Probability in the Engineering and Informational Sciences, page 1 of 42, 2021.
doi:10.1017/S0269964821000048

SCALING PROPERTIES OF QUEUES WITH TIME-VARYING LOAD PROCESSES: EXTENSIONS AND APPLICATIONS

Rein Vesilo
School of Engineering, Macquarie University, Sydney, Australia
E-mail: rein.vesilo@mq.edu.au

Mor Harchol-Balter
Department of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Alan Scheller-Wolf
Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

New computing and communications paradigms will result in traffic loads in information server systems that fluctuate over much broader ranges of time scales than current systems. In addition, these fluctuation time scales may only be indirectly known or even be unknown. However, we should still be able to accurately design and manage such systems. This paper addresses this issue: we consider an $M/M/1$ queueing system operating in a random environment (denoted $M/M/1(R)$) that alternates between HIGH and LOW phases, where the load in the HIGH phase is higher than in the LOW phase. Previous work on the performance characteristics of $M/M/1(R)$ systems established fundamental properties of the shape of performance curves. In this paper, we extend monotonicity results to include convexity and concavity properties, provide a partial answer to an open problem on stochastic ordering, develop new computational techniques, and include boundary cases and various degenerate $M/M/1(R)$ systems. The basis of our results are novel representations for the mean number in system and the probability of the system being empty. We then apply these results to analyze practical aspects of system operation and design; in particular, we derive the optimal service rate to minimize mean system cost and provide a bias analysis of the use of customer-level sampling to estimate time-stationary quantities.

Keywords: cubic polynomial, $M/M/1$ single-server queue, monotonicity, random environment, sampling, time varying load

1. INTRODUCTION

With the explosion of new computing and communications paradigms such as big data, fog computing, Internet-of-Things, machine-to-machine communication, and mobile edge computing and caches, system traffic loads will fluctuate over a much broader range of time

0269-9648/19 $25.00

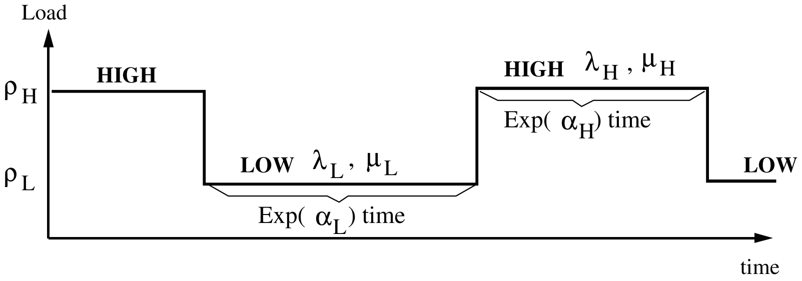

FIGURE 1. Illustration of the $M/M/1(R)$ model.

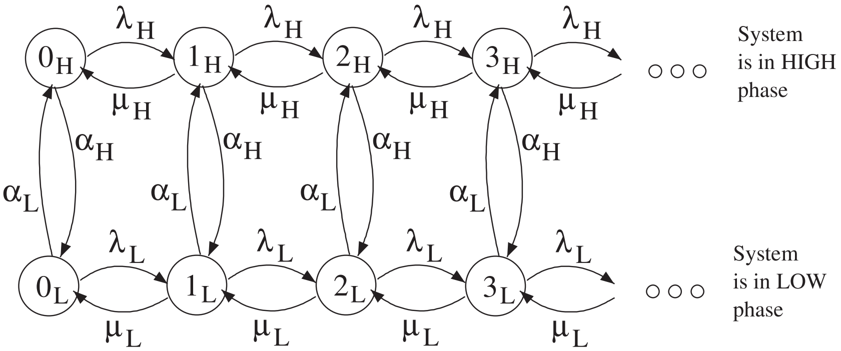

FIGURE 2. Markov chain for the $M/M/1(R)$ model.

scales than extant systems. At times, the load may vary extremely rapidly, while on other occasions, the fluctuations may be slow, with many possible rates in between. In some cases, the fluctuation time scales may only be known indirectly, while in other cases, they may not even be known at all. Under such circumstances, many performance analysis tools will struggle to maintain accuracy—in particular, performance estimates using an $M/M/1$ queue model operating under the average arrival and service rates may be grossly in error—yet we still need to be able to accurately design and operate systems for these circumstances.

The most common model of a single-server system under time-varying load conditions is an extension of the classic $M/M/1$ queue to an $M/M/1$ queue "operating in a random environment," denoted $M/M/1(R)$, illustrated in Figure 1. The $M/M/1(R)$ system consists of a single-server queue with an infinite waiting area in which arriving customers, belonging to a single class, are served in first-come first-served (FCFS) order. The system alternates between two phases: LOW and HIGH. Times spent in the LOW and HIGH phases are assumed to be mutually independent exponentially distributed random variables with rate parameters α_L and α_H, respectively. The arrival rates of customers are λ_L and λ_H, in the LOW and HIGH phases, respectively; the service rates are μ_L and μ_H, respectively. The load quantities are defined by $\rho_L = \lambda_L/\mu_L$ and $\rho_H = \lambda_H/\mu_H$. The load in the HIGH phase, ρ_H, is assumed to be larger than the load in the LOW phase, ρ_L, that is, $\rho_L \leq \rho_H$, but no particular relationship between λ_L and λ_H or between μ_L and μ_H is assumed. The Markov Chain associated with our $M/M/1(R)$ model is given in Figure 2.

There has been considerable research on performance measures for the $M/M/1(R)$ system. The paper Yechiali and Naor [34] was one of the earliest to address an $M/M/1(R)$ system, using a semi-infinite two-dimensional continuous-time Markov chain; this and subsequent papers have produced accurate results for quantities such as the mean number in system[1] (and hence, via Little's Law, mean response time) and the probability of the system is empty. However, there are only a few papers that provide qualitative insight into how the performance of the $M/M/1(R)$ system varies as a function of its primitives, which is a fundamental concern to us here.

One such paper, Gupta *et al.* [17], introduced a scaling parameter (a function of α_H and α_L) to vary the rate of switching between phases and examined the shape of the performance curves as a function of this rate. It found that the response time varies monotonically between the two extreme cases of switching rate zero and switching rate infinity. Furthermore, the effect of the rate of fluctuation on the mean response time is determined by the relative "slack" in each phase ($\mu_L - \lambda_L$ and $\mu_H - \lambda_H$). It also proved that the number in system at the end of a HIGH phase is stochastically larger than the number at the end of a LOW phase and the number in a stationary "average" system. It left as an open problem possible stochastic monotonicity of the number in system at the end of a HIGH phase as the phase switching rate increases.

While certainly an important contribution, Gupta *et al.* [17] leave a number of questions about $M/M/1(R)$ systems unanswered. We complement Gupta *et al.* [17] with the following contributions:

1. We sharpen the monotonicity results of Gupta *et al.* [17] by showing that the mean number in system at the end of a HIGH phase is a convex function of the scaling parameter, while the mean number in system at the end of a LOW phase can be concave, or convex, or a constant, depending on the load in the HIGH phase. We also show that the time-stationary response time can be either a convex or a concave function of the switching rate depending on the slack in the LOW and HIGH phases.
2. Gupta *et al.* [17] treat the cases of the HIGH phase being underloaded and overloaded separately and do not consider the boundary case of $\rho_H = 1$ at all. We derive a unified expression for all of these cases, providing a more holistic analysis.
3. We provide an answer to the open question posed by Gupta *et al.* [17] regarding whether the number of customers in the system at the end of a HIGH phase is stochastically monotonic with the switching rates. We show that this is not always the case, but on a positive note, we provide sufficient conditions for it to be true.
4. Gupta *et al.* [17] are silent on how the analysis of the $M/M/1(R)$ queue can be used in system evaluation and design. We present two applications of our results to this end. The first considers the problem of sampling—since neither the arrival nor departure processes of an $M/M/1(R)$ system are Poisson in general, sampling at customer epochs will introduce bias. We obtain expressions for the degree of bias in sampling and provide expressions to correct for biasing errors. The second application is on the optimal design of an $M/M/1(R)$ service system: Assuming service rate costs are proportional to the service rate and holding time costs are proportional to time in system, we show that the optimal service rate balances "slack" (as opposed to load) across phases.
5. Gupta *et al.* [17] do not address either "degenerate" systems (such as an $M/M/1$ queue with interrupted arrivals, or service, or both) or "generalized" $M/M/1(R)$ systems (such as an $M/M/c$ system operating in a random environment, or an $M/M/1$ queue with disasters). We explicitly consider these systems, both focusing and generalizing the $M/M/1(R)$ analysis to these and other important models.

[1] Results are in terms of roots of cubic equations; this will be discussed later.

As in Yechiali and Naor [34] and Gupta *et al.* [17], much of our analysis relies on the solution of the equations for the z-transform of the number in system at the end of the HIGH and LOW phases, which in general is a cubic polynomial. Our results are facilitated by our broader investigation of the roots of this polynomial: Instead of focusing on one root of the cubic as these works did, we exploit all the roots of the cubic, thereby obtaining simpler expressions which we can leverage for our results. And, as the solution of this cubic is crucial to our analysis, we also provide a new computational method for obtaining the roots of the $M/M/1(R)$ cubic polynomial.

An overview of the paper is as follows. Section 2 provides an overview of related work on different techniques for analyzing $M/M/1(R)$ systems. Section 3 presents the background on generating function techniques needed to analyze the $M/M/1(R)$ system. Section 4 presents our new representations used to obtain the mean number in system and the probability of an empty system that are the foundation of our results. Section 5 presents the results on the sensitivity and convexity/concavity properties of performance measures. Section 6 presents the results that give a partial answer to the stochastic monotonicity of the number in system at the end of a HIGH phase. Section 7 presents the two applications of our results: the allocation of the service rate to minimize cost, and the analysis of sampling bias. Section 8 presents the new method of computing roots. Section 9 shows how degenerate $M/M/1(R)$ systems can be analyzed and shows how the techniques developed for the $M/M/1(R)$ system can be applied to the analysis of more sophisticated types of systems operating in a random environment. The conclusion is given in Section 10.

2. RELATED WORK

Analysis of time-varying $M/M/1$ queues and related models has been of long-standing interest. We present a short review of work most relevant for this paper; for a more comprehensive coverage of the related literature, we refer the reader to Gupta *et al.* [17]. The most relevant papers to the current paper are Yechiali and Naor [34] and Gupta *et al.* [17] where the use of generating functions leading to a cubic polynomial is developed and applied. Other papers that involve generating functions and transforms include Neuts [23] and Çinlar [13] for queues with Poisson arrivals and semi-Markovian service times; Çinlar [12] for queues with semi-Markovian arrivals and exponential service times; Arjas [6] where a Wiener–Hopf factorization is obtained for MAP arrivals and general service; and Purdue [26] for $M/M/1$ queues in a Markovian environment. Matrix analytic methods have proven to be effective tools for analyzing time varying queues. The broad class of models include those with Markovian Arrival Process (MAP), phase-type service, and Neuts processes. Seminal papers include those by Neuts [24] and Ramaswami [27]; texts include Neuts [25] and Latouche and Ramaswami [21].

Analysis of the $M/M/1$ with time varying load is closely tied with the transient analysis of the $M/M/1$ queue. Early work on transient analysis either used generating functions to obtain state probabilities in terms of incomplete Bessel functions (see [8]), or expressed the transient probabilities as solutions to integral equations of the Volterra type (see [10] or [14]). A comprehensive analysis of the transient $M/M/1$ queue is given in Abate and Whitt [1], Abate and Whitt [3] and Abate and Whitt [2] using techniques that include transforms, space-time scaling, reflected Brownian motion, and heavy traffic limits. The paper Abate *et al.* [5] provides a decomposition of the $M/M/1$ probability transition functions. Numerical methods for state probabilities in terms of Bessel functions are given by Sharma [28], Conolly and Langaris [15], Sharma and Bunday [29] and Tarabia [31]. Other computational techniques are given in Abate and Whitt [4] using numerical integration, in van de Coevering

[32] using trigonometric integrals, and in Bolot and Shankar [9] using optimal least-squares approximations. The derivation of the Bessel function form of the state probabilities for the $M/M/1$ using sample path techniques is given by Baccelli and Massey [7].

Queues with time-varying arrival rates are closely related to our work. However, we assume a random environment whereas much of the time-varying arrival literature assumes inhomogeneous but non-random environment processes. A wide ranging review of such queues is given in Whitt [33] with additional reviews referenced therein. A specific recent application in the healthcare area of queues with time-varying arrival rates is in Chan *et al.* [11], which models patient inspections before discharge. An example illustrating the breadth of techniques used to analyze time varying systems is Hampshire *et al.* [19], that uses fluid and diffusion limits to analyze transient sojourn times in time-varying processor sharing systems.

3. PROBLEM FORMULATION AND PRELIMINARY RESULTS

In this section, we present our model, and then derive some previously known quantities (see [17,34], for example) that we will use in the balance of our paper.

The $M/M/1(R)$ system consists of a single-server queue with an infinite waiting area in which arriving customers, belonging to a single class, are served in FCFS order. The system operates in an environment that alternates between two phases: LOW and HIGH. Times spent in the LOW and HIGH phases are assumed to be mutually independent exponentially distributed random variables with rate parameters α_L and α_H, respectively. The arrival rates of customers are λ_L and λ_H, in the LOW and HIGH phases, respectively; the service rates are μ_L and μ_H, respectively. The load quantities are defined by $\rho_L = \lambda_L/\mu_L$ and $\rho_H = \lambda_H/\mu_H$. The load in the HIGH phase, ρ_H, is assumed to be greater than or equal to the load in the LOW phase, ρ_L, that is, $\rho_L \leq \rho_H$ (ties are broken arbitrarily), but no particular relationship between λ_L and λ_H or between μ_L and μ_H is assumed. To avoid trivial cases, it will be assumed that neither $\lambda_L = \lambda_H = 0$ nor $\mu_L = \mu_H = 0$ apply, and that both $\alpha_L > 0$ and $\alpha_H > 0$. For consistent labeling of states, it is assumed that if the service rate in one of the phases is zero, then that phase is defined to be the HIGH phase, even if the arrival rate in that phase is zero. With this labeling, it can always be assumed that $\mu_L > 0$.

3.1. Generating Functions

The state probabilities for the $M/M/1(R)$ system can be obtained using the generating function method, which was used by Yechiali and Naor [34] to originally analyze this system. We begin with the generating function of the number in system of a transient $M/M/1$ queue, with arrival rate λ and service rate μ, at a random time. Let $N(t)$ be the number in the system at time t and define $p_n(t) = \Pr(N(t) = n)$. Following Bailey [8] defines the generating function: $\Pi(z,t) = \sum_{n=0}^{\infty} z^n p_n(t)$. Using standard techniques, $\Pi(z,t)$ can be found to satisfy the equation:

$$\frac{\partial}{\partial t}\Pi(z,t) = \Pi(z,t)\left[-(\lambda+\mu) + \lambda z + \frac{\mu}{z}\right] - \mu\left(\frac{1}{z} - 1\right)p_0(t). \tag{1}$$

The Laplace Transform of $\Pi(z,t)$ (denoted by $\widehat{\Pi}(z,s)$) is given by

$$\widehat{\Pi}(z,s) = \frac{\mu(1-z)\hat{p}_0(s) - z\Pi(z,0)}{\lambda z^2 - (s+\lambda+\mu)z + \mu}, \tag{2}$$

where $\hat{p}_0(s)$ is the Laplace transform of $p_0(t)$.

Now consider the number of customers in system, $N(T)$, in a transient $M/M/1$ system at a random time T, where T has an exponential probability distribution with rate α. Define the probability distribution of $N(T)$ by $p_{\alpha,n} \equiv \Pr(N(T)=n)$. It is straightforward to show that the generating function of $N(T)$, denoted by $\widehat{\Pi}_\alpha(z) \equiv \mathrm{E}(z^{N(T)}) = \sum_{n=0}^{\infty} z^n p_{\alpha,n}$, is given by

$$\widehat{\Pi}_\alpha(z) = \alpha\hat{\Pi}(z,\alpha) = \frac{\alpha z\Pi(z,0) - \mu(1-z)\pi_0}{\alpha z - \mu(1-z) + \lambda z(1-z)}, \tag{3}$$

in which $\Pi(z,0)$ is the probability generating function of the number in the system at time 0 and $\pi_0 \equiv P(N(T)=0)$ is the probability that the system is empty at time T.

We now apply (3) to the $M/M/1(R)$ system. Let $N_L(t)$ denote the number in the system t time units after the beginning of a generic LOW phase, given that the phase has not transitioned during these t time units. We define $N_H(t)$, respectively. In particular, $N_L(0)$ and $N_H(0)$ denote the number in the system at the very beginning of the generic LOW and HIGH phases, respectively. Let T_L and T_H denote generic random variables for the duration of the LOW and HIGH phases, respectively. The generating functions for $N_L \equiv N_L(T_L)$ and $N_H \equiv N_H(T_H)$ are given by the generating function, (3), for a transient $M/M/1$ system with the appropriate substitution of primitive parameters for each phase. Also, from the continuity of distribution functions at state transitions, it follows that $N_L(0) \overset{d}{=} N_H(T_H)$ and $N_H(0) \overset{d}{=} N_L(T_L)$. Hence, after defining $\widehat{\Pi}_L \equiv \widehat{\Pi}_{\alpha_L}$ and $\widehat{\Pi}_H \equiv \widehat{\Pi}_{\alpha_H}$, we obtain

$$\widehat{\Pi}_L(z) = \frac{z\alpha_L\widehat{\Pi}_H(z) - \mu_L(1-z)\pi_{0L}}{\alpha_L z - \mu_L(1-z) + \lambda_L z(1-z)}, \tag{4}$$

$$\widehat{\Pi}_H(z) = \frac{z\alpha_H\widehat{\Pi}_L(z) - \mu_H(1-z)\pi_{0H}}{\alpha_H z - \mu_H(1-z) + \lambda_H z(1-z)}, \tag{5}$$

where $\pi_{0L} \equiv P(N_L(T_L)=0)$ and $\pi_{0H} \equiv P(N_H(T_H)=0)$ are the probabilities of the system being empty at the end of a LOW phase and a HIGH phase, respectively. Solving for $\widehat{\Pi}_H(z)$ and $\widehat{\Pi}_L(z)$ gives

$$\widehat{\Pi}_L(z) = \frac{(\mu_L\alpha_H\pi_{0L} + \alpha_L\mu_H\pi_{0H})z - \mu_L\pi_{0L}(1-z)(\mu_H - \lambda_H z)}{D_0(z)}, \tag{6}$$

$$\widehat{\Pi}_H(z) = \frac{(\mu_H\alpha_L\pi_{0H} + \alpha_H\mu_L\pi_{0L})z - \mu_H\pi_{0H}(1-z)(\mu_L - \lambda_L z)}{D_0(z)}, \tag{7}$$

where

$$\begin{aligned} D_0(z) &= (\alpha_H(\mu_L - \lambda_L z) + \alpha_L(\mu_H - \lambda_H z))z - (1-z)(\mu_L - \lambda_L z)(\mu_H - \lambda_H z) \\ &= \lambda_L\lambda_H z^3 - z^2(\alpha_H\lambda_L + \alpha_L\lambda_H + \lambda_L\lambda_H + \mu_L\lambda_H + \mu_H\lambda_L) \\ &\quad + z(\alpha_H\mu_L + \alpha_L\mu_H + \mu_L\lambda_H + \mu_H\lambda_L + \mu_L\mu_H) - \mu_L\mu_H. \end{aligned} \tag{8}$$

Eq. (8) is formally a cubic polynomial, but depending on the parameter choices, it may reduce into a quadratic or even a linear polynomial. This occurs, for example, if $\lambda_H = 0$ or $\lambda_L = 0$, in which case the leading coefficient becomes zero; if $\mu_H = 0$, in which case the constant term becomes zero; or if the variable z factors out from the numerator and the denominator. For the remainder of this section, we consider only the non-degenerate case of $\mu_L, \mu_H, \lambda_L, \lambda_H > 0$ with unequal load in the phases, that is, $\rho_H > \rho_L$ (various degenerate cases are examined in Section 9.1).

