Adv. Appl. Prob. **20,** 145–178 (1988)
Printed in N. Ireland
© *Applied Probability Trust* 1988

TRANSIENT BEHAVIOR OF THE $M/M/1$ QUEUE VIA LAPLACE TRANSFORMS

JOSEPH ABATE* AND
WARD WHITT,** *AT & T Bell Laboratories*

Abstract

This paper shows how the Laplace transform analysis of Bailey (1954), (1957) can be continued to yield additional insights about the time-dependent behavior of the queue-length process in the $M/M/1$ model. A transform factorization is established that leads to a decomposition of the first moment as a function of time into two monotone components. This factorization facilitates developing approximations for the moments and determining their asymptotic behavior as $t \to \infty$. All descriptions of the transient behavior are expressed in terms of basic building blocks such as the first-passage-time distributions. The analysis is facilitated by appropriate scaling of space and time so that regulated or reflected Brownian motion (RBM) appears as the special case in which the traffic intensity ρ equals the critical value 1. An operational calculus is developed for obtaining $M/M/1$ results directly from corresponding RBM results as well as vice versa. The analysis thus provides useful insight about RBM approximations for queues.

APPROACH TO STEADY STATE; RELAXATION TIMES; BIRTH-AND-DEATH PROCESSES; QUEUES; BROWNIAN MOTION; FIRST-PASSAGE TIMES; BUSY PERIOD; MOMENTS

1. Introduction

Bailey (1954), (1957), (1964) showed that the transient behavior of the queue-length process in the $M/M/1$ model can be described using double transforms (with respect to space and time). This analysis is described in many textbooks and is now quite familiar. In this paper we show that this analysis can be continued to obtain additional interesting and useful descriptions of the transient behavior. Our overall goal is to develop a better understanding of the transient behavior of queues and related stochastic flow systems, so that we can provide relatively simple descriptions suitable for engineering purposes.

Here are the main ideas. First, many transient results for the queue-length process can be obtained quite easily without deriving or applying the entire probability transition function (the usual first step). Second, a special role is played by the zero initial condition; it can be exploited to obtain useful results even for the case of general initial conditions (a new factorization in Section 2 with applications to moments in Sections 5 and 6). Third, appropriate scaling of time and space

Received 12 August 1986; revision received 2 December 1986.

* Postal address: AT & T Bell Laboratories, LC 2W-E06, 184 Liberty Corner Road, Warren, NJ 07060, USA.

** Postal address: AT & T Bell Laboratories, MH 2C-178, Murray Hill, NJ 07974, USA.

(especially time) reveals the close connection to regulated or reflecting Brownian motion (RBM); with the scaling it is easy to see that the $M/M/1$ queue-length process is the discrete analog of RBM. Finally, many descriptions of the transient behavior can be expressed solely in terms of relatively simple building blocks such as first-passage-time distributions. In fact, many characteristics can be expressed solely in terms of the probability transition function in the associated unrestricted process on the integers (with the barrier at zero removed).

Our work here complements Abate and Whitt (1987a, b) henceforth referred to as AWa and AWb, in which we describe the transient behavior of RBM and the $M/M/1$ queue-length process. Many result here are analogs of the transform results for RBM in Sections 1.3, 4, 5 and 9 of AWa. Some proofs here also represent alternate derivations of $M/M/1$ results in AWb. As shown in these previous papers, these results provide a basis for developing simple approximations to describe the moments of the queue length as functions of time. Our approach is to decompose the moments into two monotone pieces that can be normalized to obtain cumulative distribution functions (c.d.f.'s). We then approximate each c.d.f. by more convenient c.d.f.'s by matching the first three moments. As a basis for this procedure, we determine simple closed-form (exact) expressions for the first three moments of these c.d.f.'s. For the first-moment function in the $M/M/1$ queue, this program is completed here in Section 6 by calculating the first three moments of the second component, the first-moment-difference c.d.f.; the other component was treated in AWb. It is often appropriate to approximate each component c.d.f. by a mixture of two exponentials, so that the overall moment with general initial condition is approximated by the linear combination of four exponentials. In many regions of interest, one exponential component dominates, so that the approach to steady state can reasonably be described by a simple exponential; e.g., (2.1) of AWb.

The approximations for the $M/M/1$ model are important not only to describe the transient behavior of $M/M/1$ models but also to approximately describe the behavior of general $GI/G/m$ models. In a forthcoming paper we apply our $M/M/1$ results for this purpose (using heavy-traffic limit theorems, as indicated in AWb). In addition to the approximations for the time-dependent moments described above, we develop relatively simple normal approximations for the time-dependent c.d.f. based on Corollary 4.2.5; see Remark 4.5. Since approximations are considered at length in AWab, we do not dwell on them here. Related work is contained in Gaver and Jacobs (1986), Kelton (1985), Kelton and Law (1985), Lee (1985), Middleton (1979), Odoni and Roth (1983), Pegden and Rosenshine (1982), Roth (1981), Lee and Roth (1986) and references cited in these sources.

In addition to developing supporting theory for simple approximations, we uncover interesting relationships about the transient behavior of the $M/M/1$ queue, e.g., Corollary 4.2.3. In fact, some of these relationships do not yet have a satisfactory simple probabilistic interpretation, e.g., Theorems 9.1–9.3. Of particular interest for understanding RBM approximations for queues is an operational calculus developed in Section 10 to obtain $M/M/1$ queue-length formulas directly from corresponding RBM formulas. Obtaining RBM formulas from $M/M/1$ queue-length formulas is achieved via heavy-traffic limit theorems; going the other way is sometimes possible because the $M/M/1$ queue-length process is the discrete analog of RBM. A brief summary of the paper appears at the end in Section 11.

We conclude this introduction by mentioning that a similar theory exists for the workload or virtual waiting-time process (to some extent, even for the $M/G/1$ model). Moreover, the autocorrelation functions of stationary versions of all the basic $M/M/1$ processes can be expressed directly in terms of the moment c.d.f.'s; see Abate and Whitt (1987c), which we abbreviate to AWc. These related results supplement Ott (1977a, b), Cox and Isham (1986) and references cited there.

2. A transform factorization

Let $Q(t)$ represent the queue length (including the customer in service, if any) at time t in the $M/M/1$ model. Without loss of generality, let the service rate be 1, so that the arrival rate coincides with the traffic intensity ρ. Assume that $\rho < 1$, so that the system is stable with $Q(t)$ converging in distribution to $Q(\infty)$ where $P(Q(\infty) = j) \equiv P_j(\infty) = (1-\rho)\rho^j$ for $j \geqq 0$. (Transient results can also be obtained for $\rho \geqq 1$, but we will only consider the case $\rho < 1$.)

We begin by establishing a factorization for the double transform of the probability mass function (p.m.f.) of $Q(t)$. This transform factorization is closely related to the stochastic process decomposition in Section 2 of AWa. The transform factorization here is the discrete analog of Theorem 9.1 of AWa for RBM. The analysis is faciliated by scaling time in a manner consistent with the heavy-traffic limit to RBM as $\rho \to 1$; see Section 2.2 of AWb. We scale time but not space here, so a further space scaling is necessary to connect the result here to RBM; see Section 10.

