$$P(X = C) = \pi_4 = 0.1476.$$

c) Each representative is busy when there are $2, 3$ or 4 jobs in the system, plus a half of the time when there is 1 job (because there is a 50% chance that the other representative handles this job). This totals

$$\pi_2 + \pi_3 + \pi_4 + 0.5\pi_1 = 0.709 \text{ or } 70.9\% \text{ of the time.}$$

■