The performance of the $M/M/1(R)$ system depends on the rate of fluctuation of the random environment process. We use the scaling parameter, $\kappa \in (0, \infty)$, to succinctly quantify this rate of fluctuation, defined by

$$\kappa \equiv \frac{\alpha_L}{\mu_L} + \frac{\alpha_H}{\mu_H}. \tag{9}$$

To show more explicitly the dependence on load in the LOW and HIGH phases, Eqs. (6) and (7) can be expressed in terms of the following parameters. Define $\tilde{\lambda}_L \equiv \lambda_L/\alpha_L$, $\tilde{\lambda}_H \equiv \lambda_H/\alpha_H$, $\tilde{\mu}_L \equiv \mu_L/\alpha_L$, and $\tilde{\mu}_H \equiv \mu_H/\alpha_H$. Define the quantities $\tau \equiv \tilde{\mu}_H/\tilde{\mu}_L$ and

$$\rho_{av} \equiv \frac{\tilde{\lambda}_L + \tilde{\lambda}_H}{\tilde{\mu}_L + \tilde{\mu}_H}. \tag{10}$$

These quantities are all well-defined under the assumptions stated above. Interpretation of these quantities will be given shortly. Under these definitions, $\kappa = (1+\tau)/\tilde{\mu}_H$. The equations for $\widehat{\Pi}_L(z)$ and $\widehat{\Pi}_H(z)$ in (6) and (7), respectively, can now be expressed as:

$$\widehat{\Pi}_L(z) = \frac{(\pi_{0L} + \tau\pi_{0H})z - (1-z)\tilde{\mu}_H\pi_{0L}(1 - z\rho_H)}{\tilde{\mu}_H D(z)}, \tag{11}$$

$$\widehat{\Pi}_H(z) = \frac{(\pi_{0L} + \tau\pi_{0H})z - (1-z)\tilde{\mu}_H\pi_{0H}(1 - z\rho_L)}{\tilde{\mu}_H D(z)}, \tag{12}$$

where

$$\begin{aligned} D(z) &= \kappa z(1 - \rho_{av} z) - (z-1)(\rho_H z - 1)(1 - \rho_L z) && (13)\\ &= \rho_L\rho_H z^3 - (\kappa\rho_{av} + \rho_L + \rho_H + \rho_L\rho_H)z^2 + (\kappa + 1 + \rho_L + \rho_H)z - 1. && (14)\end{aligned}$$

We can interpret $\tilde{\lambda}_L$ ($\tilde{\lambda}_H$) and $\tilde{\mu}_L$ ($\tilde{\mu}_H$) as the mean number of arrivals and the mean number of potential service completions in a LOW (HIGH) phase, respectively. The quantity τ represents the ratio of the average number of potential customer completions in a HIGH phase to that in a LOW phase. For the $M/M/1(R)$ system, the average number of arrivals and the average number of potential completions in a LOW-HIGH cycle are given by $\lambda_{av} = \lambda_L/\alpha_L + \lambda_H/\alpha_H$ and $\mu_{av} = \mu_L/\alpha_L + \mu_H/\alpha_H$, respectively, and so the quantity ρ_{av} equals the long-term system load, λ_{av}/μ_{av}. Using the strong law of large numbers, the system stability condition is $\rho_{av} < 1$ (the details are not given here). The relationship between ρ_{av}, ρ_L, and ρ_H is given by:

$$(1+\tau)\rho_{av} = \rho_L + \tau\rho_H. \tag{15}$$

Observe that, given ρ_L, ρ_H, and ρ_{av}, we can solve for τ using:

$$\tau = \frac{\rho_{av} - \rho_L}{\rho_H - \rho_{av}}. \tag{16}$$

Thus, τ is a measure of the load imbalance between phases: if $\tau = 1$, then $\rho_{av} - \rho_L = \rho_H - \rho_{av}$; for any given $\alpha_L, \alpha_H > 0$, $\rho_L \geq 0$, and $\rho_H > 0$, we can make ρ_{av} assume any value between ρ_L to ρ_H by varying τ from 0 to ∞ (by varying the service rate ratio μ_H/μ_L).

Using ρ_{av}, we can define an "average $M/M/1$ system"—a conventional $M/M/1$ system with input rate λ_{av}, service rate μ_{av}, and load ρ_{av}. The mean number in system for this $M/M/1$ system is $\rho_{av}/(1-\rho_{av}) \equiv \mathrm{E}N_{av}$.

We define scaling in such a way so as to be able to compare the performance of the $M/M/1(R)$ system against the performance of the average $M/M/1$ system as the rate of the environment process changes. In general, there are six degrees of freedom, corresponding to the primitives λ_L, λ_H, μ_L, μ_H, α_L, and α_H. For a given scenario, we keep λ_L, λ_H, μ_L, and μ_H constant, leaving two degrees of freedom, corresponding to α_L and α_H. Solving for α_L and α_H in (9) and (10) gives

$$\alpha_L = \frac{\mu_L(\rho_{av} - \rho_L)\kappa}{\rho_H - \rho_L} \quad \text{and} \quad \alpha_H = \frac{\mu_H(\rho_H - \rho_{av})\kappa}{\rho_H - \rho_L}.$$

To vary α_L and α_H, we keep ρ_{av} constant and change κ. Keeping ρ_{av} constant implies that τ is also a constant, meaning that for given μ_L and μ_H, the ratio α_L/α_H remains fixed. We collectively refer to these assumptions as our *scaling assumptions*.

$D(z)$ has a root r_1 such that $0 < r_1 \leq 1$. (This can be shown by means of Rouché's Theorem but techniques using analytic geometry can also be used.) The other two roots of $D(z)$ are both greater than 1. Denote these two roots by r_2 and r_3, with $r_2 < r_3$. The intervals within which the roots are located can be readily shown to be given by:

$$\rho_H < 1: \quad r_1 \in (0,1), \quad r_2 \in (1/\rho_H, 1/\rho_{av}), \quad r_3 > 1/\rho_L, \tag{17}$$

$$\rho_H \geq 1: \quad r_1 \in (0,1/\rho_H), \quad r_2 \in (1, 1/\rho_{av}), \quad r_3 > 1/\rho_L. \tag{18}$$

3.2. State Probabilities

Denote the state probabilities for N_L and N_H by $p_{nL} \equiv \Pr(N_L = n)$ and $p_{nH} \equiv \Pr(N_H = n)$, respectively. Since $\widehat{\Pi}_L(z)$ and $\widehat{\Pi}_H(z)$ are generating functions of proper probability functions both are analytic functions on the unit disk, $|z| < 1$. Thus, any poles must be greater than one and must also be roots of $D(z)$, that is, the poles are at r_2 and r_3. Consider $\widehat{\Pi}_L(z)$ first. Taking a partial fraction expansion of its generating function, and noting that the order of the numerator is less than the order of the denominator, $D(z)$, gives

$$\widehat{\Pi}_L(z) = \frac{a_L}{1 - z/r_2} + \frac{b_L}{1 - z/r_3},$$

for some constants a_L and b_L. Since the probability generating function of a discrete random variable with probability mass function $\Pr(X = n) = a\xi^n$ is $a/(1 - z\xi)$, the state probabilities for N_L can be obtained as

$$p_{nL} = a_L\rho_2^n + b_L\rho_3^n, \tag{19}$$

where $\rho_2 \equiv 1/r_2$ and $\rho_3 \equiv 1/r_3$. Note that, $\rho_3 < \rho_2 < 1$. Solving for a_L and b_L gives

$$a_L = \frac{(1-\rho_2)[(1-\rho_3) - \pi_{0L}]}{(\rho_2 - \rho_3)}, \quad b_L = \frac{(1-\rho_3)[(1-\rho_2) - \pi_{0L}]}{(\rho_3 - \rho_2)}. \tag{20}$$

Similarly, the state probabilities for N_H are given by:

$$p_{nH} = a_H\rho_2^n + b_H\rho_3^n, \tag{21}$$

where

$$a_H = \frac{(1-\rho_2)[(1-\rho_3) - \pi_{0H}]}{(\rho_2 - \rho_3)}, \quad b_H = \frac{(1-\rho_3)[(1-\rho_2) - \pi_{0H}]}{(\rho_3 - \rho_2)}. \tag{22}$$

4. NEW REPRESENTATION FOR PERFORMANCE MEASURES

In this section, we begin by deriving novel expressions for the phase-dependent performance quantities $\mathrm{E}N_L$, $\mathrm{E}N_H$, π_{0L}, and π_{0H}, in Theorem 4.1. Since the sequences $\{\rho_2^n\}$ and $\{\rho_3^n\}$ $(n = 0, 1, 2, \ldots)$ are geometric, the expectations of N_L and N_H are given by:

$$\mathrm{E}N_L = \frac{a_L\rho_2}{(1-\rho_2)^2} + \frac{b_L\rho_3}{(1-\rho_3)^2}, \tag{23}$$

$$\mathrm{E}N_H = \frac{a_H\rho_2}{(1-\rho_2)^2} + \frac{b_H\rho_3}{(1-\rho_3)^2}. \tag{24}$$

Substituting the expressions for a_L and b_L given by (20), and for a_H and b_H given by (22), respectively, into these equations gives

$$\mathrm{E}N_L = [(1-\rho_2\rho_3) - \pi_{0L}]\psi, \tag{25}$$

$$\mathrm{E}N_H = [(1-\rho_2\rho_3) - \pi_{0H}]\psi, \tag{26}$$

where

$$\psi \equiv \frac{1}{(1-\rho_2)(1-\rho_3)}. \tag{27}$$

The main theorem for this section (Theorem 4.1) derives expressions for $\mathrm{E}N_L$, $\mathrm{E}N_H$, π_{0L}, and π_{0H} that are cast in terms of expressions for the difference $\mathrm{E}N_H - \mathrm{E}N_L$, ψ and the difference $\pi_{0L} - \pi_{0H}$.

Theorem 4.1: *For a stable $M/M/1(R)$ system:*

$$EN_L = \frac{\rho_{av}}{1-\rho_{av}} - \frac{\tau(1-\rho_H)(EN_H - EN_L)}{(1+\tau)(1-\rho_{av})}, \tag{28}$$

$$EN_H = \frac{\rho_{av}}{1-\rho_{av}} + \frac{(1-\rho_L)(EN_H - EN_L)}{(1+\tau)(1-\rho_{av})}, \tag{29}$$

and

$$\pi_{0L} = 1 - \rho_{av} + \frac{\tau(\pi_{0L} - \pi_{0H})}{1+\tau}, \tag{30}$$

$$\pi_{0H} = 1 - \rho_{av} - \frac{\pi_{0L} - \pi_{0H}}{1+\tau}, \tag{31}$$

where $EN_H - EN_L$, ψ, and $\pi_{0L} - \pi_{0H}$ are given by

$$EN_H - EN_L = \frac{1}{1-\rho_H r_1} - \frac{1}{1-\rho_L r_1} = \frac{r_1\rho_H}{1-\rho_H r_1} - \frac{r_1\rho_L}{1-\rho_L r_1}, \tag{32}$$

$$\psi = \frac{1-r_1}{\kappa r_1(1-\rho_{av})} = \frac{(1-\rho_{av}r_1)}{(1-\rho_{av})(1-\rho_H r_1)(1-\rho_L r_1)}, \tag{33}$$

$$\pi_{0L} - \pi_{0H} = \frac{EN_H - EN_L}{\psi}. \tag{34}$$

Proof: A sketch of the proof is as follows. A representation for $\mathrm{E}N_H - \mathrm{E}N_L$ is obtained by subtracting (25) from (26). Expressions for π_{0L} and π_{0H} are obtained in terms of r_1 using the analytic properties of the generating functions $\widehat{\Pi}_L(z)$ and $\widehat{\Pi}_H(z)$. The resulting expressions are then manipulated to obtain the final results by exploiting the relationships between the roots of the cubic, $D(z)$, and the coefficients of the cubic. The details of the proof are given in Appendix A. ■

Observe that each of the expressions for π_{0L}, π_{0H}, $\mathrm{E}N_L$, and $\mathrm{E}N_H$ in Theorem 4.1 is composed of two terms: a first term that is the corresponding quantity for the *average* $M/M/1$ system, and then plus or minus a second correction term that is the product of a constant (dependent on other system primitives but independent of κ), and either $\pi_{0L} - \pi_{0H}$ or $\mathrm{E}N_H - \mathrm{E}N_L$, depending on the particular expression.

Using (16), we can express the results in Theorem 4.1 in alternative form as follows:

$$\mathrm{E}N_L = \frac{\rho_{av}}{1-\rho_{av}} - \frac{(\rho_{av}-\rho_L)(1-\rho_H)}{(\rho_H-\rho_L)(1-\rho_{av})}(\mathrm{E}N_H - \mathrm{E}N_L), \tag{35}$$

$$\mathrm{E}N_H = \frac{\rho_{av}}{1-\rho_{av}} + \frac{(\rho_H-\rho_{av})(1-\rho_L)}{(\rho_H-\rho_L)(1-\rho_{av})}(\mathrm{E}N_H - \mathrm{E}N_L), \tag{36}$$

and

$$\pi_{0L} = 1-\rho_{av} + \frac{\rho_{av}-\rho_L}{\rho_H-\rho_L}(\pi_{0L}-\pi_{0H}), \tag{37}$$

$$\pi_{0H} = 1-\rho_{av} - \frac{\rho_H-\rho_{av}}{\rho_H-\rho_L}(\pi_{0L}-\pi_{0H}). \tag{38}$$

These equations show the impact of load imbalance; in particular, that as ρ_{av} approaches ρ_L, then $\mathrm{E}N_L$ and π_{0L} approach the corresponding values for the average system, and that as ρ_{av} approaches ρ_H, then $\mathrm{E}N_H$ and π_{0H} approach the corresponding values for the average system.

4.1. Unit Load in the HIGH Phase ($\rho_H = 1$)

A gap in the analysis of Gupta *et al.* [17] is the case $\rho_H = 1$, which is at the boundary between the HIGH state being overloaded and underloaded. For an $M/M/1$ system, this would lead to an unstable system. However, for the $M/M/1(R)$ system, the results in Theorem 4.1 reveal some unexpected behavior. As $\kappa \to 0$, intuition may lead us to expect that $\mathrm{E}N_H$ becomes unbounded since the HIGH phase is unstable and the HIGH phase duration increases as $\kappa \to 0$. Applying Theorem 4.1 confirms this. Setting $\rho_H = 1$ in (36) shows that $\mathrm{E}N_H$ is found to be

$$\mathrm{E}N_H = \frac{\rho_{av}}{1-\rho_{av}} + \frac{(1-\rho_L)(\mathrm{E}N_H - \mathrm{E}N_L)}{(1+\tau)(1-\rho_{av})} = \frac{\rho_{av}}{1-\rho_{av}} + \mathrm{E}N_H - \mathrm{E}N_L.$$

Now, from (32), $\mathrm{E}N_H - \mathrm{E}N_L = 1/(1-r_1) - 1/(1-\rho_L r_1)$, giving

$$\mathrm{E}N_H = \frac{\rho_{av}}{1-\rho_{av}} + \frac{1}{1-r_1} - \frac{1}{1-\rho_L r_1}.$$

Since $r_1 \to 1$ $(\kappa \to 0)$ for $\rho_H = 1$, $\mathrm{E}N_H \to \infty$. Surprisingly, in contrast, upon setting $\rho_H = 1$ in (35), $\mathrm{E}N_L$ is found to always have the same value, $\rho_{av}/(1-\rho_{av})$, regardless of the scaling value κ.

4.2. Time-Stationary Quantities

The mean time stationary time in system, $\mathrm{E}W$, is of particular interest to customers, while the time stationary probability of the system being empty, denoted by π_0, is of interest mainly to system operators. Using Little's law, $\mathrm{E}W = \mathrm{E}N/\lambda_{av}$, where $\mathrm{E}N$ is the mean time stationary number in system. Since the environment process is an alternating renewal process, the renewal-reward technique can be used to obtain the time stationary generating function for the number in system, denoted by $\widehat{\Pi}(z)$, given by (see [17])

$$\widehat{\Pi}(z) = \frac{\frac{\widehat{\Pi}_L(z)}{\alpha_L} + \frac{\widehat{\Pi}_H(z)}{\alpha_H}}{\frac{1}{\alpha_L} + \frac{1}{\alpha_H}}. \tag{39}$$

Using this equation, the time-stationary quantities $\mathrm{E}N$ and π_0 are given by the following corollary.

Corollary 4.1:

$$EN = \frac{\rho_{av}}{1-\rho_{av}} + \frac{(EN_H - EN_L)[\alpha_L(1-\rho_L) - \alpha_H\tau(1-\rho_H)]}{(\alpha_L+\alpha_H)(1+\tau)(1-\rho_{av})}, \tag{40}$$

$$\pi_0 = 1 - \rho_{av} + \frac{\alpha_L\alpha_H r_1(1-\rho_{av})(\rho_H - \rho_L)}{(\alpha_L+\alpha_H)(1+\tau)(1-\rho_{av}r_1)}\left[\frac{\tau}{\alpha_L} - \frac{1}{\alpha_H}\right]. \tag{41}$$

Proof: The proof is in Appendix A. ■

Note, in the $M/M/1(R)$ system, that although stability is determined by the requirement that $\rho_{av} < 1$, π_0 is not a simple function of the system primitives, and $\pi_0 \neq 1 - \rho_{av}$ (unlike in an $M/M/1$ system).