Let $P_{ij}(t)$ be the probability transition function of the $M/M/1$ queue-length process with *scaled time*, i.e.,

$$P_{ij}(t) = P(Q(2(1-\rho)^{-2}t) = j \mid Q(0) = i) \tag{2.1}$$

and let $\bar{P}_i(z, s)$ be the double transform, defined by

$$\bar{P}_i(z, s) = \sum_{j=0}^{\infty} z^j \hat{P}_{ij}(s) \quad \text{and} \quad \hat{P}_{ij}(s) = \int_0^{\infty} \exp(-st) P_{ij}(t)\, dt. \tag{2.2}$$

The standard formula for $\bar{P}_i(z, s)$ due to Bailey (1954) that is given in the textbooks, e.g., p. 8 of Prabhu (1965) and p. 79 of Cohen (1982), is (after adjustment for the scaling)

$$\bar{P}_i(z, s) = \frac{2\theta^2 z^{i+1} - (1-z)\hat{P}_{i0}(s)}{\rho(z - z_1)(z_2 - z)} \tag{2.3}$$

using the *definitions*

$$\begin{aligned} &\theta=(1-\rho)/2, \qquad \Psi=[1+2(1-\theta)s+(\theta s)^2]^{\frac{1}{2}}, \\ (2.4) \qquad &r_1=\Psi+(1-\theta s), \qquad r_2=\Psi-(1-\theta s), \\ &\rho z_1=1-\theta r_1 \quad \text{and} \quad \rho z_2=1+\theta r_2 \end{aligned}$$

and the *relations*

$$(2.5) \qquad r_1r_2=2s, \qquad \rho z_1z_2=1 \quad \text{and} \quad \rho(1-z_1)(z_2-1)=2\theta^2 s.$$

The functions $z_1\equiv z_1(s)$ and $z_2\equiv z_2(s)$ are the two roots of the basic quadratic equation $\rho z^2-(1+\rho+2\theta^2 s)z+1=0$, as can be verified by elementary algebra. Additional discussion of the time scaling appears in the Appendix.

Since the denominator of (2.3) has only one zero in the unit circle (by Rouché) and the double transform needs to be analytic in the unit circle, the numerator of (2.3) must be 0 when $z=z_1$, so that

$$(2.6) \qquad \hat{P}_{i0}(s)=\frac{2\theta^2 z_1^{i+1}}{1-z_1}=\frac{\theta r_1 z_1^i}{s}$$

and

$$(2.7) \qquad \bar{P}_i(z,s)=\frac{2\theta^2 z^{i+1}-(1-z)\theta r_1 z_2^i s^{-1}}{\rho(z-z_1)(z_2-z)}.$$

This is where the standard transform argument stops. For example, following Bailey (1954), Cohen (1982) next expands (2.7) in powers of z to identify $\hat{P}_{ij}(s)$ as the coefficient of z^j, which yields an expression $\hat{P}_{ij}(s)$ as the sum of six terms, (4.28) on p. 80, each of which can be inverted to obtain the standard expression for $P_{ij}(t)$ involving modified Bessel functions of the first kind. This approach is of course effective, but as an alternative we suggest continuing to work with the double transform $\bar{P}_i(z,s)$. In particular, we find that $\bar{P}_i(z,s)$ can be factored in a useful way. (This is not difficult to check; we frequently omit routine algebraic proofs.)

Theorem 2.1. The double transform (2.7) can be factored as $\bar{P}_i(z,s)=\bar{P}_0(z,s)\bar{Q}_i(z,s)$, where

$$(2.8) \qquad \bar{P}_0(z,s)=\frac{z_2-1}{s(z_2-z)}=\frac{(1-\rho z_1)}{s(1-\rho z_1 z)}$$

and

$$(2.9) \qquad \bar{Q}_i(z,s)=\frac{(1-z_1)z^{i+1}-(1-z)z_1^{i+1}}{(z-z_1)}$$

with $\bar{P}_0(1,s)=s^{-1}$ and $Q_i(1,s)=1$.

Theorem 2.1 does not seem to help determine the transition function $P_{ij}(t)$ for all i and j, but as in Section 9 of AWa, we can obtain the time-transformed moments

$$(2.10) \qquad \hat{m}_k(s,i)=\int_0^\infty \exp(-st)m_k(t,i)\,dt$$

where

$$m_k(t, i) = E([Q(2(1-\rho)^{-2}t)]^k \mid Q(0) = i) \tag{2.11}$$

by differentiating $\bar{P}_i(z, s)$ with respect to z and setting $z = 1$. The factorization leads to the moment decomposition

$$m_k(t, i) = m_k(t, 0) + d_k(t, i) \tag{2.12}$$

where $d_k(t, i) = m_k(t, i) - m_k(t, 0)$. From the resulting transforms, which we will examine in Sections 5 and 6, or from Theorems 7.3, 11.1 and 1.3 of AWa, we know that $m_k(t, 0)$ and $d_k(t, i)$ are *monotone functions* of t for all i and $k = 1$ and 2. This decomposition is particularly convenient for describing the moment functions $m_k(t, i)$. For example, the decomposition yields a much easier derivation of the asymptotic behavior of $m_1(t, i)$ as $t \to \infty$ than the standard direct method; compare p. 180 of Cohen (1982) with Corollaries 5.2.3 and 6.1.2 and Theorems 3.1 and 6.2 here. This simplification occurs because the components $m_1(t, 0)$ and $d_1(t, i)$ of $m_1(t, i)$ can be very simply expressed in terms of the busy-period distribution. To set the stage, we treat the busy period and related first-passage times next.

3. First-passage times

As with RBM, everything can be expressed in terms of the first-passage times. For the $M/M/1$ queue, this primarily means the busy-period distribution. Let T_{ij} be the (time-scaled) first-passage time from i to j in the $M/M/1$ queue and let $f(t; i, j)$ be its density and $F(t; i, j)$ its c.d.f. Let $\hat{f}$ and $\hat{F}$ denote the corresponding Laplace transforms with respect to time. Let $b(t)$ be the density of the (time-scaled) busy period T_{10} and let $\hat{b}(s)$ be its Laplace transform. Let $B(t)$ be the associated c.d.f. and $\hat{B}(s)$ its transform. Let $f \sim g$ mean $f(t)/g(t) \to 1$ as $t \to \infty$. Let $\tau = (1 + \sqrt{\rho})^2/2$ be the *time-scaled relaxation time*; $I_i(v)$ the modified Bessel function (p. 377 of Abramowitz and Stegun (1972)), $v = t\theta^{-2}\rho^{\frac{1}{2}}$ and

$$L(t, \rho) = (2\pi\rho^{\frac{3}{2}}t^3)^{-\frac{1}{2}} \exp(-t/\tau). \tag{3.1}$$

The form (3.1) is convenient for seeing the connection to RBM; note that $\tau \equiv (1 + \sqrt{\rho})^2/2 \to 2$, the relaxation time for RBM, as $\rho \to 1$ and $L(t, 1) = t^{-1}\gamma(t)$ where $\gamma(t)$ is the gamma density with mean 1 and shape parameter $\frac{1}{2}$, as in (4.2) of AWa; $\gamma(t)$ itself coincides with the density of BM (unregulated) at 0 starting at 0, i.e., $\gamma(t) = (2\pi t)^{-\frac{1}{2}} \exp(-t/2)$. Also $L(t, 1)$ is the limit of $x^{-1}f(t; x, 0)$ as $x \to 0$, where $f(t; x, 0)$ is the inverse Gaussian first-passage-time density in (1.5) of AWa.

The following theorem expresses well-known results of Bailey (1957).

Theorem 3.1 (Bailey). (a) $\hat{f}(s; i, 0) = \hat{b}(s)^i = z_1^i$;
(b) $f(t; i, 0) = (i/t)\rho^{-\frac{1}{2}} \exp(-t/\tau)[\exp(-v)I_i(v)] \sim i\theta\rho^{-([i-1]/2)}L(t, \rho)$;
(c) $1 - F(t; i, 0) \sim \tau i\theta\rho^{-([i-1]/2)}L(t, \rho)$.

Proof. Part (a) comes from Section 3 of Bailey (1954) or Section 6 of Bailey

(1957); ξ in (39) becomes z_1. The idea is to carry out an analysis as in (2.1)–(2.7) using a modified process that ceases when it reaches the origin. However, a separate argument is really not needed; we can directly apply $\hat{P}_{i0}(s)$ in (2.6). From first principles (to go from i to 0 you must do so for a first time), $P_{i0}(t) = \int_0^t f(u; i, 0)P_{00}(t-u)\,du$, so that $\hat{P}_{i0}(s) = \hat{P}_{00}(s)\hat{f}(s; i, 0)$, from which we immediately get (a). It is elementary that T_{i0} can be represented as the sum of i i.i.d. copies of T_{10}. The first part of (b) is (40) on p. 332 of Bailey (1957) in our time scale. For the second part of (b), use 9.7.1 on p. 377 of Abramowitz and Stegun (1972): $e^{-v}I_i(v) \sim (2\pi v)^{-\frac{1}{2}}$ independent of i. For (c), apply an asymptotic theorem for integration, e.g., p. 17 of Erdélyi (1956), using $\int_t^\infty L(x, \rho)\,dx \sim \tau L(\tau, \rho)$, which we prove below in Corollary 5.2.3.

