Asian European Journal of Probability and Statistics
Vol. 1, No-1, (2024), pp. 29-44

Limiting Behavior of Two M/M/1 Queues Sharing a Common Waiting Room

William Boler[a], William Henby[b], Jyotirmoy Sarkar[c], Gregory Vaughan[d]
[a,d]*Purdue University at Indianapolis*, [b,c]*Indiana University Indianapolis*

ARTICLE HISTORY
Compiled June 14, 2024

Received 26 December 2023; Accepted 18 April 2024

ABSTRACT
Customers arrive independently to receive service at either of two businesses, each managed by a single server, according to Poisson processes with different rates. The customers' service times are independent and exponentially distributed with server-specific rates. The customers of both businesses wait at a limited-capacity common service facility. When new arrivals to either business find the facility full, they depart for good without waiting. But if the facility is not full, customers are admitted anytime during the business hours each day and are served by their respective servers even if that takes them past the closing time.

We study the limiting behavior of these two M/M/1 queuing systems sharing a common facility with capacity constraint N to answer these questions relevant to the servers: (Q1) What percentage of customers does each server lose? (Q2) What percentage of time does each server remain idle during regular business hours? (Q3) At closing time, how many customers are waiting to be served by each server?

KEYWORDS
Memoryless property; semi-Markov process; embedded Markov chain; stationary distribution; expected sojourn time

1. Introduction

Because real estate is so expensive in a crowded city, some businesses may choose to team up to share the same facility and lower their cost of operation. The businesses need designated areas and specialized instruments to operate, but customers can wait in a common area equipped with essential amenities but with limited capacity. For example, the following businesses may agree to share the same facility: an insurance agent and a mortgage agent, a travel agent and a financial advisor, a psychologist and a therapist, or a lawyer and a real estate broker. Such cooperation between professions that are socioeconomically compatible yet non-competitive is likely to be mutually beneficial and sustainable.

Businesses contemplating sharing a common facility with capacity constraint N would like answers to these questions: (Q1) What percentage of customers does each server lose? (Q2) What percentage of time does each server remain idle during regular

CONTACT Jyotirmoy Sarkar[c]. Email: jsarkar@iu.edu

business hours? (Q3) At closing time, how many customers are waiting to be served by each server?

The paper is organized as follows: In Section 2, we model the stochastic evolution of the arrival, wait, service, and departure of customers in the service facility in terms of a continuous-time stochastic process (CTSP). In Section 3, we derive expressions of the limiting distributions of the CTSP. Section 4 numerically evaluates limiting distributions. Section 5 presents an alternative numerical computation. Section 6 answers the questions listed in the previous paragraph. Section 7 concludes the paper with a summary and some directions for future research.

2. Description of the CTSP

As time progresses, customers arrive at a service station with a common waiting area to receive service from Business A or B (but not both). In the next three subsections: (1) we shall model the arrival times and the service times; (2) we shall describe the state of the stochastic process according to the numbers of customers who came to server A and server B, respectively; and (3) we shall identify an embedded discrete-time stochastic process (DTSP) by focusing on the epochs when a customer arrives or when a customer's service is completed and the customer leaves the facility. This last step explains the CTSP as a semi-Markov process.

2.1. Modeling inter-arrival times and service times

Suppose that businesses A and B share the same facility. Customers to these businesses arrive at the facility independently according to Poisson processes with rates λ_1 and λ_2, respectively. In particular, the inter-arrival times between successive arrivals to business A are independent exponential(λ_1) variables (with mean $1/\lambda_1$), and the inter-arrival times to business B are independent exponential(λ_2) variables. Also, assume that the service times of business A are independent exponential(μ_1) variables, service times of business B are independent exponential(μ_2) variables, and these two sequences of service times are independent of each other.

Recall that the exponential distribution has the memoryless property: no matter how much time has already elapsed, the remaining time is still exponentially distributed with the same parameter! That is, if X has exponential(λ) distribution, then $P\{X > t + s|X > s\} = P\{X > t\}$ for all $s, t > 0$. Also, if X_1 and X_2 are independent exponential variables with rates λ_1 and λ_2 respectively, then $Z = \min\{X_1, X_2\}$ is an exponential($\lambda_1 + \lambda_2$) variable, and $P\{Z = X_1\} = \lambda_1/(\lambda_1 + \lambda_2) = 1 - P\{Z = X_2\}$.

As a consequence of these properties of independent exponential variables, not only at the epoch of arrival or the epoch of departure (immediately after service is over) of a customer but also at *any time* the future prospect of the evolution of the process has the same distribution as that at the latest arrival or departure epoch. This future prospect changes only at the epoch of the next arrival or departure. The duration until the next arrival or departure depends only on the present state defined by the pair of numbers of customers in the service facility either being served or waiting to be served by the two servers (or businesses).

2.2. State space and transition rates

Suppose that businesses A and B share the same waiting room with capacity $N \geq 1$. For convenience, imagine that there are N chairs available — chairs for clients being served, chairs for clients waiting to be served, and empty chairs, if any. The state space is given by

$$\mathcal{S} = \{(i, j) : 0 \leq i, j \leq i + j \leq N\} \tag{1}$$

consisting of $1 + 2 + 3 \ldots + N + (N + 1) = (N + 1)(N + 2)/2 = \binom{N+2}{2}$ states. For convenience of sorting the states in a well-defined order, we also label state (i, j) as

$$l = l(i, j) = \binom{i + j + 1}{2} + j + 1 = \binom{i + j + 2}{2} - i. \tag{2}$$

For $N = 1$, only one business can operate at a time, and if so [that is, if the state is $(1, 0)$ or $(0, 1)$], all arrivals to either business are lost for good, but when both businesses are idling [or the state is $(0, 0)$], a new arrival to either business can enter the facility and receive service from the intended server immediately.

For $N = 2$, the following situations are possible: (1) both businesses are operating simultaneously [state $(1, 1)$]; (2) one business is operating with another client waiting for the same business [state $(2, 0)$ or $(0, 2)$]; (3) one business is operating with no one waiting [state $(1, 0)$ or $(0, 1)$]; (4) both businesses are idling [state $(0, 0)$]. In cases (1) and (2), all new arrivals to either business are lost for good because, finding no available seat in the waiting room, they go to a competitor to receive service. In cases (3) and (4), any new arrival to either server can enter the service facility and she will receive service immediately if her server is free, or join the queue if her server is serving a previously arriving customer.

