CONTRIBUTED PAPER

TRANSIENT BEHAVIOR OF THE M/M/1 QUEUE: STARTING AT THE ORIGIN

Joseph ABATE and Ward WHITT

AT&T Bell Laboratories, U.S.A. *

Received 5 May 1986
(Revised 15 October 1986)

Abstract

This paper presents some new perspectives on the time-dependent behavior of the M/M/1 queue. The factorial moments of the queue length as functions of time when the queue starts empty have interesting structure, which facilitates developing approximations. Simple exponential and hyperexponential approximations for the first two moment functions help show how the queue approaches steady state as time evolves. These formulas also help determine if steady-state descriptions are reasonable when the arrival and service rates are nearly constant over some interval but the process does not start in steady state.

Keywords

Transient behavior; approach to steady state; relaxation times; birth-and-death process; queues; Brownian motion; first passage times; busy period; complete monotonicity.

1. Introduction

This paper describes the evolution of the classical M/M/1 queue. Of course the M/M/1 model has been studied extensively and much is known about its time-dependent or transient behavior, e.g., chapter II.2 of Cohen [5] and sect. 1.2 of Prabhu [23], but we believe that there is more to discover. Our goal is to obtain simple approximations and structural theorems that expose the essential nature of the transient behavior.

We focus on $Q(t)$, the queue length (including the customer in service) at time t in the M/M/1 queue. This is a birth-and-death process on the nonnegative integers with constant arrival (birth) rate λ and constant service (death) rate μ.

* Current addresses of the authors: JOSEPH ABATE, AT&T Bell Laboratories, LC 2W-E06, 184 Liberty Corner Road, Warren, NJ 07060, U.S.A.
WARD WHITT, AT&T Bell Laboratories, MH 2C-178, 600 Mountain Avenue, Murray Hill, NJ 07974, U.S.A.

We fix the measuring units for time by setting $\mu = 1$. Then the traffic intensity $\rho = \lambda/\mu$ is just λ. We assume that $\rho < 1$, so that the system is stable; i.e., $Q(t)$ converges in distribution as $t \to \infty$ to a random variable $Q(\infty)$ with a geometric distribution, i.e., $P(Q(\infty) = k) = (1-\rho)\rho^k$, $k \geq 0$. We want to understand how $Q(t)$ approaches this steady-state limit; e.g., we want to describe the moments of $Q(t)$ as functions of time.

One way that has been proposed to obtain more easily comprehensible descriptions of the transient behavior is to consider limit theorems describing the asymptotic behavior as $t \to \infty$. This approach is typically discussed under the name *relaxation times*; see Blanc [4], Cohen [5], Keilson [15] and references cited there. One of our goals is to investigate the quality of these asymptotic approximations. While these approximations do give a rough idea about the transient behavior, unfortunately they do not seem to be very accurate. Roth [24], Odoni and Roth [22], Lee [17] and Lee and Roth [18] have done empirical investigations of the transient behavior of various GI/G/1 queues, which indicate that the limit theorems related to the relaxation time provide only crude lower bounds on the rate of approach to steady state. Our analysis supports their conclusion.

In this paper, we consider the stochastic process $\{Q(t): t \geq 0\}$ under the *special initial condition* $Q(0) = 0$. We are primarily interested in the conditional moments $m_k(t) \equiv m_k(t, 0) \equiv E(Q(t)^k \mid Q(0) = 0)$ as functions of time. We are mostly interested in the case $k = 1$, but we also consider general k to some extent. We want to understand how these moments approach their steady-state limits. For example, we want a relatively simple expression for the time required for $m_k(t, 0)$ to reach 99% of the steady-state limit $m_k(\infty) \equiv E(Q(\infty)^k)$.

Restricting attention to the special case $Q(0) = 0$ is important because $m_k(t, 0)$ is *increasing* in t. In contrast, it not difficult to prove that $m_1(t, n) \equiv E(Q(t) \mid Q(0) = n)$ is *initially decreasing* in t for all $n > 0$; see Kelton and Law [16] for typical numerical values. In fact, there are three possible shapes for the first-moment function $m_1(t, n)$. First, $m_1(t, n)$ is increasing in t for all t if and only if $n = 0$. Second, for sufficiently large n (depending on ρ), the function $m_1(t, n)$ is decreasing in t for all t. Otherwise, $m_1(t, n)$ is initially decreasing and then increasing in t. We call the smallest n such that $m_1(t, n)$ is decreasing in t for all t the *critical damping level* and denote it by n_d. The various possibilities are described in fig. 1. Note that $n_d > m_1(\infty)$; i.e., $m_1(t, n)$ can start above the steady-state limit $m_1(\infty)$, fall below it and then approach it from below. (In sect. 8 of Abate and Whitt [3] we apply results in van Doorn [28] to *prove* that $m_1(t, n)$ has this shape.)

In order to understand the moment functions with the general initial condition $Q(0) = n$, it is useful to decompose the moment function into two parts

$$m_k(t, n) = m_k(t, 0) + [m_k(t, n) - m_k(t, 0)] \tag{1.1}$$

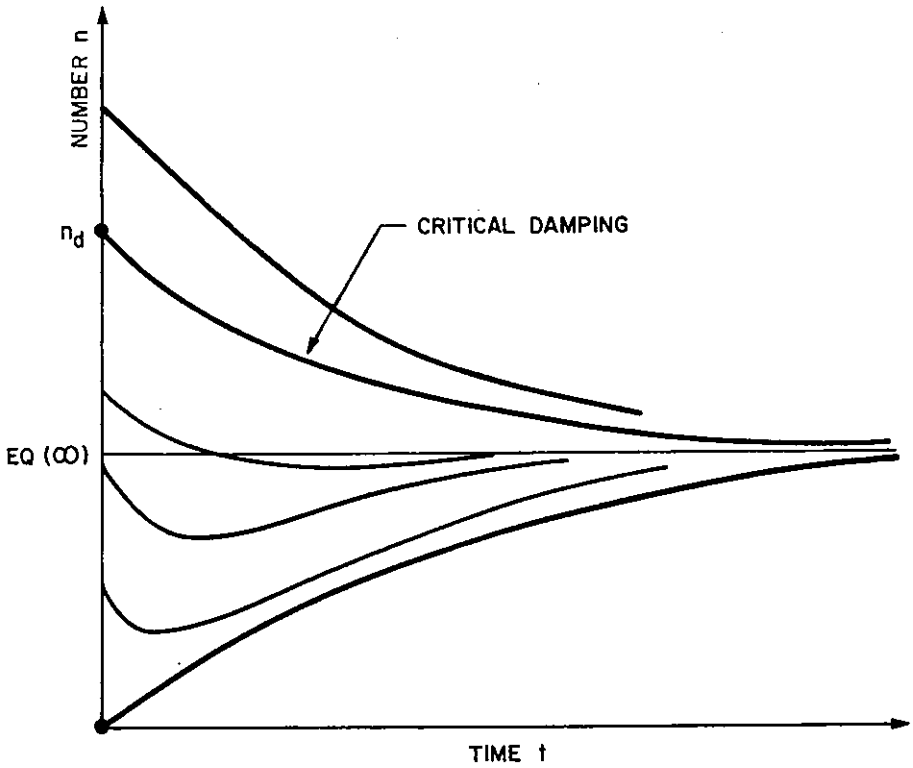

Fig. 1. $E(Q(t) \mid Q(0) = n)$ as a function of n and t.

because for $k = 1$ and 2 (but not $k \geq 3$) formula (1.1) represents $m_k(t, n)$ as the *sum of two monotone functions*. This is established in theorems 7.2 and 11.3 of Abate and Whitt [2] and in sects. 5 and 6 of Abate and Whitt [3]. We thus can apply similar techniques to approximate both parts. We treat the case of a general initial condition in Abate and Whitt [3]; here we focus on the case $Q(0) = 0$.

This paper is related to Abate and Whitt [1,2], in which we describe the transient behavior of regulated Brownian motion (RBM). The results here for the M/M/1 queue are discrete analogs of our RBM results. In turn, the RBM results can be obtained from the M/M/1 results in the limit as the traffic intensity ρ approaches 1. The results here will also be applied in a subsequent paper to develop approximate descriptions of the transient behavior of other GI/G/1 queues. For the GI/G/1 queue, we apply heavy-traffic limit theorems in Iglehart and Whitt [13] establishing convergence of the normalized GI/G/1 queue-length process to RBM as the traffic intensity ρ approaches the critical value 1. We could then, in the spirit of Gaver [10], directly apply descriptions of the transient behavior of RBM in Abate and Whitt [1,2], but instead we apply the heavy-traffic limit theorem *twice*, first to approximate the GI/G/1 queue-length process by

RBM and, second, to approximate RBM by the M/M/1 queue-length process studied here. The general interarrival-time and service-time distributions of the original GI/G/1 model enter in the approximations only via their first two moments. The second moments alter the M/M/1 approximations only by producing a transformation of the time scale, consistent with the empirically derived relaxation time (9) of Odoni and Roth [22].

