Short Review

So, we have the following parameters and random variables that describe the performance of a stationary queuing system.

$$\begin{aligned}
\lambda_A &= \frac{E\big(A(t)\big)}{t} = \text{arrival rate} \\
\lambda_S &= \text{service rate} \\
\mu_A &= 1/\lambda_A = \text{mean interarrival time} \\
\mu_S &= 1/\lambda_S = \text{mean service time} \\
r &= \lambda_A/\lambda_S = \mu_S/\mu_A = \textbf{utilization} \text{ (arrival-to-service ratio)} \\
X_s(t) &= \text{number of jobs receiving service at time } t \\
X_w(t) &= \text{number of jobs waiting in a queue} \\
X(t) &= X_s(t) + X_w(t) = \text{total number of jobs in the system at time } t \\
S &= \text{service time for a job} \\
W &= \text{waiting time for a job} \\
R &= S + W = \textbf{response time} \text{ for a job} \\
&= \text{total time a job spends in the system, from arrival to departure}
\end{aligned}$$

2 Little's Law

This is one of the most important results in queuing theory. It was first established and used by Philip. M. Morse and other researchers in the 1950's. In 1954, Morse published it, but was not able to prove it, so he challenged his readers to find a situation where it did not hold. **John D. C. Little**, Professor Emeritus at the MIT Sloan School of Management (since 1962), proved it in 1961. Later, in the 1990's and 2000's there were more developments and versions both in theory and in practice.

Little's Law gives a simple relationship between the expected number of jobs, the expected response time, and the arrival rate. It is valid for any stationary queuing system.

Proposition 2.1 (Little's Law).

$$E(X) = \lambda_A E(R). \tag{2.1}$$

Proof. We make a diagram (see Figure 1), with time t on the x-axis and number of arrivals $A(t)$ on

the y-axis. Each job is represented by a rectangle with height 1 and length stretching between its arrival and its departure time.

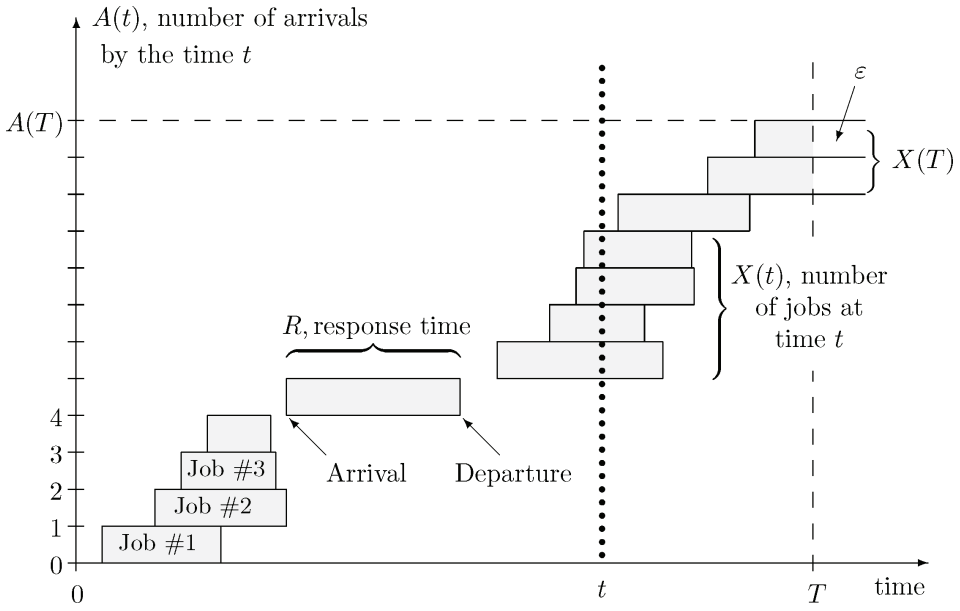

Fig. 1: Illustration of Littles's Law

We will compute the shaded area in two ways, geometrically (with areas) and analytically (with integrals).