Corollary 3.1.1. In terms ρ and $\theta = (1-\rho)/2$, (a) the first five moments of the busy period T_{10} are $m_1 = m_2 = \theta$, $m_3 = 3\theta(1-\theta)$, $m_4 = 3\theta[1 + 3\rho + \rho^2]$ and $m_5 = 15\theta(1-\theta)[1 + 5\rho + \rho^2]$ and (b) the first three moments of T_{i0} are $m_1 = i\theta$, $m_2 = i\theta([i-1]\theta + 1)$ and $m_3 = i\theta[(i\theta)^2 + 3i\theta(1-\theta) + 3\rho + 2\theta^2]$.

In fact, *all* moments of the $M/M/1$ busy period T_{10}, and thus all moments of T_{i0}, can be found from a basic recursion due to Riordan (1962). (Riordan's result (70) on p. 107 is stated only for the conditional waiting time in the $M/M/1$ model with LCFS (last-come first-served) discipline. It is necessary to observe that this coincides with the ordinary $M/M/1$ busy-period c.d.f.; see Remark 4.1 below.)

Theorem 3.2 (Riordan). The moments m_n of T_{10} satisfy the recurrence relation

$$m_{n+2} = (2n+1)(1-\theta)m_{n+1} - (n^2-1)\theta^2 m_n$$

for $m_0 = 1$ and $m_1 = \theta$.

Remark 3.1. An explicit expression for the moments is given on pp. 231–233 of Takács (1967). However, it is not necessarily more useful than the recursion above.

An important role will be played by the equilibrium time to emptiness $T_{\varepsilon 0}$, which is the first-passage time to 0 starting with the equilibrium geometric stationary distribution. (See Corollaries 4.1.1 and 5.2.1.) Let $f_{\varepsilon 0}(t)$ and $F_{\varepsilon 0}(t)$ be the density and c.d.f. of $T_{\varepsilon 0}$. This involves a slight abuse of terminology because $T_{\varepsilon 0}$ has an atom with mass $(1-\rho)$ at 0, so that it does not have a proper density. The transform $\hat{f}_{\varepsilon 0}(s)$ should thus be interpreted as the Laplace–Stieltjes transform $\int_{0-}^\infty \exp(-st)\,dF_{\varepsilon 0}(t)$. The following result follows easily from Theorem 3.1.

Theorem 3.3. The equilibrium time to emptiness $T_{\varepsilon 0}$ has Laplace–Stieljes transform

$$\hat{f}_{\varepsilon 0}(s) = \sum_{i=0}^{\infty} (1-\rho)\rho^i z_1^i = \frac{1-\rho}{1-\rho z_1} = \frac{2}{r_1} = \frac{r_2}{s} = \frac{r_2+2}{r_1+s},$$

moments $m_{\varepsilon 1} = \rho/2$ and $m_{\varepsilon 2} = \rho(1-\theta)$ and squared coefficient of variation $c_\varepsilon^2 = (m_{\varepsilon 2} - m_{\varepsilon 1}^2)/m_{\varepsilon 1}^2 = 1 + 2/\rho$.

Corollary 3.3.1. As $\rho \to 1$, $T_{\varepsilon 0}$ converges in distribution to the equilibrium time to emptiness for RBM, which has transform $2/r_1$ where r_1 is the RBM quantity; see Corollary 1.3.1 and (1.10) of AWa and Section 10 here.

Following Bailey (1957), we can also describe the first-passage time upward from 0.

Theorem 3.4 (Bailey). For the time-scaled $M/M/1$ queue,

$$\hat{f}(s; 0, j) = \frac{r_1 + r_2}{r_1 z_1^j + r_2 z_2^j} = \frac{(r_1 + r_2)(\rho z_1)^j}{r_1 \rho^j z_1^{2j} + r_2} \tag{3.2}$$

for r_1, r_2, z_1 and z_2 in (2.4), so that

$$E(T_{0j}) = 2^{-1}[\rho^{-j} - 1 - 2\theta j] \tag{3.3}$$

and

$$\mathrm{Var}\,(T_{0j}) = 4^{-1}[\rho^{-2j} - 1 - 4j\theta(1-\theta) + 4\rho^{-j}(\rho - 2\theta j) - 4\rho]. \tag{3.4}$$

Proof. Formula (45) on p. 333 of Bailey (1957) becomes (3.2) by using $\rho z_1 = 1 - \theta r_1$ and $\rho z_2 = 1 + \theta r_2$ in (2.4). Bailey's (46) is (3.3). We know of no reference for (3.4), which is obtained by further differentiation and algebra.

Note that $E(T_{0j}) \to 0$ in (3.3) as $\rho \to 1$. This is a consequence of the time scaling (2.1). As a direct consequence of Theorems 3.1 and 3.4, we obtain the corresponding results for RBM, originally due to Darling and Siegert (1953). The time scaling helps make the connections clear.

Corollary 3.4.1 (Darling and Siegert). For RBM,

$$\hat{f}(s; x, 0) = \exp(-x r_2) \tag{3.5}$$

and

$$\hat{f}(s; 0, x) = \frac{r_1 + r_2}{r_1 \exp(-x r_2) + r_2 \exp(x r_1)} \tag{3.6}$$

for r_1 and r_2 in Section 1.3 of AWa, so that

$$E(T_{x0}) = x, \qquad E(T_{0x}) = 2^{-1}[\exp(2x) - 1 - 2x], \qquad \mathrm{Var}\,(T_{x0}) = x, \tag{3.7}$$

$$\mathrm{Var}\,(T_{0x}) = 4^{-1}[\exp(4x) - 1 - 4x + 4\exp(2x)(1-2x) - 4]. \tag{3.8}$$

Proof. By the heavy-traffic limit theorem in Iglehart and Whitt (1970) or Stone (1963) and the continuous mapping theorem (Theorem 5.1 of Billingsley (1968)) to treat the first-passage-time functional (Section 7 of Whitt (1980)), $T_{[x\theta^{-1}],[y\theta^{-1}]}$ for $M/M/1$ converges in distribution to T_{xy} for RBM as $\rho \to 1$. By direct calculation, $z_1^{[\theta^{-1}x]} \to \exp(-x r_2)$ and $z_2^{[\theta^{-1}x]} \to \exp(x r_1)$ as $\rho \to 1$, where r_1 and r_2 for RBM are the obvious limits from (2.4) as $\rho \rightarrow 1$. The moments converge too by an additional uniform integrability argument; p. 32 of Billingsley (1968).

Formula (3.6) is a variant of (9–138) on p. 358 of Heyman and Sobel (1982). Their counterpart to (3.7), (9–140) on p. 359, seems to be in error, however. We know of no reference for the variance in (3.8). For further discussion about the relation between $M/M/1$ and RBM, see Section 10.

Corollary 3.4.1 is of interest not only as a description of RBM but also for the convergence of $M/M/1$ as $\rho \rightarrow 1$ described in the proof. We can also apply Theorem 3.1 to obtain another heavy-traffic limit theorem for the first-passage-time distributions. Of course, for $i = 1$ this is the busy-period distribution. (We are unaware of any previous heavy-traffic limit theorem for a busy-period distribution.)

Theorem 3.5. For each positive t and i,

$$\text{(a)} \quad \lim_{\rho \rightarrow 1} \theta^{-1} f(t; i, 0) = i(2\pi t^3)^{-\frac{1}{2}} \exp(-t/2) \equiv iL(t, 1)$$

and

$$\text{(b)} \quad \lim_{\rho \rightarrow 1} (1 - \rho)^{-1}[1 - F(t; i, 0)] = i(t^{-\frac{1}{2}} \phi(t^{\frac{1}{2}}) - [1 - \Phi(t^{\frac{1}{2}})]).$$

Proof. (a) Apply Theorem 3.1(b) noting that $v = t\theta^{-2}\rho^{\frac{1}{2}} \rightarrow \infty$ as $\theta \rightarrow 0$ as well as when $t \rightarrow \infty$. (b) Apply part (a) together with the Lebesgue dominated convergence theorem; from 9.7.1 on p. 377 of Abramowitz and Stegun (1972), $\sqrt{v}\, e^{-v} I_i(v) \leqq (2\pi)^{-\frac{1}{2}}$ for all sufficiently large v. By differentiating the limit in (b), it is easy to verify that it is the integral over (t, ∞) of the limit in (a).