Figure 1 shows the state space and the transition rates for $N = 5$ as an illustration. Readers should study it carefully so that they can restrict it or generalize it to all $N \geq 1$. It also classifies the states into one of seven categories according as the total transition rate in effect in that state.

A thick arc represents the arrival of a customer (a right arc or a top arc indicates that the newly arriving customer needs service from either business A or B, respectively). A thin arc represents the departure of a customer (a left arc or a down arc indicates that a customer's service has been completed by business A or B, respectively). Note also that the states are classified into seven categories (different shades of gray) according as the total transition rate in effect is

$$\begin{aligned} &a = \lambda_1 + \lambda_2, \quad b = \lambda_1 + \lambda_2 + \mu_1, \quad c = \lambda_1 + \lambda_2 + \mu_2, \\ &d = \lambda_1 + \lambda_2 + \mu_1 + \mu_2, \quad e = \mu_1 + \mu_2, \quad \mu_1, \quad \mu_2. \end{aligned} \tag{3}$$

We hope an attentive reader can restrict or generalize Figure 1 to any $N \geq 1$. The resultant CTSP is denoted by the symbol $2(\mathrm{M/M/1})/N$, representing a paired M/M/1 queuing systems with shared capacity constraint N. The goal is to find the limiting probability θ_l that the stochastic process is in state $l = (i, j) \in \mathcal{S}$.

We found this problem posed in [5] as Exercise 5.24. To the best of our knowledge the solution is not published anywhere.

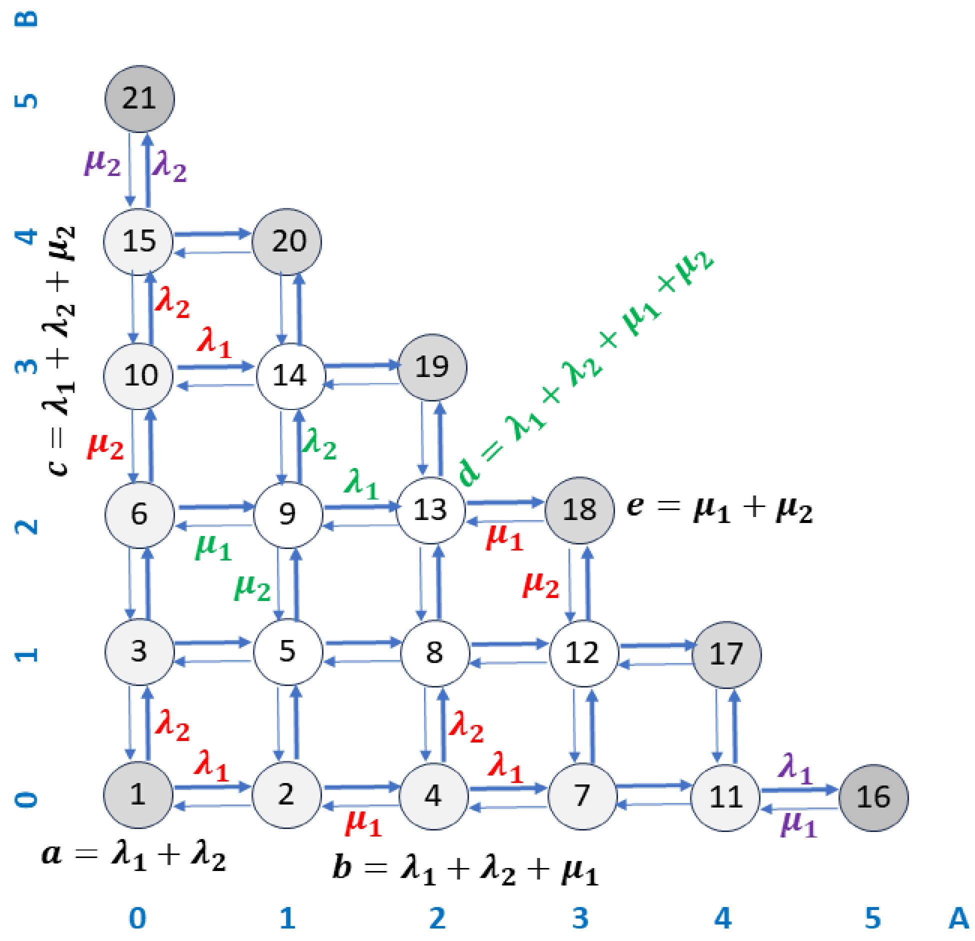

Figure 1. States and transition rates when capacity is $N = 5$.

2.3. *The CTSP is a semi-Markov process*

First, let us focus on the epochs of transitions from one state to another when a customer arrives or departs. The corresponding DTSP is a Markov chain because the transition probabilities depend only on the current state (explained in the next paragraph) and not on the history of how the process arrived at the current state. Second, the sojourn time in each state has a distribution dependent on the current state (but not on the next state).

For example, for $N = 5$, the sojourn time in state 5, has the same distribution as that of $\min\{X_1, X_2, Y_1, Y_2\}$ where X_k has exponential(λ_k) distribution and Y_k has exponential(μ_k) distribution and all four random variables are independent. Likewise, the sojourn time in state 6 has the same distribution as that of $\min\{X_1, X_2, Y_2\}$. The sojourn time in state 16 is exponential(μ_1).