The rest of this paper is organized as follows. In sect. 2 we develop our approximations and in sect. 3 we establish supporting theory. In sect. 3 we consider the *normalized factorial moments* of $(Q(t) \mid Q(0) = 0)$. Recall that, for any positive integer r, the r^{th} factorial moment of random variable X is $E[X(X-1)\cdots(X-r+1)]$; we denote it by $\phi_r(X)$. Of course, the first factorial moment is just the ordinary first moment; the factorial moments become significant when $r \geq 2$. When $Q(0) = 0$, it turns out that the normalized r^{th} factorial moment

$$H_r(t) = \phi_r(Q(t) \mid Q(0) = 0)/\phi_r(Q(\infty)), \quad t \geq 0, \tag{1.2}$$

is increasing in t, so that we can regard $H_r(t)$ as a cdf for each r. As we saw for RBM in Abate and Whitt [1], this probabilistic view pays handsome dividends. In theorem 3.1 we show that the factorial-moment cdf's can be identified with negative-binomial mixtures of convolutions of the M/M/1 busy-period cdf. The first-moment cdf $H_1(t)$ thus coincides with the equilibrium residual-life distribution associated with the M/M/1 busy-period cdf. In theorem 3.2 we show that the r^{th} factorial-moment cdf $H_r(t)$ is the r-fold convolution of the first-factorial moment cdf $H_1(t)$. As a consequence, we see that the $(r+1)^{\text{st}}$ factorial moment approaches steady state more slowly than the r^{th} factorial moment for each r. We also easily obtain the moments of the factorial-moment cdf's $H_r(t)$, so that we can approximate these factorial-moment cdf's by other cdf's by simply matching moments.

We conclude in sect. 4 by briefly discussing approximations for second and higher moments. Since the r^{th} factorial-moment cdf $H_r(t)$ is the r-fold convolution of the first-moment cdf $H_1(t)$, we can approximate $H_r(t)$ by the r-fold convolution of the cdf used to approximate $H_1(t)$. We show that this procedure works pretty well for $H_2(t)$ in the range of t of primary interest, but a direct hyperexponential fit to the first three moments of $H_2(t)$ performs even better.

2. Approximations

2.1. A SIMPLE EXPONENTIAL APPROXIMATION

We propose a simple exponential approximation for the first-moment function, namely,

$$m_1(\infty) - m_1(t) \equiv EQ(\infty) - E(Q(t) \mid Q(0) = 0) \approx b(\rho)\, e^{-t/a(\rho)} \tag{2.1}$$

for t sufficiently large, where

$$a(\rho) = 2(1-\rho)^{-2}c(\rho), \qquad b(\rho) = \rho(1-\rho)^{-1}(1+2c(\rho))^{-1}$$
$$c(\rho) = \left(\frac{2+\rho}{4}\right)\left(1+\left[1-\left(\frac{2}{2+\rho}\right)^2\right]^{1/2}\right) = \frac{2+\rho+[5-(1-\rho)(5+\rho)]^{1/2}}{4}. \tag{2.2}$$

(Later we will discuss the further simplifications $c(\rho) \approx (1+\rho^{1/2})^2/(2+\rho^{1/4})$ and $c(\rho) \approx (1+\rho^{1/2})^2/3$ for ρ not too small.) Approximation (2.1) is appropriate if t and ρ are not too small. For example, (2.1) tends to be good if $\rho \geq 0.50$ and $t \geq 2(1-\rho)^{-2}$.

Let $t_p \equiv t_p(\rho)$ be the time required for $E(Q(t) \mid Q(0)=0)$ to first be (and remain) within $100p\%$ of the steady-state limit $m_1(\infty)$. From (2.1), we obtain $b(\rho)e^{-t_p}a(\rho) \approx pm_1(\infty)$, so that

$$t_p(\rho) \approx a(\rho)\log[(1-\rho)b(\rho)/\rho p] = 2(1-\rho)^{-2}[c(\rho)\log(1/[p(1+2c(\rho))])] \tag{2.3}$$

for p sufficiently small, e.g., $p \leq 0.20$. The term $a(\rho)$ in (2.1)–(2.3) is our proposed *approximate relaxation time*.

Approximation (2.3) is especially convenient for quickly seeing how $t_p(\rho)$ depends on the parameters p and ρ. We can easily see the notorious increase of $t_p(\rho)$ as ρ increases. For example, from (2.2) and (2.3), we obtain $c(0.5) = 1.00$ and $t_{0.01}(0.5) \approx 28$, but $c(0.9) = 1.25$ and $t_{0.01}(0.9) \approx 838$. The first-order behavior is captured by the time-scaling $2(1-\rho)^{-2}$, which assumes the values 8 and 200 for $\rho = 0.5$ and 0.9. Formula (2.3) provides a refinement.

Approximations (2.1) and (2.2) agree closely with approximations proposed by Odoni and Roth [22] and Lee and Roth [18], based on statistical analysis of data obtained from an algorithm to solve the state equations (the Runge-Kutta method) in order to describe transient behavior of various GI/G/1 queues with Markovian structure (e.g., $M/E_k/1$, $E_k/M/1$ and $M/H_k/1$). The interest here thus is largely in the approach, which is very different. We obtain (2.1) *theoretically* without reference to data or numerical methods. It is of course important that the results match the data. Moreover, given that the Odoni-Roth-Lee approximations fit the data, it is appropriate that our approximations are similar. Note, however, that Odoni, Roth and Lee focus on the number in queue excluding the customer in service, instead of the number in system. We believe that the number in system is more appropriate for direct approximation because then $m_1(\infty) - m_1(t)$ is log-convex; see corollary 3.3.2; otherwise, it is not. In a forthcoming paper we generate an approximation for the probability of emptiness, $P_{00}(t)$ which can be combined with (2.1) to yield an approximation for the expected number in queue, $m_1(t) - 1 + P_{00}(t)$. We also include a proof there that

the expected number in queue, excluding the customer in service, is *not* log-convex for all ρ sufficiently small. (This follows easily from corollary 4.2.3 of Abate and Whitt [3].)

2.2. SCALING SPACE AND TIME

A useful initial step is to scale space and time to identify and isolate the extreme limiting behavior as $\rho \to 1$ and as $\rho \to 0$. Obviously, $EQ(\infty) \to \infty$ as $\rho \to 1$ and $EQ(\infty) \to 0$ as $\rho \to 0$. A good way to scale space and time is in a way so that nondegenerate limits occur as $\rho \to 1$ and as $\rho \to 0$. In particular, we scale space and time so that canonical RBM (with drift coefficient -1 and diffusion coefficient 1) treated in Abate and Whitt [1] is obtained in the limit as $\rho \to 1$. From Stone [25] or Iglehart and Whitt [13], it is known that the family of stochastic processes $\{2^{-1}(1-\rho)Q(t2(1-\rho)^{-2}) : t \geq 0\}$ indexed by ρ converges in distribution to canonical RBM as $\rho \to 1$. In particular, the dominant effect of ρ in the approach to steady state when ρ is near 1 is captured by this heavy-traffic limit theorem. This effect is represented by the *time scaling* $2(1-\rho)^{-2}$ and the *space scaling* $2(1-\rho)^{-1}$. Thus, we express the relaxation time $a(\rho)$ in (2.2) as $a(\rho) = 2(1-\rho)^{-2}c(\rho)$, so that it only remains to determine $c(\rho)$. It can be seen that $c(\rho)$ is increasing in ρ with $c(0) = 0.50$ and $c(1) = 1.31$. (See table 4.) To exploit this scaling, we define the moment cdf's (cumulative distribution functions)

$$H_{\rho k}(t) = \frac{E\left(Q\left(t2(1-\rho)^{-2}\right)^k\right)}{E\left(Q(\infty)^k\right)}, \quad t \geq 0. \tag{2.4}$$

The criterion of convergence to canonical RBM as $\rho \to 1$ and the normalization by the steady-state limits lead to well-defined unique scalings of space and time in (2.4). The scaling yields nondegenerate cdf's both as $\rho \to 1$ and as $\rho \to 0$. (Both limits are established with transforms in corollary 5.2.2 of Abate and Whitt [3].) For related discussions of scaling, see sect. 3.5 of Mori [19] and sect. 5.2 of Newell [21]. Just as with RBM, $E(Q(t)^k \mid Q(0) = 0)$ is increasing in t, so that $H_{\rho k}(t)$ is a legitimate cdf for each k. (See theorem 1.2 of Abate and Whitt [1] or corollarly 3.1.1 here.)

