A Single-Server Queue with Markov Modulated Service Times *

Noah Gans

OPIM Department
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104-6340
gans@wharton.upenn.edu

Yong-Pin Zhou

Dept. of Management Science
School of Business Administration
University of Washington, Seattle
Seattle, WA 98195-3200
yongpin@u.washington.edu

Abstract

We study a queueing system with a Poisson arrival process and Markov modulated, exponential service requirements. For a modulating Markov Chain with two states, we show that the distribution of the number-in-system is a superposition of two matrix-geometric series and provide a simple algorithm for computing the rate and coefficient matrices. These results hold for both finite and infinite waiting space systems, as well as for cases in which eigenvalues of the rate matrices' characteristic polynomials have multiplicity grater than one.

We make the conjecture that the Markov-modulated system performs better than its M/G/1 analogue if and only if the switching probabilities between the two states satisfy a simple condition. We give an intuitive argument to support this conjecture.

Key words: queues, Markov modulated, matrix-geometric method.

*Research supported by the Wharton Financial Institutions Center and by NSF Grant SBR-9733739

1 Overview

Consider the following single-server queue: the arrival process is Poisson; service times are exponentially distributed; and the service discipline is first-come, first-served (FCFS). However, the rates of these exponential service times are determined by an underlying Markov chain, and transitions of the Markov chain take place at service completions.

The Markov chain has m states. If the new state of the Markov chain is i, $1 \leq i \leq m$, then the rate of next exponential service time will be μ_i. We will call this system with Markov-regulated Markovian services an M/MM/1 queue.

Our interest in this type of queueing system comes from the study of service systems with human servers. Employee learning and turnover cause the sequence of service-time distributions to exhibit systematic non-stationarities: as an employee learns, his or her service speed increases; when an employee turns over, s/he is usually replaced by a new person with lower service speeds. We wish to understand the effect of employee learning and turnover on measures of system performance such as average waiting time and queue length.

We model employee learning and turnover as transitions through states of a Markov chain. After each service an employee may learn and advance to a higher skill level with a pre-specified probability. After each service an employee may also turn over with another pre-specified probability, in which case s/he is replaced by a new employee at the lowest skill level. Skill levels correspond to states of the Markov chain and the Markov chain modulates the service-time distribution. In the simplest case, when there is only one employee, the human server queueing system becomes an M/MM/1 system.

In addition to modelling server "learning and turnover", the M/MM/1 queue may be used to model a processor in a data network. The processor works at a constant speed but processes jobs from several sources. The aggregate arrival process is a stationary Poisson process, but the source from which a particular job comes (the job "type") is determined by an underlying Markov chain. Jobs from different sources carry with them exponentially distributed amounts of work with different means.

When the waiting space is infinite, the dynamics of the two systems are equivalent. When there is a finite limit on the waiting space, however, the behavior of the two systems differs. In the data-processing model, arriving jobs that are lost still generate transitions of the modulating Markov chain, and changes in the service-time distribution from one job to the next depend on whether or not the waiting space is full. Alternatively, in the human-server model it is service completions that generate transitions of the modulating Markov chain, and these transitions are unaffected by lost arrivals.

Using a matrix difference equation approach, we are able to obtain a complete characterization of the system's behavior when the Markov chain has two states ($m = 2$). In this case, we can also use closed-form solutions to the resulting cubic equations to obtain *exact* solutions for the computation of required rate coefficient matrices in the numerical study. Our analysis yields the following results.

We obtain traditional measures of queueing performance for this M/MM/1 system: the distribution of the number of customers in the system and, in turn, the system utilization, the average number in the system, the average waiting time in queue and in the system. In the case of systems with finite waiting rooms we also obtain the loss probability.

More fundamentally we show that, for systems with either infinite or finite waiting spaces, the steady-state distribution of the number of customers in the system can be represented as the superposition of two matrix-geometric series: $X_n = (R_1^n K_1 + R_2^n K_2) X_0$. Here R_1 and R_2 are two square matrices and X_n is the vector of steady-state system probabilities for states which have n customers in the system.

Moreover, our analysis develops explicit, computable analytical expressions for both the rate and coefficient matrices of the geometric series. Thus, for the case of a 2-state Markov chain, we obtain an *efficient computational procedure* for calculating the steady-state distribution of the number-in-system for M/MM/1 systems with both finite and infinite waiting rooms. At the end of this paper, we also discuss how this procedure may be extended to M/MM/1 systems whose underlying Markov chain has $m \geq 3$ states.

For the infinite waiting space system, we compare the M/MM/1 model with an analogous M/G/1 model with the same arrival rate and the same first two moments of service time. Through numerical examples we show that the M/G/1 system, which has independent service times, does not necessarily out-perform the M/MM/1 system with correlated service times. When the transition probabilities of the modulating Markov chain are invariant across states, the M/MM/1 system is equivalent to an M/H_2/1 system, and therefore it has the same expected backlog as its M/G/1 analogue. When the modulating Markov chain's transition probabilities out of the current state fall below these M/H_2/1 transition probability levels, however, numerical results show that M/MM/1 performance suffers. Conversely, when the transition probabilities out of the current state exceed these levels, then the expected backlog in the M/MM/1 system is smaller than in the M/H_2/1 system. In the finite waiting space case, loss probabilities of the M/MM/1 system and its M/G/1 analogue exhibit the same pattern.

This numerical evidence leads us to believe that the pattern of observed differences between the M/MM/1 system and its M/G/1 analogue is provably true. We give an intuitive argument to support this conjecture.

2 Literature Review

The M/MM/1 system is a special case of a "Quasi Birth and Death" (QBD) process. QBD processes can be used to model a wide variety of stochastic systems, in particular many telecommunications systems. For background and examples, see Neuts [5] and Servi [6].

Neuts's [5] seminal work characterizes QBD systems with countable state spaces as having, when a certain boundary condition holds, a steady-state distribution of the number-in-system that can be described as a single, matrix-geometric series: $X_n = R^n X_0$. The rate matrix R may be difficult to calculate, however, and the required boundary condition that R must satisfy is difficult to verify.

For finite QBD processes with a limit of N in the system, Naoumov [4] develops results that are similar to ours. Its determination of the rate matrices, R_1 and R_2, requires the computation of two infinite series of (recursively defined) matrices, however. Hence the calculation of its solution is approximate and may be computationally intensive.

Mitrani and Chakka [3] is the paper closest to ours. Using a spectral expansion method that is similar to the approach used in the current paper, it shows that the steady-state distribution of the number-in-system has mixed-geometric solutions. The paper's general results appear to be broader than ours, applying to cases in which $m \geq 3$.

However, the paper does not directly address the case in which some eigenvalues of the characteristic matrix polynomial have multiplicity higher than 1. While (as [3] points out) this does not appear to be a practical problem, it is both interesting and important theoretically: without it, the treatment of the problem and the characterization of its solution are incomplete. This case is also technically difficult to analyze.

In this paper we offer a constructive characterization of the rate matrices that complements the approach use by Mitrani and Chakka [3]. Our approach allows us to address the uncovered case in which the eigenvalues of the characteristic matrix polynomial have multiplicity higher than 1.

Furthermore, it offers computational advantages over the approach laid out in [3]: the mixed-matrix geometric form of our solution is more compact; and, because it retains all of the eigenvalue-eigenvector information, our solution allows for straightforward calculation of higher moments of the queueing system performance. (For details, see §4.) Therefore, our solution procedure is more straightforward and efficient numerically.

Thus, for M/MM/1 systems with $m = 2$, we develop a characterization of system performance that represents a link between Neuts's single-geometric-series characterization of an infinite QBD processes and Naoumov's dual-geometric-series characterization of finite QBD systems. We offer a unified approach and a single characterization of system performance that covers both the finite and countable-state-space cases. Moreover, its constructive characterization complements Mitrani and Chakka's work and addresses cases in which there are duplicated eigenvalues.

The rest of the paper is organized as follows. In §3.1-§3.2 we give a complete solution to the steady-state probability distribution of the number-in-system of an M/MM/1 system. Then in §3.3 we compute important queueing performance measures, such as average number in the system. In §4 we analyze the finite waiting space queueing system, M/MM/1/N. In §5 we present numerical analyses which compare both the infinite and finite systems to their analogues that have *i.i.d.* service times. Finally, in §6 we discuss possible extensions of our results.

3 M/MM/1 queueing system solution

In the following analysis, the Markov chain that modulates the service-time distribution has $m = 2$ states. We denote the two states of the Markov chain as fast, F, and slow, S.

Jobs arrive according to a Poisson process of rate λ, and service times are exponentially distributed.

When the Markov chain is in state F, the server works at a rate of μ_F, and when the Markov chain is in state S, the server works at rate $\mu_S < \mu_F$. When the server is in state F and completes a service it remains fast with probability p_{FF} and becomes slow with probability $p_{FS} = 1 - p_{FF}$. Similarly, when the server is in state S and completes a service, it remains slow with probability p_{SS} and becomes fast with probability $p_{SF} = 1 - p_{SS}$.