5. SENSITIVITY OF PERFORMANCE AND CONVEXITY/CONCAVITY PROPERTIES

Most of the performance quantities in an $M/M/1(R)$ system are known to be monotonic functions of κ: Under our scaling assumptions, some quantities are always increasing with κ, others are always decreasing with κ and some may either increase or decrease monotonically, depending on the particular values of system primitives. Such analysis was conducted by Gupta *et al.* [17]. The benefit of monotonicity properties is that they enable predictable variations of behavior to occur as fluctuation rate varies. However, system operators may require more detailed information; in particular, they may wish to know the sensitivity of performance measure variation with κ. This section derives such sensitivity results. In so doing, the section also shows that performance curves are, in fact, convex or concave, depending on the particular performance metric being considered.

The following derivative results for the roots of the cubic can be used to establish monotonicity and convexity/concavity properties of variables.

Lemma 5.1: *Let r_1, r_2, and r_3 be the roots of the cubic (14). Taking derivatives of these with respect to κ:*

$$r_1' = \frac{r_1(\rho_{av}r_1 - 1)}{\rho_L\rho_H(r_1-r_3)(r_1-r_2)}, \tag{42}$$

$$r_2' = \frac{r_2(\rho_{av}r_2 - 1)}{\rho_L\rho_H(r_2 - r_3)(r_2 - r_1)}, \tag{43}$$

$$r_3' = \frac{r_3(\rho_{av}r_3 - 1)}{\rho_L\rho_H(r_3 - r_2)(r_3 - r_1)}. \tag{44}$$

Since $\rho_{av}, r_1 < 1$, $r_1 < r_3$, $r_2 < 1/\rho_{av}$, $r_3 > 1/\rho_{av}$, and $r_1 < r_2$, it follows that $r_1' < 0$, $r_2' > 0$, and $r_3' > 0$.

Proof: The proof is in Appendix B. ■

We now establish the convexity of r_1 as a function of κ.

Lemma 5.2: *r_1 is a decreasing convex function of κ.*

Proof: From Lemma 5.1 and (A.7),

$$r_1' = \frac{r_1(\rho_{av}r_1 - 1)}{\rho_L\rho_H(r_1 - r_3)(r_1 - r_2)} = \frac{(r_1 - 1)(\rho_H r_1 - 1)(\rho_L r_1 - 1)}{\rho_L\rho_H\kappa(r_2 - r_1)(r_3 - r_1)}.$$

The denominator is an increasing function of κ since, by Lemma 5.1, $r_2 - r_1$, $r_3 - r_1$ and, trivially, κ are all increasing functions of κ. Since the numerator is a cubic with roots at 1, $1/\rho_H$, and $1/\rho_L$, the numerator is negative but decreases in magnitude to 0 ($\kappa \to \infty$), for $r_1 \le \min(1, 1/\rho_H)$. Hence, r_1' is negative and decreases in magnitude as κ increases, proving that r_1 is a decreasing convex function of κ. ■

To obtain monotonicity *and* convexity and concavity properties of the phase-related mean number in system, we proceed by considering $\mathrm{E}N_H - \mathrm{E}N_L$.

Theorem 5.1: *For all $\kappa > 0$, $EN_H - EN_L$ is a decreasing convex function of κ.*

Proof: Begin by writing

$$\mathrm{E}N_H - \mathrm{E}N_L = \frac{1}{1 - \rho_H r_1} - \frac{1}{1 - \rho_L r_1} = \frac{1/\rho_H}{1/\rho_H - r_1} - \frac{1/\rho_L}{1/\rho_L - r_1}.$$

We now present the following intermediate lemma. Given $0 < a < b$ define

$$g(x) = \frac{a}{a - x} - \frac{b}{b - x}. \tag{45}$$

■

Lemma 5.3: *For $x \in (0, a)$, (i) $g(x) > 0$, (ii) $g'(x) > 0$, and (iii) $g''(x) > 0$.*

Proof: The proof is in Appendix C. ■

To prove $EN_H - EN_L$ is decreasing in κ, differentiate $EN_H - EN_L$ to obtain

$$\frac{d}{d\kappa}(EN_H - EN_L) = \frac{d}{dr_1}(EN_H - EN_L)\frac{dr_1}{d\kappa}. \tag{46}$$

Applying Lemma 5.3, with $g(x) = EN_H - EN_L$, $a = \min(1, 1/\rho_H)$, $b = 1/\rho_L$, and $x = r_1$, gives $d(EN_H - EN_L)/dr_1 > 0$, while Lemma 5.1 gives $dr_1/d\kappa < 0$, proving that $EN_H - EN_L$ is decreasing.

To prove $EN_H - EN_L$ *is convex, differentiate* $EN_H - EN_L$ *a second time to obtain*

$$\frac{d^2}{d\kappa^2}(EN_H - EN_L) = \frac{d^2}{dr_1^2}(EN_H - EN_L)(\frac{dr_1}{d\kappa})^2 + \frac{d}{dr_1}(EN_H - EN_L)\frac{d^2 r_1}{d\kappa^2}.$$

Applying Lemma 5.3 gives $\mathrm{d}^2(EN_H - EN_L)/dr_1^2 > 0$ *and* $\mathrm{d}(EN_H - EN_L)/dr_1 > 0$*, and applying Lemma 5.2 gives* $\mathrm{d}^2 r_1/d\kappa^2 > 0$*, and so* $\mathrm{d}^2(EN_H - EN_L)/d\kappa^2 > 0$*; proving that* $EN_H - EN_L$ *is convex.*

REMARK 5.1: *The specific rate of change of* $EN_H - EN_L$ *with* κ *can be obtained by using (46). For a given value of* κ*, the corresponding value of* r_1 *can be obtained by root solving. Then, (42) gives* $dr_1/d\kappa$ *in terms of* r_1*,* r_2*, and* r_3*;* r_2 *and* r_3 *can be obtained from* r_1 *by solving a quadratic equation (see (A.9) in Appendix A); while* $\mathrm{d}(EN_H - EN_L)/d\kappa$ *can be obtained by applying (C.1) in the proof of Lemma 5.3 in Appendix C. The rate of change for* EN_L *and* EN_H *can be obtained by differentiating (28) and (29), respectively.*

The following corollary is now obtained.

COROLLARY 5.1: *For all* $\kappa > 0$*, (i)* EN_H *is a decreasing convex function of* κ *and (ii) for* $\rho_H < 1$*,* EN_L *is an increasing concave function of* κ*, for* $\rho_H > 1$*,* EN_L *is a decreasing convex function of* κ*, and for* $\rho_H = 1$*,* EN_L *is a constant function.*

PROOF: These results following directly from Theorem 4.1 and Theorem 5.1. (i) Considering $\mathrm{E}N_H$ first, given in (29), it is the sum of a $\rho_{av}/(1-\rho_{av})$ plus a positive constant, $1-\rho_H$, times a decreasing convex function, proving the result. (ii) Examination of $\mathrm{E}N_L$ in (28) shows it to be equal to a $\rho_{av}/(1-\rho_{av})$ minus a constant, $1-\rho_H$, times a decreasing convex function. Depending on ρ_H, the constant is positive, zero, or negative for $\rho_H < 1$, $\rho_H > 1$, or $\rho_H = 1$, respectively, proving the result. ■

An implication of this corollary is that both $\mathrm{E}N_H$ and $\mathrm{E}N_L$ are more sensitive to changes in κ for smaller values of κ than for larger values of κ.

6. STOCHASTIC MONOTONICITY OF N_H

Mean values provide a first-order basis for comparing the performance of two systems. A more detailed comparison is possible using stochastic ordering concepts. For example, systems can be compared on the basis of tail probabilities of the number in system.

Recall that a random variable X is less than or equal to a random variable Y in the sense of stochastic ordering, denoted $X \leq_{\mathrm{st}} Y$, if $\Pr(X > x) \leq \Pr(Y > x)$ for all x.

Given a random variable L that is a function of κ we say L is stochastically increasing (decreasing) in κ if for $\kappa_2 > \kappa_1$, we have $L(\kappa_1) \leq_{\mathrm{st}} (\geq_{\mathrm{st}}) L(\kappa_2)$. In previous work, Gupta *et al.* [17] showed that $N_L \not\leq_{\mathrm{st}} N_A$ and $N_L \leq_{\mathrm{st}} N_H$. This section examines the open problem in Gupta *et al.* [17] of whether N_H is stochastically monotone with κ. First, we provide some insight into this problem. From (19) and (21), it follows that

$$\Pr(N_L > n) = \sum_{i=n+1}^{\infty} \Pr(N_L = i) = \frac{a_L \rho_2^{n+1}}{1-\rho_2} + \frac{b_L \rho_3^{n+1}}{1-\rho_3}, \tag{47}$$

$$\Pr(N_H > n) = \sum_{i=n+1}^{\infty} \Pr(N_H = i) = \frac{a_H \rho_2^{n+1}}{1-\rho_2} + \frac{b_H \rho_3^{n+1}}{1-\rho_3}. \tag{48}$$

Focusing on N_H, we see from (48) that $\Pr(N_H > n)$ is dominated by the first term as $n \to \infty$, that is, $\Pr(N_H > n) \sim a_H \rho_2^{n+1}/(1-\rho_2)$. Suppose for the moment that a_H is a positive constant, independent of κ. Then, for fixed n,

$$\frac{d}{d\kappa}\frac{a_H \rho_2^{n+1}}{1-\rho_2} = \frac{d}{d\rho_2}\frac{a_H \rho_2^{n+1}}{1-\rho_2}\frac{d\rho_2}{d\kappa} = \frac{a_H((1-\rho_2)(n+1)+\rho_2)\rho_2^n}{(1-\rho_2)^2}\frac{d\rho_2}{d\kappa}.$$

Since $d\rho_2/d\kappa < 0$, this expression would be negative and so, in the asymptotic regime, $\Pr(N_H > x)$ would be decreasing with κ. However, for stochastic monotonicity, we need to consider all $n \geq 0$, as well as take into consideration that a_H varies with κ and include the term $b_H \rho_3^{n+1}/(1-\rho_3)$, which complicates the determination of conditions for N_H to be stochastically decreasing.

Despite these difficulties, it will shown, that provided the loads $\rho_L, \rho_H, \rho_{av}$ satisfy certain conditions, then N_H does stochastically decrease with κ locally, by which we mean that there exists an open interval containing κ such that for any $\kappa_2 > \kappa_1$ in that interval $N_H(\kappa_2) \leq_{\text{st}} N_H(\kappa_1)$.

THEOREM 6.1: *N_H is stochastically decreasing locally in κ if*

$$(1-\rho_{av} r_2)(r_3 - r_1) > (1-\rho_L r_2)(r_2 - r_1). \tag{49}$$

PROOF: A sketch of the proof is as follows. An expression for $\Pr(N_H > n)$ is first derived from (48) by substituting in the expressions for a_H and b_H given in (22). The resulting expression contains the term $1 - \rho_3 - \pi_{0H}$. The proof then shows that under the theorem conditions, this term is decreasing, and that this is sufficient for N_H to be locally stochastically decreasing with κ. The details of the proof are given in Appendix D. ■

Taking the limit $\kappa \to 0$ in (49), for the case of $\rho_H < 1$, the theorem gives the requirement

$$\left(1 - \frac{\rho_{av}}{\rho_H}\right)\left(\frac{1}{\rho_L} - 1\right) > \left(1 - \frac{\rho_L}{\rho_H}\right)\left(\frac{1}{\rho_H} - 1\right).$$

If this is satisfied, then it is also true in a neighborhood of 0 since the functions involved are continuous. For example, if $\tau = 1$, $\rho_H = 0.9$, and $\rho_L = 0.1$, the condition is satisfied but if $\tau = 1$, $\rho_H = 0.3$, and $\rho_L = 0.2$, then it is not satisfied. For the case $\rho_H > 1$, the corresponding condition for $\kappa \to 0$ becomes

$$(1-\rho_{av})\left(\frac{1}{\rho_L} - \frac{1}{\rho_H}\right) > (1-\rho_L)\left(1 - \frac{1}{\rho_H}\right),$$

which likewise holds for some parameter values but not others.

For large values of κ observe that $1-\rho_{av}r_2 \to 0$ but $r_3 \to \infty$ ($\kappa \to \infty$). To circumvent this problem, multiply (49) by $\rho_{av}r_3 - 1$ and rearrange to give the requirement:

$$(1-\rho_{av}r_2)(\rho_{av}r_3-1) > (1-\rho_L r_2)(r_2-r_1)\frac{\rho_{av}r_3-1}{r_3-r_1}.$$

Using (A.2) to replace the left-hand side of this gives

$$\frac{(1-\rho_{av})(\rho_H-\rho_{av})(\rho_{av}-\rho_L)}{\rho_L\rho_H(1-\rho_{av}r_1)} > (1-\rho_L r_2)(r_2-r_1)\frac{\rho_{av}r_3-1}{r_3-r_1}.$$

Taking limits as $\kappa \to \infty$ gives (since $r_3 \to \infty$)

$$\frac{(1-\rho_{av})(\rho_H-\rho_{av})(\rho_{av}-\rho_L)}{\rho_L\rho_H} > \left(1-\frac{\rho_L}{\rho_{av}}\right).$$

If this is satisfied, then it is also true for κ large enough since the functions involved are continuous. For example, if $\tau = 1$, $\rho_H = 0.9$, and $\rho_L = 0.1$, the condition is satisfied but if $\tau = 1$, $\rho_H = 0.9$, and $\rho_L = 0.2$, then it is not satisfied.

7. PRACTICAL APPLICATION OF RESULTS

In this section, we consider two practical applications of the results of our paper.

7.1. Server Rate Allocation

In this section, we use the expression for $\mathrm{E}N_H - \mathrm{E}N_L$ given in (32) to optimize the performance of a service center. The approach we follow parallels that given in Section 1.1 of Stidham [30] for optimizing service rate in an $M/M/1$ system. Before proceeding to the analysis of the cost minimization problem, we first review the concept of slack, that will required in that analysis.

7.1.1. Slack The following is based on Gupta *et al.* [17], except we use our novel representations of $\mathrm{E}N$ and π_0 to obtain the results. Define the slack values in the LOW and HIGH states by $\mu_L - \lambda_L$ and $\lambda_H - \mu_H$, respectively. By rearranging (40), we can express

$$\mathrm{E}N = \frac{\rho_{av}}{1-\rho_{av}} + D\left[\frac{r_1\rho_H}{1-r_1\rho_H} - \frac{r_1\rho_L}{1-r_1\rho_L}\right],$$

where

$$D = (\alpha_L/\mu_L)\frac{[(\mu_L-\lambda_L)-(\mu_H-\lambda_H)]}{(\alpha_L+\alpha_H)(1+\tau)(1-\rho_{av})}.$$

This shows that if $\mu_L - \lambda_L < \lambda_H - \mu_H$, then $\lim_{\kappa\to 0}\mathrm{E}N < \lim_{\kappa\to\infty}\mathrm{E}N$; and if $\mu_L - \lambda_L > \lambda_H - \mu_H$, then $\lim_{\kappa\to 0}\mathrm{E}N > \lim_{\kappa\to\infty}\mathrm{E}N$.

In the case of π_0, rearranging (41) gives

$$\pi_0 = 1-\rho_{av} + \frac{\alpha_L r_1(1-\rho_{av})(\rho_H-\rho_L)}{\mu_L(\alpha_L+\alpha_H)(1+\tau)(1-\rho_{av}r_1)}[\mu_H-\mu_L].$$

In this case, if $\mu_H > \mu_L$, then π_0 is a decreasing function of κ (since $r_1/(1-\rho_{av}r_1)$ is a decreasing function), that is, the server utilization increases as κ increases; and the converse applies if $\mu_H < \mu_L$.

The application of the representations in Theorem 4.1 gives the new result that customers and the system operator see different behavior as κ increases. From (34), $\pi_{0L} - \pi_{0H} = (\mathrm{E}N_H - \mathrm{E}N_L)/\psi$. Since $1/\psi = (1-\rho_2)(1-\rho_3)$, it can be shown that $d(1/\psi)/d\kappa > 0$. Hence, $\pi_{0L} - \pi_{0H}$ equals the product of the decreasing function $\mathrm{E}N_H - \mathrm{E}N_L$ and the increasing function $1/\psi$, and so it follows that $\pi_{0L} - \pi_{0H}$ approaches 0 more slowly than $\mathrm{E}N_H - \mathrm{E}N_L$ approaches 0. This means that as the fluctuation rate increases, when the mean number in system is considered, the system approaches the performance of the average system faster than when server utilization is considered.

7.1.2. Cost Minimization Assume a cost of c units per unit rate of server capacity, and that a customer incurs a holding cost of h units per unit time in the system. The total mean cost is given by

$$\begin{aligned} C(\mu_L,\mu_H) &= c\mu_{av} + h\mathrm{E}N = c\mu_{av} + h\frac{\rho_{av}}{1-\rho_{av}} + hD(\mathrm{E}N_H - \mathrm{E}N_L) \\ &= c\mu_{av} + h\frac{\lambda_{av}}{\mu_{av}-\lambda_{av}} + hD\frac{r_1\lambda_H}{\mu_H - r_1\lambda_H} - hD\frac{r_1\lambda_L}{\mu_L - r_1\lambda_L} \\ &= c\frac{\frac{\mu_L}{\alpha_L}+\frac{\mu_H}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}} + h\frac{\lambda_{av}}{\frac{\frac{\mu_L}{\alpha_L}+\frac{\mu_H}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}} - \lambda_{av}} + hD\frac{r_1\lambda_H}{\mu_H - r_1\lambda_H} - hD\frac{r_1\lambda_L}{\mu_L - r_1\lambda_L}. \end{aligned} \quad \mathbf{(50)}$$

The goal is to minimize cost. Differentiating partially with respect to μ_L and μ_H, equating to zero to find stationary points and simplifying shows (see Appendix E) the only solution is $D = 0$, that is, $\mu_L - \lambda_L = \mu_H - \lambda_H$, and

$$c - h\frac{\lambda_{av}}{\left(\frac{\frac{\mu_L}{\alpha_L}+\frac{\mu_H}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}} - \lambda_{av}\right)^2} = 0,$$

yielding

$$\frac{\frac{\mu_L}{\alpha_L}+\frac{\mu_H}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}} = \lambda_{av} + \sqrt{\frac{h\lambda_{av}}{c}}. \quad \mathbf{(51)}$$

We can interpret this as follows. First ensure that service rates are constrained such that the slack in both phases is equal—from (40), this gives the mean number in system as $\rho_{av}/(1-\rho_{av})$; then, just as in the $M/M/1$ case, the average service rate is then chosen to minimize the average delay—this gives (51).