After the normalization in (2.4), the exponential approximation (2.1) for the first moment becomes

$$1 - H_{\rho 1}(t) \approx (1 + 2c(\rho))^{-1} \, e^{-t/c(\rho)}, \quad t \geq 1, \tag{2.5}$$

for $c(\rho)$ in (2.2). (In scaled time, the term $2(1-\rho)^{-2}$ is omitted from (2.3).) The function $c(\rho)$ in (2.2) and (2.5) is the *approximate scaled relaxation time*. For $\rho = 1$, (2.5) reduces to the simple exponential approximation for canonical RBM in (1.3) of Abate and Whitt [1].

2.3. AN H_2 FIT USING THREE MOMENTS

We obtain the simple exponential approximations (2.1) and (2.5) by fitting an H_2 (hyperexponential) cdf to the first three moments of the cdf $H_{\rho 1}(t)$. The complementary cdf of an H_2 distribution has the form

$$1 - H(t) = p_1 \, e^{-t/\tau_1} + p_2 \, e^{-t/\tau_2}, \quad t \geq 0, \tag{2.6}$$

where $p_1 + p_2 = 1$ and $\tau_1 \leq \tau_2$. The second exponential component with mean τ_2 is dominant for large t since $\tau_1 < \tau_2$. This general moment-fitting approach is appropriate when we want to obtain a good fit for relatively large t, which is our goal here. The full H_2 approximating complementary cdf $1 - \tilde{H}_{\rho 1}(t)$ we obtain is

$$1 - \tilde{H}_{\rho 1}(t) = 2c(\rho)[1 + 2c(\rho)]^{-1} \, e^{-4c(\rho)t} + [1 + 2c(\rho)]^{-1} \, e^{-t/c(\rho)}. \tag{2.7}$$

(Note that $c(\rho) \geq 1/4c(\rho)$, so that the second term is dominant for large t.) This H_2 approximation tends to be pretty good for all ρ when $t \geq 0.25$. We obtain the simple exponential approximation (2.6) from (2.7) by simply ignoring the first term in (2.7). The H_2 approximation (2.7) is in general uniformly more accurate than the simplification (2.5); (2.5) is introduced only to obtain a simple easily comprehensible description, e.g., (2.3). If ρ is not too small, then (2.5) and (2.7) are nearly equivalent for $t \geq 1$ because then the first exponential term in (2.7) tends to be negligible. However, as $\rho \to 0$, $c(\rho) \to 0.50$ and $1/4c(\rho) \to 0.50$, so that the two component means coincide. Thus the reduction of the H_2 cdf to one component exponential by simply eliminating the other component will not work when ρ is too small. Then we propose the exponential approximation $1 - H_{\rho 1}(t) \approx e^{-t/c(\rho)}$. As $\rho \to 0$, the relaxation time is still described by $c(\rho)$, but the appropriate multiplier changes. As ρ decreases, the multiplier should increase from $[1 + 2c(\rho)]^{-1}$ to 1.

To carry out this procedure, of course we must be able to determine the first three moments of $H_{\rho 1}(t)$ in (2.4) and fit these moments to the standard H_2 parameters. It is significant that we not only obtain the numerical values of the H_2 cdf for each ρ, but we also obtain the moments and the H_2 parameters in closed form as explicit functions of ρ. We apply the method described in sect. 5 of Abate and Whitt [1]. The second H_2 fitting method in (5.7) there is especially useful, because for the M/M/1 model here the parameter γ there turns out to be 2, independent of ρ.

In support of the exponential approximation (2.5) and the H_2 approximation (2.7), we prove that the first-moment cdf $H_{\rho 1}(t)$ in (2.4) is actually a mixture of exponentials. (See corollary 3.3.1.) This guarantees that an H_2 fit is possible and shows that it should perform reasonably well. It is obviously significant that, in addition to having the same first three moments, the actual moment cdf $H_{\rho 1}(t)$ in (2.4) has a shape similar to the H_2 approximation in (2.7).

2.4. NUMERICAL COMPARISONS

We do numerical comparisons by inverting the Laplace transform for the scaled M/M/1 queue, as described in sect. 4.4 of Abate and Whit [1]. Numerical comparisons appear in tables 1–3. Table 1 displays the actual complementary cdf $1 - H_{\rho 1}(t)$ in (2.4); table 2 displays the approximating complementary H_2 cdf $1 - \tilde{H}_{\rho 1}(t)$ in (2.7); and table 3 displays the simple exponential approximation

Table 1
Numerical values for the first-moment complementary cdf $1 - H_{\rho 1}(t)$ defined in (2.4), obtained by Laplace transform inversion. (Time has been scaled.)

Time t	Traffic intensity				
	$\rho = 0$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.90$	$\rho = 1.00$
0.5	0.368	0.306	0.291	0.284	0.280
1.0	0.135	0.144	0.148	0.150	0.151
1.5	0.050	0.077	0.084	0.088	0.090
2.0	0.018	0.043	0.051	0.055	0.057
3.0	0.025	0.0151	0.0202	0.0230	0.0247
4.0	0.00034	0.0057	0.0087	0.0104	0.0115
7.0	0.00000	0.00032	0.00078	0.00113	0.00146

Table 2
Numerical values for the H_2 approximation $1 - \tilde{H}_{\rho 1}(t)$ in (2.7) of the first-moment complementary cdf $1 - H_{\rho 1}(t)$ in (2.4). (Time has been scaled.)

Time t	Traffic intensity					
	$\rho = 0$	$\rho = 0.05$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.90$	$\rho = 1.00$
0.5	0.368	0.366	0.292	0.265	0.250	0.241
1.0	0.135	0.141	0.135	0.134	0.133	0.132
1.5	0.050	0.058	0.076	0.083	0.087	0.088
2.0	0.018	0.024	0.045	0.054	0.058	0.060
3.0	0.0025	0.0048	0.0166	0.0226	0.0259	0.0279
4.0	0.00034	0.00099	0.0061	0.0095	0.0117	0.0130
7.0	0.00000	0.00001	0.00030	0.00072	0.00106	0.00131
H_2 Parameters	$\rho = 0$	$\rho = 0.05$	$\rho = 0.50$	$\rho = 0.75$	$\rho = 0.90$	$\rho = 1.00$
p_1	0.50	0.556	0.667	0.698	0.714	0.724
p_2	0.50	0.444	0.333	0.302	0.286	0.276
τ_1	0.50	0.400	0.250	0.216	0.200	0.191
τ_2	0.50	0.625	1.000	1.159	1.250	1.309
m_1	0.50	0.50	0.50	0.50	0.50	0.50
c^2	1.00	1.10	2.00	2.50	2.80	3.00
r	0.500	0.444	0.333	0.302	0.286	0.276

Table 3
Numerical values for the simple exponential approximation (2.5) of the first-moment complementary cdf $1 - H_{\rho 1}(t)$ in (2.4). (Time has been scaled.)

Time t	Traffic intensity					
	$\rho = 0$	$\rho = 0.05$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.9$	$\rho = 1.0$
0.5	0.184	0.200	0.202	0.196	0.192	0.188
1.0	0.068	0.090	0.123	0.127	0.129	0.129
1.5	0.025	0.040	0.074	0.083	0.086	0.088
2.0	0.0091	0.018	0.045	0.054	0.058	0.060
3.0	0.0012	0.0037	0.0166	0.0226	0.0259	0.0279
4.0	0.00017	0.00074	0.0061	0.0095	0.0117	0.0130
7.0	0.00000	0.00001	0.00030	0.00072	0.00106	0.00131

(2.5). (As we indicated above, better approximations for small ρ are possible with $e^{-t/c(\rho)}$ instead of (2.5).)

The case $\rho = 1$ in table 1 provides the exact values for RBM, as in table 1 of Abate and Whitt [1]. With the normalization in (2.4), the complementary cdf for RBM ($\rho = 1$) is a pretty good approximation for the other cases when ρ is not very small and t is not very large, e.g., for $\rho \geq 0.75$ and $t \leq 2.0$.

As indicated above, the simple exponential approximation (2.5) is essentially the same as the H_2 approximation (2.7) for $t \geq 1$, providing that ρ is not too small. In fact, both the true cdf (2.4) and the H_2 approximation in (2.7) approach a simple exponential with mean 1/2 as $\rho \to 0$, so that the H_2 approximation (2.7) is asymptotically correct as $\rho \to 0$. The values for $\rho = 0$ in table 3 are exactly one-half of the values for $\rho = 0$ in table 2 because only one of the two identical exponential components is counted in table 3. (The H_2 approximation is meaningful and nontrivial as $\rho \to 0$ because the scaling leads to a nondegenerate limit.)