3.1 The steady-state probability distribution.

We let $P_{S,n}, n = 0, 1, ...$ denote the steady-state probability that the server is slow and there are n jobs in the system. Similarly, $P_{F,n}$ denotes the steady-state probability that the server is fast and there are n jobs in the system.

Figure 1: State-transition diagram of the Continuous Time Markov Chain

The state-transition equations of the M/MM/1 system's associated Continuous Time Markov Chain (CTMC) are presented below. The corresponding state-transition diagram can be found in Figure 1.

For $n = 0$

$$\lambda P_{S,0} = \mu_S p_{SS} P_{S,1} + \mu_F p_{FS} P_{F,1} \tag{1}$$

$$\lambda P_{F,0} = \mu_S p_{SF} P_{S,1} + \mu_F p_{FF} P_{F,1}, \tag{2}$$

and for $n \geq 1$,

$$(\mu_S + \lambda) P_{S,n} = \lambda P_{S,n-1} + \mu_S p_{SS} P_{S,n+1} + \mu_F p_{FS} P_{F,n+1} \tag{3}$$

$$(\mu_F + \lambda) P_{F,n} = \lambda P_{F,n-1} + \mu_S p_{SF} P_{S,n+1} + \mu_F p_{FF} P_{F,n+1}. \tag{4}$$

We can present the balance equations in a matrix-vector notation. Let

$$X_n = \begin{pmatrix} P_{S,n} \\ P_{F,n} \end{pmatrix}, \qquad A = \begin{pmatrix} \mu_S p_{SS} & \mu_F p_{FS} \\ \mu_S p_{SF} & \mu_F p_{FF} \end{pmatrix}, \qquad B = \begin{pmatrix} \lambda + \mu_S & 0 \\ 0 & \lambda + \mu_F \end{pmatrix},$$

$C = \lambda A^{-1}$, and $D = A^{-1} B$. Then the balance equations (1)–(4) become

$$X_1 = C X_0. \tag{5}$$

$$X_{n+2} - D X_{n+1} + C X_n = 0, \qquad \forall n \geq 0 \tag{6}$$

We note that when $p_{SF} + p_{FS} = 1$ ($p_{SF} = p_{FF}$, $p_{SS} = p_{FS}$), the service times become *i.i.d.* hyper-exponential random variables. In this case, the M/MM/1 system becomes an M/H_2/1 system. Furthermore, if either p_{SF} or p_{FS} is zero, then in the steady-state, the system operates as an M/M/1 queue. Since both systems have been studied (see Kleinrock [1] for an example), in this paper we will focus on the case in which $p_{SF} + p_{FS} \neq 1$ and $p_{SF} \cdot p_{FS} \neq 0$.

Given the representation (5) and (6), we are ready to state our main result.

Theorem 1 *When $p_{SF} + p_{FS} \neq 1$ (i.e. $p_{SF} \neq p_{FF}, p_{SS} \neq p_{FS}$) and $p_{SF} \cdot p_{FS} \neq 0$, the solution to (6) and (5) is of the form*

$$X_n = (R_1^n K_1 + R_2^n K_2) X_0, \tag{7}$$

where $R_1, R_2, K_1,$ and K_2 are such that

$$R_i^2 - DR_i + C = 0 \qquad i = 1, 2 \tag{8}$$

$$K_1 + K_2 = I \tag{9}$$

$$R_1 K_1 + R_2 K_2 = C. \tag{10}$$

Once the matrices R_1, R_2, K_1, and K_2 satisfying (8)-(10) are found, $\{X_n\}_{n=0}^{\infty}$ as defined by (7) is clearly a solution to (6). Moreover, given X_0, (5) and (6) uniquely determine all other probabilities X_n, $\forall n > 0$. So it suffices to prove the existence of a solution of the form (7) such that (8)-(10) are satisfied.

We constructively prove the existence of R_1, R_2, K_1, and K_2. For the cases in which eigenvalues of R_1 and R_2 all have multiplicity of 1, Mitrani and Chakka [3] have more general results. The important theoretical result of this paper is the thorough investigation of the cases in which eigenvalues have higher multiplicities.

An outline of the proof of Theorem 1 is as follows. For R_1 and R_2 to satisfy (8), their eigenvalues must satisfy a quadratic equation similar to (8). Their eigenvectors can be obtained from the equation as well. When linearly independent eigenvectors are found (e.g. when all of the eigenvalues are distinct), the rate matrices R_1 and R_2 can be easily constructed from these eigenvectors and eigenvalues. When the eigenvalues have multiplicity of more than one and the eigenvectors are linearly dependent, however, we must reconstruct R_1 and R_2 from their Jordan forms, along with the corresponding linearly-independent vectors (which are derived from the eigenvectors). The full proof of Theorem 1 is quite long and technical, and we present it in Appendix A.

3.2 Complete Solution of the Steady State Probability Distribution

From Theorem 1, we see that, once we know $X_0 = (P_{S,0}, P_{F,0})'$, then all the other probabilities can be obtained from equation (7). The following two propositions provide two independent equations to determine $P_{S,0}$ and $P_{F,0}$ and, in turn, the entire probability distribution. Their proofs can be found in Appendix B.

Proposition 1

(i) The long-run average service time of the M/MM/1 system is $1/\mu$ where

$$\frac{1}{\mu} = \frac{p_{FS}}{p_{SF}+p_{FS}}\frac{1}{\mu_S} + \frac{p_{SF}}{p_{SF}+p_{FS}}\frac{1}{\mu_F}. \tag{11}$$

(ii) Let $\rho = \lambda/\mu$. When $\rho < 1$, the system is stable, and ρ is the long-run proportion of time the system is busy.

(iii) When $\rho \geq 1$, the system is unstable.

Proposition 1 provides the first equation relating $P_{S,0}$ and $P_{F,0}$:

$$P_{S,0} + P_{F,0} = 1 - \rho. \tag{12}$$

To provide the second equation, we use the fact that probabilities sum to one. Let $(a_M, b_M) = (1,1)\sum_{n=0}^{M}(R_1^n K_1 + R_2^n K_2)$. Then

$$1 = (1,1)\sum_{n=0}^{\infty} X_n = \lim_{M\to\infty}[(a_M, b_M)X_0]. \tag{13}$$

Arrange γ_1, γ_2, γ_3, and γ_4 in descending order with regard to their absolute values (or, in the case of complex numbers, modulus). Let the corresponding vectors be $V_1 = (v_{11}, v_{12})'$, $V_2 = (v_{21}, v_{22})'$, $V_3 = (v_{31}, v_{32})'$, and $V_4 = (v_{41}, v_{42})'$. Since one is an eigenvalue, we must have $|\gamma_1| \geq 1$.

For the following discussion, we will assume $R_1 = (V_1, V_2)\begin{pmatrix} \gamma_1, & 0 \\ 0, & \gamma_2 \end{pmatrix}(V_1, V_2)^{-1}$ and $R_2 = (V_3, V_4)\begin{pmatrix} \gamma_3, & 0 \\ 0, & \gamma_4 \end{pmatrix}(V_3, V_4)^{-1}$. Other cases are similar.

Now denote the coefficient matrices K_1 and K_2 by

$$K_1 = \begin{pmatrix} K_1(1,1), & K_1(1,2) \\ K_1(2,1), & K_1(2,2) \end{pmatrix}, \quad K_2 = \begin{pmatrix} K_2(1,1), & K_2(1,2) \\ K_2(2,1), & K_2(2,2) \end{pmatrix},$$

and

$$\alpha_1 = \frac{(K_1(1,1)v_{22} - K_1(2,1)v_{21})(v_{11}+v_{12})}{v_{11}v_{22} - v_{12}v_{21}}, \qquad \beta_1 = \frac{(K_1(1,2)v_{22} - K_1(2,2)v_{21})(v_{11}+v_{12})}{v_{11}v_{22} - v_{12}v_{21}},$$

$$\alpha_2 = \frac{(K_1(2,1)v_{11} - K_1(1,1)v_{12})(v_{21}+v_{22})}{v_{11}v_{22} - v_{12}v_{21}}, \qquad \beta_2 = \frac{(K_1(2,2)v_{11} - K_1(1,2)v_{12})(v_{21}+v_{22})}{v_{11}v_{22} - v_{12}v_{21}},$$

$$\alpha_3 = \frac{(K_2(1,1)v_{42} - K_2(2,1)v_{41})(v_{31}+v_{32})}{v_{31}v_{42} - v_{32}v_{41}}, \qquad \beta_3 = \frac{(K_2(1,2)v_{42} - K_2(2,2)v_{41})(v_{31}+v_{32})}{v_{31}v_{42} - v_{32}v_{41}},$$

$$\alpha_4 = \frac{(K_2(2,1)v_{31} - K_2(1,1)v_{32})(v_{41}+v_{42})}{v_{31}v_{42} - v_{32}v_{41}}, \qquad \beta_4 = \frac{(K_2(2,2)v_{31} - K_2(1,2)v_{32})(v_{41}+v_{42})}{v_{31}v_{42} - v_{32}v_{41}}.$$

Then

$$(a_M, b_M) \quad = \quad (1,1)\left[(\sum_{n=0}^{M} R_1^n)K_1 + (\sum_{n=0}^{M} R_2^n)K_2\right]$$

$$
\begin{aligned}
&= (1,1)\Bigg[(V_1,V_2)\begin{pmatrix} \sum_{n=0}^{M}\gamma_1^n, & 0 \\ 0, & \sum_{n=0}^{M}\gamma_2^n \end{pmatrix}(V_1,V_2)^{-1}K_1 \\
&\qquad + (V_3,V_4)\begin{pmatrix} \sum_{n=0}^{M}\gamma_3^n, & 0 \\ 0, & \sum_{n=0}^{M}\gamma_4^n \end{pmatrix}(V_3,V_4)^{-1}K_2\Bigg] \\
&= \left(\sum_{i=1}^{4}\alpha_i E_{M,i}, \sum_{i=1}^{4}\beta_i E_{M,i}\right), \qquad (14)
\end{aligned}
$$

where $E_{M,i} = \sum_{n=0}^{M}\gamma_i^n$, and α_i, β_i, $i = 1, 2, 3, 4$, are constants as defined before.