Given the transition rates in effect in each state $l = (i, j)$, the transition probabilities for the DTSP are found by dividing each rate by the total rate. This is because the actual transition is determined by the minimum of several independent exponential variables with the given rates. Once the transition has happened, by the memoryless property of exponential variables, the next transition is determined by the (possibly new) transition rates in effect in the new state. The unique values of all transition probabilities are defined by the following constants

$$
\begin{array}{llll}
a_1 = \lambda_1/a, & a_2 = \lambda_2/a; & & \\
b_1 = \lambda_1/b, & b_2 = \lambda_2/b, & b_3 = \mu_1/b; & \\
c_1 = \lambda_1/c, & c_2 = \lambda_2/c, & c_3 = \mu_2/c; & \\
d_1 = \lambda_1/d, & d_2 = \lambda_2/d, & d_3 = \mu_1/d, & d_4 = \mu_2/d; \\
e_1 = \mu_1/e, & e_2 = \mu_2/e. & &
\end{array}
\tag{4}
$$

From the definitions of a–e given in (3), it follows that

$$1 = a_1 + a_2 = b_1 + b_2 + b_3 = c_1 + c_2 + c_3 = d_1 + d_2 + d_3 + d_4 = e_1 + e_2.$$

For example, for $N = 1$ and $N = 2$, the transition rates are (we omit 0's so that we can better focus on the non-zero rates only)

$$
R_1 = \left[\begin{array}{c|cc}
 & \lambda_1 & \lambda_2 \\
\hline
\mu_1 & & \\
\mu_2 & &
\end{array}\right]
\quad \text{and} \quad
R_2 = \left[\begin{array}{c|cc|ccc}
 & \lambda_1 & \lambda_2 & & & \\
\hline
\mu_1 & & & \lambda_1 & \lambda_2 & \\
\mu_2 & & & & \lambda_1 & \lambda_2 \\
\hline
 & \mu_1 & & & & \\
 & \mu_2 & \mu_1 & & & \\
 & & \mu_2 & & &
\end{array}\right]
$$

whence, dividing each row by the sum of entries in that row, the transition probability matrices become

$$
P_1 = \left[\begin{array}{c|cc}
 & a_1 & a_2 \\
\hline
1 & & \\
1 & &
\end{array}\right]
\quad \text{and} \quad
P_2 = \left[\begin{array}{c|cc|ccc}
 & a_1 & a_2 & & & \\
\hline
b_3 & & & b_1 & b_2 & \\
c_3 & & & & c_1 & c_2 \\
\hline
 & 1 & & & & \\
 & e_2 & e_1 & & & \\
 & & 1 & & &
\end{array}\right]
$$

We intentionally skip the 10×10 transition matrix P_3 for $N = 3$ so that the reader can verify their understanding after studying the pattern of non-zero rates in the entire

15×15 transition probability matrix P_4 given by

$$P_4 = \left[\begin{array}{c|cc|ccc|cccc|ccccc}
 & a_1 & a_2 & & & & & & & & & & & & \\ \hline
b_3 & & & b_1 & b_2 & & & & & & & & & & \\
c_3 & & & & c_1 & c_2 & & & & & & & & & \\ \hline
 & b_3 & & & & & b_1 & b_2 & & & & & & & \\
 & d_4 & d_3 & & & & & d_1 & d_2 & & & & & & \\
 & & c_3 & & & & & & c_1 & c_2 & & & & & \\ \hline
 & & & b_3 & & & & & & & b_1 & b_2 & & & \\
 & & & d_4 & d_3 & & & & & & & d_1 & d_2 & & \\
 & & & & d_4 & d_3 & & & & & & & d_1 & d_2 & \\
 & & & & & c_3 & & & & & & & & c_1 & c_2 \\ \hline
 & & & & & & 1 & & & & & & & & \\
 & & & & & & e_2 & e_1 & & & & & & & \\
 & & & & & & & e_2 & e_1 & & & & & & \\
 & & & & & & & & e_2 & e_1 & & & & & \\
 & & & & & & & & & 1 & & & & &
\end{array}\right] \tag{5}$$

Readers will do well to note the pattern in the transition matrix P_4: Read it as a 5×5 block matrix with the diagonal consisting of square matrices of sizes $1 \times 1, 2 \times 2, \ldots, 5 \times 5$ with all entries zero. The four cells just above the diagonal have matrices of sizes $1 \times 2, 2 \times 3, 3 \times 4, 4 \times 5$ with step-by-step right-staggered rows

$$(a_1, a_2); (b_1, b_2), (c_1, c_2); (b_1, b_2), (d_1, d_2), (c_1, c_2); (b_1, b_2), (d_1, d_2), (d_1, d_2), (c_1, c_2).$$

The four cells just below the diagonal have matrices of sizes $2 \times 1, 3 \times 2, 4 \times 3, 5 \times 4$ with step-by-step right-staggered rows

$$b_3, c_3; b_3, (d_4, d_3), c_3; b_3, (d_4, d_3), (d_4, d_3), c_3; 1, (e_2, e_1), (e_2, e_1), (e_2, e_1), 1.$$

Having observed the pattern in P_4, the attentive reader should be able to construct P_N for all $N \geq 1$.

For any $N \geq 1$, there are finitely many states (to be exact $\binom{N+2}{2}$ states) that communicate with one another (that is, there is a positive probability of going from one state to the other in finitely many steps). Therefore, the DTSP is irreducible (belongs to one single communication class).

Readers will benefit from examples of other CTSPs treated as semi-Markov processes found in [4].

3. Limiting Proportion of Time Spent in Each State

We shall use the ergodic theorem of a semi-Markov process (see Theorem 4.8.3 of [5], for example) whose proof relies on the strong law of large numbers (see [1]) and the strong law for renewal processes (see [8]). These are typically taught in an introductory graduate course in probability theory or stochastic processes.

Theorem 3.1. *Suppose that a semi-Markov process is irreducible with successive returns to state l having a non-lattice distribution. Let π be the row-vector of the stationary probability distribution for the embedded DTSP satisfying $\pi P = \pi$ and let ν be* *the row-vector of the expected sojourn times in the various states. Then the limiting proportion of time the CTSP spends in state l (or the limiting probability that the CTSP will be found in state l) is given by*

$$\theta_l = \frac{\pi_l \, \nu_l}{\sum_k \pi_k \, \nu_k}. \tag{6}$$

Recall that each state l is classified into one of seven categories according as the total transition rate is $a, b, c, d, e, \mu_1, \mu_2$. Accordingly, the sojourn time in state l being the minimum of several independent exponential variables, the expected sojourn time ν_l is the reciprocal of the total transition rate in effect in state l.

In view of Theorem 3.1, our remaining task is to find the stationary distribution π of the DTSP, or any arbitrary multiple of π. We should emphasize that it does not matter how we discover π since once proposed, we can check if it satisfies $\pi \, P = \pi$. If so, the uniqueness theorem (see, for example, Theorem 4.3.3 of [5]) guarantees that there is no other stationary distribution.