In addition to the numerical values for the H_2 approximation, table 2 displays the H_2 parameters. (We determine *all* moments of $H_{\rho k}(t)$ in (2.4) for *all* k in sec. 3; see corollaries 3.1.3 and 3.2.2.) For the i^{th} component exponential, the mean is τ_i and the probability is p_i. Also given are the overall mean $m_1 = 1/2$, the squared coefficient of variation (the variance divided by the square of the mean) $c^2 = 1 + 2\rho$, and the parameter $r = p_1\tau_1/(p_1\tau_1 + p_2\tau_2)$ giving the proportion of the mean in the component with the smaller mean. In this case, r turns out to be just p_2. The parameters m_1, c^2 and r are the exact values for $H_{\rho 1}(t)$ determined by the first three moments; see sect. 5 of Abate and Whitt [1].

From the numerical values in tables 1–3, we see that the simple exponential approximation (2.5) when ρ is not too small and the H_2 approximation (2.7) more generally provide satisfactory descriptions of the approach to steady state for times of primary interest, in particular, for $t \geq 1$ or $H_{\rho 1}(t) \geq 0.85$. However,

Table 4
A comparison of the scaled relaxation time $c(\rho)$ in (2.2) with other candidate expressions.

Traffic intensity ρ	3-moment H_2 fit $c(\rho)$ in (2.2)	Asymptotic relaxation time Cohen [5] $\frac{(1+\sqrt{\rho})^2}{2}$	Odoni and Roth [22] $\frac{(1+\sqrt{\rho})^2}{2.8}$	Newell [21] $\frac{1+\rho}{2}$	Gaver and Jacobs [11] $\sqrt{2(1+\rho)}$	Proposed simplification of $c(\rho)$ $\frac{(1+\sqrt{\rho})^2}{(2+\rho^{1/4})}$
0.0	0.50	0.50	0.36	0.50	1.41	0.50
0.1	0.69	0.87	0.62	0.55	1.48	0.68
0.2	0.78	1.05	0.75	0.60	1.55	0.79
0.3	0.86	1.20	0.86	0.65	1.61	0.88
0.4	0.93	1.33	0.95	0.70	1.67	0.95
0.5	1.00	1.46	1.04	0.75	1.73	1.03
0.6	1.07	1.57	1.12	0.80	1.79	1.09
0.7	1.13	1.69	1.20	0.85	1.84	1.16
0.8	1.19	1.79	1.28	0.90	1.90	1.22
0.9	1.25	1.90	1.36	0.95	1.95	1.28
1.0	1.31	2.00	1.43	1.00	2.00	1.33

the approximate scaled relaxation time $c(\rho)$ in (2.2) has the drawback that it is a rather complex expression. It would be nice if $c(\rho)$ could be replaced by a simple function of ρ related to the relaxation time arising in the asymptotic theory, as on p. 180 of Cohen [5]. Since the unscaled relaxation time in Cohen (1982) is $(1-\sqrt{\rho})^{-2}$, the associated scaled relaxation time is $(1+\sqrt{\rho})^2/2$. (Recall that $(1-\sqrt{\rho})^2(1+\sqrt{\rho})^2 = (1-\rho)^2$.) Similarly, the approximate scaled relaxation time of Odoni and Roth [22] for the expected number in queue is $(1+\sqrt{\rho})^2/2.8$. As a simple approximation of this kind, we propose $c(\rho) \approx (1+\sqrt{\rho})^2/(2+\rho^{1/4})$ or, for ρ not too small, $(1+\sqrt{\rho})^2/3$. (The Odoni and Roth [22] value is apparently somewhat larger because they are aiming for an approximate upper bound. They are also treating a different process.) The various candidates are compared in table 4. One additional candidate $(1+\rho)/2$ comes form (5.6) on p. 151 of Newell [21]. Another $\sqrt{2(1+\rho)}$ comes from (3.10) of Gaver and Jacobs [11]. The value ρ itself is suggested in (1) of Morse [20]. Note that the differences here only concern the factor $c(\rho)$. All methods agree on the dominant factor $(1-\rho)^{-2}$ used in the scaling (2.4). In other words, when we focus on $c(\rho)$, we are looking for a second-order refinement.

As long as ρ is not too small, we conclude, in agreement with Odoni and Roth [22], that the *real* relaxation time for times of primary interest is about 2/3 of the relaxation time from the asymptotic theory. Equivalently, in the regions of primary interest the first moment approaches steady state at a rate about 1.5 times faster than predicted by the inverse of the asymptotic relaxation time. We

Table 5
A comparison of approximations of the complementary cdf $1 - H_{\rho 1}(t)$ in (2.4) for the case $\rho = 0.8$.

Time t	Exact (2.4)	Hyperexponentials			Asymptotic as $t \to \infty$ (2.8)
		$\gamma = -8$	$\gamma = 2$ (2.7)	$\gamma = 0.67$	
0.00	1.000	1.000	1.000	1.000	
0.25	0.444	0.421	0.454	0.528	
0.50	0.289	0.319	0.260	0.292	
1.00	0.149	0.183	0.134	0.110	0.87
2.00	0.052	0.060	0.055	0.037	0.176
3.00	0.0212	0.0198	0.0238	0.0198	0.055
4.00	0.0092	0.0065	0.0103	0.0114	0.0204
5.00	0.0042	0.0021	0.0044	0.0066	0.0084
6.00	0.0019	0.0007	0.0019	0.0038	0.0036
7.00	0.0009	0.0002	0.0008	0.0022	0.0017

have related our approximate scaled relaxation time $c(\rho)$ to the asymptotic approximation $\tau \equiv \tau(\rho) \equiv (1 + \sqrt{\rho})^2/2$ in Cohen [5]. For the scaled M/M/1 system, the full limit as $t \to \infty$ is

$$1 - H_{\rho 1}(t) \sim \tau^2 \left(2\pi\rho^{3/2} t^2\right)^{-1/2} e^{-t/\tau} \tag{2.8}$$

where $f(t) \sim g(t)$ means $f(t)/g(t) \to 1$ as $t \to \infty$. As $\rho \to 1$, this limit approaches the limit for RBM in the corollary 1.1.2(a) of Abate and Whitt [1]. Numerical comparisons for the case $\rho = 1$ are contained in table 3 there. Numerical comparisons for the case $\rho = 0.8$ are contained in table 5 here. As before, the H_2 approximation (2.7) and the exponential simplification (2.5) provide an order-of-magnitude improvement over (2.8) in the quality of the approximation (Also included in table 5 is a comparison with other H_2 approximations having the same first two moments but different parameter γ from (5.7) of Abate and Whitt [1]. This shows that the three-moment fit is important. In fact, using two-moment H_2 fits, the approximate relaxation time (dominant exponential mean) can assume any value greater than 1 when $\rho = 1$ by choosing an appropriate third moment.

3. Factorial-moment CDFs

In this section we present the probabilistic characterization of the normalized factorial-moment function $H_r(t)$ in (1.2) and describe some consequences. (Note that in this section we are not using the scaling (2.4).) Our results are discrete analogs of corresponding results for RBM in Abate and Whitt [1].

REMARK 3.1

Related results can be obtained for the moments of the work in system at time t, say $W(t)$, using the relation $W(t) = \sum_{i=1}^{Q(t)} v_i$ where v_i are the exponential service times; e.g.,

$$EW(t) = EQ(t) \text{ and } E\left[W(t)^2\right] = E\left[Q(t)^2\right] + EQ(t).$$

For the M/M/1 queue, the pmf (probability mass function) of the steady-state queue length $Q(\infty)$ is *geometric*. The r-fold convolution of this geometric distribution is *negative binomial*. It has density

$$g_r(k) = \binom{r+k-1}{k}(1-\rho)^r\rho^k, \quad k = 0, 1, 2, \ldots \tag{3.1}$$

In a sequence of Bernoulli trials with probability $1-\rho$ of success, the probability that the r^{th} failure occurs at trial $k+r+1$ is $g_r(k)$; see pp. 164, 268 of Feller [9].

As is commonly done with discrete distributions, we work with *factorial moments*. Let $\phi_r(X) \equiv \phi_r(p)$ denote the r^{th} factorial moment of a random variable X with pmf p on the nonnegative integers, defined by

$$\phi_r(X) = E\left[X(X-1)\cdots(X-r+1)\right] = \sum_{k=r}^{\infty} \frac{k!}{(k-r)!} p_k. \tag{3.2}$$

Factorial moments are obtained directly by differentiating the probability generating function

$$\psi_X(s) \equiv E(s^X) = \sum_{k=0}^{\infty} s^k P(X=k).$$

In particular, if $\psi_x^{(r)}(s)$ denotes the r^{th} derivative of $\psi_x(s)$, then $\psi_x^{(r)}(1) = \phi_r(X)$.