Proposition 2 *If $\rho < 1$, then the Markov process is ergodic and there exists a positive probability vector X_0 such that equation (13) is satisfied. Moreover, either*

(i) (a_M, b_M) does not converge to finite (a, b) but the ratio a_M/b_M converges to a constant, K, and

$$\frac{P_{F,0}}{P_{S,0}} = -\lim_{M\to\infty}\frac{a_M}{b_M} = -K, \qquad (15)$$

(ii) or, (a_M, b_M) converges to finite vector (a, b) and

$$aP_{S,0} + bP_{F,0} = 1. \qquad (16)$$

To prevent divergence, eigenvalues of the rate matrices must be restricted to the inside of the unit disk. The fact that the spectral radii of the rate matrices in our system could be no smaller than one provides us with equation (15) or (16), a second equation that we seek.

3.3 Average waiting time and queue length.

Once we know the complete distribution of the number in the system, we can compute all the important queueing measures - average number in the system, average queue length, average waiting time in the system, and average waiting time in queue. In fact we only need to compute any one of the four. The others follow easily from Little's Law and $W_s = W_q + 1/\mu$. We will focus on finding the average number in the system.

Because $X_n = (R_1^n K_1 + R_2^n K_2)X_0$, $\forall n$, we can let L denote the long-run average number in the system, so that $L = (1,1)\sum_{n=0}^{\infty} n(R_1^n K_1 + R_2^n K_2)X_0$.

We can find L from the following two equations. Let $G = \sum_{n=0}^{\infty} nP_{S,n}$ and $H = \sum_{n=0}^{\infty} nP_{F,n}$, then $L = G + H$. There are many ways to find G and H, including differentiation of the moment generating functions. The following are just two examples.

$$
\begin{aligned}
(\mu_S - \lambda)G + (\mu_F - \lambda)H &= \lambda \qquad (17) \\
\mu_S p_{SF} G - \mu_F p_{FS} H &= \frac{p_{FS}}{p_{SF} + p_{FS}}\cdot\frac{\lambda^2}{\mu_S} + \lambda P_{S,0} - \frac{\lambda p_{FS}}{(p_{SF} + p_{FS})} \qquad (18)
\end{aligned}
$$

We can also directly compute L from the matrices R_1, K_1, R_2, K_2 - which we have already obtained when determining X_0. This method will be particularly useful in the finite waiting space case. So we will defer the discussion till then.

Detailed derivation of (17) and (18) can be found in Appendix C.

4 M/MM/1/N queueing system solution

In many applications, there is a physical limitation on the waiting space and the system loss probability is of primary concern. In the following analysis we assume a limited capacity of N in the system, and any job that arrives when there are already N jobs in the system is lost.

We use the same $P_{S,n}$ and $P_{F,n}$ notation. The new balance equations are as follows:

$$\begin{aligned}
\lambda P_{S,0} &= \mu_S p_{SS} P_{S,1} + \mu_F p_{FS} P_{F,1} \\
\lambda P_{F,0} &= \mu_S p_{SF} P_{S,1} + \mu_F p_{FF} P_{F,1}, \\
(\mu_S + \lambda) P_{S,n} &= \lambda P_{S,n-1} + \mu_S p_{SS} P_{S,n+1} + \mu_F p_{FS} P_{F,n+1} && (19) \\
(\mu_F + \lambda) P_{F,n} &= \lambda P_{F,n-1} + \mu_S p_{SF} P_{S,n+1} + \mu_F p_{FF} P_{F,n+1}, \quad 1 \le n < N && (20) \\
\mu_S P_{S,N} &= \lambda P_{S,N-1} && (21) \\
\mu_F P_{F,N} &= \lambda P_{F,N-1}. && (22)
\end{aligned}$$

Again we need two equations to solve for $P_{S,0}$ and $P_{F,0}$. The first comes from the solution of (19) and (20). As in the infinite waiting space case, we know that there exist R_1, R_2, K_1, and K_2 such that (8)-(10) are satisfied and for all n,

$$X_n = (R_1^n K_1 + R_2^n K_2) X_0. \tag{23}$$

In particular

$$X_{N-1} = (R_1^{N-1} K_1 + R_2^{N-1} K_2) X_0 \quad \text{and} \quad X_N = (R_1^N K_1 + R_2^N K_2) X_0. \tag{24}$$

This, together with (21) and (22), implies

$$\begin{pmatrix} \mu_S & 0 \\ 0 & \mu_F \end{pmatrix} (R_1^N K_1 + R_2^N K_2) X_0 = \lambda (R_1^{N-1} K_1 + R_2^{N-1} K_2) X_0.$$

So

$$\left[(R_1 - J) R_1^{N-1} K_1 + (R_2 - J) R_2^{N-1} K_2 \right] X_0 = 0 \tag{25}$$

where $J = \begin{pmatrix} \lambda/\mu_S & 0 \\ 0 & \lambda/\mu_F \end{pmatrix}$ provides us with the first equation we need.

The second equation is obtained from the normalization condition that the probabilities sum to one. In this finite waiting space case, we do not have the problem of divergence. Therefore the second equation is quite straightforward:

$$(1,1) \sum_{n=0}^{N} X_n = 1 \qquad \Rightarrow \qquad (1,1) \sum_{n=0}^{N} (R_1^n K_1 + R_2^n K_2) X_0 = 1. \tag{26}$$

We will use the following algebraic identities to facilitate the computation of $\sum R^n$ and $\sum n R^n$.

Let

$$f_1(x, N) = \sum_{n=0}^{N} x^n = \frac{1 - x^{N+1}}{1 - x},$$

$$f_2(x,N) = \sum_{n=0}^{N} nx^n = \frac{x - x^{N+1}(1+N-Nx)}{(1-x)^2},$$

$$g_1(x,N) = \sum_{n=1}^{N} nx^{n-1} = \frac{1 - x^{N}(1+N-Nx)}{(1-x)^2},$$

and

$$g_2(x,N) = \sum_{n=1}^{N} n^2 x^{n-1} = \frac{1 + x - x^N(1+2N+N^2+x-2Nx-2N^2x+N^2x^2)}{(1-x)^3},$$

then

- if $R = (V_1, V_2)\begin{pmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{pmatrix}(V_1,V_2)^{-1}$, then $R^n = (V_1,V_2)\begin{pmatrix} \gamma_1^n & 0 \\ 0 & \gamma_2^n \end{pmatrix}(V_1,V_2)^{-1}$,

$$\sum_{n=0}^{N} R^n = (V_1,V_2)\begin{pmatrix} f_1(\gamma_1,N) & 0 \\ 0 & f_1(\gamma_2,N) \end{pmatrix}(V_1,V_2)^{-1},$$

and

$$\sum_{n=0}^{N} nR^n = (V_1,V_2)\begin{pmatrix} f_2(\gamma_1,N) & 0 \\ 0 & f_2(\gamma_2,N) \end{pmatrix}(V_1,V_2)^{-1};$$

- if $R = (V_1, U_1)\begin{pmatrix} \gamma & 1 \\ 0 & \gamma \end{pmatrix}(V_1,U_1)^{-1}$, then $R^n = (V_1,U_1)\begin{pmatrix} \gamma^n & n\gamma^{n-1} \\ 0 & \gamma^n \end{pmatrix}(V_1,U_1)^{-1}$,

$$\sum_{n=0}^{N} R^n = (V_1,U_1)\begin{pmatrix} f_1(\gamma,N) & g_1(\gamma,N) \\ 0 & f_1(\gamma,N) \end{pmatrix}(V_1,U_1)^{-1},$$

and

$$\sum_{n=0}^{N} nR^n = (V_1,U_1)\begin{pmatrix} f_2(\gamma,N) & g_2(\gamma,N) \\ 0 & f_2(\gamma,N) \end{pmatrix}(V_1,U_1)^{-1}.$$

Note that all the γ_i's and V_i's have already been obtained in the process of computing R_1 and R_2. So the above computations are straightforward. Using these identities, we can simplify (25) and (26) and quickly compute X_0. After that, we can calculate X_n for all n via (23). The other queueing measures follow from straightforward computation and will not be presented here. Again, we will take advantage of the fact that we have already obtained all the eigenvalues and eigenvectors in computing X_0 to facilitate these computations. For example, to find the long-run average number-in-system, we directly compute $L = \sum_{n=0}^{N} nX_n$, using the identities above concerning $\sum nR^n$.