The form of the limit in Theorem 3.5(b) will be explained by Corollary 4.2.3 and Section 10. For a closely related result, see Corollary 5.2.3.

4. The probability mass function

In this section we develop new representations for the p.m.f. $P_{ij}(t)$ in (2.1), which have special appeal when $i = 0$ or $j = 0$. We show how $P_{ij}(t)$ can be represented solely in terms of the first-passage-time densities $f(t; i, 0)$ and $f(t; 0, j)$ described in Section 3.

A key property of RBM is that the time-transformed density of the state at time t starting at the origin has a simple exponential form; (2.11) of Gaver (1968) and (1.8) of AWa. It is significant that a discrete analog, a simple geometric form, holds for the $M/M/1$ queue. Our new geometric representation for $\hat{P}_{0j}(s)$ can be obtained directly from Theorem 2.1 by algebraic manipulations, but it can also be obtained from a fundamental relationship that is *valid for any reversible Markov process*; see Keilson (1979) or Kelly (1979). Let $*$ denote convolution.

Theorem 4.1. $f_{\varepsilon 0}(t) * P_{0j}(t) = P_j(\infty) F(t; j, 0)$.

Proof. Note that $\{Q(t): t \geqq 0\}$ starting in equilibrium is a reversible process. The left side is the probability of hitting 0 somewhere in the interval $(0, t)$ and ending up at j at time t, starting in equilibrium. The right side is the equilibrium probability of initially being at j and hitting 0 somewhere in the interval $(0, t)$. By reversibility, these two probabilities are equal.

With transforms, the convolution in Theorem 4.1 can be represented as simple product, so that we can divide by the transform of $f_{\varepsilon 0}(t)$ to obtain an expression for the transform of $P_{0j}(t)$.

Corollary 4.1.1. $\hat{P}_{0j}(s) = P_j(\infty)\hat{F}(s; j, 0)/\hat{f}_{\varepsilon 0}(s)$.

We apply Theorem 2.1 (alternatively Theorem 4.1 and Corollary 3.1.2) to obtain the following expressions for $\hat{P}_{0j}(s)$.

Theorem 4.2. The double transform $\bar{P}_0(z, s)$ can be represented as

$$\bar{P}_0(z, s) = \frac{1-\rho z_1}{s(1-\rho z_1 z)} = \frac{\theta r_1}{s(1-(1-\theta r_1)z)} = \left(\frac{\theta r_1}{s}\right)\sum_{j=0}^{\infty}(1-\theta r_1)^j z^j \tag{4.1}$$

which can be immediately inverted to express the time-transformed p.m.f. as (the simple geometric form being the first relation)

$$\begin{aligned}\hat{P}_{0j}(s) &= \left(\frac{\theta r_1}{s}\right)(1-\theta r_1)^j = \left(\frac{\theta r_1}{s}\right)(\rho z_1)^j = [(1-\rho)\rho^j]\left(\frac{r_1}{2s}\right)z_1^j \\ &= P_j(\infty)\hat{P}_{00}(s)\hat{f}(s; j, 0)/P_0(\infty) = \rho^j\frac{[1-\rho z_1]}{s}\hat{f}(s; j, 0) \qquad (4.2)\\ &= \rho^j[\hat{F}(s; j, 0) - \rho\hat{F}(s; j+1, 0)] = \rho^j\hat{F}(s; j, 0) - \rho^{j+1}\hat{F}(s; j+1, 0).\end{aligned}$$

From the geometric form we can easily go the steady-state limit as $t \to \infty$.

Corollary 4.2.1. The steady-state limit is

$$P_j(\infty) \equiv \lim_{t\to\infty} P_{0j}(t) = \lim_{s\to 0} s\hat{P}_{0j}(s) = \lim_{s\to 0} \theta r_1(1-\theta r_1)^j = 2\theta(1-2\theta)^j = (1-\rho)\rho^j.$$

It is significant that the final form of (4.2) can be immediately inverted.

Corollary 4.2.2. $P_{0j}(t) = \rho^j F(t; j, 0) - \rho^{j+1}F(t; j+1, 0)$.

The case $j = 0$ in Corollary 4.2.2 is particularly interesting.

Corollary 4.2.3. $P_{00}(t) = 1 - \rho B(t)$.

Remarks 4.1. Surprisingly, Corollary 4.2.3 is not well known, although it is implicit in (27) of Bailey (1954) and appears at the bottom of p. 25 of Beneš (1963). Kosten (1973) established the connection for the transforms in (5.3.25) on p. 85, but he only relates $P_{00}(t)$ to the conditional expected waiting time with the last-come first-served discipline. (Recall the comment about Theorem 3.2.) This in turn is connected to the busy period on p. 786 of Riordan (1961), p. 221 of Cooper (1972) and p. 438 of Cohen (1982). Corollary 4.2.3 was also discovered independently by our colleagues A. Kumar and W. S. Wong (1987).

4.2. If we only wanted Corollary 4.2.3, then a very quick proof is possible. Indeed, given (2.6) and Theorem 3.1(a), Corollary 4.2.3 is already contained in the relation $\rho z_1 = 1 - \theta r_1$ in (2.4).

4.3. D. R. Smith (personal communication) has provided the following probabilistic proof of Corollary 4.2.3. (A similar argument using reversibility yields Corollary 4.2.2.) Consider the $M/M/1$ model in equilibrium and note that the flow of probability mass from state 0 to all other states in the time interval $[0, t]$ equals the flow in the other direction in the same time interval. The first is obviously $P_0(\infty)(1 - P_{00}(t))$, while the second can be shown to be $(1 - P_0(\infty))B(t)P_0(\infty)$. The first term in the second expression is the equilibrium probability of starting above 0. The rest is the equilibrium conditional probability of being at 0 at time t given that the initial state is greater than 0. The conditioning means that there is at least one customer in the system at time 0. In order to be at 0 at time t, the busy period generated by the first customer must end before time t. It ends at time s with density $b(s)$. At the instant the busy period ends, the additional customers begin to receive service, but the number of additional customers has the equilibrium distribution, so that the probability of being at 0 at times s and t is $P_0(\infty)$. Integrating from 0 to t yields the second expression.

4.4. Erik van Doorn (personal communication) has observed that Corollary 4.2.3 also can be derived from (5.4) on p. 98 and p. 103 of Karlin and McGregor (1958). From the perspective of the spectral representation, $P_{00}(t)$ plays a central role because it is the Laplace–Stieltjes transform of the spectral measure.

From Corollary 4.2.2 and Theorem 3.1(c), we immediately obtain the asymptotic behavior of $P_{0j}(t)$, as given in (4.34) on p. 84 of Cohen (1982).

Corollary 4.2.4.

$$\begin{aligned} P_j(\infty) - P_{0j}(t) &= \rho^j[1 - F(t; j, 0)] - \rho^{j+1}[1 - F(t; j+1, 0)] \\ &\sim \tau\theta\rho^{j/2}L(t, \rho)[j\rho^{\frac{1}{2}} - (j+1)\rho] \end{aligned}$$

for $L(t, \rho)$ in (3.1) and τ the time-scaled relaxation time.

Corollary 4.2.4 indicates (but does not prove) that $P_{0j}(t)$ is strictly increasing in t for all t if and only if $j\rho^{\frac{1}{2}} \geqq (j+1)\rho$ or, equivalently, if and only if $j \geqq \sqrt{\rho}/(1 - \sqrt{\rho}) = (1 + \rho^{-\frac{1}{2}})m_1(\infty)$ where $m_1(\infty) = \rho/(1-\rho)$ is the steady-state mean. We prove this in Section 8. The criterion is equivalent to $j\theta \geqq (1 + \sqrt{\rho})\sqrt{\rho}/2$, which reduces to the criterion $x \geqq 1$ established for RBM in Theorem 1.9 of AWa as $\rho \to 1$ with $j\theta \to x$. (For $M/M/1$ we scaled time but not space.)