We also work with the *tail probability function* associated with a pmf p, defined by

$$\bar{p}_k = p_{k+1} + p_{k+2} + \cdots, \quad k = 0, 1, 2, \ldots \tag{3.3}$$

Since the probability generating functions

$$\psi_{\bar{p}}(s) = \sum_{k=0}^{\infty} s^k \bar{p}_k \text{ and } \psi_p(s) = \sum_{k=0}^{\infty} s^k p_k$$

are related by $\psi_{\bar{p}}(s) = [1 - \psi_p(s)]/(1-s)$, the factorial moments of p and $\bar{p}$ are related by

$$\sum_{k=r}^{\infty} \frac{k!}{(k-r)!} p_k = r \sum_{k=r-1}^{\infty} \frac{k!}{(k-r+1)!} \bar{p}_k, \tag{3.4}$$

e.g.,

$$\sum_{k=1}^{\infty} k p_k = \sum_{k=0}^{\infty} \bar{p}_k \text{ and } \sum_{k=2}^{\infty} k(k-1) p_k = 2 \sum_{k=1}^{\infty} k \bar{p}_k;$$

see p. 265 of Feller [9].

It is also convenient that the factorial moments of the geometric pmf have a very simple form, namely,

$$\phi_r(Q(\infty)) = \sum_{k=r}^{\infty} \frac{k!}{(k-r)!}(1-\rho)\rho^k = r!\rho^r(1-\rho)^{-r}; \tag{3.5}$$

see (14) on p. 126 of Johnson and Kotz [14].

Let T_i be the first passage time to 0 from state i in the M/M/1 queue, which has the distribution of the i-fold convolution of the busy period T_1.

THEOREM 3.1

For each positive integer r, the normalized r^{th} factorial moment of $(Q(t) \mid Q(0) = 0)$ in (1.2) can be represented as a negative-binomial mixture of first-passage-time cdf's; in particular,

$$H_r(t) \equiv \frac{\phi_r(Q(t) \mid Q(0) = 0)}{\phi_r(Q(\infty))} = \sum_{k=0}^{\infty} g_r(k) P(T_{k+r} \le t) \tag{3.6}$$

where $g_r(k)$ is the negative binomial pmf in (3.1), $\phi_r(Q(\infty))$ is the r^{th} factorial moment of the geometric distribution in (3.5) and T_k is the first passage time from k to 0, which is distributed as the k-fold convolution of the distribution of the M/M/1 busy period T_1.

Proof

By (3.4),

$$\phi_r(Q(t) \mid Q(0) = j) = r \sum_{k=r-1}^{\infty} \frac{k!}{(k-r+1)!} P(Q(t) > k \mid Q(0) = j) \tag{3.7}$$

for all positive integers r and j. Under the special case $Q(0) = 0$, $Q(t)$ is equal in distribution to $M(t)$, the maximum of the unrestricted birth-and-death process on the integers with birth rates $\lambda_n = \lambda$ and death rates $\mu_n = \mu$ for all n, say $X(t)$; i.e., for each t,

$$Q(t) \stackrel{\text{d}}{=} M(t) = \max\{ X(s) : 0 \le s \le t \}, \tag{3.8}$$

where $\stackrel{\text{d}}{=}$ means equal in distribution; p. 11 of Prabhu [23]. Let T_{ij} be the first-passage time to j from i in the unrestricted process $X(t)$. By the familiar inverse relation between $M(t)$ and T_{0j}, $P(M(t) \ge j) = P(T_{0j} \le t)$ for all nonnegative t and j. Finally, by the reversibility of birth-and-death processes,

$$P(T_{0j} \le t) = \rho^j P(T_{j0} \le t) \tag{3.9}$$

for all nonnegative j and t, as shown in theorem 1.4 of Abate and Whitt [1], following Doney [8]. (Note that the first-passage time T_{j0} for the unrestricted

process is identical to the first passage time T_j defined previously for the restricted process.) Combining these relations, we obtain

$$\phi_r(Q(t) \mid Q(0)=0) = r \sum_{k=r-1}^{\infty} \frac{k!}{(k-r+1)!} P(M(t) \geq k+1)$$

$$= r \sum_{k=r-1}^{\infty} \frac{k!}{(k-r+1)!} P(T_{0,k+1} \leq t)$$

$$= r \sum_{k=r-1}^{\infty} \frac{k!}{(k-r+1)!} \rho^{k+1} P(T_{k+1,0} \leq t)$$

so that, by (3.5),

$$H_r(t) = \frac{\phi_r(Q(t) \mid Q(0)=0)}{\phi_r(Q(\infty))} = \sum_{k=r-1}^{\infty} \binom{k}{r-1} (1-\rho)^r \rho^{k-r+1} P(T_{k+1,0} \leq t)$$

$$= \sum_{n=0}^{\infty} \binom{n+r-1}{r-1} (1-\rho)^r \rho^n P(T_{n+r,0} \leq t)$$

with the last step based on a change of variables $n = k-r+1$. □

COROLLARY 3.1.1

For each r, $\phi_r(Q(t) \mid Q(0)=0)$ is nondecreasing in t.

The ordinary r^{th} moment $E[Q(t)^r]$ can be expressed as a linear combination of the first r factorial moments with nonnegative coefficients, namely, the Stirling numbers of the second kind; p. 4 of Johnson and Kotz [14]. Consequently, we have an analog of corollary 3.1.1 for the ordinary moments. (This is also a consequence of stochastic order results in van Doorn [28].)

COROLLARY 3.1.2

For each r, $E[Q(t)^r]$ is nondecreasing in t.

Theorem 3.1 takes a particularly simple form when $r=1$. Since the first factorial moment is just the first moment, we also refer to $H_1(t)$ as the first-moment cdf. Let $B_e(t)$ be the equilibrium-excess or stationary-residual-life cdf associated with the busy period cdf $B(t) = P(T_1 \leq t)$, defined as usual by

$$B_e(t) = \int_0^t [1-B(u)]\,du \Big/ \int_0^{\infty} [1-B(u)]\,du = (1-\rho)\int_0^t [1-B(u)]\,du; \quad (3.10)$$

(p. 28 of Cox [6]).

COROLLARY 3.1.3

The first-moment cdf $H_1(t)$ in (1.2) coincides with the cdf of the equilibrium time to emptiness conditional on not starting empty, i.e., the first passage time to 0 starting in steady state conditional on a positive starting state, which in turn coincides with the busy-period equilibrium-excess cdf, i.e.,

$$H_1(t) = \sum_{k=1}^{\infty} P(Q(\infty) = k \mid Q(\infty) > 0) P(T_k \le t) = B_e(t), \quad t \ge 0.$$

As a consequence of corollary 3.1.3, we can express the moments of $H_1(t)$ in terms of the moments of the busy-period cdf $B(t)$, e.g., p. 149 of Cox and Smith [7]. Let $m_n(G)$ be the n^{th} moment of a cdf G. We use the fact that for any cdf G on $[0, \infty)$

$$m_n(G_e) = m_{n+1}(G)/m_1(G)(n+1); \tag{3.11}$$

see p. 64 of Cox [6]. The n^{th} moment of $B(t)$ for any n is easily obtained from Riordan's recursion in theorem 3.2 of Abate and Whitt [3].

COROLLARY 3.1.4

For the M/M/1 queue, $m_n(H_1(t)) = (1-\rho)m_{n+1}(B(t))/(n+1)$.

We can also relate the higher factorial-moment cdf's $H_r(t)$ to the first-moment cdf $H_1(t)$ is a simple way.

THEOREM 3.2

For each r, the r^{th} factorial-moment cdf $H_r(t)$ in (3.6) is the r-fold convolution of $H_1(t)$.

Proof

Since $g_r(k)$ is the r-fold convolution of $g_1(k)$ and $P(T_{k+r} \le t)$ is the convolution of $P(T_k \le t)$ and $P(T_r \le t)$, the convolution of $H_r(t)$ and $H_1(t)$ is

$$\begin{aligned}
\int_0^t H_r(t-s)\, \mathrm{d}H_1(s) &= \int_0^t \sum_{k=0}^{\infty} g_r(k) P(T_{k+r} \le t-s) \sum_{j=0}^{\infty} g_1(j)\, \mathrm{d}P(T_{1+j} \le s) \\
&= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} g_r(k) g_1(j) \int_0^t P(T_{k+r} \le t-s)\, \mathrm{d}P(T_{1+j} \le s) \\
&= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} g_r(k) g_1(j) P(T_{k+j+r+1} \le t) \\
&= \sum_{m=0}^{\infty} \sum_{j=0}^{m} g_r(m-j) g_1(j) P(T_{m+r+1} \le t) \\
&= \sum_{m=0}^{\infty} g_{r+1}(m) P(T_{m+r+1} \le t) = H_{r+1}(t). \quad \square
\end{aligned}$$

The representation in theorem 3.2 is convenient for comparing the approach to steady state of the factorial moments $\phi_r(Q(t) \mid Q(0) = 0)$ for different r. Intuitively, we would expect that higher factorial moments approach steady state more slowly than lower factorial moments. Theorem 3.2 makes this property easy to express and establish. We use the notion of stochastic order; Stoyan [26].