Because arrivals are Poisson, the PASTA property for continuous-time Markov chains (see Wolff [7], for example) implies that the loss probability equals the probability that there are N in the system: $P_{S,N} + P_{F,N} = (1,1)X_N$.

Remark 1 Naoumov [4] proves that, in finite QBD systems, the steady-state distribution of the number-in-system may be described as the superposition of two matrix-geometric series: $X_n = R^n a + S^{N-n} b$. Here a and b are vectors that satisfy certain boundary conditions. While this solution form holds for $m \geq 3$, the calculation of the two matrices, R and S, necessitates the computation of two infinite series of (recursively defined) matrices. In particular,

$$R = \lim_{k\to\infty} R_k, \qquad \text{where} \qquad R_0 = 0, \qquad R_{k+1} = R_k^2 - (D - I)R_k + C$$
$$S = \lim_{k\to\infty} S_k, \qquad \text{where} \qquad S_0 = 0, \qquad S_{k+1} = CS_k^2 - (D - I)S_k + I$$

Therefore in the case of a 2-state M/MM/1 system, we extend his results by providing more properties of the rate matrices and by providing a computationally more efficient procedure.

5 Numerical analysis

In this section, we study the performance difference between the M/MM/1 system and an analogous M/G/1 system in which service times have the same first two moments as those in the M/MM/1 system but are *i.i.d.*

5.1 Infinite waiting space: the M/MM/1 system.

We first consider the case of systems with infinite waiting spaces. The Pollaczek-Khintchine formula implies that the average queue length of an M/G/1 system depends on the service time distribution only through the first two moments (for example, see Wolff [7, page 385]). Therefore, without loss of generality, when calculating queueing measures such as the average queue length, we can assume that the M/G/1 system has *i.i.d.* hyper-exponential (H_2) service times, with $p_{FS}/(p_{SF} + p_{FS})$ fraction of the services being slow and $p_{SF}/(p_{SF} + p_{FS})$ fraction being fast.

The cases we study include a wide variety of scenarios: high/low system utilization, high/medium/low switching probabilities, and combinations of these. In Table 1 we report the average queue length in these systems, and we observe two interesting phenomena from these results.

First, when $\mu_S < \lambda$ and p_{SF} is very small (at the same time p_{FS} cannot be very large as otherwise the system may be unstable), the expected queue length in the M/MM/1 system is much larger than that in the M/H$_2$/1 system. This is not surprising: when p_{SF} is small, once the server becomes slow it tends to stay slow for a long time; if at the same time $\mu_S < \lambda$, then the queue length grows very quickly. In the corresponding M/H$_2$/1 system, however, the *i.i.d.* service times prevent this from happening, and the system backlog fluctuates less. Neuts [5, page 266, Example 2] observes similar numerical phenomenon as well.

Second, when $p_{SF} + p_{FS} > 1$, the expected backlog in the M/H$_2$/1 queue actually exceeds that for the M/MM/1 system. This phenomenon is somewhat unexpected because one would normally think that the serial correlations among the modulated service times would cause the M/MM/1 system to have a worse performance than the M/H$_2$/1 system.

As we noted before Theorem 1, however, the M/H$_2$/1 system is in fact an M/MM/1 system with switching probabilities $(p'_{SF}, p'_{FS}) = \frac{1}{p_{SF}+p_{FS}}(p_{SF}, p_{FS})$ where $p'_{SF} + p'_{FS} = 1$. So, the comparison in Table 1 is

Table 1: M/MM/1/∞ vs M/G/1/∞

λ	μ_S	μ_F	p_{SF}	p_{FS}	ρ	Q M/MM/1	Q M/H_2/1	Q % Diff.
10	8	50	0.6	0.3	0.550	1.262	1.217	-3.60%
			0.30	0.60	0.900	10.80	10.55	-2.34%
			0.90	0.15	0.350	0.389	0.396	1.77%
			0.15	0.90	1.100	–*	–*	–*
			0.15	0.15	0.725	4.596	2.914	-36.61%
			0.55	0.55	0.725	2.836	2.914	2.73%
			0.90	0.90	0.725	2.518	2.914	15.74%
10	12.5	50	0.60	0.30	0.400	0.409	0.400	-2.20%
			0.30	0.60	0.600	1.115	1.100	-1.39%
			0.90	0.15	0.286	0.174	0.176	1.00%
			0.15	0.90	0.714	1.934	1.940	0.30%
			0.15	0.15	0.500	0.867	0.680	-21.56%
			0.55	0.55	0.500	0.669	0.680	1.67%
			0.90	0.90	0.500	0.619	0.680	9.82%

* Unstable.

equivalent to the comparison between an M/MM/1 system with switching probabilities (p_{SF}, p_{FS}) and an M/MM/1 system with switching probabilities (p'_{SF}, p'_{FS}).

When $p_{SF} + p_{FS} > 1$, $p'_{SF} < p_{SF}$ and $p'_{FS} < p_{FS}$. From the intuition obtained in the first observation, we conjecture that because the M/MM/1 system representing the M/G/1 analogue has smaller switching probabilities, the underlying Markov Chain tends to stay in both states longer and therefore the system performance is actually worse than the original M/MM/1 system. Conversely, when $p_{SF} + p_{FS} < 1$, the M/G/1 system performs better.

Conjecture 1 *The long-run average queue length of the M/MM/1 system is smaller than that of its M/G/1 analogue when $p_{SF} + p_{FS} > 1$, larger when $p_{SF} + p_{FS} < 1$, and the same when $p_{SF} + p_{FS} = 1$.*

We will not attempt to prove the conjecture in this paper. The numerical results in Table 1, however, show that this conjecture holds for a wide variety of examples.

Most significantly, we find concrete examples to show that the system performance (average queue length, average number in the system, average waiting time in queue, and average waiting time in the system) of the M/MM/1 system is not necessarily worse or better than its analogous M/G/1 system. As the conjecture states, the difference appears to depend on the switching probabilities.

5.2 The M/MM/1/N System.

We next compare results for systems with finite waiting spaces. Table 2 reports results that are computed for the same set of parameters as those in Table 1. The difference here is that there is an $N = 7$ limit on the waiting space. In addition, we also compare the loss probabilities here.

Table 2: M/MM/1/N vs M/G/1/N when N=7

λ	μ_S	μ_F	p_{SF}	p_{FS}	Q % Diff.	P{Loss} M/MM/1/7	P{Loss} $M/H_2/1/7$	P{Loss} % Diff.
10	8	50	0.6	0.3	-1.93%	3.11%	2.93%	-5.76%
			0.30	0.60	-0.44%	11.05%	10.85%	-1.79%
			0.90	0.15	1.28%	0.78%	0.82%	4.49%
			0.15	0.90	0.01%	17.92%	17.96%	0.22%
			0.15	0.15	-13.35%	9.08%	6.13%	-32.49%
			0.55	0.55	0.98%	5.93%	6.13%	3.28%
			0.90	0.90	5.50%	5.06%	6.13%	21.00%
10	12.5	50	0.60	0.30	-1.76%	0.52%	0.49%	-6.28%
			0.30	0.60	-0.86%	1.91%	1.86%	-2.70%
			0.90	0.15	0.86%	0.14%	0.14%	4.15%
			0.15	0.90	0.14%	3.37%	3.39%	0.47%
			0.15	0.15	-14.82%	1.63%	1.02%	-37.67%
			0.55	0.55	1.20%	0.98%	1.02%	4.17%
			0.90	0.90	7.13%	0.79%	1.02%	27.98%

Note that Conjecture 1 not only holds in this finite-waiting-space for the expected queue length, it also holds for the loss probabilities. The intuition provided in the previous section also appears to apply here.

Note also that the relative difference in loss probabilities between the M/MM/1 system and its $M/H_2/1$ analogue is magnitudes higher than the difference in expected queue length in all the cases. This suggests that while an M/G/1 approximation may perform well in terms of the expected queue length, it may not be a good approximation in terms of real loss probability.

6 Conclusion

Mitrani and Chakka [3] show that the mixed-geometric solution form holds for $m \geq 3$ as well. However, they focus only on the cases in which all eigenvalues have multiplicity of 1, and if some eigenvalue has multiplicity greater than 1, they assume that linearly independent eigenvectors always exist. We believe that our analysis and procedures can be extended to $m \geq 3$. In particular, our Jordan-form approach should remain valid in the cases where there are duplicate eigenvalues, though there are several difficulties: 1) high dimensionality of the matrices; 2) lack of closed-form solution to high degree polynomial equation (27); and 3) difficulty in numerically inverting large matrices. Nevertheless, there are numerical procedures for finding roots to high-degree polynomial equations and, with the fast-increasing available computing power, even large matrices can be inverted relatively quickly.