Clearly, for $N = 1$, we can verify that $\pi_1 \propto (1, a_1, a_2)$ satisfies $\pi_1 \, P_1 = \pi_1$. Hence,

$$\theta \propto (1/a, a_1/\mu_1, a_2/\mu_2) \propto (1, \lambda_1/\mu_1, \lambda_2/\mu_2).$$

For $N = 2$, if we guess the stationary distribution is of the form $\pi_2 \propto (*, 1, x, *, *, *)$, then by setting $\pi_2 \, P_2 = \pi_2$, we can fill in the unspecified values to obtain

$$\pi_2 \propto (b_3 + c_3 x, 1, x, b_1, b_2 + c_1 x, c_2 x)$$

Next, setting the third element of π_2, we get $x = a_2(b_3 + c_3 x) + e_1(b_2 + c_1 x) + c_2 x$, or

$$x = \frac{a_2 b_3 + b_2 e_1}{1 - a_2 c_3 - c_1 e_1 - c_2}.$$

Or, setting the second element of π_2, we get $1 = a_1(b_3 + c_3 x) + b_1 + (b_2 + c_1 x) e_2$, or

$$x = \frac{1 - a_1 b_3 - b_1 - b_2 e_2}{a_1 c_3 + c_1 e_2}.$$

To verify that the above two expressions of x are identical, we can check that the denominators are identical and so are the numerators. That is,

$$1 - a_2 c_3 - c_1 e_1 - c_2 = 1 - c_3 + a_1 c_3 - c_1 + c_1 e_2 - c_2 = a_1 c_3 + c_1 e_2,$$

and

$$1 - a_1 b_3 - b_1 - b_2 e_2 = 1 - b_3 + a_2 b_3 - b_1 - b_2 + b_2 e_1 = a_2 b_3 + b_2 e_1.$$

To develop insight into the solution for $N \geq 3$, we wish to study in the next two subsections two special cases — (1) $\lambda_1 = \lambda_2 = \mu_1 = \mu_2$, and (2) $\lambda_1 = \lambda_2$, $\mu_1 = \mu_2$. Thereafter, the solutions to the special cases will inspire us to conjecture the solution to the general case, which we will verify to be true.

3.1. Stationary distribution when $\boldsymbol{\lambda_1 = \lambda_2 = \mu_1 = \mu_2}$

Without loss of generality, let $\lambda_1 = \lambda_2 = \mu_1 = \mu_2 = 1$. Then the DTSP reduces to a symmetric random walk on the vertices of a graph where each vertex is a state and an edge joins two vertices if and only if a direct (one-step) transition is possible between them. The random walk is symmetric because, from any vertex, the DTSP is equally likely to move to any one of the adjacent vertices. In this special case of a symmetric random walk, the stationary distribution is given by a theorem found in Lovasz [2].

Theorem 3.2. *For a symmetric random walk on the vertices of a finite graph, the stationary distribution has probabilities proportional to the degrees of the vertices.*

In view of Theorem 3.2, for $N \leq 4$, the stationary distributions are respectively proportional to

$$\begin{aligned}
&(2; 1, 1),\\
&(2; 3, 3; 1, 2, 1),\\
&(2; 3, 3; 3, 4, 3; 1, 2, 2, 1),\\
&(2; 3, 3; 3, 4, 3; 3, 4, 4, 3; 1, 2, 2, 2, 1).
\end{aligned}$$

We invite the reader to write down the stationary distribution for $N = 5$ following the above pattern.

Since the expected sojourn time in each state is the reciprocal of the corresponding element in π, the long-run proportions of time spent in various states, θ, are given by a *discrete uniform distribution* — a pleasantly surprising result — reminiscent of a similar result in [6] for a symmetric random walk on the vertices of a polygon.

3.2. Stationary distribution when $\boldsymbol{\lambda_1 = \lambda_2}$ and $\boldsymbol{\mu_1 = \mu_2}$

Without loss of generality, let $\lambda_1 = \lambda_2 = \rho$ and $\mu_1 = \mu_2 = 1$. For $N \leq 4$, the stationary distributions are respectively proportional to

$$\begin{aligned}
&(2; 1, 1),\\
&(2; 1 + 2\rho, 1 + 2\rho; \rho, 2\rho, \rho),\\
&(2; 1 + 2\rho, 1 + 2\rho; \rho(1 + 2\rho), 2\rho(1 + \rho), \rho(1 + 2\rho); \rho^2, 2\rho^2, 2\rho^2, \rho^2),\\
&(2; 1 + 2\rho, 1 + 2\rho; \rho(1 + 2\rho), 2\rho(1 + \rho), \rho(1 + 2\rho);\\
&\quad \rho^2(1 + 2\rho), 2\rho^2(1 + \rho), 2\rho^2(1 + \rho), \rho^2(1 + 2\rho); \rho^3, 2\rho^3, 2\rho^3, 2\rho^3, \rho^3).
\end{aligned} \tag{7}$$

This we know by checking that $\pi P = \pi$. Again, we invite the reader to write down the stationary distribution for $N = 5$ following the above pattern.

For $N = 1$, the total transition rates in the three states are $(2\rho, 1, 1)$; for $N = 2$, they are $(2\rho, 1 + 2\rho, 1 + 2\rho, 1, 2, 1)$; and for $N \geq 3$, the total transition rates in the seven categories of states are respectively, $(2\rho, 1 + 2\rho, 1 + 2\rho, 2(1 + \rho), 1, 2, 1)$, the corresponding expected sojourn times are element-wise reciprocals. Hence, for $N \leq 4$, the proportions of time spent in various states (after multiplying all entries by ρ) are respectively proportional to

$$
\begin{aligned}
&(1; \rho, \rho),\\
&(1; \rho, \rho; \rho^2, \rho^2, \rho^2),\\
&(1; \rho, \rho; \rho^2, \rho^2, \rho^2; \rho^3, \rho^3, \rho^3, \rho^3),\\
&(1; \rho, \rho; \rho^2, \rho^2, \rho^2; \rho^3, \rho^3, \rho^3, \rho^3; \rho^4, \rho^4, \rho^4, \rho^4, \rho^4).
\end{aligned} \tag{8}
$$

Thus, the limiting proportion of time spent in state $(i + j)$ is proportional to ρ^{i+j}, a geometric progression — again, a pleasantly surprising result — reminiscent of a similar result in [6] for an asymmetric random walk on the vertices of a polygon. Of course, as $\rho \to 1$, the solutions in this subsection approach those in the previous subsection.