COROLLARY 3.2.1

For all $r \geq 1$ and $t > 0$, $H_r(t) \geq H_{r+1}(t)$; i.e., the r^{th} factorial-moment cdf's are stochastically increasing in r.

Theorems 3.1 and 3.2 also facilitate calculating the moments of the factorial-moment cdf's.

COROLLARY 3.2.2

The factorial-moment cdf moments are

$$m_{rj} \equiv m_j(H_r(t)) = \int_0^\infty t^j \, \mathrm{d}H_r(t) = \sum_{k=0}^{\infty} g_r(k) E\left(T_{k+r}^j\right); \tag{3.12}$$

e.g.,

$$m_{r1} = \frac{r}{(1-\rho)^2}, \qquad m_{r2} = \frac{r^2 + r(2\rho+1)}{(1-\rho)^4},$$
$$m_{r3} = \frac{r\left[6\rho^2 + 12\rho + 2\right] + 3r^2(2\rho+1) + r^3}{(1-\rho)^6}. \tag{3.13}$$

Proof

Note that

$$m_{r1} = \sum_{k=0}^{\infty} \binom{k+r-1}{k} (1-\rho)^r \rho^k \left[(k+r)/(1-\rho)\right]$$
$$= \left[1/(1-\rho)\right]\left(r\rho/(1-\rho) + r\right) = r/(1-\rho)^2.$$

For $j = 2$, start with $r = 1$. By theorem 3.1,

$$m_{12} = \sum_{k=0}^{\infty} g_1(k) E\left(T_{k+1}^2\right) = \sum_{k=0}^{\infty} (1-\rho)\rho^k \left(\mathrm{Var}(T_{k+1}) + \left[E(T_{k+1})\right]^2\right)$$
$$= \sum_{k=0}^{\infty} (1-\rho)\rho^k \left([k+1]E\left(T_1^2\right) + k(k+1)\left[E(T_1)\right]^2\right)$$
$$= \frac{E\left(T_1^2\right)}{1-\rho} + \frac{2\rho\left[E(T_1)\right]^2}{(1-\rho)^2} = \frac{2}{(1-\rho)^4} + \frac{2\rho}{(1-\rho)^4} = \frac{2(1+\rho)}{(1-\rho)^4},$$

so that

$$m_{12} = m_{11}^2 = \frac{2(1+\rho)}{(1-\rho)^4} - \frac{1}{(1-\rho)^4} = \frac{2\rho+1}{(1-\rho)^4}.$$

By theorem 3.2,

$$m_{r2} - m_{r1}^2 = \frac{r(2\rho+1)}{(1-\rho)^4},$$

so that

$$m_{r2} = \frac{r(2\rho+1)}{(1-\rho)^4} + \frac{r^2}{(1-\rho)^4} = \frac{r^2+r(2\rho+1)}{(1-\rho)^4}.$$

For $j \geq 3$, we work with the cumulants (or semi-invariants) because the j^{th} cumulant of a k-fold convolution of a distribution is just k times the j^{th} cumulant of the distribution; p. 20 of Johnson and Kotz [14]. Let $\beta_j(T)$ be the j^{th} cumulant of T and let β_{rj} be the j^{th} cumulant of the cdf $H_r(t)$ with moments m_{rj}. We apply known relations connecting the moments to the cumulants. (The second and third cumulants are just the second and third central moments [about the mean].) By theorem 3.1,

$$\begin{aligned}
m_{13} &= \sum_{k=0}^{\infty} g_1(k) E\left(T_{k+1}^3\right) \\
&= \sum_{k=0}^{\infty} (1-\rho)\rho^k \left(\beta_3(T_{k+1}) + 3\beta_1(T_{k+1})\beta_2(T_{k+1}) + \left[\beta_1(T_{k+1})\right]^3\right) \\
&= \sum_{k=0}^{\infty} (1-\rho)\rho^k \left([k+1]\beta_3(T_1) + 3(k+1)^2\beta_1(T_1)\beta_2(T_1)\right. \\
&\quad \left. + (k+1)^3\left[\beta_1(T_1)\right]^3\right) \\
&= \frac{\beta_3(T_1)}{1-\rho} + \frac{3(\rho+1)}{(1-\rho)^2}\beta_1(T_1)\beta_2(T_1) + \frac{(\rho^2+4\rho+1)}{(1-\rho)^3}\left[\beta_1(T_1)\right]^3 \\
&= \frac{E\left(T_1^3\right) - 3E(T_1)E\left(T_1^2\right) + 2\left[E(T_1)\right]^3}{1-\rho} \\
&\quad + \frac{(3\rho+3)E(T_1)\left(E\left(T_1^2\right) - \left[E(T_1)\right]^2\right)}{(1-\rho)^2} + \frac{(\rho^2+4\rho+1)}{(1-\rho)^3}\left[E(T_1)\right]^3 \\
&= \frac{6+6\rho-6(1-\rho)+2(1-\rho)^2}{(1-\rho)^6} + \frac{(3\rho+3)(2-(1-\rho))}{(1-\rho)^6} + \frac{\rho^2+4\rho+1}{(1-\rho)^6} \\
&= \frac{6\left(\rho^2+3\rho+1\right)}{(1-\rho)^6},
\end{aligned}$$

so that

$$\beta_{13} = m_{13} - 3m_{11}m_{12} + 2m_{11}^3$$

$$= \frac{6\rho^2 + 18\rho + 6}{(1-\rho)^6} - \frac{3(1 + (2\rho + 1))}{(1-\rho)^6} + \frac{2}{(1-\rho)^6}$$

$$= \frac{6\rho^2 + 12\rho + 2}{(1-\rho)^6}.$$

Hence,

$$\beta_{r3} = \frac{r[6\rho^2 + 12\rho + 2]}{(1-\rho)^6}$$

and

$$m_{r3} = \beta_{r3} + 3\beta_{r1}\beta_{r2} + \beta_{r1}^3$$

$$= \frac{r[6\rho^2 + 12\rho + 2]}{(1-\rho)^6} + \frac{3r}{(1-\rho)^2}\left(\frac{r^2 + r(2\rho + 1) - r^2}{(1-\rho)^4}\right) + \frac{r^3}{(1-\rho)^6}$$

$$= \frac{r[6\rho^2 + 12\rho + 2] + 3r^2(2\rho + 1) + r^3}{(1-\rho)^6}. \quad \square$$

From corollary 3.2.2 we obtain the first three moments of $H_1(t)$ in (1.2) by setting $r = 1$. Since $c_1^2 \equiv (m_{12} - m_{11}^2)/m_{11}^2 = 1 + 2\rho \geq 1$ and $m_{13}m_{11}/m_{12}^2 = (3/2)(1 + [\rho/(1+\rho)]) \geq 3/2$, it is always possible to fit an H_2 distribution to the first three moments of $H_1(t)$; see sect. 5 of Abate and Whitt [1]. (Since $c_2^2 \equiv (m_{22} - m_{21}^2)/m_{21}^2 = \rho + 1/2$ by corollary 3.2.2, this is *not* the case for the second-factorial-moment cdf $H_2(t)$ for all ρ.)

We can also use corollary 3.2.2 to calculate the moments of the time-scaled ordinary-moment cdf's $H_{\rho k}(t)$ in (2.4). For example it is easy to see that

$$H_{\rho 2}(t) = \alpha H_1\left(t2(1-\rho)^{-2}\right) + (1-\alpha)H_2\left(t2(1-\rho)^{-2}\right) \tag{3.14}$$

where $\alpha = EQ(\infty)/E[Q(\infty)^2] = (1-\rho)/(1+\rho)$. As a consequence of (3.14), note that $H_{\rho 2}(t)$ coincides with $H_{\rho 1}(t)$ when $\rho = 0$ and $H_{\rho 2}(t)$ coincides with $H_2(t2(1-\rho)^{-2})$ when $\rho = 1$.