References

[1] Leonard Kleinrock. *Queueing Systems*, volume 1: Theory. John Wiley & Sons, 1975.

[2] R.M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Mathematical Proceedings of the Cambridge Philosophical Society*, 58(3):497–520, 1965.

[3] Isi Mitrani, and Ram Chakka. Spectral expansion solution for a class of Markov models: application and comparison with the matrix-geometric method. *Performance Evaluation*, 23:241–260, 1995.

[4] Valeri Naoumov. Matrix-multiplicative approach to quasi-birth-and-death processes analysis. In Srinivas R. Chakravarthy and Attahiru S. Alfa, editors, *Matrix-Analytic Methods in Stochastic Models*, pages 87–106, New York, 1996. Marcel Dekker Inc.

[5] Marcel F. Neuts. *Matrix-Geometric Solutions in Stochastic Models : An Algorithmic Approach*. Number 2 in Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 1981.

[6] L. D. Servi. Algorithmic solutions to recursively tridiagonal linear equations with application to multi-dimensional birth-death processes. Working paper, GTE Laboratories Incorporated, Waltham, MA 02254 USA, 1999.

[7] Ronald W. Wolff. *Stochastic Modeling and the Theory of Queues*. Prentice-Hall Inc., Englewood Cliffs, New Jersey 07632, 1989.

A Proof of Theorem 1

Lemma 1 *Suppose R satisfies (8), then if γ is its eigenvalue, it satisfies*

$$\det\left(\gamma^2 I - \gamma D + C\right) = 0. \tag{27}$$

Proof Let γ and V be such that $RV = \gamma V$. Then $R^2 - DR + C = 0$ implies $(R^2 - DR + C)V = 0$. Therefore

$$(\gamma^2 I - \gamma D + C)V = 0 \tag{28}$$

Since V is non-zero, this implies (27). □

Lemma 1 shows that the eigenvalues and eigenvectors of any solution to (8) satisfy (27) and (28). Moreover, it shows that they can be directly computed from (27) and (28). The following two propositions show how to construct the two solutions to (8), $R_{1,2}$, based on the solutions to (27) and (28).

Since $p_{SF} + p_{FS} \neq 1$, there are four roots to equation (27): γ_1, γ_2, γ_3, and γ_4. Let V_1, V_2, V_3, and V_4 be the corresponding vectors given by (28).

Proposition 3 *If V_i and V_j are linearly independent, then $R = (V_i, V_j)\begin{pmatrix} \gamma_i & 0 \\ 0 & \gamma_j \end{pmatrix}(V_i, V_j)^{-1}$ is a solution to (8).*

Proof It can be verified as follows:

$$\begin{aligned}(R^2 - DR + C)(V_i, V_j) &= (V_i, V_j)\begin{pmatrix} \gamma_i^2 & 0 \\ 0 & \gamma_j^2 \end{pmatrix} - D(V_i, V_j)\begin{pmatrix} \gamma_i & 0 \\ 0 & \gamma_j \end{pmatrix} + C(V_i, V_j) \\ &= (\gamma_i^2 V_i, \gamma_j^2 V_j) - D(\gamma_i V_i, \gamma_j V_j) + C(V_i, V_j) \\ &= \mathbf{0}\end{aligned}$$

from (28). Therefore $(R^2 - DR + C) = 0$, as (V_i, V_j) is invertible. □

If a solution to (8), R, is non-diagonalizable, then let $\hat{\gamma}$ be its multiple eigenvalues. Since clearly $R \neq \hat{\gamma} I$, R can be transformed into a Jordan form: $\exists (V, U)$ such that $R = (V, U)\begin{pmatrix} \hat{\gamma} & 1 \\ 0 & \hat{\gamma} \end{pmatrix}(V, U)^{-1}$, i.e.,

$$RV = \hat{\gamma} V \tag{29}$$

$$RU = V + \hat{\gamma} U. \tag{30}$$

The following proposition shows that the inverse is also true.

Proposition 4 *If $\hat{\gamma}$ is a multiple root of (27) and V is its corresponding solution in (28), then there exists a U, linearly independent of V, such that $R = (V, U)\begin{pmatrix} \hat{\gamma} & 1 \\ 0 & \hat{\gamma} \end{pmatrix}(V, U)^{-1}$ is a solution to (8).*

Proof We prove that the required vector, U, can be found via the following equation,

$$(\hat{\gamma}^2 I - D\hat{\gamma} + C)U = -(2\hat{\gamma}I - D)V, \tag{31}$$

and that it always exists. To do this, we note that

$$2\gamma I - D = \frac{d}{d\gamma}(\gamma^2 I - D\gamma + C).$$

Therefore if we denote $\gamma^2 I - D\gamma + C$ by $\begin{pmatrix} w_1(\gamma), & w_2(\gamma) \\ w_3(\gamma), & w_4(\gamma) \end{pmatrix}$, then $-(2\gamma I - D) = \begin{pmatrix} -w_1^{'}(\gamma), & -w_2^{'}(\gamma) \\ -w_3^{'}(\gamma), & -w_4^{'}(\gamma) \end{pmatrix}$.

Furthermore, $\det(\gamma^2 I - D\gamma + C) = w_1(\gamma)w_4(\gamma) - w_2(\gamma)w_3(\gamma)$. $\hat{\gamma}$ being a multiple solution to (27) implies that:

$$\det(\gamma^2 I - D\gamma + C)|_{\gamma=\hat{\gamma}} = w_1(\hat{\gamma})w_4(\hat{\gamma}) - w_2(\hat{\gamma})w_3(\hat{\gamma}) = 0, \tag{32}$$

and

$$\left.\frac{d\det(\gamma^2 I - D\gamma + C)}{d\gamma}\right|_{\gamma=\hat{\gamma}} = w_1^{'}(\hat{\gamma})w_4(\hat{\gamma}) + w_1(\hat{\gamma})w_4^{'}(\hat{\gamma}) - w_2^{'}(\hat{\gamma})w_3(\hat{\gamma}) - w_2(\hat{\gamma})w_3^{'}(\hat{\gamma}) = 0. \tag{33}$$

Moreover, V being a solution to (28) means $(\hat{\gamma}^2 - D\hat{\gamma} + C)V = 0$; i.e. if we denote $V = (v_1, v_2)^{'}$, then

$$\begin{pmatrix} w_1(\hat{\gamma}), & w_2(\hat{\gamma}) \\ w_3(\hat{\gamma}), & w_4(\hat{\gamma}) \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0. \tag{34}$$

Without loss of generality, from (34) we can assume that

$$w_3(\hat{\gamma}) = cw_1(\hat{\gamma}), \quad w_4(\hat{\gamma}) = cw_2(\hat{\gamma}), \quad v_1 = -w_2(\hat{\gamma}), \quad v_2 = w_1(\hat{\gamma}),$$

where c is a constant.

Then (33) reduces to:

$$\begin{aligned} 0 &= w_1^{'}(\hat{\gamma})w_4(\hat{\gamma}) + w_1(\hat{\gamma})w_4^{'}(\hat{\gamma}) - w_2^{'}(\hat{\gamma})w_3(\hat{\gamma}) - w_2(\hat{\gamma})w_3^{'}(\hat{\gamma}) \\ &= w_1^{'}(\hat{\gamma})(-cv_1) + v_2 w_4^{'}(\hat{\gamma}) - w_2^{'}(\hat{\gamma})(cv_2) - (-v_1)w_3^{'}(\hat{\gamma}). \end{aligned}$$

Hence

$$w_3^{'}(\hat{\gamma})v_1 + w_4^{'}(\hat{\gamma})v_2 = c(w_1^{'}(\hat{\gamma})v_1 + w_2^{'}(\hat{\gamma})v_2), \tag{35}$$

and (31) becomes:

$$\begin{pmatrix} w_1(\hat{\gamma}), & w_2(\hat{\gamma}) \\ cw_1(\hat{\gamma}), & cw_2(\hat{\gamma}) \end{pmatrix} U = \begin{pmatrix} -(w_1^{'}(\hat{\gamma})v_1 + w_2^{'}(\hat{\gamma})v_2) \\ -c(w_1^{'}(\hat{\gamma})v_1 + w_2^{'}(\hat{\gamma})v_2) \end{pmatrix}. \tag{36}$$

Since these two equations are linearly dependent, a non-trivial solution U always exists for (31).