7.2. Sampling Methods

Sampling of system state can be used for various purposes. According to the PASTA property (Poisson Arrivals See Time Averages), the mean number in system as determined by sampling the system number at epochs of an independent Poisson process equals the time average number in system. Applying Little's law then gives the mean time in system as $\mathrm{E}W = \mathrm{E}N/\lambda_{av}$. If the input process to the $M/M/1(R)$ system is Poisson, that is, $\lambda_L = \lambda_H$, the actual arrivals can be used to do the sampling. If not, then sampling at customer arrival and departure epochs (which is common) may yield biased estimates. In this section, we derive expressions for the amount of bias that can occur, how bias is affected by fluctuation rate and determine how to correct for the bias.

In the following, it is assumed that the LOW phase is sampled by an independent Poisson process of rate β_L and that the HIGH phase is sampled by an independent Poisson process of rate β_H. This form of sampling method will be referred to as state-dependent sampling.

Different forms of sampling can be obtained depending on how β_L and β_H are chosen. Regardless of the specific choices, the sampling produces a discrete-time series of samples that, in general, is denoted by $\{N_i^{(S)}\}$, where i is an integer. From this can be derived a (discrete)-time-stationary sample random variable denoted by $N^{(S)}$ whose probability generating function is $\widehat{\Pi}^{(S)}(z) \equiv \mathrm{E}(z^{N^{(S)}})$. The following theorem relates $\widehat{\Pi}^{(S)}(z)$ to $\widehat{\Pi}_L(z)$ and $\widehat{\Pi}_H(z)$.

Theorem 7.1: *The (discrete)-time stationary probability generating function of the state-dependent sampling sequence is given by*

$$\widehat{\Pi}^{(S)}(z) = \frac{\frac{\beta_L}{\alpha_L}\widehat{\Pi}_L(z) + \frac{\beta_H}{\alpha_H}\widehat{\Pi}_H(z)}{\frac{\beta_L}{\alpha_L} + \frac{\beta_H}{\alpha_H}}. \tag{52}$$

Proof: The proof of the theorem, given in Appendix F, is based on the application of the renewal-reward theorem for random sequences. ■

7.2.1. Sampling at Arrival Epochs Setting $\beta_L = \lambda_L$ and $\beta_H = \lambda_H$ in Theorem 7.1 gives sampling at potential arrival epochs. However, we assume the potential arrival epochs used for sampling coincide with the actual arrivals, in which case sampling is performed at arrival epochs.

Lemma 7.1: *(i) The (discrete)-time stationary probability generating function of the state-dependent sampling sequence at arrival epochs is given by*

$$\widehat{\Pi}_A(z) = \frac{\rho_L\widehat{\Pi}_L(z) + \tau\rho_H\widehat{\Pi}_H(z)}{\rho_L + \tau\rho_H}. \tag{53}$$

(ii) The probability that the system is empty as seen by an arriving customer is given by

$$\pi_{0A} = \frac{\rho_L\pi_{0L} + \tau\rho_H\pi_{0H}}{\rho_L + \tau\rho_H} = 1 - \rho_{av} - \frac{\tau r_1(1-\rho_{av})(\rho_H - \rho_L)^2}{(1+\tau)(\rho_L + \tau\rho_H)(1-\rho_{av}r_1)}.$$

(iii) The mean number in the system at arrival sampling instances is given by

$$EN_A = \frac{\rho_L EN_L + \tau\rho_H EN_H}{\rho_L + \tau\rho_H} = \frac{\rho_{av}}{1-\rho_{av}} + \frac{\tau(\rho_H - \rho_L)(EN_H - EN_L)}{(1+\tau)^2\rho_{av}(1-\rho_{av})}.$$

Proof: (i) Setting $\beta_L = \lambda_L$ and $\beta_H = \lambda_H$ in (52) gives

$$\widehat{\Pi}_A(z) = \frac{\frac{\lambda_L}{\alpha_L}\widehat{\Pi}_L(z) + \frac{\lambda_H}{\alpha_H}\widehat{\Pi}_H(z)}{\frac{\lambda_L}{\alpha_L} + \frac{\lambda_H}{\alpha_H}} = \frac{\frac{\lambda_L}{\mu_L}\widehat{\Pi}_L(z) + \frac{\lambda_H}{\mu_H}\frac{\mu_H\alpha_L}{\mu_L\alpha_H}\widehat{\Pi}_H(z)}{\frac{\lambda_L}{\mu_L} + \frac{\lambda_H}{\mu_H}\frac{\mu_H\alpha_L}{\mu_L\alpha_H}}$$
$$= \frac{\rho_L\widehat{\Pi}_L(z) + \tau\rho_H\widehat{\Pi}_H(z)}{\rho_L + \tau\rho_H}.$$

(ii) Setting $z = 0$ in (53) gives

$$\pi_{0A} = \widehat{\Pi}_A(0) = \frac{\rho_L\pi_{0L} + \tau\rho_H\pi_{0H}}{\rho_L + \tau\rho_H}.$$

Using the expressions for π_{0L} and π_{0H} in (A.14) and (A.15), in Lemma A.3, respectively, gives

$$\pi_{0A} = 1 - \rho_{av} + \frac{\rho_L\tau r_1(1-\rho_{av})(\rho_H - \rho_L)}{(1+\tau)(\rho_L + \tau\rho_H)(1 - \rho_{av}r_1)}$$
$$- \frac{\tau\rho_H r_1(1-\rho_{av})(\rho_H - \rho_L)}{(1+\tau)(\rho_L + \tau\rho_H)(1 - \rho_{av}r_1)},$$

which simplifies to the desired result.

(iii) Taking the derivative of $\widehat{\Pi}_A(z)$ in (53) and setting $z = 1$ gives

$$\mathrm{E}N_A = \widehat{\Pi}'_A(1) = \frac{\rho_L\mathrm{E}N_L + \tau\rho_H\mathrm{E}N_H}{\rho_L + \tau\rho_H}.$$

The result now follows straightforwardly, using the expressions for $\mathrm{E}N_L$ and $\mathrm{E}N_H$ in (28) and (29), in Theorem 4.1, respectively. ■

This lemma shows that $\mathrm{E}N_A$ is a decreasing function of κ since $\mathrm{E}N_H - \mathrm{E}N_L$ is a decreasing function of κ.

To provide a unified analysis of the effect of sampling, we introduce the general quantity $\gamma > 0$ that enables us to write

$$\mathrm{E}N_L = -\frac{\gamma(\mathrm{E}N_H - \mathrm{E}N_L)}{1+\gamma} + \frac{\mathrm{E}N_H + \gamma\mathrm{E}N_L}{1+\gamma}, \tag{54}$$

$$\mathrm{E}N_H = \frac{\mathrm{E}N_H - \mathrm{E}N_L}{1+\gamma} + \frac{\mathrm{E}N_L + \gamma\mathrm{E}N_H}{1+\gamma}, \tag{55}$$

$$\mathrm{E}N = \frac{\frac{\mathrm{E}N_H - \mathrm{E}N_L}{\alpha_H(1+\gamma)} - \frac{\gamma(\mathrm{E}N_H - \mathrm{E}N_L)}{\alpha_L(1+\gamma)}}{\frac{1}{\alpha_H} + \frac{1}{\alpha_L}} + \mathrm{E}N_\gamma = \frac{\mathrm{E}N_H - \mathrm{E}N_L}{1+\gamma}\left[\frac{\frac{1}{\alpha_H} - \frac{\gamma}{\alpha_L}}{\frac{1}{\alpha_H} + \frac{1}{\alpha_L}}\right] + \mathrm{E}N_\gamma, \tag{56}$$

where $\mathrm{E}N_\gamma = (\mathrm{E}N_H + \gamma \mathrm{E}N_L)/(1+\gamma)$. For sampling at arrival times, set $\gamma = \tau\rho_H/\rho_L = (\lambda_H/\alpha_H)/(\lambda_L/\alpha_L)$, for which $\mathrm{E}N_\gamma = \mathrm{E}N_A$, to give

$$\begin{aligned}
\mathrm{E}N_L &= -\frac{(\tau\rho_H/\rho_L)(\mathrm{E}N_H - \mathrm{E}N_L)}{1+\tau\rho_H/\rho_L} + \mathrm{E}N_A,\\
\mathrm{E}N_H &= \frac{\mathrm{E}N_H - \mathrm{E}N_L}{1+\tau\rho_H/\rho_L} + \mathrm{E}N_A,\\
\mathrm{E}N &= \frac{\mathrm{E}N_H - \mathrm{E}N_L}{1+\tau\rho_H/\rho_L}\left[\frac{\frac{1}{\alpha_H} - \frac{\lambda_H\alpha_L}{\lambda_L\alpha_H\alpha_L}}{\frac{1}{\alpha_H}+\frac{1}{\alpha_L}}\right] + \mathrm{E}N_A\\
&= \frac{\mathrm{E}N_H - \mathrm{E}N_L}{(1+\tau\rho_H/\rho_L)\alpha_H}\left[\frac{1-\frac{\lambda_H}{\lambda_L}}{\frac{1}{\alpha_H}+\frac{1}{\alpha_L}}\right] + \mathrm{E}N_A.
\end{aligned}$$

Hence, if $\lambda_H > \lambda_L$, $\mathrm{E}N_A$ overestimates $\mathrm{E}N$, and vice versa. The final equation also gives the correction factor needed to retrieve $\mathrm{E}N$ from $\mathrm{E}N_A$.

7.2.2. Sampling at Potential Service Completions

Setting $\beta_L = \mu_L$ and $\beta_H = \mu_H$ in Theorem 7.1 gives sampling at potential service completion epochs. This form of sampling can be seen as a form of exit polling. (One can use actual departures as sampling points; and when the system becomes empty sample at epochs of two internally generated Poisson processes of rate μ_L and μ_L, for LOW and HIGH phases, respectively.)

LEMMA 7.2: *(i) The (discrete)-time stationary probability generating function of the state-dependent sampling sequence at potential service completion epochs is given by*

$$\widehat{\Pi}_S(z) = \frac{\widehat{\Pi}_L(z) + \tau\widehat{\Pi}_H(z)}{1+\tau}.$$

(ii) The probability that the system is empty at a potential service completion is given by

$$\pi_{0S} = \frac{\pi_{0L} + \tau\pi_{0H}}{1+\tau} = 1 - \rho_{av} = \frac{(1-\rho_L) + \tau(1-\rho_H)}{1+\tau}.$$

(iii) The mean number in the system at potential service sampling instances can be obtained by differentiating the generating functions in Lemma 7.1.

$$EN_S = \frac{EN_L + \tau EN_H}{1+\tau} = \frac{\rho_{av}}{1-\rho_{av}} + \frac{\tau(\rho_H - \rho_L)(EN_H - EN_L)}{(1+\tau)^2(1-\rho_{av})}.$$

PROOF: (i) Setting $\beta_L = \mu_L$ and $\beta_H = \mu_H$ in (52) gives the result along similar lines to the proof of Lemma 7.1(i).

(ii) The derivation of π_{0S} follows immediately from Lemma A.2.

(iii) The proof of the result for $\mathrm{E}N_S$ is similar to the proof of Lemma 7.1(iii). ■

This lemma shows that $\mathrm{E}N_S$ is a decreasing function of κ since $\mathrm{E}N_H - \mathrm{E}N_L$ is a decreasing function of κ.

Setting $\gamma = \tau = (\mu_H/\alpha_H)/(\mu_L/\alpha_L)$ in (54)—(56), respectively, gives

$$\begin{aligned}
\mathrm{E}N_L &= -\frac{\tau(\mathrm{E}N_H - \mathrm{E}N_L)}{1+\tau} + \mathrm{E}N_S,\\
\mathrm{E}N_H &= \frac{\mathrm{E}N_H - \mathrm{E}N_L}{1+\tau} + \mathrm{E}N_S,\\
\mathrm{E}N &= \frac{\mathrm{E}N_H - \mathrm{E}N_L}{1+\tau}\left[\frac{\frac{1}{\alpha_H} - \frac{\mu_H\alpha_L}{\mu_L\alpha_H\alpha_L}}{\frac{1}{\alpha_H}+\frac{1}{\alpha_L}}\right] + \mathrm{E}N_S\\
&= \frac{\mathrm{E}N_H - \mathrm{E}N_L}{(1+\tau)\alpha_H}\left[\frac{1-\frac{\mu_H}{\mu_L}}{\frac{1}{\alpha_H}+\frac{1}{\alpha_L}}\right] + \mathrm{E}N_S.
\end{aligned}$$

Hence, if $\mu_H > \mu_L$, $\mathrm{E}N_S$ overestimates $\mathrm{E}N$, and vice versa.

7.2.3. Discussion

1. Constant rate sampling, that is, $\beta_L = \beta_H = \mathit{constant}$ (e.g., 1), gives expressions equal to time-stationary quantities.
2. For sampling at potential service completions, it is observed that irrespective of the value of the scaling parameter κ, the probability of the system being empty at a potential service completion is constant. This is a particular instance of the following result for any stable queueing system without loss: Rate of arrivals = Probability of busy at a potential service completion × rate of potential service completions.[2]
3. Comparing $\mathrm{E}N_A$ with $\mathrm{E}N_s$, we see that $\mathrm{E}N_A > \mathrm{E}N_S$. That is, Arrivals see on average more in the system than do potential completions. To see why this is the case consider a coupled system where actual departures occur at potential service completion instances. This is probabilistically the same as the original system. Let the mean number in the system at departures be denoted by $\mathrm{E}N_D$. Now, departures see the same distribution as arrivals, which is true for general single-server system where both arrivals and departures occur singly, so that $\mathrm{E}N_A = \mathrm{E}N_D$. In the coupled system, every departure corresponds to a potential service completion but only service completions when the number in system $N > 0$ correspond to a departure. Potential service completions when $N = 0$ are included in $\mathrm{E}N_S$ but not $\mathrm{E}N_D$. Thus, $\mathrm{E}N_S < \mathrm{E}N_D = \mathrm{E}N_A$.

7.2.4. Examples The examples in this section illustrate how the time stationary and sampled mean number in system vary with κ. Within each example given, the service rates μ_L and μ_H are kept fixed (although μ_L and μ_H are different for each example) while the phase switching rates α_L, α_H are varied, keeping the ratio α_L/α_H constant. To achieve this, define $\delta \equiv \mu_H/\mu_L$ and $\sigma \equiv \alpha_H/\alpha_L$. Here, δ represents the ratio of the service rate in the HIGH phase to that in the LOW phase; and σ represents the ratio of the average time spent in the LOW phase to the average time spent in the HIGH phase. This gives $\tau = \delta/\sigma$ and

$$\kappa = \frac{\alpha_L}{\mu_L}\left(1 + \frac{\sigma}{\delta}\right).$$

In this case, the parameter κ is proportional to α_L.

[2] We thank the anonymous Editorial Board Member for pointing out this simple relation.

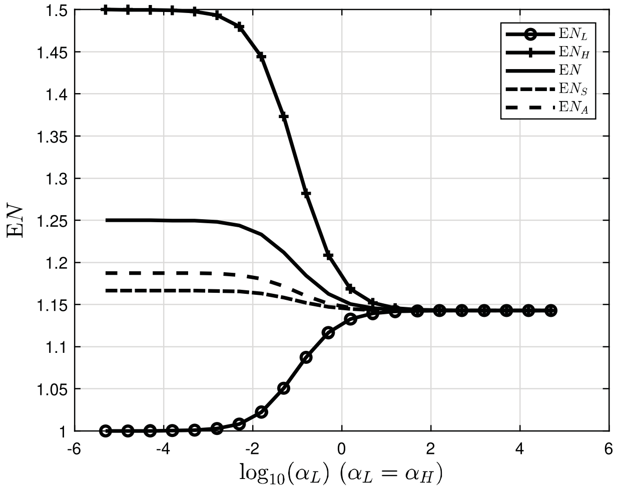

FIGURE 3. Plot of the mean number in the system, EN, where EN is decreasing.

For both examples, $\rho_L = 0.5$ and $\rho_H = 0.6$. Phase rates are set equal: $\alpha_L = \alpha_H$, so that $\delta = 1$, and α_L is varied. We plot EN, EN_L, EN_H, EN_S, and EN_A against $\log_{10}(\alpha_L)$. We use α_L on the horizontal axis rather than κ because the examples use different values of μ_L and μ_H. The first system (Figure 3) uses the service rates: $\mu_L = 2$ and $\mu_H = 1$. This gives the parameter values: $\lambda_L = 1$, $\lambda_H = 0.6$, $\rho_{av} = 0.53333$, and $\tau = 0.5$. The slack condition satisfied is $\mu_L - \lambda_L = 1 > \mu_H - \lambda_H = 0.4$ and EN is decreasing with α_L (and also κ). Observe in both this example and the next that EN_H, EN_S, and EN_A are always decreasing with α_L, whereas EN_L is increasing with α_L. The second system (Figure 4) uses the service rates: $\mu_L = 1$ and $\mu_H = 2$. This gives the parameter values: $\lambda_L = 0.5$, $\lambda_H = 1.2$, $\rho_{av} = 0.566667$, and $\tau = 2$. The slack condition satisfied is $\mu_L - \lambda_L = 0.5 < \mu_H - \lambda_H = 0.8$ and EN is increasing with α_L. The examples demonstrate how it is possible that two systems can have the same load conditions and phase processes; yet divergent behavior for EN is possible by varying other parameters— in this case, the service rates μ_L and μ_H—because of different values of slack in the two states. The examples also show that in some cases (e.g., Figure 3), sampling at arrival times and potential completion times can both *underestimate* EN, whereas in other cases (e.g., Figure 4), sampling at arrival times and potential completion times can both *overestimate* EN.