COROLLARY 3.2.3

The k^{th} moment $m_{\rho 2k}$ of $H_{\rho 2}(t)$ in (2.4) is $2^{-k}(1-\rho)^{2k}[\alpha m_{1k} + (1-\alpha)m_{2k}]$ for $\alpha = (1-\rho/(1+\rho)$; e.g., the first three moments are

$$m_{\rho 21} = \frac{(1+3\rho)}{2(1+\rho)}, \quad m_{\rho 22} = \frac{(1+6\rho+3\rho^2)}{2(1+\rho)}, \quad m_{\rho 23} = \frac{3(1+10\rho+14\rho^2+3\rho^3)}{4(1+\rho)}.$$

From corollary 3.2.3, we see that unlike $H_2(t)$, $H_{\rho 2}(t)$ admits a hyperexponential match to its first three moments, because

$$c_{\rho 2}^2 \equiv \frac{m_{\rho 22} - m_{\rho 21}^2}{m_{\rho 21}^2} = \frac{(1+8\rho+9\rho^2+6\rho^3)}{(1+6\rho+9\rho^2)} \geq 1 \tag{3.15}$$

and

$$\frac{m_{\rho 23} m_{\rho 21}}{m_{\rho 22}^2} = \left(\frac{3}{2}\right)\frac{(1+13\rho+44\rho^2+45\rho^3+9\rho^4)}{(1+12\rho+42\rho^2+36\rho^3+9\rho^4)} \geq \frac{3}{2}. \tag{3.16}$$

For $\rho = 1$, $c_{\rho 2}^2 = 3/2$ and $(m_{\rho 23} m_{\rho 21}/m_{\rho 22}^2) = 42/25$ in agreement with previous results for RBM in corollary 1.3.4 of Abate and Whitt [1]. Similarly, the moments $m_{\rho 2k}$ match the RBM results after introducing the space scaling. For $\rho = 0$, the moments match a simple exponential. Corollary 5.2.7 of Abate and Whitt [3] establishes that $H_{\rho 2}(t)$ in fact converges in distribution to this exponential as $\rho \to 0$.

From theorem 1.7 of Abate and Whitt [1], we know that $H_{\rho k}(t)$ is a mixture of exponentials, i.e., has a completely monotone density $h_{\rho k}(t)$, for $k = 1$ and 2 (but not $k = 3$) when $\rho = 1$. (Recall that a density $h(t)$ is *completely monotone* if derivatives $h^{(n)}(t)$ exist and $(-1)h^{(n)}(t) \geq 0$ for all n and t; see p. 66 of Keilson [15]. This is equivalent to $h(t)$ being a mixture of exponential densities). Such complete monotonicity clearly also holds when $\rho = 0$. We now investigate the remaining cases $0 < \rho < 1$. From (3.15) and (3.16), we expect similar positive results for all ρ, but surprisingly a negative result for the case $k = 2$ is provided by the following corollary.

COROLLARY 3.2.4

As $t \to 0$,

$$H_{\rho 2}(t) \sim \frac{2t}{1+\rho} - \frac{2(1-2\rho)t^2}{(1+\rho)(1-\rho)^2}$$

and

$$h_{\rho 2}(t) \sim \frac{2}{1+\rho} - \frac{4(1-2\rho)t}{(1+\rho)(1-\rho)^2},$$

so that derivative of the density satisfies $h'_{\rho 2}(0) > 0$ for $1/2 < \rho < 1$.

Proof

Let $\hat{H}_{\rho 2}(s)$, $\hat{H}^c_k(s)$ and $\hat{h}_k(s)$ be the Laplace transforms of the time-scaled complementary cdf's $1 - H_{\rho 2}(t)$ and $1 - H_2(t2(1-\rho)^{-2})$ and the associated density, respectively. From (3.14),

$$\hat{H}^c_{\rho 2}(s) = \frac{\theta}{\omega}\hat{H}^c_1(s) + \frac{\rho}{s\omega}\left[1 - \left(\hat{h}_1(s)\right)^2\right]$$

where $\theta = (1-\rho)/2$ and $\omega = (1+\rho)/2$. Applying corollary 3.1.4, we can expand these transforms in powers of s for s near 0 to obtain

$$\hat{H}^c_1(s) = \frac{1}{s} - \frac{1}{\theta s^2} + \frac{1}{2\theta^3 s^3} + o(s^3)$$

and

$$\hat{h}_1(s) = \frac{1}{\theta s} - \frac{1}{2\theta^3 s^2} + o(s^2).$$

Thus

$$\hat{H}^c_{\rho 2}(s) = \frac{1}{s} - \frac{1}{\omega s^2} + \frac{1-2\rho}{2\theta^2 \omega s^3} + o(s^3)$$

from which we obtain our first asymptotic relation by inverting. □

We have shown that $h_{\rho 2}(t)$ is *not* completely monotone for $1/2 < \rho < 1$. We have not determined what happens for $0 < \rho < 1/2$.

CONJECTURE 3.1

$h_{\rho 2}(t)$ is completely monotone for $0 \le \rho \le 1/2$.

We now prove that $H_1(t)$ in (1.2) is actually a mixture of exponentials. We first establish this property for the M/M/1 busy-period distribution. This is in fact a general result for first passage times to neighboring states in birth-and-death processes; see pp. 40, 67 of Keilson [15]; we give a direct proof.

THEOREM 3.3

The M/M/1 busy-period density $b(t)$ is completely monotone.

Proof

Use the representation of $b(t)$ in terms of a gamma density, (36) or p. 116 of Takács [27] or (4) of Heyman [12], giving

$$b(t) = \sum_{n=1}^{\infty} (n!)^{-1} e^{-\rho t}(\rho t)^{n-1} g_n(t; 1), \quad t \geq 0, \tag{3.17}$$

where

$$g_n(t; \alpha) = \alpha^n t^{n-1} e^{-\alpha t}/(n-1)!, \quad t \geq 0,$$

which is equivalent to

$$b(t) = \sum_{n=1}^{\infty} a_n g_n(t; 1+\rho) \tag{3.18}$$

for $a_n = \rho^{n-1}/(1+\rho)^n n! > 0$. Differentiating with respect to t in (3.18) yields

$$b'(t) = \sum_{n=1}^{\infty} a_n' g_n(t; 1+\rho)$$

where $a_n' = [-(1+\rho) + \rho/(n+1)]a_n$. Hence, by induction, the n^{th} derivative satisfies

$$b^{(n)}(t) = \sum_{n=1}^{\infty} [-(1+\rho) + \rho/(n+1)]^n a_n g_n(t; 1+\rho).$$

Since $-(1+\rho) + \rho/(n+1) < 0$, $(-1)^n b^{(n)}(t) > 0$ for all n and t. □

We apply corollary 3.1.3 to obtain the following desired conclusion from theorem 3.3.

COROLLARY 3.3.1

The first-moment cdf $H_1(t)$ in (1.2), which coincides with the M/M/1 busy-period stationary-excess cdf $B_e(t)$ in (3.10), has a completely monotone density.

Proof

Note that $B_e^{(n+2)}(t) = -(1-\rho)b^{(n)}(t)$, where the subscript indicates the order of the derivative. □

The complete monotonicity implies other important structure.

COROLLARY 3.3.2

The complementary cdf $1 - H_1(t)$ and the density $h_1(t)$ are log-convex.

4. Approximations for higher moments

Let $\tilde{H}_r(t)$ and $\tilde{H}_{\rho r}(t)$ denote approximations for the r^{th} factorial moment cdf $H_r(t)$ in (1.2) and the r^{th} moment cdf $H_{\rho r}(t)$ in (2.4), respectively. Recall that $H_r(t) = H_{\rho r}(t)$, $t \geq 0$, for all r when $\rho = 1$ and for all ρ when $r = 1$, but not otherwise.

Since $H_r(t)$ is the r-fold convolution of $H_1(t)$ by theorem 3.2, we can use the r-fold convolution of $\tilde{H}_{\rho 1}(t)$ in (2.7), say $\tilde{H}_r(t)$, to approximate $H_r(t)$ for $r \geq 2$. Since $\tilde{H}_{\rho 1}(t)$ has the same first three moments as $H_1(t)$, obviously $\tilde{H}_r(t)$ has the same first three moments as $H_r(t)$. Moreover, since $\tilde{H}_{\rho 1}(t)$ coincides with $H_1(t)$ for $\rho = 0$, $\tilde{H}_{\rho r}(t)$ also coincides with $H_r(t)$ for $\rho = 0$. Since $H_1(t)$ is exponential when $\rho = 0$, $H_r(t)$ is Erlang (E_r), the r-fold convolution of an exponential, when $\rho = 0$.

In the case $r = 2$, we easily obtain the approximating convolution cdf $\tilde{H}_2(t)$ from (2.7). We then apply (3.14) to convert it into an approximation $\tilde{H}_{\rho 2}(t)$ for the ordinary second-moment function:

$$1 - \tilde{H}_{\rho 2}(t) = [A(\rho) + B(\rho)t]\ e^{-4c(\rho)t} + [1 - A(\rho) + D(\rho)t]\ e^{-t/c(\rho)} \tag{4.1}$$

where

$$A(\rho) = \frac{16c(\rho)^4 - 4c(\rho)^2 - 4c(\rho)}{16c(\rho)^4 + 16c(\rho)^3 - 4c(\rho) - 1} \tag{4.2}$$

$$B(\rho) = 16c(\rho)^3 \text{ and } D(\rho) = \left[c(\rho)\left(1 + 2c(\rho)^2\right)\right]^{-1}. \tag{4.3}$$

Numerical comparisons with exact values of $H_{\rho 2}(t)$ for the cases $\rho = 0.5$ and $\rho = 1.0$ are contained in tables 6 and 7. For $\rho < 1$, the exact values come from Laplace transform inversion as in sect. 2.4. For $\rho = 1$, the exact values come from theorem 1.1 and table 5 of Abate and Whitt [1]. For t of primary interest, e.g., $2 \leq t \leq 9$ where $0.001 \leq 1 - H_2(t) \leq 0.15$, approximation (4.1) performs well. In fact, paralleling (2.1) and (2.5), the second component $[1 - A(\rho) + D(\rho)t]\ e^{-t/c(\rho)}$ alone performs well in this region. As expected, for small t approximation (4.1) does not perform well.