Now suppose, by contradiction, that U and V are linearly dependent. Then $U = c_0 V$ for some constant c_0. From (34) this means that if we denote U by $(u_1, u_2)^{'}$, then

$$w_1(\hat{\gamma})u_1 + w_2(\hat{\gamma})u_2 = 0, \quad w_3(\hat{\gamma})u_1 + w_4(\hat{\gamma})u_2 = 0,$$

and therefore from (36)

$$w_1^{'}(\hat{\gamma})u_1 + w_2^{'}(\hat{\gamma})u_2 = 0, \quad w_3^{'}(\hat{\gamma})u_1 + w_4^{'}(\hat{\gamma})u_2 = 0.$$

That is, $(\hat{\gamma}^2 I - D\hat{\gamma} + C)U = 0$ and $(2\hat{\gamma}I - D)U = 0$. Pre-multiplying both equations by A we get $(A\hat{\gamma}^2 - B\hat{\gamma} + \lambda I)U = 0$ and $\frac{d\det(A\gamma^2 - B\gamma + \lambda I)}{d\gamma}\Big|_{\gamma=\hat{\gamma}} U = 0$. Because

$$(A\gamma^2 - B\gamma + \lambda I) = \begin{pmatrix} \mu_S p_{SS}\gamma^2 - (\lambda + \mu_S)\gamma + \lambda, & \mu_F p_{FS}\gamma^2 \\ \mu_S p_{SF}\gamma^2, & \mu_F p_{FF}\gamma^2 - (\lambda + \mu_F)\gamma + \lambda \end{pmatrix},$$

this would imply $(\mu_S p_{SS}\hat{\gamma}^2 - (\lambda+\mu_S)\hat{\gamma} + \lambda)u_1 + \mu_F p_{FS}\hat{\gamma}^2 u_2 = 0$ and $(2\mu_S p_{SS}\hat{\gamma} - \lambda - \mu_S)u_1 + 2\mu_F p_{FS}\hat{\gamma}u_2 = 0$. This means $\hat{\gamma} = \frac{2\lambda}{\lambda+\mu_F}$. Similarly we can show $\hat{\gamma} = \frac{2\lambda}{\lambda+\mu_S}$, and, in turn, that $\mu_S = \mu_F$. Contradiction.

Thus U and V are linearly independent. If we let $R = (V,U)\begin{pmatrix} \hat{\gamma} & 1 \\ 0 & \hat{\gamma} \end{pmatrix}(V,U)^{-1}$, then (29) and (30) hold. Moreover,

$$\begin{aligned}(R^2 - DR + C)(V,U) &= R(\hat{\gamma}V, V + \hat{\gamma}U) - D(\hat{\gamma}V, V + \hat{\gamma}U) + C(V,U) \\ &= (\hat{\gamma}^2 V, 2\hat{\gamma}V + \hat{\gamma}^2 U) - D(\hat{\gamma}V, V + \hat{\gamma}U) + C(V,U) \\ &= ((\hat{\gamma}^2 I - D\hat{\gamma} + C)V, (\hat{\gamma}^2 I - D\hat{\gamma} + C)U + (2\hat{\gamma}I - D)V) \\ &= \mathbf{0}.\end{aligned}$$

Therefore $(R^2 - DR + C) = 0$, as (V,U) is invertible. □

Now, define

$$\rho = \lambda\left[\left(\frac{p_{FS}}{p_{SF} + p_{FS}}\right)\frac{1}{\mu_S} + \left(\frac{p_{SF}}{p_{SF} + p_{FS}}\right)\frac{1}{\mu_F}\right]. \tag{37}$$

Lemma 2

1. *When $\rho \neq 1$, one and only one of the four γ's is 1. Furthermore, the other three eigenvalues cannot all be the same, and, none equals 0.*

2. *The eigenvector corresponding to the eigenvalue 1 is $(\mu_F p_{FS}, \mu_S p_{SF})'$, and it is linearly independent of the eigenvectors of other eigenvalues.*

3. *If $\gamma_i = \gamma_j$, then V_i and V_j are linearly dependent.*

4. *If $\gamma_i \neq \gamma_j$ and V_i and V_j are linearly dependent, $1 \leq i \neq j \leq 4$, then $\gamma_i\gamma_j$ is an eigenvalue of C.*

5. *If V_i, V_j, and V_k are linearly dependent, $1 \leq i \neq j \neq k \leq 4$, then γ_i, γ_j, and γ_k cannot be all distinct.*

Proof

Part 1. Once we substitute $\gamma = 1$ into (27), it is straightforward to verify that the determinant of the resultant matrix is zero.

As a result, (27), or equivalently, $\det(A\gamma^2 - B\gamma + \lambda I) = 0$ can be simplified to

$$\begin{aligned}0 = {} & (\mu_S\mu_F(p_{SS}p_{FF} - p_{SF}p_{FS}))\gamma^4 - ((\lambda+\mu_S)\mu_F p_{FF} + (\lambda+\mu_F)\mu_S p_{SS})\gamma^3 \\ & +(\lambda(\mu_S p_{SS} + \mu_F p_{FF}) + (\lambda+\mu_S)(\lambda+\mu_F))\gamma^2 - (2\lambda^2 - \lambda\mu_S - \lambda\mu_F)\gamma + \lambda^2\end{aligned}$$

$$
\begin{aligned}
&= (\gamma-1)[\mu_S\mu_F(p_{SS}+p_{FF}-1)\gamma^3-(\lambda\mu_S p_{SS}+\lambda\mu_F p_{FF}+\mu_S\mu_F)\gamma^2 \\
&\qquad +\lambda(\lambda+\mu_S+\mu_F)\gamma-\lambda^2].
\end{aligned}
$$

Obviously 0 cannot be a root because $\lambda^2 \neq 0$. Now suppose we have another root that is 1. Then we would have

$$
\mu_S\mu_F(p_{SS}+p_{FF}-1)-(\lambda\mu_S p_{SS}+\lambda\mu_F p_{FF}+\mu_S\mu_F)+\lambda(\lambda+\mu_S+\mu_F)-\lambda^2=0, \tag{38}
$$

which amounts to

$$
1=\lambda\frac{\mu_S p_{SF}+\mu_F p_{FS}}{\mu_S\mu_F(p_{SF}+p_{FS})}, \tag{39}
$$

i.e. $\rho=1$, a contradiction.

Now suppose the other three eigenvalues are the same, $\gamma^{'}$. Then

$$
\begin{aligned}
3\gamma^{'} &= \frac{\lambda\mu_S p_{SS}+\lambda\mu_F p_{FF}+\mu_S\mu_F}{\mu_S\mu_F(p_{SS}+p_{FF}-1)}, && (40)\\
3\gamma^{'2} &= \frac{\lambda(\lambda+\mu_S+\mu_F)}{\mu_S\mu_F(p_{SS}+p_{FF}-1)}, && (41)\\
\gamma^{'3} &= \frac{\lambda^2}{\mu_S\mu_F(p_{SS}+p_{FF}-1)}.
\end{aligned}
$$

(40) and (41) imply

$$
(\lambda\mu_S p_{SS}+\lambda\mu_F p_{FF}+\mu_S\mu_F)^2 = 3\lambda(\lambda+\mu_S+\mu_F)\mu_S\mu_F(p_{SS}+p_{FF}-1)
$$

i.e.

$$
\begin{aligned}
0 &= \lambda^2[\mu_S^2 p_{SS}^2+\mu_F^2 p_{FF}^2+2\mu_S\mu_F p_{SS}p_{FF}-3\mu_S\mu_F(p_{SS}+p_{FF}-1)] && (42)\\
&+ \lambda[2\mu_S^2\mu_F p_{SS}+2\mu_S\mu_F^2 p_{FF}-3(\mu_S+\mu_F)\mu_S\mu_F(p_{SS}+p_{FF}-1)]+\mu_S^2\mu_F^2.
\end{aligned}
$$

To show contradiction, we now prove that (42), as a quadratic equation of λ, has no real roots. That is, the discriminant is negative:

$$
\begin{aligned}
0 &> [2\mu_S^2\mu_F p_{SS}+2\mu_S\mu_F^2 p_{FF}-3(\mu_S+\mu_F)\mu_S\mu_F(p_{SS}+p_{FF}-1)]^2 \\
&\quad -4[\mu_S^2 p_{SS}^2+\mu_F^2 p_{FF}^2+2\mu_S\mu_F p_{SS}p_{FF}-3\mu_S\mu_F(p_{SS}+p_{FF}-1)]\mu_S^2\mu_F^2 \\
&= 3(p_{SS}+p_{FF}-1)\mu_S^2\mu_F^2\{\mu_S^2[-3p_{FS}-p_{SS}]+\mu_F^2[-3p_{SF}-p_{FF}]+\mu_S\mu_F[2(p_{SS}+p_{FF}-1)]\}
\end{aligned}
$$

But $(p_{SS}+p_{FF}-1)>0$, due to (41), and the coefficients of the quadratic terms in the parenthesis are negative. Thus we, again, only need to prove that the discriminant is negative:

$$
\begin{aligned}
0 &> 4(p_{SS}+p_{FF}-1)^2-4(-3p_{FS}-p_{SS})(-3p_{SF}-p_{FF}) \\
&= -16p_{SF}p_{SS}-16p_{FS}p_{FF}-32p_{SF}p_{FS},
\end{aligned}
$$

which is clear. Therefore, we have proved that (40) and (41) are contradictory. As a result, the other three eigenvalues cannot all be the same.