We can also apply Corollary 4.2.2 to obtain simple expressions for the complementary c.d.f. and the first moment starting at 0, defined in (2.11). Part (c) below constitutes an alternative proof of Corollary 3.2.1 of AWb.

Corollary 4.2.5. The complementary c.d.f. and moments can be expressed as

$$\text{(a)} \qquad \sum_{j=n}^{\infty} P_{0j}(t) = \rho^n F(t; n, 0) = \sum_{j=n}^{\infty} P_j(\infty) F(t; n, 0)$$

$$\text{(b)} \qquad m_1(t, 0) = \sum_{n=1}^{\infty} \sum_{j=n}^{\infty} P_{0j} = \sum_{n=1}^{\infty} \rho^n F(t; n, 0),$$

and

$$\text{(c)} \qquad m_1(t, 0)/m_1(\infty) = \sum_{n=1}^{\infty} (1-\rho)\rho^{n-1} F(t; n, 0)$$

$$= \sum_{n=1}^{\infty} P(Q(\infty) = n \mid Q(\infty) > 0) P(T_{n0} \leqq t).$$

Remarks 4.5. Corollary 4.2.5(a) is the basis for a normal approximation for the normalized complementary c.d.f. Since the first-passage time T_{n0} is distributed as the sum of n i.i.d. copies of T_{10} with finite moments as in Corollary 3.1.1, T_{n0} is asymptotically normally distributed as $n \to \infty$. Thus, for n not too small, we obtain the approximation

$$P(Q(t) > n \mid Q(0) = 0) \approx P(Q(\infty) > n) \Phi([t - n\theta]/\sqrt{n\theta(1-\theta)})$$

with time scaling (2.1) and

$$P(Q(t) > n \mid Q(0) = 0) \approx P(Q(\infty) > n) \Phi([t - n(1-\rho)^{-1}]/\sqrt{n(1+\rho)/2(1-\rho)})$$

without time scaling. (Refinements also follow by applying the refinements to the central limit theorem; e.g., Chapter XVI of Feller (1971).) These normal approximations can be viewed as large-deviation results (asymptotically correct as $n \to \infty$); see Asmussen and Thorisson (1986) and (8.3.4) of Siegmund (1985). Corollary 4.2.5(a) provides a remarkably simple derivation for this case.

4.6. Corollaries 3.4.1 and 4.2.5 together imply corresponding results and associated normal approximations for RBM. Direct derivations appear in Section 1.7 of AWa.

Let $I_0(t)$ be the cumulative idle time in $[0, t]$ starting at the origin. Corollary 4.2.3 yields an interesting characterization of its mean. For any density $g(t)$ on $[0, \infty)$ with mean m_1 and Laplace transform $\hat{g}(s)$, let $g_e(t)$ be the associated stationary-excess (or equilibrium-residual-life) density with transform $\hat{g}_e(s) = [1 - \hat{g}(s)]/m_1 s$.

Corollary 4.2.6. The expected cumulative idle time in $[0, t]$, first without time scaling, is

$$E[I_0(t)] = \int_0^t P_{00}(u)\,du = \int_0^t [1 - \rho B(u)]\,du = (1-\rho)t + m_1(\infty)B_e(t)$$
$$= 2\theta t + m_1(\infty)B_e(t)$$

and, second with time scaling, is

$$E[I_0(t)] = 2\theta t + 2\theta^2 m_1(\infty)B_e(t).$$

Proof. Apply Corollary 4.2.3 using $B_e(t) = (1-\rho)\int_0^t [1 - B(u)]\,du$ without time scaling.

To treat the general case $P_{ij}(t)$, we apply reversibility again via the basic relation

$$(4.3) \qquad P_i(\infty)P_{ij}(t) = P_j(\infty)P_{ji}(t)$$

which for $M/M/1$ becomes $P_{ij}(t) = \rho^{j-i}P_{ji}(t)$. We can immediately combine (4.3) and Corollary 4.2.4 to describe $P_{i0}(t)$.

Corollary 4.2.7.

$$(1-\rho) - P_{i0}(t) = [1 - F(t; i, 0)] - \rho[1 - F(t, i+1, 0)]$$
$$\sim \tau\theta\rho^{-(i/2)}L(t, \rho)[i\rho^{\frac{1}{2}} - (i+1)\rho].$$

For $0 \leqq i < j$, we also have

$$(4.4) \qquad P_{0j}(t) = f(t; 0, i) * P_{ij}(t) = \int_0^t f(s; 0, i)P_{ij}(t-s)\,ds$$

because to get from 0 to j the process might pass through i, and to pass through i it must do so for a first time. As a consequence, we have the following expression for $\hat{P}_{ij}(s)$.

Theorem 4.3. (a) For $0 < i < j$,

$$\hat{P}_{ij}(s) = \hat{P}_{0j}(s)/\hat{f}(s; 0, i) = \frac{\rho^j\hat{F}(s; j, 0) - \rho^{j+1}\hat{F}(s; j+1, 0)}{\hat{f}(s; 0, i)};$$

(b) For $i > j$,

$$\hat{P}_{ij}(s) = \rho^{j-i}\hat{P}_{ji}(s) = \rho^{j-i}\hat{P}_{0i}(s)/\hat{f}(s; 0, j) = \frac{\rho^j\hat{F}(s; i, 0) - \rho^{j+1}\hat{F}(s; i+1, 0)}{\hat{f}(s; 0, j)}.$$

It remains to determine if our approach has anything special to offer for describing $P_{ij}(t)$ in the time domain and its asymptotic behavior as $t \to \infty$ when neither $i = 0$ nor $j = 0$; so far, the analysis we would suggest is essentially the same as on p. 82 of Cohen (1982) or p. 12 of Prabhu (1965). The difficulty in treating $P_{ij}(t)$ perhaps helps explain why the first moment function is so much easier to treat via the decomposition (2.12). We do apply Theorem 4.3 here in Section 10 to relate $M/M/1$ to RBM.

5. Moment functions starting at the origin

We now discuss the moment functions and associated moment c.d.f.'s; for additional discussion and motivation see AWab. Let $m_k(t, i)$ be the time-scaled kth moment function in (2.11) and let $\hat{m}_k(s, i)$ be the associated Laplace transform with respect to time in (2.10). Let $m_{(k)}(t, i)$ be the associated kth factorial moment function, i.e., $m_{(k)}(t, i) = E[X(X-1)\cdots(X-k+1)]$ where

$$X = (Q(2(1-\rho)^{-2}t) \mid Q(0) = i)$$

and let $\hat{m}_{(k)}(s, i)$ be the associated Laplace transform. We introduce these factorial moments because they have nice structure, as illustrated by Corollary 5.2.1. below. Additional insight into this structure is provided by the probabilistic proofs in AWab.

As in AWb, we first focus on the case $i = 0$, but here we also treat the general case in Section 6. Let $H_k(t)$ be the kth factorial-moment c.d.f. defined by $H_k(t) = m_{(k)}(t, 0)/m_{(k)}(\infty)$ with density $h_k(t)$ and associated transforms $\hat{H}_k(s)$ and $\hat{h}_k(s)$. (In Corollary 5.2.1 below we prove that $H_k(t)$ is a c.d.f. for each k.) We can obtain $\hat{m}_1(s, 0)$ from $\bar{P}_0(z, s)$ in (2.9) by differentiating, i.e.

$$\hat{m}_1(s, 0) = \frac{\partial}{\partial z}\bar{P}_0(z, s)\big|_{z=1} = \frac{\rho z_1}{s(1-\rho z_1)}. \tag{5.1}$$

We then obtain the Laplace transform $\hat{H}_1(s)$ by scaling space; i.e., let $\hat{H}_1(s) = 2\theta\hat{m}_1(s, 0)/\rho$. Finally, we can combine (5.1), (2.4) and (2.5) to obtain the corresponding transform $\hat{h}_1(s)$ of the first-moment density $h_1(t)$.