Geometrically: add the areas of the rectangles. For job $\#k$, the area of the rectangle is its base length (since its height is 1), so the difference between departure and arrival times, i.e. *response time*:

$$\text{Area (rectangle } k) \;=\; \text{Departure time} \;-\; \text{Arrival time} \;=\; R_k.$$

By the time T ($k = \overline{1, A(T)}$), there are $A(T)$ arrivals. Among them, $X(T)$ jobs remain in the system at time T. *Not all* of these jobs are completed by time T, a portion of them will be completed *after* time T, call that portion ε. Then, the total shaded area is

$$\text{Shaded area} \;=\; \sum_{k=1}^{A(T)} R_k - \varepsilon.$$

Analytically: recall from Calculus that every area can be computed by integration of the *cross-section*. So we let t run from 0 to T and integrate the cross-section of the shaded region at t. As seen on the picture, the length of this cross-section is $X(t)$, the number of jobs in the system at time t. Hence,

$$\text{Shaded area} \;=\; \int_0^T X(t)\,dt.$$

So, we have

$$\int_0^T X(t)\,dt \;=\; \sum_{k=1}^{A(T)} R_k - \varepsilon. \tag{2.2}$$

Take expectations on both side, divide by T and then let $T \to \infty$. We compute separately the LHS (left-hand side) and RHS of (2.2). Recall from Calculus that the *mean value* of a function f on an interval $[a, b]$ is defined as

$$\frac{1}{b-a}\int_a^b f(x)dx.$$

So, in (2.2), we have

$$\begin{aligned}
\text{LHS} &= \lim_{T\to\infty} \frac{1}{T} E\left(\int_0^T X(t)\,dt\right) \\
&= \lim_{T\to\infty} E\left(\frac{1}{T}\int_0^T X(t)\,dt\right) \\
&= \lim_{T\to\infty} E(X) \;=\; E(X).
\end{aligned}$$

On the other side, we have

$$\begin{aligned}
\text{RHS} &= \lim_{T\to\infty} \frac{1}{T} E\left(\sum_{k=1}^{A(T)} R_k - \varepsilon\right) \\
&= \lim_{T\to\infty} \left(\frac{1}{T} E\Big(\sum_{k=1}^{A(T)} R_k\Big) - \frac{\varepsilon}{T}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{E(A(T))} E\big(R_k\big) - 0 \\
&= \lim_{T\to\infty} \frac{E\big(A(T)\big)}{T} E(R) \;=\; \lambda_A E(R),
\end{aligned}
$$

since $\lambda_A = \dfrac{E\big(A(T)\big)}{T}$.
Thus, we have

$$
E(X) \;=\; \lambda_A E(R).
$$

□

Example 2.2. A person walks into a bank at 10:00 a.m. He counts a total of 10 customers in the bank and assumes that this is the typical, average number. He also notices that, on average, new customers walk in every 2 minutes. When should he expect to finish his business and leave the bank?

Solution. The average number of customers in the bank, i.e. the *expected number of jobs* in the system, is

$$
E(X) \;=\; 10.
$$

On average, new customers walk in every 2 minutes, that is the *mean interarrival time*,

$$
\begin{aligned}
\mu_A &= 2 \text{ minutes, so} \\
\lambda_A &= 1/\mu_A \;=\; 1/2 \text{ / minute.}
\end{aligned}
$$

Then the amount of time he is expected to spend in the bank, i.e. the *expected response time* is, by Little's Law,

$$
E(R) \;=\; \frac{1}{\lambda_A} E(X) \;=\; \mu_A E(X) \;=\; 20 \text{ minutes.}
$$

Thus, he should expect to leave at 10:20. ■

Remark 2.3.

1. Little's Law is universal, it applies to any *stationary* queuing system *and* even to the system's components, the queue and the servers.

Thus, we can immediately deduce the equations for the number of waiting jobs,

$$E(X_w) \;=\; \lambda_A E(W),$$

and for the number of jobs currently receiving service,

$$E(X_s) \;=\; \lambda_A E(S).$$

Note that the *same* arrival rate, λ_A, applies to the components, as for the entire queuing system.

2. Looking at the second equation above, $E(S)$ is the expected or the *mean* service time, i.e. μ_S. So, we have

$$E(X_s) \;=\; \lambda_A \cdot \mu_S \;=\; \frac{\lambda_A}{\lambda_S} \;=\; r,$$

so we just obtained another important definition of *utilization*, which also justifies its name.