Part 2. If we let $\gamma = 1$, it is straightforward to verify that $(\mu_F p_{FS}, \mu_S p_{SF})'$ is a solution to (28). Moreover, if we fix $V = (\mu_F p_{FS}, \mu_S p_{SF})'$ in (28), then we get the following two equations:

$$\begin{aligned} \mu_S \mu_F p_{FS} \gamma^2 - (\lambda + \mu_S)\mu_F p_{FS}\gamma + \lambda \mu_F p_{FS} &= 0 \\ \mu_S \mu_F p_{SF} \gamma^2 - (\lambda + \mu_F)\mu_S p_{SF}\gamma + \lambda \mu_S p_{SF} &= 0 \end{aligned}$$

The first equation has two roots: 1 and λ/μ_S, and the second equation has two roots: 1 and λ/μ_F. Because $\mu_S < \mu_F$, 1 is then the only solution.

Part 3. By contradiction, suppose $\gamma_i = \gamma_j = \gamma$ and V_i and V_j are linearly independent. Then due to Proposition 3, $R = (V_i, V_j)\begin{pmatrix} \gamma & 0 \\ 0 & \gamma \end{pmatrix}(V_i, V_j)^{-1}$ is a solution to (8). Note, however, that here $R = \gamma I$. But this is impossible because there exists no γ such that $\gamma^2 I - \gamma D + C = \mathbf{0}$.

Part 4. If V_i and V_j are linearly dependent then $V_i = cV_j$ where c is a constant. Therefore,

$$(\gamma_i^2 I - D\gamma_i + C)V_i = 0 \tag{43}$$

$$(\gamma_j^2 I - D\gamma_j + C)V_j = 0 \Longleftrightarrow (\gamma_j^2 I - D\gamma_j + C)V_i = 0. \tag{44}$$

Multiplying (43) by γ_j and (44) by γ_i and taking the difference, we have

$$\begin{aligned} (\gamma_i\gamma_j(\gamma_i - \gamma_j)I + (\gamma_j - \gamma_i)C)V_i &= 0 \\ (\gamma_i\gamma_j I - C)V_i &= 0, \end{aligned}$$

since $\gamma_i \neq \gamma_j$. Therefore $\gamma_i\gamma_j$ is an eigenvalue of C.

Part 5. Suppose, by contradiction, that γ_i, γ_j, and γ_k are distinct. Then from Lemma 4, $\gamma_i\gamma_j$, $\gamma_i\gamma_k$ and $\gamma_j\gamma_k$ are all eigenvalues of C. Furthermore, if γ_i, γ_j, and γ_k are distinct and non-zero, then these three eigenvalues are distinct as well. But C has at most two distinct eigenvalues, a contradiction. □

The following proposition uses the results of Lemma 2 and Propositions 3 and 4 to provide a procedure for determining solution to (7)-(10). Thus it provides a constructive proof of Theorem 1. Without loss of generality, we can let $\gamma_1 = 1$.

Proposition 5

1. *Let $\gamma_i, V_i, i = 1, 2, 3, 4,$ be given by (27) and (28), and let $\gamma_1 = 1$. There are two possibilities:*

   (a) *Suppose there exist a pair of linearly independent vectors in V_2, V_3 and V_4, say V_3 and V_4. There are two possible cases. In the first, γ_2 is different from both γ_3 and γ_4, then $R_1 = (V_1, V_2)\begin{pmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{pmatrix}(V_1, V_2)^{-1}$ and $R_2 = (V_3, V_4)\begin{pmatrix} \gamma_3 & 0 \\ 0 & \gamma_4 \end{pmatrix}(V_3, V_4)^{-1}$ are both solutions to (8). In the other case, $\gamma_2 = \gamma_3$ or γ_4. Without loss of generality, let $\gamma_2 = \gamma_3$. Then $R_1 = (V_1, V_4)\begin{pmatrix} \gamma_1 & 0 \\ 0 & \gamma_4 \end{pmatrix}(V_1, V_4)^{-1}$ and $R_2 = (V_2, U_2)\begin{pmatrix} \gamma_2 & 1 \\ 0 & \gamma_2 \end{pmatrix}(V_2, U_2)^{-1}$ are both solutions to (8), where U_2 is found via Proposition 4 (equation (31)).*

(b) *Suppose V_2, V_3 and V_4 are pair-wise linearly dependent, then γ_2, γ_3, and γ_4 can be neither all distinct nor all the same. Suppose $\gamma_3 = \gamma_4 = \gamma \neq \gamma_2$. Then, $R_1 = (V_1, V_2)\begin{pmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{pmatrix}(V_1, V_2)^{-1}$ and $R_2 = (V_3, U_3)\begin{pmatrix} \gamma & 1 \\ 0 & \gamma \end{pmatrix}(V_3, U_3)^{-1}$ are both solutions to (8), where U_3 is found via Proposition 4.*

2. *R_1 and R_2, as constructed in (1a) and (1b), have no common eigenvalues.*

3. *Let R_1 and R_2 be given in (1a) and (1b). Then there exist K_1 and K_2 such that (9) and (10) are satisfied.*

Proof

Part (1a). We note that in the latter case, V_1 and V_4 are linearly independent according to part 2 of Lemma 2. Part (1a) then follows from Proposition 3 in the former case and Propositions 3 and 4 in the latter case.

Part (1b). We note that parts 1 and 5 of Lemma 2 together show that γ_2, γ_3, and γ_4 can be neither all distinct nor all the same. The rest follows again from Propositions 3 and 4.

Part 2. From our construction of R_1 and R_2 in (1a) and (1b), it is clear that they have no common eigenvalues in all cases. We note that, in the latter case of (1a), $1 = \gamma_1 \neq \gamma_2$ and $\gamma = \gamma_3 \neq \gamma_4$ from part 3 of Lemma 2.

Part 3. The following two lemmas are used in the proof.

Lemma 3 *Let R be a 2×2 matrix with distinct eigenvalues γ_1 and γ_2 and corresponding eigenvectors V_1 and V_2. Then for any vector $V \neq \mathbf{0}$, if $R^2 V = cV$ for a non-zero constant c, then $c = \gamma_i^2$ ($i = 1$ or 2), and $RV = \gamma_i V$ for the same i.*

Proof of Lemma 3 Since γ_1 and γ_2 are distinct, V_1 and V_2 are linearly independent, and V can be expressed as a linear combination of V_1 and V_2: $V = c_1 V_1 + c_2 V_2$. Then $R^2 V = cV$ implies $c_1\gamma_1^2 V_1 + c_2 \gamma_2^2 V_2 = c(c_1 V_1 + c_2 V_2)$, $i = 1$ or 2.

Again, since V_1 and V_2 are linearly independent, this implies $c_1\gamma_1^2 = c_1 c$ and $c_2\gamma_2^2 = c_2 c$. Because c_1 and c_2 cannot both be 0, if $c_1 \neq 0$, then $c = \gamma_1^2$, $c_2 = 0$, and $V = c_1 V_1$; else if $c_2 \neq 0$, then $c = \gamma_2^2$, $c_1 = 0$, and $V = c_2 V_2$. □

Lemma 4 *$R_1 - R_2$ is invertible.*

Proof of Lemma 4 Suppose, by contradiction, that $R_1 - R_2$ is non-invertible. Then there exists $V \neq \mathbf{0}$ such that $R_1 V = R_2 V$. This, together with (8), implies that $R_1^2 V = R_2^2 V$, and hence $R_1 R_2 V = R_2 R_1 V$.

Now from (8), we have:

$$\begin{aligned} [R_1^2 - DR_1 + C]R_2V - [R_2^2 - DR_2 + C]R_1V &= \mathbf{0} \\ R_1(R_1R_2)V - R_2(R_2R_1)V &= \mathbf{0} \\ (R_1 - R_2)(R_1R_2V) &= \mathbf{0}. \end{aligned}$$

Since $R_1 \neq R_2$, the dimension of solution space of $(R_1 - R_2)X = \mathbf{0}$ is at most one. Because V and R_1R_2V are both solutions, we have $R_1R_2V = c_1V$ for some constant c_1. Moreover $R_1V = R_2V$ and $R_1R_2V = c_1V$ imply $R_1^2V = c_1V$, and hence $R_2^2V = c_1V$. Without loss of generality, let R_1 be the one-matrix solution of (8) with one as its eigenvalue. Then from Lemma 1, R_1 has two distinct eigenvalues. Since c_1 is an eigenvalue of R_1^2 and V its corresponding eigenvector, Lemma 3 implies that V is an eigenvector of R_1: $R_1V = \gamma V$. This implies $R_2V = \gamma V$ as well. But R_1 and R_2 do not have common eigenvalues according to part 2 of Proposition 5. This is a contradiction. □

Proof of part 3 of Proposition 5 From (9) and (10), $(R_1-R_2)K_1 = C-R_2$ and $(R_1-R_2)K_2 = R_1-C$. Then by Lemma 4, $K_1 = (R_1-R_2)^{-1}(C-R_2)$, $K_2 = (R_1-R_2)^{-1}(R_1-C)$ is a solution to (8), (9), and (10). □

This concludes the proposition's proof. □

B Proof of Propositions 1 and 2

Proof of Proposition 1 The transition probability matrix of the Embedded Markov Chain (EMC) at service completion epochs is $P = \begin{pmatrix} p_{SS} & p_{SF} \\ p_{FS} & p_{FF} \end{pmatrix}$, with the steady-state distribution $\pi = (\pi_S, \pi_F) = \left(\frac{p_{FS}}{p_{SF}+p_{FS}}, \frac{p_{SF}}{p_{SF}+p_{FS}}\right)$ such that $\pi = \pi P$. Note that (π_S, π_F) are also the long-run proportion of slow and fast services. More specifically, if we let $m(n)$ be the number of slow services in the first n services the server provides, then $\lim_{n\to\infty} m(n) = \infty$, $\lim_{n\to\infty} \frac{m(n)}{n} = \pi_S$ and $\lim_{n\to\infty} \frac{n-m(n)}{n} = \pi_F$ with probability one.