8. SIMPLE AND EXACT CLOSED-FORM COMPUTATION OF ROOTS

The paper by Gupta *et al.* [17] develops various quadratic approximations to the cubic to enable easier computation of performance measures using simple tools such as spreadsheets. However, this quadratic approximation has a lower bound on error, and the error is largest in the intermediate region. In this paper, we develop a computational approach that can also be used with simple tools that gives complete accuracy, down to the level of numerical precision of the tool. In particular, our approach for computing the roots of the cubic $D(z)$ is obtained by developing a new representation of the cubic polynomial that shows that the three roots of $D(z)$ are positioned at equally spaced locations across a single period of a suitably scaled and vertically shifted cosine wave. All that is required of the computation tool, in addition to basic arithmetic functions, are accurate implementations of the cosine

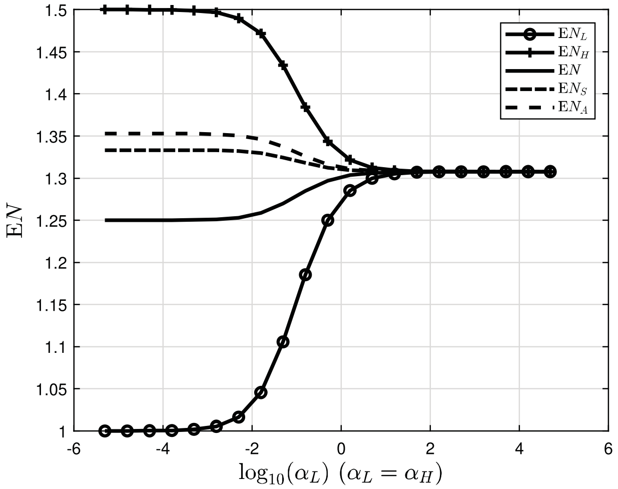

FIGURE 4. Plot of the mean number in the system, EN, where EN is increasing.

function and the inverse cosine function. As an intermediate result, we also show that the roots are the eigenvalues of a suitable constructed 3×3 matrix.

The new method for obtaining the roots of the cubic polynomial (13) is based on the method described in Day and Romero [16] for finding the roots of a polynomial using Chebyshev polynomials whereby the roots of (13) are obtained as the eigenvalues of a matrix. In general, we seek to obtain the roots of an nth order polynomial $p(z)$ in an interval $[a, b]$. Given the inner product $\langle \cdot, \cdot \rangle_r$ defined by

$$\langle f, g \rangle_r = \int_a^b f(z) g(z) r(z)\, dz,$$

where f and g are functions defined on $[a, b]$ and r is a suitable weight function, $\{\phi_i(z)\}$ is a set of orthogonal polynomials with respect to this inner product if $\langle \phi_i, \phi_j \rangle_r = 0$ for $i \neq j$.

Assume $p(z)$ can be expressed as the following weighted sum with coefficients γ_i:

$$p(z) = \sum_{i=0}^{n} \gamma_i \phi_i(z).$$

Suppose that $\{\phi_i(z)\}$ satisfy the following recurrence relation for some constants $h_{i,j}$

$$z\phi_{n-1}(z) = \sum_{i=0}^{n} h_{i,n-1} \phi_i(z)$$

and define the $n \times n$ matrix $\mathbf{H} = [h_{i,j}]_{0 \leq i,j \leq n-1}$. Define the following vectors: $\mathbf{f}_n(z) = (\phi_0(z), \ldots, \phi_{n-1}(z))^T$ and $\mathbf{c} = (\gamma_0, \ldots, \gamma_{n-1})^T$ so that $p(z)$ can be expressed as

$$p(z) = \mathbf{f}_n(z)^T \mathbf{c} + \gamma_n \phi_n.$$

Theorem 2.3 of Day and Romero [16] shows that the roots of $p(z)$ are the eigenvalues of the nonstandard companion matrix

$$\mathbf{B}_n = \mathbf{H}_n - \frac{h_{n,n-1}}{\gamma_n} \mathbf{c} \mathbf{e}_{n-1}^T,$$

where $\mathbf{e}_{n-1}$ is the $n \times 1$ column vector $\mathbf{e}_{n-1} = (0, \ldots, 0, 1)^T$.

This technique is applied to the polynomial (13) to obtain the root r_1 in the interval $[-1,1]$ by using Chebyshev polynomials $T_n(z)$ as the set of orthogonal polynomials, that is, $\phi_i(z)=T_i(z)$. This gives

$$p(z)=\sum_{i=0}^{n}\gamma_i T_i(z). \tag{57}$$

The nth Chebyshev polynomial is defined by

$$T_n(z)=\cos(n\cos^{-1}(z)),$$

where the first four Chebyshev polynomials are

$$\begin{aligned} &T_0(z)=1, \quad T_1(z)=z,\\ &T_2(z)=2z^2-1, \quad T_3(z)=4z^3-3z. \end{aligned}$$

Chebyshev polynomials satisfy the following recursion ($k\geq 1$)

$$zT_k(z)=\frac{1}{2}T_{k-1}(z)+\frac{1}{2}T_{k+1}(z)$$

from which is obtained that $h_{0,1}=1/2$, $h_{1,0}=1$, $h_{i,i+1}=h_{i+1,i}=1/2$ $(i=1,\ldots,n-1)$. Thus, for $n=3$

$$\mathbf{H}_3=\begin{pmatrix}0 & 1/2 & 0\\ 1 & 0 & 1/2\\ 0 & 1/2 & 0\end{pmatrix}$$

and

$$\mathbf{B}_3=\begin{pmatrix}0 & 1/2 & 0\\ 1 & 0 & 1/2\\ 0 & 1/2 & 0\end{pmatrix}-\frac{h_{3,2}}{\gamma_3}\begin{pmatrix}0 & 0 & \gamma_0\\ 0 & 0 & \gamma_1\\ 0 & 0 & \gamma_2\end{pmatrix}=\begin{pmatrix}0 & 1/2 & -\dfrac{\gamma_0}{2\gamma_3}\\ 1 & 0 & 1/2-\dfrac{\gamma_1}{2\gamma_3}\\ 0 & 1/2 & -\dfrac{\gamma_2}{2\gamma_3}\end{pmatrix}.$$

Matching the coefficients of z in (13) and with those in (57), the coefficients γ_i can be determined to be

$$\begin{aligned} \gamma_3 &= \frac{\rho_L\rho_H}{4},\\ \gamma_2 &= -\frac{\kappa\rho_{av}+\rho_L+\rho_H+\rho_L\rho_H}{2},\\ \gamma_1 &= \kappa+1+\rho_L+\rho_H+\frac{3\rho_L\rho_H}{4},\\ \gamma_0 &= -\frac{\kappa\rho_{av}+\rho_L+\rho_H+\rho_L\rho_H+2}{2}. \end{aligned}$$

This yields the roots of (13) as the eigenvalues of

$$\mathbf{B}_3=\begin{pmatrix}0 & 1/2 & \dfrac{\kappa\rho_{av}+\rho_L+\rho_H+\rho_L\rho_H+2}{\rho_L\rho_H}\\ 1 & 0 & -\dfrac{\kappa+1+\rho_L+\rho_H+\frac{\rho_L\rho_H}{2}}{\frac{\rho_L\rho_H}{2}}\\ 0 & 1/2 & \dfrac{\kappa\rho_{av}+\rho_L+\rho_H+\rho_L\rho_H}{\rho_L\rho_H}\end{pmatrix}.$$

An exact solution for the eigenvalues λ_i of a 3×3 matrix A can be found using the following approach.[3] Define

$$q = \mathrm{tr}(A)/3 \quad \text{and} \quad p = \mathrm{tr}((A - qI)^2/6)^{1/2}$$

and define the matrix B by $B = (A - qI)/p$. The eigenvalues of A are then given by $\lambda_k = p\beta_k + q$, $k = 0, 1, 2$, where

$$\beta_k = 2\cos\left(\frac{1}{3}\cos^{-1}(\det(B)/2) + \frac{2k\pi}{3}\right).$$

To apply the above, let $A = \mathbf{B}_3$. The roots occur equally spaced along the x-axis of a vertically shifted and scaled cosine wave. The cosine function and the inverse cosine function can be evaluated by efficient numerical algorithms. Alternatively, these functions can be precomputed and loaded into a lookup table, whose accuracy is determined by the amount of memory available. Either way, the accuracy of computation is determined by the amount of resources available.

9. EXAMPLES OF OTHER TYPES OF SYSTEMS

9.1. Degenerate $M/M/1(R)$ Systems

If one or more of λ_L, λ_H, μ_L, μ_H are zero or $\rho_L = \rho_H$, then (6) and (7) simplify, leading to what may be called degenerate $M/M/1(R)$ systems, which in some cases represent systems that have useful practical application.

If $\rho_L = 0$, then arrivals occur only in the HIGH phase. This is denoted as a system with ON–OFF arrival rates. In this case, (11) and (12) become

$$\widehat{\Pi}_L(z) = \frac{(\pi_{0L} + \tau\pi_{0H})z - \tilde{\mu}_L(1-z)\pi_{0L}(\tau - z\tilde{\rho}_{HL})}{(1+\tau)(1-\rho_{av}z)z - \tilde{\mu}_L(z-1)(\tilde{\rho}_{HL}z - \tau)},$$

$$\widehat{\Pi}_H(z) = \frac{(\pi_{0L} + \tau\pi_{0H})z - \tilde{\mu}_L(1-z)\tau\pi_{0H}}{(1+\tau)(1-\rho_{av}z)z - \tilde{\mu}_L(z-1)(\tilde{\rho}_{HL}z - \tau)},$$

where $\tilde{\rho}_{HL} \equiv \tilde{\lambda}_H/\tilde{\mu}_L$. The denominator $D(z) = -((1+\tau)\rho_{av} + \tilde{\mu}_H\tilde{\rho}_{HL})z^2 + ((1+\tau) + \tilde{\mu}_L\tau + \tilde{\mu}_L\tilde{\rho}_{HL})z - \tau\tilde{\mu}_L$ has two roots only one of which is greater than one: $r_1 < 1$, $r_2 > 1$.

Using a similar approach, other degenerate models are:

- Interrupted service model: In this model, $\mu_H = 0$ and no service is provided in the HIGH phase, although arrivals may still occur in that phase.
- Inactive state model: In this model, $\lambda_H = 0$ and $\mu_H = 0$, and so there are no arrivals or service conducted during the HIGH phase.
- Alternating phase model: In this model, $\lambda_L = 0$, $\mu_H = 0$, and so arrivals occur only in the HIGH phase and service occurs only in the LOW phase. In this case, $\rho_L = 0$ and $\tau = 0$.

In each of these systems, there is only one root greater than one, that we denote by r_2. The techniques to deal with one root will be similar to those used in the non-degenerate $M/M/1(R)$ system. The starting equations can be obtained setting $\rho_3 = 0$ into the non-degenerate system equations (25)–(27). We omit the details.

[3] This method of solution was given in https://en.wikipedia.org/wiki/Eigenvalue_algorithm; however, we have been unable to locate a journal reference for this.

One final degenerate case occurs when $\rho_H = \rho_L$ but $\mu_L \neq \mu_H$. Setting $\rho_L = \rho_H$ in (32) gives $\mathrm{E}N_H - \mathrm{E}N_H = 0$. From (28), this implies $\mathrm{E}N_L = \mathrm{E}N_H = 1/(1-\rho_{av})$. Hence, the mean number in system remains the same regardless of the value of the fluctuation parameter, κ, and equals the mean number in the average $M/M/1$ system with average load.

9.2. More Sophisticated Time-Varying Systems

In this section, we examine more sophisticated systems that may be analyzed using the same approach as used for the $M/M/1(R)$ system and discuss what additional complexities may be involved and how these may be addressed. The main assumption is that the system operates in a random environment that is in one of two phases that we notionally call "LOW" and "HIGH" and the time spent in each phase is given by two independent exponentially distributed random variables. It is assumed that at the end of each phase, the number in system can be represented using two random variables N_L and N_H defined on the space of natural numbers whose probability generating functions $\widehat{\Pi}_L(z)$ and $\widehat{\Pi}_H(z)$. For notational continuity, $\widehat{\Pi}_L(z)$ will be called the equation for the LOW phase and $\widehat{\Pi}_H(z)$ the equation for the HIGH phase; although in the general case, LOW and HIGH phases may have different interpretations.

The quantity of interest is the number in system and it is assumed that at a phase change, the number in system does not change. As in the $M/M/1(R)$ system, it will be assumed that the system can be described by two *coupled equations* having a common denominator polynomial, $V(z)$, and numerator polynomials, $U_L(z)$ and $U_H(z)$, respectively:

$$\widehat{\Pi}_L(z) = \frac{U_L(z)}{V(z)}, \quad \widehat{\Pi}_H(z) = \frac{U_H(z)}{V(z)}.$$

The approach used to obtain these equations is to generalize the structure of the expressions in the $M/M/1(R)$ case and assume that we can express $\widehat{\Pi}_L(z)$ and $\widehat{\Pi}_H(z)$ as

$$\widehat{\Pi}_L(z) = \frac{a(z) - pz\widehat{\Pi}_H(z)}{b(z)}, \quad \widehat{\Pi}_H(z) = \frac{c(z) - qz\widehat{\Pi}_L(z)}{d(z)},$$

where p, q are constants and $a(z), b(z), c(z), d(z)$ are polynomials. Solving for $\widehat{\Pi}_L(z)$and $\widehat{\Pi}_H$ gives

$$\widehat{\Pi}_L(z) = \frac{a(z)d(z) - pzc(z)}{b(z)d(z) - pqz^2}, \tag{58}$$

$$\widehat{\Pi}_H(z) = \frac{c(z)b(z) - qza(z)}{b(z)d(z) - pqz^2}. \tag{59}$$

This gives $V(z) = b(z)d(z) - pqz^2$, and $U_L(z)$ and $U_H(z)$ are similarly matched.

9.3. Catastrophe Queue in a Random Environment

The first example is of an $M/M/1$ queue with catastrophes operating in a random environment. First, we review results for a *homogeneous* single-server $M/M/1$ system that is cleared of all customers at the renewal epochs of a Poisson process, called the catastrophe process. The same notation is used as in Section 3 and the rate of the catastrophe process is assumed to be ω. The differential equation for $\Pi(z,t)$ for this homogeneous catastrophe system is given by Kumar and Arivudainambi [20]:

$$\frac{\partial}{\partial t}\Pi(z,t) = \Pi(z,t)\left[-(\lambda+\mu+\omega)+\lambda z+\frac{\mu}{z}\right] - \mu\left(\frac{1}{z}-1\right)p_0(t)+\omega.$$

The Laplace Transform of $\Pi(z,t)$ is found to be

$$\widehat{\Pi}(z,s) = \frac{\mu(1-z)\hat{p}_0(s) - z\Pi(z,0) - \omega z/s}{\lambda z^2 - (s+\lambda+\mu+\omega)z+\mu},$$

where $\hat{p}_0(s)$ is the Laplace transform of $p_0(t)$.

Suppose now a catastrophe queue operates in an alternating two-phase random environment similar to the $M/M/1(R)$ system, governed by α_L and α_H. Following Section 3, the generating functions for the number in the system at the end of the LOW and HIGH phases are given by

$$\widehat{\Pi}_L(z) = \frac{\mu_L(1-z)\pi_{0L} - \alpha_L z\Pi_L(z,0) - \omega_L z}{\lambda_L z^2 - (\alpha_L+\lambda_L+\mu_L+\omega_L)z+\mu_L},$$
$$\widehat{\Pi}_H(z) = \frac{\mu_H(1-z)\pi_{0H} - \alpha_H z\Pi_H(z,0) - \omega_H z}{\lambda_H z^2 - (\alpha_H+\lambda_H+\mu_H+\omega_H)z+\mu_H},$$

in which $\Pi_L(z,0) = \widehat{\Pi}_H(z)$ and $\Pi_H(z,0) = \widehat{\Pi}_L(z)$ because of continuity of generating functions at phase transitions. Using (58) gives for the LOW phase (the HIGH phase is similar),

$$\widehat{\Pi}_L(z) = \frac{(\mu_L(1-z)\pi_{0L} - \omega_L z)(\lambda_H z^2 - (\alpha_H+\lambda_H+\mu_H+\omega_H)z+\mu_H) - \alpha_L z(\mu_H(1-z)\pi_{0H} - \omega_H z)}{(\lambda_L z^2 - (\alpha_L+\lambda_L+\mu_L+\omega_L)z+\mu_L)(\lambda_H z^2 - (\alpha_H+\lambda_H+\mu_H+\omega_H)z+\mu_H) - \alpha_L\alpha_H z^2}. \tag{60}$$

In general, the denominator is a fourth-order polynomial; unlike the $M/M/1(R)$ system, $z-1$ will not be a factor in the denominator. We sketch a numerical method for dealing with this situation. The quartic in the denominator in (60) has four roots r_0, r_1, r_2, r_3 that can be found using numerical methods. Numerical tests we conducted showed that two roots are in the interval $(0,1)$, that we denote r_0 and r_1 with $r_0 < r_1$, and two roots are greater than 1 that we denote r_2 and r_3 with $r_2 < r_3$. We do not go into detailed examination of roots here. For the generating function to be analytic, r_0 and r_1 must be roots of the numerator in (60). Setting z equal to r_0 and r_1 gives two linear equations in π_{0L} and π_{0H} that can be solved for those values. Thus, we have a system with two roots $1 < r_2 < r_3$ with the values of π_{0L} and π_{0H} known. As an example, for a system with the parameters $\lambda_L = 0.5$, $\lambda_H = 0.9$, $\mu_L = \mu_H = 1$, $\alpha_L = \alpha_H = 0.1$, $\omega_L = \omega_H = 0.01$, the quartic in the denominator had the roots $r_0 = 0.6936$, $r_1 = 0.9793$, $r_2 = 1.3581$, and $r_3 = 2.4313$, giving $\pi_{0L} = 0.4090$ and $\pi_{0H} = 0.2390$.