Theorem 5.1.

$$\hat{h}_1(s) = \frac{2\theta z_1}{1-\rho z_1} = \frac{r_1}{r_1+s} = \frac{2}{2+r_2} = \frac{2}{r_1+2\theta s} = \frac{2z_1}{r_1}.$$

We can also easily describe all the factorial moment functions using Theorem 4.2. (Alternatively, we could differentiate further in (5.1).) Let $(n)_k = n(n-1)\cdots(n-k+1)$.

Theorem 5.2.

$$\hat{m}_{(k)}(s, 0) = \sum_{n=0}^{\infty}(n)_k\hat{P}_{0n}(s) = \frac{k!}{s}\left(\frac{\rho}{2\theta}\right)^k\left(\frac{2z_1}{r_1}\right)^k.$$

Since $m_{(k)}(\infty) = k!\,(\rho/2\theta)^k$, see (3.5) of AWh. ʷe immediately get expressions for all factorial-moment c.d.f.'s and densities too. We apply Theorem 3.1(a) and Corollary 3.1.2 (plus elementary algebra) to get a simple connection to the busy-period distribution.

Corollary 5.2.1.

$$\hat{h}_k(s) = \left(\frac{2z_1}{r_1}\right)^k = [\hat{f}_{\varepsilon 0}(s)\hat{b}(s)]^k = \hat{b}_e(s)^k = \hat{h}_1(s)^k.$$

Corollary 5.2.1 implies that $h_1(t)$ is simultaneously $b_e(t)$ and the convolution of $f_{\varepsilon 0}(t)$ and $b(t)$, so that it is a proper density and we obtain new proofs of Corollaries 3.1.1 and 3.1.2 of AWb. Moreover, Corollary 5.2.1 shows that $h_k(t)$ is the k-fold convolution of $h_1(t)$, thus providing a new proof of Theorem 3.2 of AWb.

Alternate proof of Corollary 5.2.1 *for* $k = 1$. For the $M/M/1$ model without time scaling, the expected queue length starting at the origin obviously coincides with the expected virtual waiting time, say $EW_0(t)$, which in turn is equal to the expected cumulative input of work in $[0, t]$ minus t plus the expected cumulative idle time in $[0, t]$. We can thus invoke Corollary 4.2.6 to obtain

$$m_1(t, 0) = EW_0(t) = \rho t - t + EI_0(t) = m_1(\infty)B_e(t).$$

We can apply Theorem 5.2 to show that the kth-factorial-moment c.d.f $H_k(t)$ converges in distribution to non-degenerate limits as $\rho \to 1$ and as $\rho \to 0$, by invoking the continuity theorem for Laplace transforms. (This justifies a claim in Section 2.2 of AWb.)

Corollary 5.2.2. (a) If $\rho \to 1$, then $\theta \to 0$ and $r_1 \to 1 + \sqrt{1 + 2s}$, so that

$$\lim_{\rho \to 1} \hat{h}_k(s) = 2^k[1 + (1 + 2s)^{\frac{1}{2}}]^{-k},$$

coinciding with the transform of the k-fold convolution of the first-moment density of RBM in (1.10) of AWa.

(b) If $\rho \to 0$, then $\theta \to \frac{1}{2}$ and $\Psi(s) \to 1 + s/2$, so that

$$\lim_{\rho \to 0} \hat{h}_k(s) = 2^k(2 + s)^{-k},$$

coinciding with the transform of the k-fold convolution of an exponential density with mean $\frac{1}{2}$.

Remark 5.1. Since $H_1(t) = B_e(t)$, Theorem 5.2.2(a) implies that $B_e(t)$, with our time scaling, converges in distribution to a proper limit as $\rho \to 1$; see also Corollary 3.3.1. This suggests that the busy-period stationary-excess c.d.f. $B_e(t)$ is in some sense more robust than the busy-period c.d.f. $B(t)$ itself, e.g., for general queues it might be easier to approximate $B_e(t)$ than $B(t)$. This observation was previously made by Delbrouck (1976); our analysis yields a possible explanation.

From Theorem 3.1(b) and Corollary 5.2.1 we can obtain the asymptotic behavior of $h_1(t)$ and $1 - H_1(t)$ as $t \to \infty$. The asymptotic relation in Corollary 5.2.3(a) below is an improvement upon the heavy-traffic limit for the busy-period c.d.f. in Theorem 3.5(b). Here we obtain the asymptotic behavior of the busy-period c.d.f. as $t \to \infty$ for each fixed $\rho < 1$. (Because of the time scaling in (2.1), Theorem 3.5 involves both $t \to \infty$ and $\rho \to 1$.) We obtain the limit in Theorem 3.5(b) from Corollary 5.2.3(a) by letting $\rho \to 1$ afterwards. The limit in Corollary 5.2.3(a) provides good approximations. (We intend to discuss approximations for busy-period distributions in another paper.)

Corollary 5.2.3. As $t \to \infty$,

$$\text{(a)}\quad h_1(t) = b_e(t) = \theta^{-1}[1 - B(t)] \sim 2\rho^{-\frac{3}{4}}[t^{-\frac{1}{2}}\phi(\sqrt{2t/\tau}) - \sqrt{2/\tau}[1 - \Phi(\sqrt{2t/\tau})]]$$
$$\sim \tau L(t, \rho)\left[1 - \frac{3}{t}\left(\frac{\tau}{2}\right) + \frac{15}{t^2}\left(\frac{\tau}{2}\right)^2 - \cdots\right]$$

and

$$\text{(b)}\quad 1 - H_1(t) = \int_t^\infty h_1(x)\,dx \sim \tau^2 L(t, \rho)\left[1 - \frac{6}{t}\left(\frac{\tau}{2}\right) + \cdots\right]$$

where τ is the time-scaled relaxation time.

Proof. (a) From Theorem 3.1(b),

$$h_1(t) = \theta^{-1}\int_t^\infty b(x)\,dx \sim \int_t^\infty L(x, \rho)\,dx = (2\pi\rho^{\frac{3}{2}})^{-\frac{1}{2}}\int_t^\infty x^{-\frac{3}{2}}\exp(-x/\tau)\,dx$$

where, after integrating by parts,

$$\int_t^\infty x^{-\frac{3}{2}}\exp(-x/\tau)\,dx = 2t^{-\frac{1}{2}}\exp(-t/\tau) - \tau^{-1}\int_t^\infty x^{-\frac{1}{2}}\exp(-x/\tau)\,dx$$
$$= 2\sqrt{2\pi}(t^{-\frac{1}{2}}\phi(\sqrt{2t/\tau}) - \sqrt{2/\tau}[1 - \Phi(\sqrt{2t/\tau})])$$
$$= 2\sqrt{2\pi}\left(\phi(\sqrt{2t/\tau})\left[\frac{1}{\sqrt{t}} - \frac{1}{\sqrt{t}}\left(1 - \frac{1}{t}\left(\frac{\tau}{2}\right) + \frac{3}{t^2}\left(\frac{\tau}{2}\right)^2 - \frac{15}{t^3}\left(\frac{\tau}{2}\right)^3 + \cdots\right)\right]\right).$$

For part (b), apply p. 17 of Erdélyi (1956), getting

$$1 - H_1(t) = \int_t^\infty h_1(x)\,dx \sim \tau\int_t^\infty L(x, \rho)\left[1 - \frac{3}{x}\left(\frac{\tau}{2}\right) + \frac{15}{x^2}\left(\frac{\tau}{2}\right)^2\right]dx$$

from which the result follows by integrating by parts twice.

Remark 5.2. Results for RBM follow from Corollary 5.2.3 by simply setting $\rho = 1$. The RBM versions of (a) and (b) are easily obtained directly from (4.3) and (4.4) of AWa or, alternatively, Corollaries 1.1 and 1.2 there.

We can also combine Theorems 3.2 and 5.2 to obtain a recurrence relation to obtain all moments of $H_1(t)$. (Specifically, we invoke Corollary 3.1.3 of AWb.) We are interested in the moments, not only as summary measures, but also to develop simple approximations by moment matching; see AWa, b, c.