Because of its superior performance for small t and its more elementary form, the direct hyperexponential approximation for $H_2(t)$ when $\rho = 1$ in Abate and Whitt [1] seems clearly preferred for RBM. However, for small ρ even a two-moment hyperexponential approximation is not available for the second-factorial-moment cdf $H_2(t)$ because as noted after corollary 3.2.2 $c^2 = \rho + 1/2$, which can be less than one.

However, as indicated after corollary 3.2.3, it is possible to fit a hyperexponential (H_2) to the first three moments of the ordinary second-moment cdf $H_{\rho 2}(t)$ in

Table 6
A comparison of approximations for the complementary second-moment cdf $1 - H_{\rho 2}(t)$ in (2.4) in the case $\rho = 0.50$ with exact values obtained from Laplace transform inversion.

Time t	Exact by numerical transform inversion	Convolution of $\tilde{H}_{\rho 1}(t)$		Direct 3-moment fit hyperexponential	
		Two terms	One term	Two terms	One term
0.01	0.987	0.990	–	0.986	–
0.1	0.871	0.893	–	0.871	–
0.5	0.512	0.511	0.374	0.515	0.318
1.0	0.284	0.270	0.241	0.282	0.206
1.5	0.165	0.160	0.154	0.163	0.133
2.0	0.099	0.100	0.098	0.098	0.086
3.0	0.038	0.040	0.040	0.038	0.036
4.0	0.015	0.016	0.016	0.015	0.015
5.0	0.0062	0.0063	0.0063	0.0064	0.0064
6.0	0.0027	0.0025	0.0025	0.0027	0.0027
7.0	0.0011	0.0010	0.0010	0.0011	0.0011
9.0	0.00022	0.00015	0.00015	0.00020	0.00020

(2.4). By corollary 5.2.7 of Abate and Whitt [3], the cdf $H_{02}(t)$ for $\rho = 0$ is exponential, so that the H_2 approximation for $H_{\rho 2}(t)$ is also exact for $\rho = 0$. Thus we have two candidate approximations for the second-moment cdf $H_{\rho 2}(t)$, which

Table 7
A comparison of approximations for the complementary second-moment cdf $1 - H_{\rho 2}(t)$ in (2.4) in the case $\rho = 1.0$ with exact values based on theorem 1.1(b) of Abate and Whitt [1].

Time t	Exact from table 5 of Abate and Whitt [1]	Convolution of $\tilde{H}_{\rho 1}(t)$		Direct 3-moment fit hyperexponential	
		Two terms	One term	Two terms	One term
0.01	0.982	0.9992	0.5406	0.980	0.497
0.1	0.858	0.941	0.509	0.819	0.468
0.5	0.540	0.525	0.392	0.542	0.358
1.0	0.333	0.298	0.281	0.324	0.257
1.5	0.216	0.203	0.201	0.209	0.184
2.0	0.144	0.144	0.144	0.141	0.132
3.0	0.068	0.073	0.073	0.069	0.068
4.0	0.033	0.037	0.037	0.035	0.035
5.0	0.016	0.018	0.018	0.018	0.018
6.0	0.0090	0.0091	0.0091	0.0092	0.0092
7.0	0.0045	0.0045	0.0045	0.0047	0.0047
9.0	0.0011	0.0011	0.0011	0.0012	0.0012

agree at the end points $\rho = 0$ (exact) and $\rho = 1$. For the case $\rho = 1/2$, both approximations are given in table 6. From (5.7) of Abate and Whitt [1], the direct H_2 fit has parameters $d^3 = 8.532$, $\gamma = 19.6$, $\alpha = 0.0191$, $r = 0.3225$, $\rho_1 = 0.5095$, $\lambda_1^{-1} = 0.5275$ and $\lambda_2^{-1} = 1.151$. (A useful sanity check for these calculations is $\lambda_1^{-1} + \lambda_2^{-1} = m_1(2 + \gamma^{-1}[c^2 - 1])$.) Unlike (2.1) and (4.1), we did not obtain the H_2 parameters as explicit expressions of ρ.

The numerical evidence indicates that both approximations perform very well, with the direct hyperexponential fit performing better than the convolution approximation; see table 6. In fact, as previously observed for RBM, unlike the approximations for $H_1(t)$, both approximations for $H_{\rho 2}(t)$ perform well for all $t > 0$.

References

[1] J. Abate and W. Whitt, Transient behavior of regulated Brownian motion, I: starting at the origin, Adv. Appl. Prob., 19 (1987), to appear.

[2] J. Abate and W. Whitt, Transient behavior of regulated Brownian motion, II: non-zero initial conditions, Adv. Appl. Prob., 19 (1987), to appear.

[3] J. Abate and W. Whitt, Transient behavior of the M/M/1 queue via Laplace transforms, Adv. Appl. Prob., 20 (1988), to appear.

[4] J.P.C. Blanc, The relaxation time of two queueing systems in series, Commun. Statist.-Stochastic Models 1 (1985) 1–16.

[5] J.W. Cohen, *The Single Server Queue*, 2nd ed. (North-Holland, Amsterdam, 1982).

[6] D.R. Cox, *Renewal Theory* (Methuen, London, 1962).

[7] D.R. Cox and W.L. Smith, *Queues* (Methuen, London, 1961).

[8] R.A. Doney, Letter to the Editor, J. Appl. Prob. 21 (1984) 673–674.

[9] W. Feller, *An Introduction to Probability Theory and its Applications*, I, 3rd ed. (Wiley, New York, 1968).

[10] D.P. Gaver, Jr., Diffusion approximations and models for certain congestion problems, J. Appl. Prob. 5 (1968) 607–623.

[11] D.P. Gaver, Jr. and P.A. Jacobs, On inference and transient response for M/G/1 models, Naval Postgraduate School, Monterey, CA, 1986.

[12] D.P. Heyman, An approximation for the busy period of the M/G/1 queue using a diffusion model, J. Appl. Prob. 11 (1974) 159–169.

[13] D.L. Iglehart and W. Whitt, Multiple channel queues in heavy traffic II: sequences, networks and batches, Adv. Appl. Prob. 2 (1970) 355–369.

[14] N.L. Johnson and S. Kotz, *Distributions In Statistics, Discrete Distributions* (Wiley, New York, 1969).

[15] J. Keilson, *Markov Chain Models - Rarity and Exponentiality* (Springer-Verlag, New York, 1979).

[16] W.D. Kelton and A.M. Law, The transient behavior of the M/M/S queue, with implications for steady-state simulation, Opns. Res. 33 (1985) 378–396.

[17] I. Lee, Stationary Markovian queueing systems: an approximation for the transient expected queue length, M.S. dissertation, Department of Electrical Engineering and Computer Science, MIT, Cambridge, 1985.

[18] I. Lee and E. Roth, Stationary Markovian queueing systems: an approximation for the transient expected queue length, unpublished paper, 1986.
[19] M. Mori, Transient behavior of the mean waiting time and its exact forms in M/M/1 and M/D/1. J. Opns. Res. Soc. Japan 19 (1976) 14–31.
[20] P.M. Morse, Stochastic properties of waiting lines, Opns. Res. 3 (1955) 255–261.
[21] G.F. Newell, *Application of Queueing Theory*, 2nd ed. (Chapman and Hall, London, 1982).
[22] A.R. Odoni and E. Roth, An empirical investigation of the transient behavior of stationary queueing systems, Opns. Res. 31 (1983) 432–455.
[23] N.U. Prabhu, *Queues and Inventories* (Wiley, New York, 1965).
[24] E. Roth, An investigation of the transient behavior of stationary queueing systems, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, 1981.
[25] C. Stone, Limit theorems for random walks, birth and death processes, and diffusion processes. Ill. J. Math. 7 (1963) 638–660.
[26] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, ed. D.J. Daley (Wiley, Chichester, 1983).
[27] L. Takacs, *Combinatorial Methods in the Theory of Stochastic Processes* (Wiley, New York, 1967).
[28] E. Van Doorn, *Stochastic Monotonicity and Queueing Applications of Birth-Death Proceses*, Lecture Notes in Statistics 4 (Springer-Verlag, New York, 1980).