9.4. $M/M/c$ in a Random Environment

The second example is an $M/M/c$ queue operating in a random environment, that we denote as $M/M/c(R)$, where the service and arrival rates vary according to a two-state continuous-time Markov chain in the same way as in the $M/M/1(R)$ system. In the non-random environment case, with fixed arrival rate λ and fixed service rate μ, the state equations for the $M/M/c$ queue, where c is the number of servers, are given by (e.g., see [22])

$$\begin{aligned}
p_0'(t) &= -\lambda p_0(t) + \mu p_1(t),\\
p_n'(t) &= -(\lambda + n\mu)p_n(t) + \lambda p_{n-1}(t) + (n+1)\mu p_{n+1}(t), \quad 1 \le n \le c-1,\\
p_n'(t) &= -(\lambda + c\mu)p_n(t) + \lambda p_{n-1}(t) + c\mu p_{n+1}(t), \quad n \ge c.
\end{aligned}$$

The generating function for the number in system satisfies

$$\frac{\partial}{\partial t}\Pi(z,t) = \left[-(\lambda + c\mu) + \frac{c\mu}{z} + \lambda z\right]\Pi(z,t) - c\mu\left(\frac{1}{z} - 1\right)\frac{q_c(t,z)}{c},$$

where

$$q_c(t,z) = \sum_{i=0}^{c-1}(c-i)p_i(t)z^i.$$

This equation is the same as (1) except that μ is replaced by $c\mu$ and $p_0(t)$ is replaced by $q_c(t,z)/c$. The Laplace Transform of $\Pi(z,t)$ is

$$\widehat{\Pi}(z,s) = \frac{\mu(1-z)\hat{q}_c(s,z) - z\Pi(z,0)}{\lambda z^2 - (s + \lambda + c\mu)z + c\mu}$$

where $\hat{q}_c(s,z)$ is the Laplace transform of $q_c(t,z)$.

For a random time T, where T has an exponential probability distribution with rate α, we have, as in Section 3, that the generating function of $N(T)$ is given by

$$\widehat{\Pi}_\alpha(z) = \mathrm{E}(z^{N(T)}) = \alpha\hat{\Pi}(z,\alpha).$$

Defining $\nu_i \equiv P(N(T) = i)$ $(i = 0, \ldots, c-1)$ gives

$$\hat{q}_c(\alpha,z) = \frac{\sum_{i=0}^{c-1}(c-i)\nu_i z^i}{\alpha}.$$

Combining the above yields

$$\widehat{\Pi}_\alpha(z) = \frac{\mu(1-z)(\sum_{i=0}^{c-1}(c-i)\nu_i) - \alpha z\Pi(z,0)}{\lambda z^2 - (\alpha + \lambda + c\mu)z + c\mu}.$$

Using (58) and (59) equations for $\widehat{\Pi}_L(z)$ and $\widehat{\Pi}_H(z)$ can be obtained, respectively. The numerators of $\widehat{\Pi}_L(z)$ and $\widehat{\Pi}_H(z)$ will have $1 - z$ as a factor, that cancels out, and the common denominator will be a cubic, similar to the $M/M/1(R)$ system, except that the service rates will be $c\mu_L$ and $c\mu_H$ in the LOW and HIGH phases, respectively. The main complications in this case will be that the order of the numerator polynomials will be larger than the order of the denominator polynomial and that we need to obtain ν_i $(i = 0, \ldots, c-1)$ for both LOW and HIGH phases. The techniques in Haghihi and Mishev [18], for example, can be used to find ν_i $(i = 0, \ldots, c-1)$ for both LOW and HIGH phases. The remaining part of the analysis will be along the same lines as the $M/M/1(R)$ system.

10. CONCLUSION

This paper examined the problem of determining the effect of varying the fluctuation rate of the underlying random environment process associated with the random two-phase $M/M/1(R)$ system on key performance quantities. It extended the results of Gupta *et al.* [17] to include boundary cases ($\rho_H = 1$) and extended monotonicity results to include convexity/concavity results for phase related and time stationary number in system. The paper achieved this by developing new representations for the mean number in system and the probability of an empty system in terms of the quantities $\mathrm{E}N_H - \mathrm{E}N_L$ and ψ that enabled simpler unified expressions to be obtained. The expressions obtained clearly showed how the system performance could be expressed as the sum of the performance of the average system plus or minus a correction term. Other results derived include: a new stochastic ordering result for $\mathrm{E}N_H$, and a new representation for the cubic, allowing new insights into root locations to be made and provided an alternative method of computing roots. The paper demonstrated the usefulness of the results obtained through two applications to practical systems: service rate optimization and sampling. The paper examined how the approach used for $M/M/1(R)$ systems could be applied, and needed to be modified, to other types of systems operating in a two-phase random environment: a range of degenerate $M/M/1(R)$ systems, an $M/M/1$ catastrophe queue, and an $M/M/c$ system. Future work would include extending these results to systems operating in random environments with more than two phases.

Acknowledgments

This research is supported by NSF-CMMI-1938909, NSF-XPS-1629444, NSF-CSR-1763701, and a Google Faculty Research Award 2020. The authors express their sincere appreciation to the anonymous Reviewer and the anonymous Editorial Board Member for their detailed reviews and insightful suggestions in creating an improved version of the paper.

References

1. Abate, J. & Whitt, W. (1987). Transient behavior of the $M/M/1$ queue: starting at the origin. *Queueing Systems* 2: 41–65.
2. Abate, J. & Whitt, W. (1988). Simple spectral representations for the $M/M/1$. *Queueing Systems* 3: 321–346.
3. Abate, J. & Whitt, W. (1988). Transient behavior of the $M/M/1$ queue via Laplace transforms. *Advances in Applied Probability* 20(1): 145–178.
4. Abate, J. & Whitt, W. (1989). Calculating time-dependent performance measures for $M/M/1$ queue. *IEEE Transactions on Communications* 37(10): 1102–1104.
5. Abate, J., Kijama, M., & Whitt, W. (1989). Decompositions of the $M/M/1$ transition function. *Queueing Systems* 9(3): 323–336.
6. Arjas, E. (1972). On the use of a fundamental identity in the theory of semi-Markov queues. *Advances in Applied Probability* 4(2): 271–284.
7. Baccelli, F. & Massey, W.A. (1989). A sample path analysis of the $M/M/1$ queue. *Journal of Applied Probability* 26(2): 418–422.
8. Bailey, N.T.J. (1954). A continuous time treatment of a simple queue using generating functions. *Journal of the Royal Statistical Society: Series B (Methodological)* 16(2): 288–291.
9. Bolot, J.-C. & Shankar, A.U. (1995). Optimal least-squares approximations to the transient behavior of the stable $M/M/1$ queue. *IEEE Transactions on Communications* 43(2/3/4): 1293–1298.
10. Champernowne, D.G. (1956). An elementary method of solution of the queueing problem with a single server and constant parameters. *Journal of the Royal Statistical Society: Series B (Methodological)* 18(1): 125–128.
11. Chan, C.W., Dong, J., & Green, L.V. (2016). Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Operations Research* 65(2): 469–495.
12. Çinlar, E. (1967). Queues with semi-Markov arrivals. *Journal of Applied Probability* 4(2): 365–379.
13. Çinlar, E. (1967). Time dependence of queues with semi-Markov services. *Journal of Applied Probability* 4(2): 356–364.

14. Clarke, A.B. (1956). A waiting line process of Markov type. *Annals of Mathematical Statistics* 27(2): 452–459.
15. Conolly, B.W. & Langaris, C. (1993). On a new formula for the transient state probabilities for $M/M/1$ queues and computational implications. *Journal of Applied Probability* 30(1): 237–246.
16. Day, D. & Romero, L. (2006). Roots of polynomials expressed in terms of orthogonal polynomials. *SIAM Journal on Numerical Analysis* 43(5): 1969–1987.
17. Gupta, V., Harchol-Balter, M., Scheller-Wolf, A., & Yechiali, U. (2006). Fundamental characteristics of queues with fluctuating load. *ACM SIGMETRICS Performance Evaluation Review* 34(1): 203–215.
18. Haghighi, A.M. & Mishev, D.P. (2013). *Difference and differential equations with applications in queueing theory*. Hoboken, New Jersey: John Wiley & Sons.
19. Hampshire, R.C., Harchol-Balter, M., & Massey, W.A. (2006). Fluid and diffusion limits for transient sojourn times of processor sharing queues with time varying rates. *Queueing Systems* 53(1–2): 19–30.
20. Kumar, B.K. & Arivudainambi, D. (2000). Transient solution of an $M/M/1$ queue with catastrophes. *Computers & Mathematics with Applications* 40: 1233–1240.
21. Latouche, G. & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling*. Philadelphia, Pennsylvania: American Statistical Association and the Society for Industrial and Applied Mathematics.
22. Medhi, J. (2003). *Stochastic models in queueing theory*, 2nd ed. Amsterdam: Academic Press.
23. Neuts, M. (1966). The single server queue with Poisson input and semi-Markov service times. *Journal of Applied Probability* 3(1): 202–230.
24. Neuts, M. (1978). The $M/M/1$ queue with randomly varying arrival and service rates. *Management Science* 15(4): 139–168.
25. Neuts, M.F. (1981). *Matrix-geometric solutions in stochastic models: an algorithmic approach*. New York: Dover Publications.
26. Purdue, P. (1974). The $M/M/1$ queue in a Markovian environment. *Operations Research* 22(3): 562–569.
27. Ramaswami, V. (1980). The N/G/1 queue and its detailed analysis. *Advances in Applied Probability* 12(1): 222–261.
28. Sharma, O.P. (1990). *Markovian queues*. Chichester: Ellis Horwood.
29. Sharma, O.P. & Bunday, B.D. (1997). A simple formula for the transient state probabilities of an $M/M/1/\infty$ queue. *Optimization* 40(1): 79–84.
30. Stidham Jr, S. (2009). *Optimal design of queueing systems*. Boca Raton, Florida: Chapman and Hall/CRC.
31. Tarabia, A.M.K. (2002). A new formula for the transient behaviour of a non-empty $M/M/1/\infty$ queue. *Applied Mathematics and Computation* 132: 1–10.
32. van de Coevering, M.C.T. (1995). Computing transient performance measures for the $M/M/1$ queue. *OR Spektrum* 17: 19–22.
33. Whitt, W. (2016). Queues with time-varying arrival rates. A bibliography. Working paper.
34. Yechiali, U. & Naor, P. (1971). Queueing problems with heterogeneous arrivals and service. *Operations Research* 19(3): 722–734.

APPENDIX A. PROOF OF THEOREM 4.1 AND COROLLARY 4.1

The proof of Theorem 4.1 is obtained by means of the sequence of results in this appendix.

Relationships Between Roots and Coefficients of $D(z)$

In general, no simple formula exists for the roots of the cubic $D(z)$. However, the coefficients of $D(z)$ express relationships between the roots that can be used to simplify equations and to explain how performance quantities depend on system parameters. To derive such relationships, first normalize $D(z)$ by dividing through by the leading coefficient. The resulting polynomial $\widetilde{D}(z)$ can be expressed as

$$\widetilde{D}(z) = \frac{D(z)}{\rho_L\rho_H} = z^3 - z^2\frac{(\kappa\rho_{av} + \rho_L + \rho_H + \rho_L\rho_H)}{\rho_L\rho_H} + z\frac{(\kappa + 1 + \rho_L + \rho_H)}{\rho_L\rho_H} - \frac{1}{\rho_L\rho_H} \tag{A.1}$$

$$\stackrel{\text{def}}{=} z^3 - C_2z^2 + C_1z - C_0. \tag{A.2}$$

Noting that $\widetilde{D}(z) = (z - r_1)(z - r_2)(z - r_3)$ gives

$$C_2 = r_1 + r_2 + r_3 = \frac{\kappa\rho_{av} + \rho_L + \rho_H + \rho_L\rho_H}{\rho_L\rho_H}, \qquad \textbf{(A.3)}$$

$$C_1 = r_1 r_2 + r_2 r_3 + r_1 r_3 = \frac{\kappa + 1 + \rho_L + \rho_H}{\rho_L\rho_H}, \qquad \textbf{(A.4)}$$

$$C_0 = r_1 r_2 r_3 = \frac{1}{\rho_L\rho_H}. \qquad \textbf{(A.5)}$$

The above together with (13) gives the identity

$$(z - r_1)(z - r_2)(z - r_3) = \frac{1}{\rho_L\rho_H}[\kappa z(1 - \rho_{av} z) - (z - 1)(\rho_H z - 1)(1 - \rho_L z)] \qquad \textbf{(A.6)}$$

that can used to obtain a number of useful relationships between roots. For example, setting $z = 1$ gives

$$(1 - r_1)(1 - r_2)(1 - r_3) = \frac{\kappa(1 - \rho_{av})}{\rho_L\rho_H}. \qquad \textbf{(A.7)}$$

Alternatively, setting $z = r_i$ $(i = 1, 2, 3)$ gives

$$0 = \kappa r_i(1 - \rho_{av} r_i) - (r_i - 1)(\rho_H r_i - 1)(1 - \rho_L r_i).$$

In particular, for $z = r_1$, the following is obtained

$$\frac{1 - r_1}{\kappa r_1} = \frac{(1 - \rho_{av} r_1)}{(1 - \rho_H r_1)(1 - \rho_L r_1)}. \qquad \textbf{(A.8)}$$

Note that given one root, for example r_1, then the other two roots can be obtained as the solution of a quadratic equation involving that root. For example, we can use (A.5) to eliminate r_3 from (A.3) to give

$$r_2 + \frac{1}{r_2(r_1\rho_L\rho_H)} = \frac{\kappa\rho_{av} + \rho_L + \rho_H + \rho_L\rho_H}{\rho_L\rho_H} - r_1, \qquad \textbf{(A.9)}$$

that leads a quadratic equation for r_2.

π_{0L} and π_{0H}

The probabilities of the system being empty at the end of a LOW and a HIGH phase, π_{0L} and π_{0H}, respectively, are given by the following theorem.

LEMMA A.1:

$$\pi_{0L} = \frac{r_1\kappa(1 - \rho_{av})}{(1 - r_1)(1 - \rho_H r_1)} = \frac{(1 - \rho_{av})(1 - \rho_L r_1)}{1 - \rho_{av} r_1}, \qquad \textbf{(A.10)}$$

$$\pi_{0H} = \frac{r_1\kappa(1 - \rho_{av})}{(1 - r_1)(1 - \rho_L r_1)} = \frac{(1 - \rho_{av})(1 - \rho_H r_1)}{1 - \rho_{av} r_1}. \qquad \textbf{(A.11)}$$

PROOF: First, consider π_{0L}. Since $\widehat{\Pi}_L(z)$ is an analytic function, r_1 must be a root of the numerator of $\widehat{\Pi}_L(z)$ (see [22]). Hence, from (11):

$$(\pi_{0L} + \tau\pi_{0H})r_1 - (1 - r_1)\tilde{\mu}_H\pi_{0L}(1 - \rho_H r_1) = 0. \tag{A.12}$$

A second equation involving π_{0L} and π_{0H} is obtained by putting $z = 1$ in (11) to give

$$\widehat{\Pi}_L(1) = 1 = \frac{\pi_{0L} + \tau\pi_{0H}}{\tilde{\mu}_H\kappa(1 - \rho_{av})}. \tag{A.13}$$

Substituting this into (A.12) gives

$$r_1\tilde{\mu}_H\kappa(1 - \rho_{av}) - (1 - r_1)\tilde{\mu}_H\pi_{0L}(1 - \rho_H r_1) = 0$$

from which is obtained the first equality in (A.10):

$$\pi_{0L} = \frac{r_1\kappa(1 - \rho_{av})}{(1 - r_1)(1 - \rho_H r_1)}.$$

Applying (A.8) then gives the second equality in (A.10). Eq. (A.11) is derived similarly. ■

Observe that applying (A.7) to (A.10) and (A.11) followed by a small amount of algebra, we can express $\pi_{0L} = (1 - \rho_2)(1 - \rho_3)/(1 - \rho_H r_1)$ and $\pi_{0H} = (1 - \rho_2)(1 - \rho_3)/(1 - \rho_L r_1)$, respectively.

The probabilities π_{0L} and π_{0H} satisfy the identities given in the following lemma.

LEMMA A.2:

$$\frac{\pi_{0L} + \tau\pi_{0H}}{1 + \tau} = 1 - \rho_{av} = \frac{(1 - \rho_L) + \tau(1 - \rho_H)}{1 + \tau}.$$

PROOF: From (A.13), $\pi_{0L} + \tau\pi_{0H} = \tilde{\mu}_H\kappa(1 - \rho_{av})$. Since (using (9)) $\kappa\tilde{\mu}_H = 1 + \tau$, the first equality follows. The second equality follows by applying (15). ■

By applying Lemma A.2, π_{0L} and π_{0H} can be expressed as deviations from the probability of finding the system empty in the averaged system, $1 - \rho_{av}$.

LEMMA A.3:

$$\pi_{0L} = 1 - \rho_{av} + \frac{\tau r_1(1 - \rho_{av})(\rho_H - \rho_L)}{(1 + \tau)(1 - \rho_{av} r_1)}, \tag{A.14}$$

$$\pi_{0H} = 1 - \rho_{av} - \frac{r_1(1 - \rho_{av})(\rho_H - \rho_L)}{(1 + \tau)(1 - \rho_{av} r_1)}. \tag{A.15}$$

Proof: Start by expressing π_{0L} as

$$\pi_{0L} = \frac{\pi_{0L} + \tau\pi_{0H} + \tau(\pi_{0L} - \pi_{0H})}{1+\tau}.$$

Taking the difference between (A.10) and (A.11) gives

$$\pi_{0L} - \pi_{0H} = \frac{r_1(1-\rho_{av})(\rho_H - \rho_L)}{1-\rho_{av}r_1}. \tag{A.16}$$

Thus, with the aid of Lemma A.2, π_{0L} can be written as

$$\pi_{0L} = \frac{(1+\tau)(1-\rho_{av}) + \frac{\tau r_1(1-\rho_{av})(\rho_H-\rho_L)}{1-\rho_{av}r_1}}{1+\tau} = 1 - \rho_{av} + \frac{\tau r_1(1-\rho_{av})(\rho_H - \rho_L)}{(1+\tau)(1-\rho_{av}r_1)}.$$

Along similar lines,

$$\begin{aligned}\pi_{0H} &= \frac{\pi_{0L} + \tau\pi_{0H} - (\pi_{0L} - \pi_{0H})}{1+\tau} = \frac{(1+\tau)(1-\rho_{av}) - \frac{r_1(1-\rho_{av})(\rho_H-\rho_L)}{1-\rho_{av}r_1}}{1+\tau} \\ &= 1 - \rho_{av} - \frac{r_1(1-\rho_{av})(\rho_H-\rho_L)}{(1+\tau)(1-\rho_{av}r_1)} = 1 - \rho_{av} - \frac{\pi_{0L} - \pi_{0H}}{1+\tau}.\end{aligned}$$

■

Bounds for π_{0L} and π_{0H}, enabling crude performance estimates for the system to be obtained, are

$$\text{(i)}\quad 1-\rho_{av} \le \pi_{0L} \le 1-\rho_L \quad\text{and}\quad \text{(ii)}\quad \max(1-\rho_H, 0) \le \pi_{0H} \le 1-\rho_{av}.$$

The lower bound in (i) is obtained from (A.10), and by noting that $\rho_{av} \ge \rho_L$:

$$\pi_{0L} = \frac{(1-\rho_{av})(1-\rho_L r_1)}{1-\rho_{av}r_1} \ge (1-\rho_{av}). \tag{A.17}$$

The upper bound in (i) follows from rewriting the right-hand side of the equality in the preceding expression as

$$\pi_{0L} = 1 - \rho_L - \frac{(\rho_{av}-\rho_L)(1-r_1)}{1-r_1\rho_{av}} \le 1 - \rho_L,$$

after observing that the final term before the inequality is non-negative. The bounds in (ii) are obtained similarly.