Corollary 5.2.4. The moments m_n of $H_1(t)$ satisfy the recurrence relation

$$m_{n+1} = \left(\frac{n+1}{n+2}\right)[(2n+1)(1-\theta)m_n - n(n-1)\theta^2 m_{n-1}]$$

for $m_0 = 1$ and $m_1 = \frac{1}{2}$; e.g., $m_2 = (1-\theta)$, $m_3 = (\frac{3}{4})[1 + 3\rho + \rho^2]$ and $m_4 = 3(1-\theta)[1 + 5\rho + \rho^2]$.

We now extend Corollary 5.2.3 to the second factorial moment function.

Corollary 5.2.5. The second factorial moment satisfies

$$\hat{h}_2(s) = \hat{h}_1(s)^2 = (2/\rho)[1 - \hat{H}_1(s) - \theta\hat{h}_1(s)]$$

so that

$$h_2(t) \sim 2(1 + \rho^{-\frac{1}{2}})h_1(t) \sim 2(1 + \rho^{-\frac{1}{2}})\tau L(t, \rho) \quad \text{and} \quad 1 - H_2(t) \sim \tau h_2(t).$$

Proof. For the first part, apply Theorem 5.1 and Corollary 5.2.1 to get

$$\hat{h}_1(s)^2 = \left(\frac{r_1}{r_1+s}\right)\left(\frac{2z_1}{r_1}\right) = \left(\frac{2}{r_1+s}\right)\left(\frac{1-\theta r_1}{\rho}\right) = \frac{2}{\rho}[1 - \hat{H}_1(s) - \theta\hat{h}_1(s)].$$

Then apply Corollary 5.2.3 to get

$$h_2(t) = \frac{2}{\rho}[1 - H_1(t) - \theta h_1(t)] \sim \frac{2}{\rho}[\tau^2 L(t, \rho) - \theta\tau L(t, \rho)]$$
$$\sim 2\tau\left(\frac{\tau - \theta}{\rho}\right)L(t, \rho) \sim 2(1 + \rho^{-\frac{1}{2}})h_1(t).$$

The argument for $1 - H_2(t)$ is just as in Corollary 5.2.3(b).

In Theorem 5.2 and the corollaries above we treated the higher factorial moments. We can apply these results to obtain descriptions of higher ordinary moments, as we illustrate for the case $k = 2$. As before, let * denote convolution. (For additional discussion, see Sections 3 and 4 of AWb.)

Corollary 5.2.6. The second-moment function satisfies

$$\hat{m}_2(s, 0) = \hat{m}_1(s, 0) + \hat{m}_{(2)}(s, 0) = \frac{\rho\hat{h}_1(s)}{2\theta s} + \frac{\rho\hat{h}_1(s)^2}{\theta s}$$
$$= \left(\frac{\rho(1-\theta)}{2\theta^2}\right)\left(\frac{\theta}{s(1-\theta)}\hat{h}_1 + \frac{\rho}{s(1-\theta)}\hat{h}_1^2\right)$$

so that

$$\frac{m_2(t, 0)}{m_2(\infty)} = \frac{\theta}{(1-\theta)}H_1(t) + \frac{\rho}{(1-\theta)}[H_1(t) * H_1(t)]. \tag{5.2}$$

We can combine Corollaries 5.2.2 and 5.2.6 to describe how the second-moment c.d.f. $m_2(t, 0)/m_2(\infty)$ behaves as $\rho \to 1$ and as $\rho \to 0$.

Corollary 5.2.7.

$$\text{(a)} \quad \lim_{\rho \to 1} \frac{\hat{m}_2(s, 0)}{m_2(\infty)} = \lim_{\rho \to 1} \hat{h}_2(s) = 4[1 + (1 + 2s)^{\frac{1}{2}}]^{-2}$$

coinciding with the transform of the two-fold convolution of the first-moment density of RBM in (1.10) of AWa.

$$\text{(b)} \quad \lim_{\rho \to 0} \frac{\hat{m}_2(s, 0)}{m_2(\infty)} = \lim_{\rho \to 0} \hat{h}_1(s) = 2(2 + s)^{-1}$$

coinciding with the transform of an exponential density with mean $\frac{1}{2}$.

By Corollary 5.2.4 above or Corollary 3.1.3 of AWb, we see that the first three moments of the c.d.f $H_1(t)$ are $m_1 = \frac{1}{2}$, $m_2 = (1 - \theta)$ and $m_3 = 3(1 - \theta)^2 + 3\rho/4$. Note that the mean m_1 is independent of ρ, but the higher moments m_2 and m_3 depend on ρ. Note that the squared coefficient of variation is $c^2 = 2\rho + 1$, so that in some sense the distribution $H_1(t)$ gets more variable (spread out) as ρ increases. In fact, from Theorem 5.1 we can establish a stochastic comparison of this kind. We use the Laplace transform ordering; p. 22 of Stoyan (1983). Let $\hat{h}_{\rho 1}(s)$ be $\hat{h}_1(s)$ for a given ρ.

Theorem 5.3. The moment c.d.f. $H_1(t)$ is monotone in ρ in the Laplace transform stochastic ordering; i.e., $\hat{h}_{\rho_1 1}(s) \leqq \hat{h}_{\rho_2 1}(s)$ for all s when $\rho_1 < \rho_2$.

Proof. By Theorem 5.1, $\hat{h}_{\rho 1}(s)$ is increasing in $r_1(s)$. By differentiating $r_1(s)$ in (2.4) with respect to ρ, we see that it is increasing in ρ.

We conjecture that $H_1(t)$ is monotone in ρ in a stronger convex stochastic ordering.

Conjecture 5.3.1. Whenever $\rho_1 < \rho_2$, $\int_0^\infty f(t) h_{\rho_1 1}(t)\, dt \leqq \int_0^\infty f(t) h_{\rho_2 1}(t)\, dt$ for all convex real-valued functions f for which the integrals are well defined.

By Corollary 3.3.1 of AWb, $H_1(t)$ is a mixture of exponential distributions. Since a simple exponential distribution is the least element in the set of mixtures of exponentials with a given mean in the convex ordering, Conjecture 5.3.1 is established for the special case in which $\rho_1 = 0$. Theorem 5.3 establishes Conjecture 5.3.1 for the subset of convex functions of the form $f(t) = \exp(-st)$. Conjecture 5.3.1 is also consistent with our numerical results: In all observed cases the c.d.f.'s crossed exactly once.

From Corollary 5.2.1, it is clear that orderings for $H_1(t)$ in Theorem 5.3 and Conjecture 5.3.1 immediately carry over to the factorial-moment c.d.f.'s. Corollary 5.2.7 suggests that something like this is also true for the ordinary second-moment c.d.f., but we have not yet been able to prove it. By (5.2), the mean of H_2t is $(1 + 3\rho)(2 + 2\rho)$.

Conjecture 5.3.2. The ordering in Conjecture 5.3.1 for all increasing convex functions applies to the second-moment c.d.f. in (5.2).

6. The first-moment-difference c.d.f.

As for RBM in Section 9 of AWa, we can also use transforms to describe the difference component $d_1(t, i)$ of the first-moment function $m_1(t, i)$ in (2.12). Let $G_1(t, i)$ be the first-moment-difference c.d.f., defined by $G_1(t, i) = i^{-1}[1 - d_1(t, i)]$, with density $g_1(t, i)$ and associated Laplace transforms $\hat{G}_1(s, i)$ and $\hat{g}_1(s, i)$.

Theorem 6.1. The time-transformed moment function satisfies

$$\hat{m}_1(s, i) = \frac{\partial}{\partial z}\bar{P}_i(z, s)\big|_{z=1} = \hat{m}_1(s, 0) + \frac{\hat{d}_1(s, i)}{s}$$

where $\hat{d}_1(s, i) = \bar{Q}_i'(1, s)$ from (2.9), so that

$$\hat{d}_1(s, i) = i + 1 - \frac{1 - z_1^{i+1}}{1 - z_1}$$

and

$$\hat{g}_1(s, i) \equiv [1 - \hat{d}_1(s, i)]/i = \frac{1}{i}\sum_{j=1}^{i} z_1^j = \frac{1}{i}\sum_{j=1}^{i} \hat{f}(s; j, 0).$$

The representation in terms of the first-passage-time transforms $\hat{f}(s; j, 0)$ facilitates calculating the moments of $G_1(t, i)$.