Definition 2.4. ***Utilization*** r *is the expected number of jobs receiving service at any given time.*

Little's Law only relates *expectations* of the number of jobs and their response time. In the remaining sections of this chapter, we evaluate the *entire distribution* of $X(t)$, which will help us compute various probabilities and expectations of interest. These quantities will describe and predict the performance of a queuing system.

Definition 2.5. *The number of jobs in a queuing system,* $X(t)$*, is called a* ***queuing process****.*

Since $X(t)$ is the *number* of jobs in the system, it is clearly a *discrete-state* stochastic process. The time set may be discrete or continuous and we will look at both cases.

In general, a queuing process is *not* a counting process because jobs arrive and depart, therefore, their number may increase and decrease, whereas any counting process is nondecreasing. However, we will use counting processes to model arrivals and service of jobs.

Another aspect is the number of servers in a queuing system, one or more. Again, we will consider both situations (in the end, even considering the case where the number of servers goes to infinity).

3 Bernoulli Single-Server Queuing Process

Definition 3.1. *A* ***Bernoulli single-server queuing process (B1SQP)*** *is a discrete-time queuing process with the following characteristics:*

- *one server;*
- *unlimited capacity;*
- *arrivals occur according to a Binomial process, and the probability of a new arrival during each frame is p_A;*
- *the probability of a service completion (and thus, a departure) during each frame is p_S, provided that there is at least one job in the system at the beginning of the frame;*
- *service times and interarrival times are independent;*
- *jobs are being serviced in the order of their arrival.*

Examples of processes modeled by a B1SQS include: customers waiting at an ATM, cars coming to a car wash or a gas station (with only one service station), documents arriving to a printer, clients calling a customer service representative, etc.

Everything we know about Binomial counting processes applies to job arrivals and to service completions, as long as there is at least one job in the system. So, we know that:
– the number of arrivals by time t, $A(t)$, is a Binomial counting process with probability p_A;
– the number of jobs being serviced at time t, $X_s(t)$, is a Binomial counting process with probability p_S (when there is at least one job in the system);
– there is a Shifted Geometric(p_A) number of frames between successive arrivals;
– there is a Shifted Geometric(p_S) number of frames between successive service completions (i.e. each service takes a $SGeo(p_S)$ number of frames);
– $p_A = \lambda_A \Delta$;
– $p_S = \lambda_S \Delta$.

Markov property

Obviously, a B1SQS is a Markov chain. Since the probabilities p_A and p_S never change, it is also a *homogeneous* Markov chain. The number of jobs in the system increases by 1 with every arrival and decreases by 1 with each departure. Conditions of a Binomial process guarantee that *at most one arrival* and *at most one departure* may occur during each frame.

The states are $\{0, 1, \ldots\}$ (number of jobs in the system). Let us find the transition probabilities.

$$\begin{aligned} p_{00} &= P\big(\text{ 0 arrivals }\big) = 1 - p_A \\ p_{01} &= P\big(\text{ 1 arrival }\big) = p_A. \end{aligned}$$

In general, for $i \geq 1$,

$$\begin{aligned}
p_{i,i-1} &= P\big(\text{ 0 arrivals and 1 departure }\big) = P\big(\{0 \text{ arrivals}\} \cap \{1 \text{ departure}\}\big) = (1-p_A)p_S \\
p_{i,i} &= P\Big(\big(\{0 \text{ arrivals}\} \cap \{0 \text{ departures}\}\big) \cup \big(\{1 \text{ arrival}\} \cap \{1 \text{ departure}\}\big)\Big) \\
&= P\big(\{0 \text{ arrivals}\} \cap \{0 \text{ departures}\}\big) + P\big(\{1 \text{ arrival}\} \cap \{1 \text{ departure}\}\big) \\
&= (1-p_A)(1-p_S) + p_A p_S \\
p_{i,i+1} &= P\big(\{1 \text{ arrival}\} \cap \{0 \text{ departures}\}\big) = p_A(1-p_S)
\end{aligned}$$

All the other transition probabilities are 0, since the number of jobs cannot change by more than 1 in any single frame.
So, the transition probability matrix is

$$P = \begin{bmatrix}
1-p_A & p_A & 0 & \dots & 0 & \dots \\
(1-p_A)p_S & \begin{matrix}(1-p_A)(1-p_S)\\ +p_A p_S\end{matrix} & p_A(1-p_S) & \dots & 0 & \dots \\
0 & (1-p_A)p_S & \begin{matrix}(1-p_A)(1-p_S)\\ +p_A p_S\end{matrix} & \dots & 0 & \dots \\
0 & 0 & (1-p_A)p_S & \dots & 0 & \dots \\
\vdots & \vdots & \vdots & & \ddots &
\end{bmatrix}, \quad (3.1)$$

an $\infty \times \infty$ tridiagonal matrix. Below, see the transition diagram.