Remark A.1: *After defining θ by*

$$\theta \equiv \frac{1-\rho_{av}}{1-\rho_{av}r_1}, \tag{A.18}$$

Theorem A.1 gives $\pi_{0L} = \theta(1-\rho_L r_1)$ and $\pi_{0H} = \theta(1-\rho_H r_1)$, from which the following is also obtained:

$$\pi_{0L} = \frac{1-r_1\rho_L}{1-r_1\rho_H}\pi_{0H}.$$

Since $\rho_L \le \rho_H$, it follows immediately that $\pi_{0L} \ge \pi_{0H}$. The above results have been obtained by Gupta et al. [17] but are included here for completeness and because they are used later.

Expectations

Expressions for the explanatory quantities $\mathrm{E}N_H - \mathrm{E}N_L$ and ψ are given by the following lemma.

Lemma A.4:

$$EN_H - EN_L = (\pi_{0L} - \pi_{0H})\psi, \tag{A.19}$$

$$= \frac{1}{1-\rho_H r_1} - \frac{1}{1-\rho_L r_1}. \tag{A.20}$$

where

$$\psi = \frac{1-r_1}{\kappa r_1(1-\rho_{av})} = \frac{(1-\rho_{av} r_1)}{(1-\rho_{av})(1-\rho_H r_1)(1-\rho_L r_1)}. \tag{A.21}$$

Proof: Subtracting (25) from (26) gives (A.19). Applying (A.7) to ψ in (27) gives

$$\psi = \frac{1-r_1}{r_1}\frac{r_1 r_2 r_3}{(1-r_1)(1-r_2)(1-r_3)} = \frac{1-r_1}{\kappa r_1(1-\rho_{av})}.$$

Applying (A.8) then gives

$$\psi = \frac{(1-\rho_{av} r_1)}{(1-\rho_{av})(1-\rho_H r_1)(1-\rho_L r_1)}.$$

Inserting into (A.19) the expression for ψ in (A.21) and the expression for $\pi_{0L} - \pi_{0H}$ in (A.16) gives

$$\begin{aligned}\mathrm{E}N_H - \mathrm{E}N_L &= \frac{r_1(1-\rho_{av})(\rho_H-\rho_L)}{1-\rho_{av} r_1}\frac{(1-\rho_{av} r_1)}{(1-\rho_{av})(1-\rho_H r_1)(1-\rho_L r_1)} \\ &= \frac{(\rho_H-\rho_L)r_1}{(1-\rho_H r_1)(1-\rho_L r_1)} = \frac{1}{1-\rho_H r_1} - \frac{1}{1-\rho_L r_1}.\end{aligned}$$

■

We now obtain an expression for $(\mathrm{E}N_L + \tau \mathrm{E}N_H)/(1+\tau)$ that will be used in the derivation of expressions for $\mathrm{E}N_L$ and $\mathrm{E}N_H$.

Lemma A.5:

$$\frac{EN_L + \tau EN_H}{1+\tau} = \frac{1}{1+\tau}\left[\frac{\rho_L(1-\rho_{av} r_1)}{(1-\rho_{av})(1-\rho_L r_1)} + \frac{\tau\rho_H(1-\rho_{av} r_1)}{(1-\rho_{av})(1-\rho_H r_1)}\right]. \tag{A.22}$$

Proof: Using (23) and (24) gives

$$\mathrm{E}N_L + \tau \mathrm{E}N_H = \frac{(1-\rho_2\rho_3) - \pi_{0L}}{(1-\rho_2)(1-\rho_3)} + \tau\frac{(1-\rho_2\rho_3) - \pi_{0H}}{(1-\rho_2)(1-\rho_3)}.$$

After some rearrangement and simplification, this becomes

$$\frac{\mathrm{E}N_L + \tau \mathrm{E}N_H}{1+\tau} = \left[(1-\rho_2\rho_3) - \frac{(\pi_{0L}+\tau\pi_{0H})}{1+\tau}\right]\psi.$$

By Theorem A.2, $(\pi_{0L}+\tau\pi_{0H})/(1+\tau) = \rho_{av}$. Thus,

$$\frac{\mathrm{E}N_L + \tau \mathrm{E}N_H}{1+\tau} = [(1-\rho_2\rho_3) - (1-\rho_{av})]\psi = [(\rho_{av} - \rho_2\rho_3]\psi.$$

Since $C_0 = r_1r_2r_3 = r_1/(\rho_2\rho_3) = 1/(\rho_L\rho_H)$ implies $\rho_2\rho_3 = r_1\rho_L\rho_H$, using (A.21) for ψ gives

$$\frac{\mathrm{E}N_L + \tau \mathrm{E}N_H}{1+\tau} = \frac{(\rho_{av} - \rho_L\rho_H r_1)(1-\rho_{av}r_1)}{(1-\rho_{av})(1-\rho_H r_1)(1-\rho_L r_1)}. \tag{A.23}$$

The following identity, where y is an arbitrary real number, is straightforward to derive.

$$(1+\tau)(\rho_{av} - \rho_L\rho_H y) = \rho_L(1-\rho_H y) + \tau\rho_H(1-\rho_L y). \tag{A.24}$$

In particular, upon setting $y = 1$,

$$(1+\tau)(\rho_{av} - \rho_L\rho_H) = \rho_L(1-\rho_H) + \tau\rho_H(1-\rho_L). \tag{A.25}$$

Rearranging this gives

$$\rho_{av} - \rho_L\rho_H r_1 = \frac{1}{1+\tau}[\rho_L(1-\rho_H r_1) + \tau\rho_H(1-\rho_L r_1)].$$

Substituting this into (A.23) gives

$$\frac{\mathrm{E}N_L + \tau \mathrm{E}N_H}{1+\tau} = \frac{1}{1+\tau}\left[\frac{\rho_L(1-\rho_{av}r_1)}{(1-\rho_{av})(1-\rho_L r_1)} + \frac{\tau\rho_H(1-\rho_{av}r_1)}{(1-\rho_{av})(1-\rho_H r_1)}\right].$$

■

$\mathrm{E}N_L$ and $\mathrm{E}N_H$ can be now be determined by the following.

Lemma A.6:

$$EN_L = \frac{\rho_{av}}{1-\rho_{av}} - \frac{\tau(1-\rho_H)(EN_H - EN_L)}{(1+\tau)(1-\rho_{av})}, \tag{A.26}$$

$$EN_H = \frac{\rho_{av}}{1-\rho_{av}} + \frac{(1-\rho_L)(EN_H - EN_L)}{(1+\tau)(1-\rho_{av})}. \tag{A.27}$$

PROOF: First express $(1+\tau)\mathrm{E}N_H$ as

$$(1+\tau)\mathrm{E}N_H = (\mathrm{E}N_H - \mathrm{E}N_L) + (\mathrm{E}N_L + \tau\mathrm{E}N_H).$$

Using the expressions for $\mathrm{E}N_L - \mathrm{E}N_H$ and $(\mathrm{E}N_L + \tau\mathrm{E}N_H)/(1+\tau)$ in (A.20) and (A.22), respectively, yields:

$$(1+\tau)\mathrm{E}N_H = \frac{1}{1-\rho_H r_1} - \frac{1}{1-\rho_L r_1} + \frac{\rho_L(1-\rho_{av}r_1)}{(1-\rho_{av})(1-\rho_L r_1)} + \frac{\tau\rho_H(1-\rho_{av}r_1)}{(1-\rho_{av})(1-\rho_H r_1)}.$$

which, after making following two expansions:

$$\frac{\rho_L(1-\rho_{av}r_1)}{(1-\rho_{av})(1-\rho_L r_1)} = \frac{\rho_{av}}{1-\rho_{av}} + \frac{\rho_L-\rho_{av}}{(1-\rho_{av})(1-\rho_L r_1)}$$

and

$$\frac{\rho_H(1-\rho_{av}r_1)}{(1-\rho_{av})(1-\rho_H r_1)} = \frac{\rho_{av}}{1-\rho_{av}} + \frac{\rho_H-\rho_{av}}{(1-\rho_{av})(1-\rho_H r_1)},$$

gives

$$\begin{aligned}(1+\tau)\mathrm{E}N_H &= \frac{1}{1-\rho_H r_1} - \frac{1}{1-\rho_L r_1} + \frac{\rho_{av}}{1-\rho_{av}} + \frac{\rho_L-\rho_{av}}{(1-\rho_{av})(1-\rho_L r_1)} + \frac{\tau\rho_{av}}{1-\rho_{av}} \\ &+ \frac{\tau(\rho_H-\rho_{av})}{(1-\rho_{av})(1-\rho_H r_1)} = \frac{(1+\tau)\rho_{av}}{1-\rho_{av}} + \frac{1-\rho_L}{1-\rho_{av}}\left[\frac{1}{1-\rho_H r_1} - \frac{1}{1-\rho_L r_1}\right].\end{aligned}$$

Using the expression for $\mathrm{E}N_H - \mathrm{E}N_L$ in Lemma A.4, this can be simplified to give

$$\mathrm{E}N_H = \frac{\rho_{av}}{1-\rho_{av}} + \frac{(1-\rho_L)(\mathrm{E}N_H - \mathrm{E}N_L)}{(1+\tau)(1-\rho_{av})}.$$

The result for $\mathrm{E}N_L$ is obtained similarly. ■

Proof of Corollary 4.1

PROOF: By differentiating (39) and setting $z=1$, $\mathrm{E}N$, is given by

$$\mathrm{E}N = \frac{\alpha_H\mathrm{E}N_L + \alpha_L\mathrm{E}N_H}{\alpha_L+\alpha_H}. \tag{A.28}$$

After substituting in (A.26) and (A.27), this gives

$$\mathrm{E}N = \frac{\rho_{av}}{1-\rho_{av}} + \frac{(\mathrm{E}N_H - \mathrm{E}N_L)[\alpha_L(1-\rho_L) - \alpha_H\tau(1-\rho_H)]}{(\alpha_L+\alpha_H)(1+\tau)(1-\rho_{av})}. \tag{A.29}$$

The probability, π_0, is found by setting $z=0$ in (39), to give

$$\pi_0 = \frac{\frac{\pi_{0L}}{\alpha_L} + \frac{\pi_{0H}}{\alpha_H}}{\frac{1}{\alpha_L} + \frac{1}{\alpha_H}}.$$

Substituting in (A.14) and (A.15) gives

$$\pi_0 = 1-\rho_{av} + \frac{\alpha_L\alpha_H r_1(1-\rho_{av})(\rho_H-\rho_L)}{(\alpha_L+\alpha_H)(1+\tau)(1-\rho_{av}r_1)}\left[\frac{\tau}{\alpha_L} - \frac{1}{\alpha_H}\right].$$

■

APPENDIX B. PROOF OF LEMMA 5.1

Proof: Start by differentiating (A.3), (A.4), and (A.5), respectively, to give

$$
\begin{aligned}
&r_1' + r_2' + r_3' = \frac{\rho_{av}}{\rho_H \rho_L},\\
&r_1'(r_2+r_3) + r_2'(r_1+r_3) + r_3'(r_1+r_2) = \frac{1}{\rho_L \rho_H},\\
&r_1' r_2 r_3 + r_2' r_1 r_3 + r_3' r_1 r_2 = 0.
\end{aligned}
$$

These equations can be expressed in matrix form as

$$Pr' = b,$$

where

$$P = \begin{pmatrix} 1 & 1 & 1 \\ r_2+r_3 & r_1+r_3 & r_1+r_2 \\ r_2 r_3 & r_1 r_3 & r_1 r_2 \end{pmatrix}, \quad r' = \begin{pmatrix} r_1' \\ r_2' \\ r_3' \end{pmatrix}, \quad b' = \begin{pmatrix} \frac{\rho_{av}}{\rho_L \rho_H} \\ \frac{1}{\rho_L \rho_H} \\ 0 \end{pmatrix}.$$

The determinant of P is given by

$$\Delta = \det(P) = (r_2 - r_3)(r_1 - r_3)(r_1 - r_2).$$

Cramer's rule can then be used to obtain r_1' as

$$r_1' = \frac{N_1}{\Delta},$$

where

$$N_1 = \begin{vmatrix} \frac{\rho_{av}}{\rho_L \rho_H} & 1 & 1 \\ \frac{1}{\rho_L \rho_H} & r_1+r_3 & r_1+r_2 \\ 0 & r_1 r_3 & r_1 r_2 \end{vmatrix} = \frac{r_1(r_2-r_3)(\rho_{av} r_1 - 1)}{\rho_L \rho_H},$$

giving

$$r_1' = \frac{r_1(\rho_{av} r_1 - 1)}{\rho_L \rho_H (r_1 - r_3)(r_1 - r_2)}.$$

r_2' and r_3' can be found similarly. ■

APPENDIX C. PROOF OF LEMMA 5.3

Proof: (i) Rewrite (45) as

$$g(x) = \frac{x(b-a)}{(a-x)(b-x)} > 0.$$

(ii) Taking the derivative of $g(x)$ gives

$$g'(x) = \frac{a}{(a-x)^2} - \frac{b}{(b-x)^2}, \tag{C.1}$$

which, after a little algebra, equals

$$g'(x) = \frac{(b-a)(ab-x^2)}{(a-x)^2(b-x)^2} > 0;$$

and so, $g(x)$ is increasing.

(iii) The second derivative of $g(x)$ equals

$$g''(x) = \frac{2a}{(a-x)^3} - \frac{2b}{(b-x)^3},$$

which after algebraic manipulation becomes

$$= \frac{2(b-a)(ab(b+a) - 3abx + x^3)}{(a-x)^3(b-x)^3}.$$

After defining $h(x) = ab(b+a) - 3abx + x^3$, we have $h(0) = ab(b+a) > 0$ and $h(a) = ab(b+a) - 3a^3b + b^3 = 2a(b-a)^3 > 0$. Thus, for $h(x)$ to have a value less than 0 for $x \in (0,a)$, then $h(x)$ must have a turning point for $x \in (0,a)$. Now, $h'(x) = -3ab + 3x^2 = 3(x^2 - ab) = 3(x - \sqrt{ab})(x + \sqrt{ab})$, and so $h(x)$ has only two turning points: One is at $x = -\sqrt{ab} < 0$ and the other is at $a < x = \sqrt{ab} < b$. Thus, $h(x)$ has no turning point for $x \in (0,a)$. Therefore, $h(x) > 0$ for $x \in (a,b)$, and so $g''(x) > 0$ for $x \in (0,a)$; proving that $g(x)$ is convex. ■

APPENDIX D. PROOF OF THEOREM 6.1

PROOF: From (48) (repeated here),

$$\Pr(N_H > n) = \frac{a_H \rho_2^{n+1}}{1-\rho_2} + \frac{b_H \rho_3^{n+1}}{1-\rho_3},$$

and, from (22) (repeated here),

$$a_H = \frac{(1-\rho_2)(1-\rho_3-\pi_{0H})}{(\rho_2-\rho_3)}, \quad b_H = \frac{(1-\rho_3)(1-\rho_2-\pi_{0H})}{(\rho_3-\rho_2)},$$

we obtain

$$\begin{aligned}\Pr(N_H > n) &= \left[\frac{(1-\rho_2)(1-\rho_3-\pi_{0H})}{\rho_2-\rho_3}\right]\frac{\rho_2^{n+1}}{1-\rho_2} + \left[\frac{(1-\rho_3)(1-\rho_2-\pi_{0H})}{\rho_3-\rho_2}\right]\frac{\rho_3^{n+1}}{1-\rho_3}\\ &= \frac{1}{\rho_2-\rho_3}[(1-\rho_3-\pi_{0H})\rho_2^{n+1} - (1-\rho_2-\pi_{0H})\rho_3^{n+1})]\\ &= \frac{1}{\rho_2-\rho_3}[(1-\rho_3)\rho_2^{n+1} - (1-\rho_2)\rho_3^{n+1} - \pi_{0H}(\rho_2^{n+1} - \rho_3^{n+1})]\\ &= \frac{1}{\rho_2-\rho_3}[(\rho_2^{n+1} - \rho_3^{n+1}) - \rho_2\rho_3(\rho_2^n - \rho_3^n) - \pi_{0H}(\rho_2^{n+1} - \rho_3^{n+1})]. \qquad \textbf{(D.1)}\end{aligned}$$

Rearrangement of (A.1) gives

$$\begin{aligned}\Pr(N_H > n) &= \frac{1}{\rho_2-\rho_3}[(\rho_2^{n+1} - \rho_3^{n+1}) - \rho_2^{n+1}\rho_3 + \rho_3^{n+1}\rho_2 - \pi_{0H}(\rho_2^{n+1} - \rho_3^{n+1})]\\ &= \frac{1}{\rho_2-\rho_3}[(\rho_2^{n+1} - \rho_3^{n+1}) - \rho_2^{n+1}\rho_3 + \rho_3^{n+2} - \rho_3^{n+2} + \rho_3^{n+1}\rho_2 - \pi_{0H}(\rho_2^{n+1} - \rho_3^{n+1})]\\ &= \frac{1}{\rho_2-\rho_3}[(\rho_2^{n+1} - \rho_3^{n+1}) - \rho_3(\rho_2^{n+1} - \rho_3^{n+1}) + \rho_3^{n+1}(\rho_2-\rho_3) - \pi_{0H}(\rho_2^{n+1} - \rho_3^{n+1})]\\ &= \frac{1}{\rho_2-\rho_3}(\rho_2^{n+1} - \rho_3^{n+1})[1-\rho_3-\pi_{0H})] + \rho_3^{n+1}\\ &= [\rho_2^n + \rho_2^{n-1}\rho_3 + \cdots + \rho_2\rho_3^{n-1} + \rho_3^n][1-\rho_3-\pi_{0H})] + \rho_3^{n+1}.\end{aligned}$$

Since ρ_2 and ρ_3 are decreasing functions of κ (because r_2 and r_3 are increasing functions), the term in the first brackets $[\cdots]$ is decreasing, as is ρ_3^{n+1}. Thus, if it can be shown that the term in the second brackets $[\cdots]$, which, using (A.11), is given by

$$F \equiv 1-\rho_3-\pi_{0H} = 1-\frac{1}{r_3}-\frac{(1-\rho_{av})(1-\rho_H r_1)}{1-\rho_{av}r_1},$$

is decreasing in κ for some set of values of ρ_L, ρ_H, and ρ_{av}, then sufficient conditions exist for N_H to be stochastically decreasing with κ.