3.3. Solution under arbitrary parameters

Recall that the vector θ of limiting proportions of time the CTSP spends in various states in $\mathcal{S}$ is proportional to the coordinate-wise product of the vector π of steady-state probabilities of the DTSP that satisfies $\pi\, P = \pi$, and the vector ν of expected sojourn times. We can directly solve for θ (without first finding π) using the matrix R of transition rates. Let us illustrate the method for $N = 4$. Generalizing to any N is easy.

Recall that π_4 is found by solving $\pi_4\, P_4 = \pi_4$. However, $P_4 = D_4^{-1} R_4$, where D_4 is a diagonal matrix with diagonal entries given by the row totals of R_4; that is,

$$D_4 = \mathrm{diag}\,(a; b, c; b, d, c; b, d, d, c; \mu_1, e, e, e, \mu_2).$$

Hence, π_4 satisfies $\pi_4\, D_4^{-1} R_4 = \pi_4$; or equivalently, $\pi_4\, D_4^{-1} R_4 = \pi_4\, D_4^{-1}\, D_4$. Next, recall that $\theta_4 \propto \pi_4\, D_4^{-1}$. Therefore, our task is to solve $\theta_4\, R_4 = \theta_4\, D_4$.

Let us revisit the special case $\lambda_1 = \lambda_2 = \rho, \mu_1 = \mu_2 = 1$. In this special case, for $N = 4$, the 15×15 transition rate matrix is

$$
R_4 = \left[\begin{array}{c|cc|ccc|cccc|ccccc}
 & \rho & \rho & & & & & & & & & & & & \\
\hline
1 & & & \rho & \rho & & & & & & & & & & \\
1 & & & & \rho & \rho & & & & & & & & & \\
\hline
 & 1 & & & & & \rho & \rho & & & & & & & \\
 & 1 & 1 & & & & & \rho & \rho & & & & & & \\
 & & 1 & & & & & & \rho & \rho & & & & & \\
\hline
 & & & 1 & & & & & & & \rho & \rho & & & \\
 & & & 1 & 1 & & & & & & & \rho & \rho & & \\
 & & & & 1 & 1 & & & & & & & \rho & \rho & \\
 & & & & & 1 & & & & & & & & \rho & \rho \\
\hline
 & & & & & & 1 & & & & & & & & \\
 & & & & & & 1 & 1 & & & & & & & \\
 & & & & & & & 1 & 1 & & & & & & \\
 & & & & & & & & 1 & 1 & & & & & \\
 & & & & & & & & & 1 & & & & &
\end{array}\right]
$$

Taking row sums of R_4, we see that D_4 is a diagonal matrix given by

$$\begin{aligned} D_4 = \operatorname{diag}(2\rho; 1+2\rho, 1+2\rho; 1+2\rho, 2(1+\rho), 1+2\rho; \\ 1+2\rho, 2(1+\rho), 2(1+\rho), 1+2\rho; 1, 2, 2, 2, 1). \end{aligned}$$

It is easy to check that

$$\theta_4 \propto (1; \rho, \rho; \rho^2, \rho^2, \rho^2; \rho^3, \rho^3, \rho^3, \rho^3; \rho^4, \rho^4, \rho^4, \rho^4, \rho^4)$$

satisfies $\theta_4 R_4 = \theta_4 D_4$.

The result for $N = 4$ easily generalizes to any N.

The insight developed in the special case prompts us to anticipate the following:

Theorem 3.3. *For the general case of arbitrary parameters $\lambda_1, \lambda_2, \mu_1, \mu_2$, letting $\rho_1 = \lambda_1/\mu_1$ and $\rho_2 = \lambda_2/\mu_2$, the proportions of times spent in various states 1 through $\binom{N+2}{2}$ of $\mathcal{S}$ are given by*

$$\theta \propto \left(1;\ \rho_1, \rho_2;\ \rho_1^2, \rho_1\rho_2, \rho_2^2;\ \rho_1^3, \rho_1^2\rho_2, \rho_1\rho_2^2, \rho_2^3;\ \rho_1^4, \rho_1^3\rho_2, \rho_1^2\rho_2^2, \rho_1\rho_2^3, \rho_2^4; \cdots\right). \tag{9}$$

Proof. First, we verify the claim for $N = 4$, say, when the transition rate matrix is

$$R_4 = \left[\begin{array}{c|cc|ccc|cccc|ccccc}
 & \lambda_1 & \lambda_2 & & & & & & & & & & & & \\ \hline
\mu_1 & & & \lambda_1 & \lambda_2 & & & & & & & & & & \\
\mu_2 & & & & \lambda_1 & \lambda_2 & & & & & & & & & \\ \hline
 & \mu_1 & & & & & \lambda_1 & \lambda_2 & & & & & & & \\
 & \mu_2 & \mu_1 & & & & & \lambda_1 & \lambda_2 & & & & & & \\
 & & \mu_2 & & & & & & \lambda_1 & \lambda_2 & & & & & \\ \hline
 & & & \mu_1 & & & & & & & \lambda_1 & \lambda_2 & & & \\
 & & & \mu_2 & \mu_1 & & & & & & & \lambda_1 & \lambda_2 & & \\
 & & & & \mu_2 & \mu_1 & & & & & & & \lambda_1 & \lambda_2 & \\
 & & & & & \mu_2 & & & & & & & & \lambda_1 & \lambda_2 \\ \hline
 & & & & & & \mu_1 & & & & & & & & \\
 & & & & & & \mu_2 & \mu_1 & & & & & & & \\
 & & & & & & & \mu_2 & \mu_1 & & & & & & \\
 & & & & & & & & \mu_2 & \mu_1 & & & & & \\
 & & & & & & & & & \mu_2 & & & & &
\end{array}\right]$$

Recall that D_4 is a diagonal matrix given by the row sums of R_4; that is,

$$D_4 = \operatorname{diag}(a; b, c; b, d, c; b, d, d, c; \mu_1, e, e, e, \mu_2).$$

Specializing (9) for $N = 4$, we have

$$\theta_4 \propto \left(1;\ \rho_1, \rho_2;\ \rho_1^2, \rho_1\rho_2, \rho_2^2;\ \rho_1^3, \rho_1^2\rho_2, \rho_1\rho_2^2, \rho_2^3;\ \rho_1^4, \rho_1^3\rho_2, \rho_1^2\rho_2^2, \rho_1\rho_2^3, \rho_2^4\right).$$

It is straight-forward to check that θ_4 satisfies $\theta_4\, R_4 = \theta_4\, D_4$.