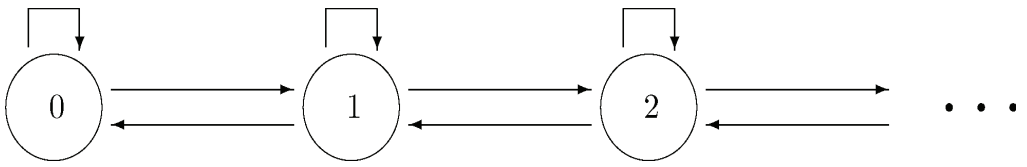

Fig. 2: Transition diagram for a B1SQS

The transition probability matrix may be used, for example, to simulate this queuing system and study its performance, as we did with general Markov chains. One can also compute k-step

transition probabilities and predict the load of a server or the length of a queue at any time in the future.

Example 3.2. Jobs (documents) are sent to a printer at the rate of 20 per hour. It takes an average of 40 seconds to print a document. Currently, the printer is printing a job, and there is another job stored in a queue. Assume a B1SQS with 20-second frames is modeling this printer.
a) Compute the probability that the printer will be idle in 2 minutes.
b) Find the expected total number of jobs in the system in 2 minutes.
c) What is the expected length of the queue in 2 minutes?
d) What is the expected waiting time for a document in 2 minutes?
e) On average, how long does it take to get the printout of a document in 2 minutes?

Solution.
First off, let us note that any printer represents a single-server queuing system, because it can process only one job at a time while other jobs are waiting in a queue.

Now, parameters are given in hours, in minutes and in seconds, so let us choose the "middle" one, i.e., express everything in minutes. We are given:

$$\begin{aligned}
\lambda_A &= 20 \text{ / hour } = 1/3 \text{ / minute, so} \\
\mu_A &= 3 \text{ minutes,} \\
\mu_S &= 40 \text{ seconds } = 2/3 \text{ minutes, so} \\
\lambda_S &= 1/\mu_S = 3/2 \text{ / minute,} \\
\Delta &= 20 \text{ seconds } = 1/3 \text{ minutes.}
\end{aligned}$$

Then

$$\begin{aligned}
p_A &= \lambda_A \Delta = 1/9, \ 1 - p_A = 8/9, \\
p_S &= \lambda_S \Delta = 1/2, \ 1 - p_S = 1/2.
\end{aligned}$$

The transition probabilities are

$$\begin{aligned}
p_{00} &= 1 - p_A = 8/9, \\
p_{01} &= p_A = 1/9, \\
p_{i,i-1} &= (1-p_A)p_S = 8/9 \cdot 1/2 = 4/9, \\
p_{i,i} &= (1-p_A)(1-p_S) + p_A p_S = 8/9 \cdot 1/2 + 1/9 \cdot 1/2 = 1/2, \\
p_{i,i+1} &= p_A(1-p_S) = 1/9 \cdot 1/2 = 1/18.
\end{aligned}$$

Hence,

$$P = \begin{bmatrix}
8/9 & 1/9 & 0 & 0 & \dots \\
4/9 & 1/2 & 1/18 & 0 & \dots \\
0 & 4/9 & 1/2 & 1/18 & \dots \\
0 & 0 & 4/9 & 1/2 & \dots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{bmatrix}.$$

Now, in $t = 2$ minutes, there are $n = \dfrac{t}{\Delta} = 6$ frames. The distribution of X after 6 frames is

$$P_6 = P_0 \cdot P^6.$$

The initial distribution (2 jobs in the system) is

$$P_0 = [0 \;\; 0 \;\; 1 \;\; 0 \;\; \dots].$$

Here is an interesting problem. How do we deal with matrix P that has *infinitely* many rows and columns? Fortunately, we only need a small portion of this matrix. In the course of 6 frames, the number of jobs in the system, $X(t)$, can change by 6 *at most* (see Figure 2), i.e. it can reach a *maximum* of 8. Thus, it is sufficient to consider the first 9 rows and 9 columns of P *only*, corresponding to states $\{0, 1, \dots, 8\}$.