Observe, in the limiting case of $\kappa \to 0$ (since $r_3 \to 1/\rho_L$ and $r_1 \to 1$), assuming $\rho_H < 1$,

$$F \to 1-\rho_L-\frac{(1-\rho_{av})(1-\rho_H)}{1-\rho_{av}} = \rho_H-\rho_L$$

and in the case of $\kappa \to \infty$ (since $r_3 \to \infty$ and $r_1 \to 0$)

$$F \to 1-(1-\rho_{av}) = \rho_{av}.$$

Thus, for F to be decreasing in κ over the entire interval $(0,\infty)$, it is required that at least $\rho_H-\rho_L \geq \rho_{av}$. However, this minimal condition is not even satisfied in many cases. For example, take $\tau=1$, $\rho_H=0.9$, $\rho_L=0.7$, $\rho_{av}=(\rho_L+\tau\rho_H)/(1+\tau)=0.8$. However, in such cases, it is still possible for F to be decreasing in portions of the interval $(0,\infty)$.

Guided by the fact that $\pi_{0H} \to 1-\rho_{av}$ $(\kappa\to\infty)$, (since $r_1 \to 0$), F can be expressed as

$$F = \rho_{av}-\frac{1}{r_3}+((1-\rho_{av})-\pi_{0H}) = \rho_{av}-\frac{1}{r_3}+A,$$

where $A = 1-\rho_{av}-\pi_{0H}$. As the constant ρ_{av} can be ignored, it is sufficient to show that $A-1/r_3$ is decreasing. Using (A.15) gives

$$\begin{aligned} A &= \frac{r_1(1-\rho_{av})(\rho_H-\rho_L)}{(1+\tau)(1-\rho_{av}r_1)} = \frac{r_1(1-\rho_{av})(\rho_H-\rho_{av})}{1-\rho_{av}r_1} \\ &= \frac{(1-\rho_{av})}{\rho_{av}}\frac{(\rho_H-\rho_{av})(\rho_{av}r_1-1+1)}{1-\rho_{av}r_1} = \frac{(1-\rho_{av})(\rho_H-\rho_{av})}{\rho_{av}}\left[-1+\frac{1}{1-\rho_{av}r_1}\right]. \end{aligned}$$

(The last equality in the first line is obtained by using (15).) This gives

$$A-\frac{1}{r_3} = -\frac{(1-\rho_{av})(\rho_H-\rho_{av})}{\rho_{av}}+\frac{(1-\rho_{av})(\rho_H-\rho_{av})}{\rho_{av}(1-\rho_{av}r_1)}-\frac{1}{r_3}.$$

As the first term is a constant, we only need to show that the second and third terms taken together are decreasing. Setting $z=1/\rho_{av}$ in (A.6) gives

$$\rho_L\rho_H\left(\frac{1}{\rho_{av}}-r_1\right)\left(\frac{1}{\rho_{av}}-r_2\right)\left(\frac{1}{\rho_{av}}-r_3\right) = -\left(\frac{1}{\rho_{av}}-1\right)\left(1-\frac{\rho_L}{\rho_{av}}\right)\left(\frac{\rho_H}{\rho_{av}}-1\right),$$

which can be rearranged as

$$-\frac{(1-\rho_{av})(\rho_H-\rho_{av})}{(1-\rho_{av}r_1)} = \frac{\rho_L\rho_H(1-\rho_{av}r_2)(1-\rho_{av}r_3)}{(\rho_{av}-\rho_L)}. \tag{D.2}$$

Therefore (using (A.5)),

$$\begin{aligned} \frac{(1-\rho_{av})(\rho_H-\rho_{av})}{\rho_{av}(1-\rho_{av}r_1)}-\frac{1}{r_3} &= -r_1r_2\rho_L\rho_H-\frac{\rho_L\rho_H(1-\rho_{av}r_2)(1-\rho_{av}r_3)}{\rho_{av}(\rho_{av}-\rho_L)} \\ &= \rho_L\rho_H\left[-r_1r_2-\frac{(1-\rho_{av}r_2)(1-\rho_{av}r_3)}{\rho_{av}(\rho_{av}-\rho_L)}\right]. \end{aligned}$$

Ignoring the positive constant factor $\rho_L\rho_H$, we need to show the following is decreasing:

$$-r_1r_2 - \frac{(1-\rho_{av}r_2)(1-\rho_{av}r_3)}{\rho_{av}(\rho_{av}-\rho_L)} = -\frac{r_1r_2\rho_{av}(\rho_{av}-\rho_L)+(1-\rho_{av}r_2)(1-\rho_{av}r_3)}{\rho_{av}(\rho_{av}-\rho_L)}$$
$$= -\frac{r_1r_2\rho_{av}^2 - r_1r_2\rho_{av}\rho_L + 1 - \rho_{av}r_2 - \rho_{av}r_3 + \rho_{av}^2r_2r_3}{\rho_{av}(\rho_{av}-\rho_L)}.$$

Since the denominator is a positive constant then what we need to show is equivalent to showing that the negative of the numerator is increasing, which, using C_1 and C_2 from (A.4) and (A.3), is equal to

$$\begin{aligned} H &\equiv r_1r_2\rho_{av}^2 - r_1r_2\rho_{av}\rho_L + 1 - \rho_{av}r_2 - \rho_{av}r_3 + \rho_{av}^2r_2r_3 \\ &= \rho_{av}^2(r_1r_2+r_2r_3) - \rho_{av}(r_2+r_3) + 1 - r_1r_2\rho_{av}\rho_L \\ &= \rho_{av}^2(C_1 - r_1r_3) - \rho_{av}(C_2 - r_1) + 1 - r_1r_2\rho_{av}\rho_L. \end{aligned}$$

Noting, again, from (A.4) and (A.3), that $(d/d\kappa)C_1 = 1/\rho_L\rho_H$ and $(d/d\kappa)C_2 = \rho_{av}/\rho_L\rho_H$, the derivative of H is given by

$$\begin{aligned} &\rho_{av}^2\frac{1}{\rho_L\rho_H} - \rho_{av}^2(r_1r_3' + r_1'r_3) - \rho_{av}\frac{\rho_{av}}{\rho_L\rho_H} + \rho_{av}r_1' - \rho_{av}\rho_L(r_1'r_2 + r_1r_2') \\ &= -\rho_{av}^2(r_1r_3' + r_1'r_3) + \rho_{av}r_1' - \rho_{av}\rho_L(r_1'r_2 + r_1r_2'). \end{aligned}$$

Taking the derivative of $C_0 = r_1r_2r_3 = 1/(\rho_L\rho_H)$ gives

$$r_1'r_2r_3 + r_1r_2'r_3 + r_1r_2r_3' = 0,$$

from which

$$\begin{aligned} r_1r_2'r_3 &= -(r_1'r_3 + r_1r_3')r_2 \\ r_1r_2r_3' &= -(r_1'r_2 + r_1r_2')r_3, \end{aligned}$$

allowing the derivative of H to be given by

$$\rho_{av}^2\frac{r_1r_2'r_3}{r_2} + \rho_{av}r_1' + \rho_L\rho_{av}\frac{r_1r_2r_3'}{r_3} = \rho_{av}r_1\left[\frac{r_1'}{r_1} + \rho_{av}r_3\frac{r_2'}{r_2} + \rho_Lr_2\frac{r_3'}{r_3}\right].$$

Therefore, since $\rho_{av}r_1 > 0$, it is required to show that

$$0 < \frac{r_1'}{r_1} + \rho_{av}r_3\frac{r_2'}{r_2} + \rho_Lr_2\frac{r_3'}{r_3}. \tag{D.3}$$

Since

$$\log(r_1r_2r_3) = \log(r_1) + \log(r_2) + \log(r_3) = \log\left(\frac{1}{\rho_L\rho_H}\right),$$

taking derivatives of this gives

$$\frac{r_1'}{r_1} + \frac{r_2'}{r_2} + \frac{r_3'}{r_3} = 0.$$

Applying this to (A.3) means it is required to show that

$$\begin{aligned} 0 &< -\frac{r_2'}{r_2} - \frac{r_3'}{r_3} + \rho_{av}r_3\frac{r_2'}{r_2} + \rho_Lr_2\frac{r_3'}{r_3} = (\rho_{av}r_3 - 1)\frac{r_2'}{r_2} + (\rho_Lr_2 - 1)\frac{r_3'}{r_3} \\ &= \frac{r_3'}{r_3}\left[(\rho_{av}r_3 - 1)\frac{r_3}{r_2}\frac{r_2'}{r_3'} + (\rho_Lr_2 - 1)\right]. \end{aligned}$$

Since $r_3, r_3' > 0$, the requirement becomes

$$0 < (\rho_{av} r_3 - 1)\frac{r_3}{r_2}\frac{r_2'}{r_3'} + (\rho_L r_2 - 1).$$

Using (43) and (44),

$$\frac{r_3}{r_2}\frac{r_2'}{r_3'} = \frac{r_3}{r_2}\frac{\frac{r_2(\rho_{av}r_2-1)}{\rho_L\rho_H(r_2-r_3)(r_2-r_1)}}{\frac{r_3(\rho_{av}r_3-1)}{\rho_L\rho_H(r_3-r_2)(r_3-r_1)}} = \frac{(\rho_{av}r_2-1)(r_3-r_1)}{(\rho_{av}r_3-1)(r_1-r_2)}.$$

Hence, the requirement for decreasing stochastic order at κ is

$$0 < (\rho_{av}r_3-1)\frac{(\rho_{av}r_2-1)(r_3-r_1)}{(\rho_{av}r_3-1)(r_1-r_2)} + \rho_L r_2 - 1 = \frac{(1-\rho_{av}r_2)(r_3-r_1)}{(r_2-r_1)} - (1-\rho_L r_2),$$

which, upon minor rearrangement, proves the theorem. ■

APPENDIX E. DERIVATION OF MINIMUM COST CONDITION

Taking partial derivatives of (50), and equating to zero gives

$$0 = \frac{\partial C(\mu_L,\mu_H)}{\partial \mu_L} = c\frac{\frac{1}{\alpha_L}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}} - h\frac{\frac{1}{\alpha_L}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}}\frac{\lambda_{av}}{\left(\frac{\frac{\mu_L}{\alpha_L}+\frac{\mu_H}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}} - \lambda_{av}\right)^2} + hD\frac{r_1\lambda_L}{(\mu_L - r_1\lambda_L)^2}$$

$$= \frac{\frac{1}{\alpha_L}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}}\left[c - h\frac{\lambda_{av}}{\left(\frac{\frac{\mu_L}{\alpha_L}+\frac{\mu_H}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}} - \lambda_{av}\right)^2}\right] + hD\frac{r_1\lambda_L}{(\mu_L - r_1\lambda_L)^2},$$

and

$$0 = \frac{\partial C(\mu_L,\mu_H)}{\partial \mu_H} c\frac{\frac{1}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}} - h\frac{\frac{1}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}}\frac{\lambda_{av}}{\left(\frac{\frac{\mu_L}{\alpha_L}+\frac{\mu_H}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}} - \lambda_{av}\right)^2} - hD\frac{r_1\lambda_H}{(\mu_H - r_1\lambda_H)^2}$$

$$= \frac{\frac{1}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}}\left[c - h\frac{\lambda_{av}}{\left(\frac{\frac{\mu_L}{\alpha_L}+\frac{\mu_H}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}} - \lambda_{av}\right)^2}\right] - hD\frac{r_1\lambda_H}{(\mu_H - r_1\lambda_H)^2}.$$

Hence,

$$\frac{\frac{1}{\alpha_L}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}}\left[c - h\frac{\lambda_{av}}{\left(\frac{\frac{\mu_L}{\alpha_L}+\frac{\mu_H}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}} - \lambda_{av}\right)^2}\right] = -hD\frac{r_1\lambda_L}{(\mu_L - r_1\lambda_L)^2}$$

$$\frac{\frac{1}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}}\left[c - h\frac{\lambda_{av}}{\left(\frac{\frac{\mu_L}{\alpha_L}+\frac{\mu_H}{\alpha_H}}{\frac{1}{\alpha_L}+\frac{1}{\alpha_H}} - \lambda_{av}\right)^2}\right] = hD\frac{r_1\lambda_H}{(\mu_H - r_1\lambda_H)^2}.$$

APPENDIX F. PROOF OF THEOREM 7.1

Proof: For simplicity of notation, instead of $\widehat{\Pi}^{(S)}(z)$ the proof is given for $\mathrm{E}N^{(S)}$. The proof for $\widehat{\Pi}^{(S)}(z)$ will be along similar lines.

$\mathrm{E}N^{(S)}$ can be found using the renewal-reward theorem applied to discrete-time processes. Consider the renewal cycle comprising of one LOW phase and the subsequent HIGH phase in the original system. Let C be the number of sample points in this cycle (note that it is possible for C to be 0). The renewal-reward theorem gives

$$\mathrm{E}N^{(S)} = \frac{\mathrm{E}(\sum_{i=1}^{C} N_i^{(S)})}{\mathrm{E}C}. \tag{F.1}$$

In the denominator, the mean number of samples during a cycle is given by

$$\mathrm{E}C = \frac{\beta_L}{\alpha_L} + \frac{\beta_H}{\alpha_H}.$$

To derive the numerator of (F.1), a similar approach to the modified system approach used in the proof of Theorem 3 of Gupta *et al.* [17] will be followed. We define the following modified systems.

Modified system 1: this system operates the same as the $M/M/1(R)$ system except that when the system switches from the HIGH phase to the LOW phase, the number of jobs at the beginning of the LOW phase is randomly sampled from the distribution of N_H. The distribution of the number in system is the same as in the $M/M/1(R)$ system because the distribution of jobs at the start of a LOW phase is the same and the system is Markovian.

Modified system 2(LOW): In this system, a governing Poisson process of rate α_L runs. With each event in the governing Poisson process, the number in the system is reset by sampling from the distribution of N_L at the start of a LOW phase in the $M/M/1(R)$ system. The system between arrivals in the governing Poisson process operates the same as in the LOW phases of the $M/M/1(R)$ system until the next event in the governing Poisson process. The system is the equivalent to the LOW phases of modified system 1 being stitched together. Modified system 2(HIGH) is defined similarly.

In the notation below, $i = 1$ denotes modified system 1, $i = 2$ denotes either modified system 2(LOW) or modified system 2(HIGH), depending on the context, and if i is absent, it denotes the original system. When considering LOW phases, let CYCLE denote a LOW phase in the context of the original system and modified system 1 and denote the time from an arrival in the governing Poisson process to the next arrival in that process in modified system 2(LOW).

For system i define:

$C_L^{(i)}$ is the number of samples in a CYCLE in system i.
$T_L^{(i)}$ be the time duration of a CYCLE.
$\{N_L^{(i,t)},\ 0 \le t < T_L^{(i)}\}$ is the number in system at time t during a CYCLE.
$N_L^{(i)}$ is the number in the system at an arbitrary time.
$\{N_{Li}^{(i,S)},\ i = 1, \ldots, C_L^{(i)}\}$ is the number in system sample values during a cycle.
$N_L^{(i,S)}$ is the value of an arbitrary sample value.

Similar notation is used for the HIGH phase and the modified system 2(HIGH).

Using this notation, (F.1) can be expressed as

$$\mathrm{E}N^{(S)} = \frac{\mathrm{E}(\sum_{i=1}^{C_L} N_{Li}^{(S)}) + \mathrm{E}(\sum_{i=1}^{C_H} N_{Hi}^{(S)})}{\mathrm{E}C}.$$

The LOW and HIGH phases are now examined separately. Consider first the LOW phase for which the following is obtained. Since the distribution at the start of a CYCLE is the same by construction, the sampling Poisson process has the same rate in each system, and the same Markov chain governs the system evolution during a CYCLE, we obtain

$$\mathrm{E}\left(\sum_{i=1}^{C_L} N_{Li}^{S}\right) = \mathrm{E}(\sum_{i=1}^{C_L^{(2)}} N_{Li}^{(2,S)}). \tag{F.2}$$

Applying the renewal-reward theorem to the modified system 2(LOW) gives

$$\frac{\mathrm{E}(\sum_{i=1}^{C_L^{(2)}} N_{Li}^{(2,S)})}{\mathrm{E}C_L^{(2)}} = \mathrm{E}N_L^{(2,S)}. \tag{F.3}$$

By applying the PASTA property to the sampling Poisson process in modified system 2(LOW), the mean number in system as seen at sampling epochs is the same as seen at an arbitrary time:

$$\mathrm{E}N_L^{(2,S)} = \mathrm{E}N_L^{(2)}.$$

Next, at a random time point, the backward recurrence time to the previous point in the governing Poisson process in modified system 2(LOW) is an exponential random variable with rate α_L. Thus,

$$\mathrm{E}N_L^{(2)} = \mathrm{E}N_L^{(2,T_L^{(2)})}.$$

(Gupta *et al.* [17] explained this using the PASTA principle.)

Since the original system, modified system 1 at the end of a LOW phase and modified system 2(LOW) at the end of a CYCLE are the same probabilistically:

$$\mathrm{E}N_L^{(2,T_L^{(2)})} = \mathrm{E}N_L^{(1,T_L^{(1)})} = \mathrm{E}N_L^{(T_L)}.$$

Combining the above gives

$$\mathrm{E}N_L^{(2,S)} = \mathrm{E}N_L^{(T_L)} = \mathrm{E}N_L,$$

where the last equality follows by definition. Also, since a CYCLE has an exponential distribution with rate α_L in each system then $\mathrm{E}C_L = \mathrm{E}C_L^{(2)}$. Hence, (F.2) and (F.3) give

$$\mathrm{E}\left(\sum_{i=1}^{C_L} N_{Li}^{(S)}\right) = \mathrm{E}N_L\mathrm{E}C_L.$$

A similar result holds for the HIGH phases:

$$\mathrm{E}N^{(S)} = \frac{\mathrm{E}N_L\mathrm{E}C_L + \mathrm{E}N_H\mathrm{E}C_H}{\mathrm{E}C} = \frac{\frac{\beta_L\mathrm{E}N_L}{\alpha_L} + \frac{\beta_H\mathrm{E}N_H}{\alpha_H}}{\frac{\mu_L}{\beta_L} + \frac{\beta_H}{\alpha_H}}.$$

■