Generalizing to any N is routine. This completes the proof. $\square$

Having found θ up to a proportionality constant, the constant of proportionality is seen to be the reciprocal of

$$\sigma_N = 1 + (\rho_1 + \rho_2) + (\rho_1^2 + \rho_1\rho_2 + \rho_2^2) + \ldots + (\rho_1^N + \rho_1^{N-1}\rho_2 + \ldots + \rho_2^N). \qquad (10)$$

In general, when $\rho_1 \neq \rho_2$, $\rho_1 \neq 1$ and $\rho_2 \neq 1$, σ_N reduces to

$$\begin{aligned}\sigma_N =& \frac{1}{\rho_1 - \rho_2}\left\{(\rho_1 - \rho_2) + (\rho_1^2 - \rho_2^2) + (\rho_1^3 - \rho_2^3) + \ldots + (\rho_1^{N+1} - \rho_2^{N+1})\right\} \\ =& \frac{1}{\rho_1 - \rho_2}\left\{\rho_1 \frac{1 - \rho_1^{N+1}}{1 - \rho_1} - \rho_2 \frac{1 - \rho_2^{N+1}}{1 - \rho_2}\right\}.\end{aligned}$$

When $\rho_1 = \rho_2 = \rho$, say, then from (10), $\sigma_N = 1 + 2\rho + 3\rho^2 + \ldots + (N+1)\rho^N$ as we have already seen in Subsection 2.2. Furthermore, when $\rho = 1$, then $\sigma_N = \binom{N+2}{2}$ as seen in Subsection 2.1. Also, when $\rho_1 \neq \rho_2 = 1$, then

$$\sigma_N = (N+1) + N\rho_1 + (N-1)\rho_1^2 + \ldots + 2\rho_1^{N-1} + \rho_1^N.$$

Etc.

Of course, without loss of generality, any one of the four parameters $(\lambda_1, \lambda_2, \mu_1, \mu_2)$ can be chosen as unity. (This is a matter of choosing the time unit.) That ought to leave three parameters arbitrary. Therefore, it is mildly surprising that θ depends on only two ratios ($\rho_1 = \lambda_1/\mu_1, \rho_2 = \lambda_2/\mu_2$). However, the stationary distribution π depends on all four parameters. In fact, using Theorem 9, we can reconstruct the stationary distribution as $\pi \propto \theta\, D_4$, or $\pi \propto$

$$\left(a;\ b\rho_1, c\rho_2;\ b\rho_1^2, d\rho_1\rho_2, c\rho_2^2;\ b\rho_1^3, d\rho_1^2\rho_2, d\rho_1\rho_2^2, c\rho_2^3;\ \mu_1\rho_1^4, e\rho_1^3\rho_2, e\rho_1^2\rho_2^2, e\rho_1\rho_2^3, \mu_2\rho_2^4\right).$$

Again, finding the stationary distribution π when N is arbitrarily large is a routine matter which we leave to the reader.

4. Numerical Evaluation and Graphical Display

For any $N \geq 1$, θ, the limiting proportion of times spent in various states can be found from (9) in Theorem 9, and thereafter the stationary distribution π can be found as $\pi \propto \theta\, D$, analogous to (3.3). However, numerical values of π and θ being tedious to read, we depict them as stick diagrams of the probability mass functions (PMF) of π and θ. For $N = 15$, if the parameter values are $\lambda_1 = \lambda_2 = \mu_1 = \mu_2 = 0.50$, we know that θ is uniformly distributed (see subsection 2.1). Figure 2 shows how the PMF changes when some parameters are changed to 0.25. As anticipated, when the arrival rates are lower than service rates, the probabilities concentrate towards the lower states (top panel), and conversely (second panel). Also, when the arrival rate of one business is lower, then the number of customers for that business are lower than that of the other business (third panel). On the other hand, if the service rate of one business is slower, then more customers for that business will remain in the queue (bottom panel).