So, we consider P_0 (and P_6) as having length 9,

$$P_0 = [0 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0]$$

and P a 9×9 matrix,

$$P \;=\; \left[\begin{array}{ccccccccc} 8/9 & 1/9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4/9 & 1/2 & 1/18 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4/9 & 1/2 & 1/18 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4/9 & 1/2 & 1/18 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4/9 & 1/2 & 1/18 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4/9 & 1/2 & 1/18 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4/9 & 1/2 & 1/18 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4/9 & 1/2 & 1/18 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4/9 & 1/2 \end{array}\right].$$

In 2 minutes (6 frames), the distribution will be

$$P_6 \;=\; P_0 \cdot P^6 \;=\; [\,0.6436 \quad 0.25 \quad 0.0799 \quad 0.0218 \quad 0.0041 \quad 0.0005 \quad 0 \quad 0 \quad 0\,].$$

Now we can answer all the questions.

a) The probability that the printer is idle after 2 minutes is the probability of 0 jobs in the system at that time, i.e.

$$P_6(0) \;=\; 0.6436.$$

b) In 2 minutes, the total number of jobs in the system, $X(2)$, has pdf

$$X\begin{pmatrix} 0 & \dots & 8 \\ & P_6 & \end{pmatrix} \;=\; \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 0.6436 & 0.25 & 0.0799 & 0.0218 & 0.0041 & 0.0005 & 0 & 0 & 0 \end{pmatrix},$$

so

$$E(X) \;=\; \sum_{k=0}^{8} kP_6(k) \;=\; 0.4944 \text{ jobs.}$$

c) Out of the X jobs in the system above, X_w jobs are waiting in a queue and X_s are being serviced. The expected length of the queue is the expected value

$$\begin{aligned} E(X_w) &= E(X - X_s) \\ &= E(X) - E(X_s). \end{aligned}$$

We have found $E(X)$, so let us turn our attention to X_s.

Since the server (printer) can process *at most* 1 job at a time, X_s is either 0 or 1, i.e. it has a Bernoulli distribution. With what parameter p? The parameter is the probability of "success", in this case, the probability that the system is working, so, *not* idle:

$$p \;=\; P(\text{printer is busy}) \;=\; 1 - P(\text{printer is idle}) \;=\; 1 - 0.6436 \;=\; 0.3564.$$

So the pdf of X_s is

$$X_s \begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

and its expected value

$$E(X_s) \;=\; p \;=\; 0.3564.$$

Then the expected queue length is

$$E(X_w) \;=\; E(X) - E(X_s) \;=\; 0.4944 - 0.3564 \;=\; 0.138 \text{ jobs.}$$

d) The expected waiting time for a document is $E(W)$. By Little's Law, we have

$$\begin{aligned} E(W) \;&=\; \frac{1}{\lambda_A} E(X_w) \;=\; \mu_A E(X_w) \\ &=\; 3 \cdot 0.138 \text{ minutes} \;=\; 0.414 \text{ minutes} \\ &=\; 24.84 \text{ seconds.} \end{aligned}$$

e) This is the expected total time the job spends in the system, i.e., the expected *response time* of a job, in 2 minutes. Again, by Little's Law, that number is

$$E(R) \;=\; \frac{1}{\lambda_A} E(X) \;=\; 3 \cdot 0.4944 \text{ minutes} \;=\; 1.4832 \text{ minutes.}$$

■

Remark 3.3.

1. A B1SQS is an *irregular* Markov chain. *Any* k-step transition probability matrix contains zeros because a k-step transition from 0 to $k+1$ is *impossible*. It requires at least $k+1$ arrivals, and this cannot happen by the conditions of the Binomial process of arrivals.

2. However, without the Binomial counting process restrictions, it can be shown that any system whose service rate exceeds the arrival rate (i.e., jobs can be served *faster* than they arrive, so there

is no overload),

$$\lambda_S > \lambda_A,$$

does have a steady-state distribution. Its computation is possible, despite the infinite dimension of P, but a little complicated. Instead, we will compute the steady-state distribution of a *continuous* queuing process, obtained by letting the frame size $\Delta \to 0$.