Let $S_1, S_2, \ldots$ denote the sequence of services provided by this server, and let
$\Omega_n = \{i : i \leq n \text{ and } S_i \text{ is a slow service}\}$, then $m(n) = |\Omega_n|$, and

$$\begin{aligned} \lim_{n\to\infty} \frac{\sum_{i=1}^n S_i}{n} &= \lim_{n\to\infty} \left(\frac{\sum_{i\in\Omega_n} S_i}{n} + \frac{\sum_{i\notin\Omega_n} S_i}{n} \right) \\ &= \lim_{n\to\infty} \left(\frac{\sum_{i\in\Omega_n} S_i}{m(n)} \cdot \frac{m(n)}{n} + \frac{\sum_{i\notin\Omega_n} S_i}{n-m(n)} \cdot \frac{n-m(n)}{n} \right) \\ &= \frac{\pi_S}{\mu_S} + \frac{\pi_F}{\mu_F} = \frac{p_{FS}}{p_{SF}+p_{FS}} \frac{1}{\mu_S} + \frac{p_{SF}}{p_{SF}+p_{FS}} \frac{1}{\mu_F} \end{aligned}$$

with probability 1. Hence the long-run average service time $1/\mu$ is defined as in (11), and it follows from (37) that $\rho = \lambda/\mu$. When $\rho < 1$, by Little's Law, ρ is the long-run average fraction of time the system is busy.

To derive stability conditions, we define (in Loynes's [2] notation) $T_1, T_2, \ldots$ to be the sequence of inter-arrival times. Moreover, define $U_n = S_n - T_n$. Then

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} U_k}{n} = \frac{1}{\mu} - \frac{1}{\lambda} \begin{cases} < 0 & \text{if } \rho < 1, \\ \geq 0 & \text{if } \rho \geq 1 \end{cases} \quad w.p.1.$$

Therefore, the system is stable when $\rho < 1$ and unstable when $\rho \geq 1$. This follows directly from Theorems 1 and 2 and Corollary 1 in Loynes [2]. □

Proof of Proposition 2 From (12), we know that $P_{S,0}+P_{F,0} > 0$ when $\rho < 1$. Suppose, by contradiction, that $P_{S,n}$ (or $P_{F,n}$) equals zero for some $n \geq 0$. Then from equations (1)-(4), $P_{S,n+1} = 0$ (or $P_{F,n+1} = 0$). Because $p_{SF} \cdot p_{FS} \neq 0$, it follows that $P_{F,n} = 0$ (or $P_{S,n} = 0$), and recursively $P_{S,k} = P_{F,k} = 0$ for all k. This contradicts $P_{S,0} + P_{F,0} > 0$. Therefore, all the probabilities ($P_{S,n},\ P_{F,n},\ \forall n \geq 0$) are positive, and the Markov process is ergodic.

Because there might be identical γ's in $\gamma_1, \gamma_2, \gamma_3, \gamma_4$, we first collect terms in (14). Then we denote by γ_i the lowest ranking γ whose α and β coefficients are not both zero.

Suppose $|\gamma_i| \geq 1$, and without loss of generality, suppose $\alpha_i \neq 0$. Then since $\lim_{M\to\infty} |E_{M,i}| = \infty$ and $a_M = \sum_{i=1}^{4} \alpha_i E_{M,i}$, this means $\lim_{M\to\infty} |a_M| = \infty$. Because $\lim_{M\to\infty} (a_M, b_M) X_0 = 1$, this also implies that $\lim_{M\to\infty} |b_M| = \infty$. Otherwise we would have $P_{S,0} = 0$, contradicting the fact that the solution to (13) is positive. So $\beta_i \neq 0$ as well.

The coefficient of the γ_i^M term in $(a_M, b_M)X_0$ is $\alpha_i P_{S,0} + \beta_i P_{F,0}$. As $M \to \infty$, this coefficient must vanish. Therefore

$$\frac{P_{F,0}}{P_{S,0}} = - \lim_{M\to\infty} \frac{a_M}{b_M} = -\frac{\alpha_i}{\beta_i}. \tag{45}$$

This corresponds to the first case in the proposition statement.

Now suppose $|\gamma_i| < 1$. Since the eigenvalues with non-zero α and β coefficients all lie within the unit disk, we have, from (13), that $\lim_{M\to\infty} (a_M, b_M) = (a, b)$ for finite (a, b). Therefore, from (13), we have equation (16). This corresponds to the second case in the proposition statement. □

C Derivation of L in the infinite waiting space case

To derive (17), we first balance flows across cuts 1 in the state-transition diagram (Figure 2). As a result we obtain the following equations:

$$\lambda(P_{S,n} + P_{F,n}) = \mu_S P_{S,n+1} + \mu_F P_{F,n+1} \quad \forall n \tag{46}$$

If we let $G = \sum_{n=0}^{\infty} n P_{S,n}$, $H = \sum_{n=0}^{\infty} n P_{F,n}$, multiply both sides of (46) by $n+1$, and sum over all n, then we obtain

Figure 2: Two cuts in the state-transition diagram

$$\begin{aligned}
\lambda\sum_{n=0}^{\infty}((n+1)P_{S,n}+(n+1)P_{F,n}) &= \mu_S\sum_{n=0}^{\infty}(n+1)P_{S,n+1}+\mu_F\sum_{n=0}^{\infty}(n+1)P_{F,n+1} \\
\lambda(G+H)+\lambda &= \mu_S G+\mu_F H.
\end{aligned}$$

This is (17), the first equation needed.

Next we rewrite (6):

$$\begin{aligned}
X_{n+2}-DX_{n+1}+CX_n &= 0, \qquad \forall n\geq 0 \\
X_1 &= CX_0.
\end{aligned} \tag{47}$$

Again, we multiply (47) by $n+1$ and sum over n from 0 to ∞, to obtain:

$$\begin{aligned}
\sum_{n=0}^{\infty}(n+2)X_{n+2}-\sum_{n=0}^{\infty}X_{n+2} &= D\sum_{n=0}^{\infty}(n+1)X_{n+1}-C\sum_{n=0}^{\infty}nX_n-C\sum_{n=0}^{\infty}X_n \\
\sum_{n=2}^{\infty}nX_n-\sum_{n=2}^{\infty}X_n &= D\sum_{n=1}^{\infty}nX_n-C\sum_{n=0}^{\infty}nX_n-C\sum_{n=0}^{\infty}X_n \\
\sum_{n=0}^{\infty}nX_n-X_1-\sum_{n=2}^{\infty}X_n &= D\sum_{n=0}^{\infty}nX_n-C\sum_{n=0}^{\infty}nX_n-C\sum_{n=0}^{\infty}X_n \\
(I-D+C)\sum_{n=0}^{\infty}nX_n &= \sum_{n=1}^{\infty}X_n-C\sum_{n=0}^{\infty}X_n.
\end{aligned} \tag{48}$$

Now if we balance the flow across cut 2 in Figure 2, then we obtain $p_{SF}\mu_S\sum_{n=1}^{\infty}P_{S,n}=p_{FS}\mu_F\sum_{n=1}^{\infty}P_{F,n}$. Moreover, $\sum_{n=1}^{\infty}(P_{S,n}+P_{F,n})=\rho$. Therefore, $\sum_{n=1}^{\infty}P_{S,n}=\frac{p_{FS}}{p_{SF}+p_{FS}}\cdot\frac{\lambda}{\mu_S}$ and $\sum_{n=1}^{\infty}P_{F,n}=\frac{p_{SF}}{p_{SF}+p_{FS}}\cdot\frac{\lambda}{\mu_F}$, so (48) becomes:

$$\begin{pmatrix} -\mu_S P_{SF} & \mu_F P_{FS} \\ \mu_S P_{SF} & -\mu_F P_{FS} \end{pmatrix}\begin{pmatrix} G \\ H \end{pmatrix}=\begin{pmatrix} \frac{P_{FS}\lambda}{P_{SF}+P_{FS}} \\ \frac{P_{SF}\lambda}{P_{SF}+P_{FS}} \end{pmatrix}-\begin{pmatrix} \frac{p_{FS}}{p_{SF}+p_{FS}}\cdot\frac{\lambda}{\mu_S}+P_{S,0} \\ \frac{p_{SF}}{p_{SF}+p_{FS}}\cdot\frac{\lambda}{\mu_F}+P_{F,0} \end{pmatrix},$$

out of which we obtain (only) one independent equation, equation (18).