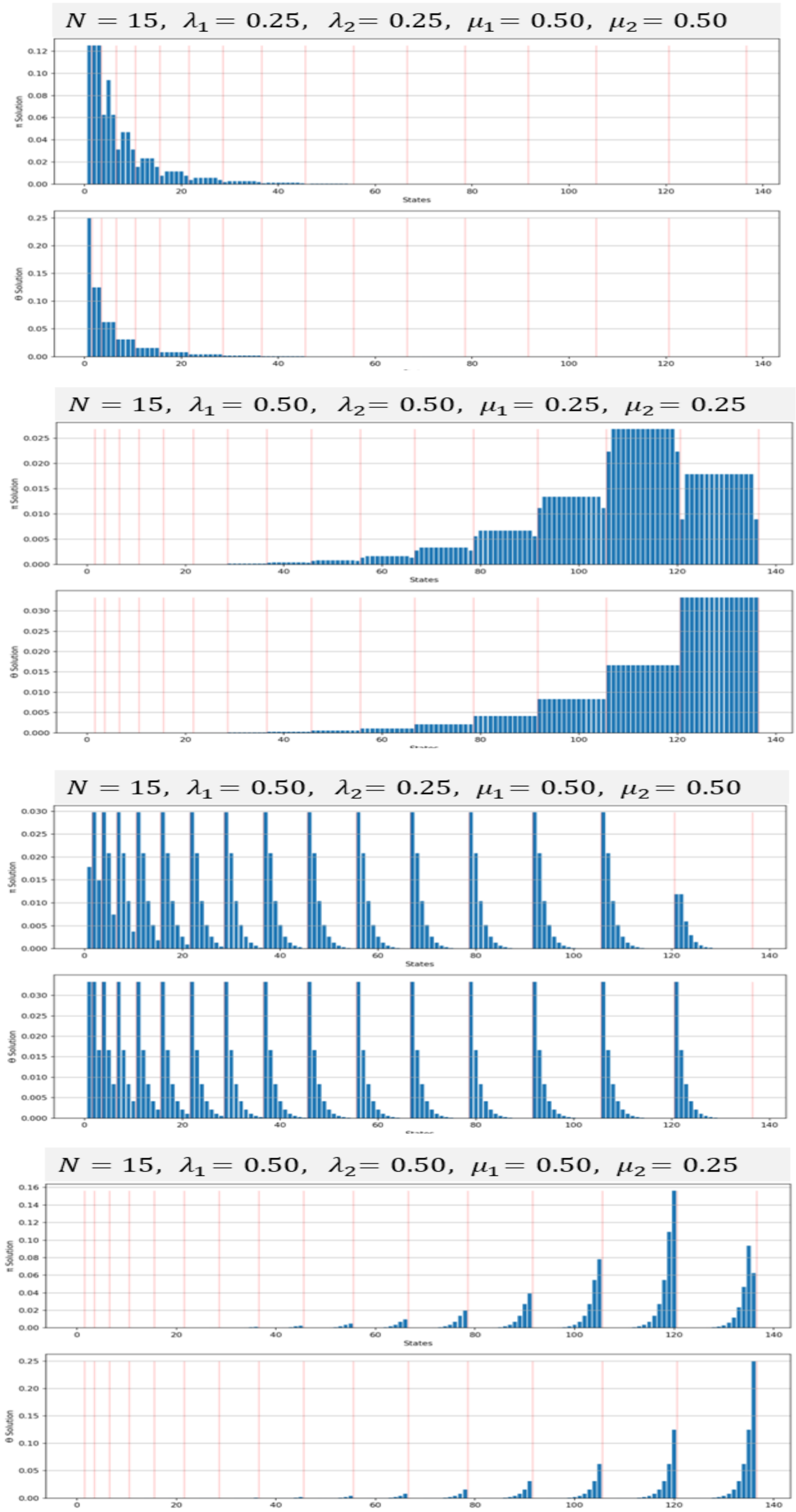

Figure 2. For some choices of parameters $(\lambda_1, \lambda_2, \mu_1, \mu_2)$, we depict the stationary distribution π_{15} and limiting time distribution θ_{15}.

5. Evaluating π Using a Numerical Method

Here we illustrate an alternative method of finding π (and hence $\theta \propto \pi D^{-1}$) via numerical computations, without invoking Theorem 9.

Suppose that $\lambda_1 = 2, \lambda_2 = 3, \mu_1 = 4, \mu_2 = 5$. Then $a = 5, b = 9, c = 10, d = 14, e = 9$. Let $N = 3$. Then the transition probability matrix is P_3, and the stationary distribution π_3 satisfies $\pi_3 P_3 = \pi_3$. Of course, then $\pi_3 P_3^n = \pi_3$ for any $n \geq 1$. We shall successively square the P_3 matrix until all 15^2 elements of $P_3^{2^k}$ and $P_3^{2^{k+1}}$ differ by no more than .0001. This criterion is satisfied for $k = 6$. Each column of $P_3^{2^6}$ consists of two distinct values — a zero and a non-zero. Indeed, π is proportional to the vector of non-zero elements in the columns of $P_3^{2^6}$ as shown in `cmax` in the R codes below. Also, noting that $\rho_1 = 2/4 = .5$ and $\rho_2 = 3/5 = .6$, indeed θ satisfies (9) as shown in `theta/theta[1]` in the R codes below. R is a freeware, see [3].

```
> round(P,4)
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9] [,10]
 [1,]    0   .4000  .6000    0      0      0      0      0      0      0
 [2,] .4444    0      0    .2222  .3333    0      0      0      0      0
 [3,] .5000    0      0      0    .2000  .3000    0      0      0      0
 [4,]    0   .4444    0      0      0      0    .2222  .3333    0      0
 [5,]    0   .3571  .2857    0      0      0      0    .1429  .2143    0
 [6,]    0     0    .5000    0      0      0      0      0    .2000  .3000
 [7,]    0     0      0      1      0      0      0      0      0      0
 [8,]    0     0      0    .5556  .4444    0      0      0      0      0
 [9,]    0     0      0      0    .5556  .4444    0      0      0      0
[10,]    0     0      0      0      0      1      0      0      0      0
> round(Q,4) # P^64
       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]  [,10]
 [1,] .3322    0      0    .1495  .2791  .2392    0      0      0      0
 [2,]    0   .2990  .3987    0      0      0    .0332  .0897  .1076  .0718
 [3,]    0   .2990  .3987    0      0      0    .0332  .0897  .1076  .0718
 [4,] .3322    0      0    .1495  .2791  .2392    0      0      0      0
 [5,] .3322    0      0    .1495  .2791  .2392    0      0      0      0
 [6,] .3322    0      0    .1495  .2791  .2392    0      0      0      0
 [7,]    0   .2990  .3986    0      0      0    .0332  .0897  .1076  .0717
 [8,]    0   .2990  .3987    0      0      0    .0332  .0897  .1076  .0718
 [9,]    0   .2990  .3987    0      0      0    .0332  .0897  .1076  .0718
[10,]    0   .2990  .3987    0      0      0    .0332  .0897  .1076  .0718
> round(cmax,4)
 [1]  0.3322 0.2990 0.3987 0.1495 0.2791 0.2392 0.0332 0.0897 0.1076  .0718
> round(pi,4) # cmax/sum(cmax)
 [1]  0.1661 0.1495 0.1993 0.0748 0.1395 0.1196 0.0166 0.0449 0.0538  .0359
> round(nu,4)
 [1]  0.2000 0.1111 0.1000 0.1111 0.0714 0.1000 0.2500 0.1111 0.1111 0.2000
> prod=pi*nu;   theta=prod/sum(prod)
> round(theta,4)
 [1]  0.2717 0.1358 0.1630 0.0679 0.0815 0.0978 0.0340 0.0408 0.0489 0.0587
> round(theta/theta[1],4) # same as (2.4)
 [1]  1.0000 0.5000 0.6000 0.2500 0.3000 0.3600 0.1250 0.1500 0.1800 0.2160
```

6. Answers to the Questions Raised in the Abstract

Having studied the limiting behavior of two M/M/1 queuing systems sharing a common facility with capacity N, we can answer some questions the two servers might have wondered about when they contemplated sharing a facility:

(Q1) What percentage of customers do the servers lose? When the system is in any one of states $\{(N-j,j) : j = 0,1,\ldots,N\}$, the arrivals to the two businesses at exponential rates λ_1 and λ_2, respectively, do not enter the waiting room and are lost for good. Therefore, each server loses the same proportion of their respective customers given by

$$\sum_{j=0}^{N} \theta(N-j,j) = \left(\rho_1^N + \rho_1^{N-1}\rho_2 + \ldots + \rho_1\rho_2^{N-1} + \rho_2^N\right)/\sigma_N. \quad (11)$$

With such perfect equality in the proportion of customers lost, the two businesses will have no axe to grind against each other.