Systems with limited capacity

As we have seen, the number of jobs in a B1SQS may potentially reach any value. However, in practice, many systems have limited resources for storing jobs. Then, there is a maximum number of jobs C that can possibly be in the system simultaneously. This number is called **capacity**. As examples, consider people going to a restaurant, cars entering a parking lot, customers going into a bank, etc.

How does the situation change for a queuing system with a limited capacity $C < \infty$? Not much, but it *does* make a difference. Up until the capacity C is reached, the system operates as before. Things change when $X = C$. At this time, the system is full, so it can accept new jobs into its queue *only* if some job departs. We have

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

This Markov chain has states $0, 1, ..., C$, its transition probability matrix is finite, and it is *regular* (any state can be reached in C steps). The transition diagram for a system with limited capacity is given in Figure 3.

Example 3.4. A customer service representative has a telephone with 2 lines, so she can talk to a customer while having another one "on hold". Suppose the representative gets an average of 10 calls per hour and the average phone conversation lasts 4 minutes. Assuming a B1SQS with 1-minute frames find the steady-state distribution and interpret it.

Solution.

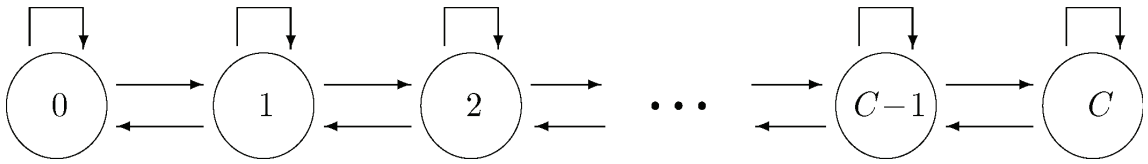

Fig. 3: Transition diagram for a B1SQS with limited capacity C

Obviously, this is a system with limited capacity $C = 2$. When the capacity is reached and someone tries to call, (s)he will get a busy signal or voice mail.
This Markov chain $X(t)$ has 3 states, $0, 1, 2$ and we have:

$$\begin{aligned} \lambda_A &= 10 \text{ / hour } = 1/6 \text{ / minute,} \\ \mu_S &= 4 \text{ minutes, so} \\ \lambda_S &= 1/\mu_S = 1/4 \text{ / minute,} \\ \Delta &= 1 \text{ minute,} \end{aligned}$$

so,

$$\begin{aligned} p_A &= \lambda_A \Delta = 1/6, \ 1 - p_A = 5/6, \\ p_S &= \lambda_S \Delta = 1/4, \ 1 - p_S = 3/4. \end{aligned}$$

The transition probability matrix, of dimensions 3×3, is

$$P = \begin{bmatrix} 1 - p_A & p_A & 0 \\ (1 - p_A)p_S & (1 - p_A)(1 - p_S) + p_A p_S & p_A(1 - p_S) \\ 0 & (1 - p_A)p_S & 1 - (1 - p_A)p_S \end{bmatrix} = \begin{bmatrix} 5/6 & 1/6 & 0 \\ 5/24 & 2/3 & 1/8 \\ 0 & 5/24 & 19/24 \end{bmatrix}.$$

The steady-state distribution is found, as usually, from the system

$$\begin{cases} \pi P &= \pi \\ \sum_{k=0}^{2} \pi_k &= 1, \end{cases}$$

which leads to

$$\begin{cases} \pi_0 - \dfrac{5}{4}\pi_1 & = & 0 \\ \dfrac{3}{5}\pi_1 - \pi_2 & = & 0 \\ \pi_0 + \pi_1 + \pi_2 & = & 1, \end{cases}$$

with solution

$$\begin{aligned} \pi_0 &= \frac{25}{57} \approx 0.439, \\ \pi_1 &= \frac{20}{57} \approx 0.351, \\ \pi_2 &= \frac{12}{57} \approx 0.21. \end{aligned}$$

Interpretation: 43.9% of the time the representative is not talking on the phone (and, implicitly, there is no one on hold), 35.1% of the time she talks to a customer, but the second line is open, and 21% of the time both lines are busy (one talking, one holding) and no new calls can get through.

■