(Q2) What percentage of time do the servers remain idle during regular business hours? Server 1 remains idle in states $\{(0,j) : j = 0,1,\ldots,N\}$ for a total proportion of time given by

$$\sum_{j=0}^{N} \theta(0,j) = \left(1 + \rho_2 + \rho_2^2 + \ldots + \rho_2^N\right)/\sigma_N = \frac{1-\rho_2^{N+1}}{\sigma_N(1-\rho_2)} \quad (12)$$

Likewise, Server 2 remains idle in states $\{(i,0) : i = 0,1,\ldots,N\}$ for a total proportion of time given by

$$\sum_{i=0}^{N} \theta(i,0) = \left(1 + \rho_1 + \rho_1^2 + \ldots + \rho_1^N\right)/\sigma_N = \frac{1-\rho_1^{N+1}}{\sigma_N(1-\rho_1)} \quad (13)$$

(Q3) At closing time, how many customers are in the facility either being served or waiting to be served by each server? The number of customers being served or waiting to be served by Server 1 is a discrete random variable W_1 with

$$P\{W_1 = k\} = \sum_{j=0}^{N-k} \theta(k,j) = \rho_1^k\,(1+\rho_2+\ldots+\rho_2^{N-k})/\sigma_N = \rho_1^k\,\frac{1-\rho_2^{N-k+1}}{\sigma_N(1-\rho_2)} \quad (14)$$

for $k = 0,1,\ldots,N$. Similarly, the number of customers being served or waiting to be served by Server 2 is a discrete random variable W_2 with

$$P\{W_2 = k\} = \sum_{i=0}^{N-k} \theta(i,k) = \rho_2^k\,(1+\rho_1+\ldots+\rho_1^{N-k})/\sigma_N = \rho_2^k\,\frac{1-\rho_1^{N-k+1}}{\sigma_N(1-\rho_1)} \quad (15)$$

for $k = 0,1,\ldots,N$. The service times being independent, the total time to serve k customers at closing time is a gamma(k,λ_h) variable, for $h = 1,2$.

7. Conclusion

We studied two intertwined M/M/1 queuing systems sharing a limited-capacity common facility. Assuming exponential inter-arrival times and exponential service times, we found the long-run proportions of time the system spends in various states defined by the pair of numbers of customers for the two servers. We have both theoretically derived and computationally evaluated the limiting results. Furthermore, for each server, we have determined the proportion of customers lost, the proportion of time spent idling, the number of remaining customers at closing time, and the additional duration to serve them.

There are several naturally anticipated extensions to this topic. For example, what if three or more businesses choose to share a common waiting room? Under similar assumptions on arrival and service time distributions, the results, though tedious, can be extended without much difficulty. Again, all servers will lose the same proportion of customers, preventing any unproductive contentious situation.

The exponential distribution, though a good starting point for developing mathematical theory, is not a perfect representation of inter-arrival times or service times in many real-life situations. We leave to future researchers to study other distributions such as gamma that may better reflect how real-world businesses operate. See [7] for an illustration of how an exponential model extends to a gamma model.

Additionally, customer demands may be better represented by non-homogeneous arrival processes. While this generalization is beyond the scope of the analytical solutions presented here, it may nonetheless be fruitful to examine simple non-homogeneous distributions in simulations in the hope of later finding a more elegant solution.

Last but not least is the extension to two or more M/M/k systems where there are $k \geq 2$ servers for each business.

Acknowledgements

The authors thank Sheldon Ross for posing the problem of two M/M/1 queues sharing a common facility with a capacity constraint. Helpful comments from a referee and guidance from the Editor-in-Chief are gratefully acknowledged.

References

[1] Bhattacharya, Rabi, Lin, Lizhen and Patrangenaru, Victor (2016). *A Course in Mathematical Statistics and Large Sample Theory*, Springer Texts in Statistics, New York, NY: Springer.

[2] Lovász, László (1993). Random Walks on Graphs: A Survey, In: *Combinatorics*, Paul Erdős is Eighty (Volume 2), Bolyai Society Mathematical Studies, Keszthely (Hungary), pp. 1–46.

[3] R Core Team (2019). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[4] Rausand, Marvin and Høyland, Arnljot (2004). *System Reliability Theory: Models, Statistical Methods, and Applications*, 2nd ed., Wiley Series in Probability and Statistics, Hoboken, NJ: Wiley-Interscience.

[5] Ross, Sheldon M. (1996). *Stochastic Processes*, 2nd ed., Wiley Series in Probability and Statistics, New York: John Wiley & Sons.

[6] Sarkar, Jyotirmoy (2006). Random walk on a polygon, In: Jiayang Sun, Vince Melfi, Anir-

ban DasGupta and Connie Page (eds.) *Recent Developments in Nonparametric Inference and Probability*, IMS Lecture Notes 50, 31–43, Institute of Mathematical Statistics, Beachwood, OH. DOI: 10.1214/074921706000000581

[7] Sarkar, Jyotirmnoy and Li, Fang (2006). Limiting average availability of a system supported by several spares and several repair facilities, *Statistics and Probability Letters* 76 (18), 1965–1974. doi.org/10.1016/j.spl.2006.04.046

[8] Whitt, Ward (2013). *Introduction to Renewal Theory*, IEOR 3106 Lecture Notes. http://www.columbia.edu/ ww2040/3106F13/lect1112.pdf