Corollary 6.1.1. The first three moments of $G_1(t, i)$ are

$$m_1 = \theta(i + 1)/2, \qquad m_2 = \frac{\theta(i + 1)}{6}[(2i + 1)\theta + 3(1 - \theta)], \qquad c^2 = \frac{1}{3} + \frac{2(1 - \theta)}{\theta(i + 1)},$$

and

$$m_3 = \frac{\theta(i + 1)}{4}[i(i + 1)\theta^2 + 2(2i + 1)\theta(1 - \theta) + 2(3\rho + 2\theta^2)].$$

As in Section 12 of AWa, Corollary 6.1.1 can be used to fit convenient approximate c.d.f.'s. For $(1 - \theta)/\theta(i + 1) \geqq \frac{1}{3}$ or $i \leqq [3(1 + \rho)/(1 - \rho)] - 1$ or $\theta i \leqq 3 - 4\theta$, $c^2 \geqq 1$ and an H_2 fit is possible; otherwise a SESE (stationary-excess of a shifted exponential) fit can be considered.

We can also combine Theorems 3.1 and 6.1 to describe the asymptotic behavior of $g_1(t, i)$ as $t \to \infty$.

Corollary 6.1.2. As $t \to \infty$,

$$ig_1(t, i) \sim \frac{\rho}{2\theta}\tau L(t, \rho)[1 - \rho^{-i/2} + i\rho^{-i/2}(\rho^{-i/2} - 1)] \quad \text{and} \quad 1 - G_1(t, i) \sim \tau g_1(t, i).$$

Finally, we can combine Corollaries 5.2.3 and 6.1.2 to obtain a description of the asymptotic behavior of the general first moment function $m_1(t, i)$. (This is to be contrasted with the detailed analysis of six terms on p. 180 of Cohen (1982).)

Theorem 6.2. As $t \to \infty$,

$$m_1(\infty) - m_1(t, i) \sim \frac{\rho^{(2-i)/2}}{2\theta} \tau^2 L(t, \rho)[1 - \theta i(2/\rho\tau)^{\frac{1}{2}}]$$

or, equivalently,

$$1 - H_1(t, i) = 1 - \frac{m_1(t, i)}{m_1(\infty)} \sim \rho^{-i/2}\tau^2 L(t, \rho)[1 - \theta i(2/\rho\tau)^{\frac{1}{2}}]$$

for all i such that $1 \neq \theta i\sqrt{2/\rho\tau}$.

As in the remark after Corollary 4.2.4, Theorem 6.2 indicates (but does not prove) that $m_1(t, i)$ is decreasing in t for all t if and only if $\theta i > \sqrt{\rho\tau/2} = [\rho(1 + \sqrt{\rho})]^{\frac{1}{2}}/2$. In other words, for $M/M/1$ the *critical damping level* in Figure 1 of AWb is evidently $i_d = \sqrt{\rho}/(1 - \sqrt{\rho}) = (1 + \rho^{-\frac{1}{2}})m_1(\infty)$. This is proved in Section 8. The normalized critical damping level $i_d/m_1(\infty)$ is thus strictly decreasing in ρ, approaching 2 as $\rho \to 1$.

7. Connections to the unrestricted process

One way to analyse the $M/M/1$ queue is to relate the transition probabilities $P_{ij}(t)$ to the associated transition probabilities, say $Q_{ij}(t)$, in the unrestricted process on all the integers obtained by removing the barrier at the origin. This approach with reflection arguments was applied by Champernowne (1956) to express $P_{00}(t)$ in terms of $\{Q_{i0}(t) : i \geqq 0\}$; see p. 13 of Prabhu (1965) and p. 17 of Conolly (1975). (This approach is also related to the Wiener–Hopf factorization, which we will not discuss; see Prabhu (1980).) We do mention that the basic functions z_1 and z_2 in (2.4) have a direct interpretation for the unrestricted process. First, z_1 is the Laplace transform of the first passage down one step for both the restricted and unrestricted processes (there is no difference). Second, $1/z_2$ is the Laplace transform of the first passage time up one step in the unrestricted process.

In our time scale, the Champernowne (1956) analysis yields

$$(7.1) \qquad P_{00}(t) = Q_{00}(t) + Q_{01}(t) + 2\theta \sum_{i=1}^{\infty} Q_{i0}(t), \qquad t \geqq 0,$$

where

$$(7.2) \quad Q_{0j}(t) = \rho^{j/2} \exp(-t/\tau) \exp(-v) I_j(v) \quad \text{and}$$

$$Q_{j0}(t) = Q_{0,-j}(t) = \rho^{-j} Q_{0j}(t), \qquad j \geqq 0,$$

with v, $I_j(v)$ and τ as in Section 3. The associated Laplace transforms are

$$(7.3) \qquad \hat{Q}_{0j}(s) = \theta\rho^j z_1^j \Psi^{-1} \quad \text{and} \quad \hat{Q}_{j0}(s) = \theta z_1^j \Psi^{-1}.$$

From these expressions, it is easy to see, for $j>0$, that $Q_{0j}(t)$ and $Q_{j0}(t)$ are directly related to the first-passage times down, i.e.,

$$(7.4) \qquad Q_{j0}(t)=(t/j)f(t;j,0), \qquad j>0,$$

while for $j=0$

$$(7.5) \qquad Q_{00}(t)=\exp(-t/\tau)\exp(-v)I_0(v) \quad \text{and} \quad \hat{Q}_{00}(s)=\theta\Psi^{-1}.$$

We remark that (7.4) can be connected to the general theory for Lévy processes; see Theorem 6, p. 81 of Prabhu (1980).

In this section we go further and show that $P_{00}(t)$ can be expressed solely in terms of $Q_{00}(t)$. We were originally motivated by the desire to obtain an $M/M/1$ analog and a better understanding of the RBM result in (4.4) of AWa, which states that

$$(7.6) \qquad h_1(t)=2\gamma(t)-\gamma_e(t), \qquad t\geqq 0,$$

where $\gamma(t)$ is the gamma density with mean 1 and shape parameter $\frac{1}{2}$, and $\gamma_e(t)=\int_0^t \gamma(u)\,du$ is the associated-stationary excess density. These objectives are met here in Theorem 7.2 and Corollary 7.2.2 below.

For the $M/M/1$ queue, it is useful to focus on the function

$$(7.7) \qquad \gamma_\rho(t)=\theta^{-1}Q_{00}(t)=\theta^{-1}\exp(-t/\tau)\exp(-v)I_0(v), \qquad t\geqq 0,$$

which from (7.5) has Laplace transform $\hat{\gamma}_\rho(s)=\Psi(s)^{-1}$. Since $Q_{00}(t)>0$, it is not difficult to see that $\gamma_\rho(t)$ is a bona fide probability density function.

Theorem 7.1. The function $\gamma_\rho(t)$ in (7.7) with transform $\hat{\gamma}_\rho(s)=\Psi^{-1}=2/(r_1+r_2)$ is a probability density function with moments $m_1=(1-\theta)$, $m_2=2^{-1}(1+4\rho+\rho^2)$ and $m_3=(3(1-\theta)/2)(1+8\rho+\rho^2)$, and $\lim_{\rho\to 1}\gamma_\rho(t)=\gamma(t)=(2\pi t)^{-\frac{1}{2}}\exp(-t/2)$.

Proof. The moments are easily deduced from the transform $\hat{\gamma}_\rho(s)=\Psi^{-1}$. The convergence as $\rho\to 1$ closely parallels Theorem 3.5(a).

Corollary 7.1.1. The expected cumulative time spent at the origin by the unrestricted process over all time is $\int_0^\infty Q_{00}(t)\,dt=\theta$.

Our main result in this section expresses $P_{00}(t)$ directly in terms of $Q_{00}(t)$ (and its derivative $Q'_{00}(t)$).

Theorem 7.2. $P_{00}(t)=2\theta+2(1-\theta)Q_{00}(t)-\int_t^\infty Q_{00}(u)\,du+\theta^2 Q'_{00}(t)$.

Proof. We start with (7.1). By transforms, it is not difficult to show that

$$(7.8) \qquad 2\theta\sum_{i=1}^{\infty} Q_{i0}(t)=\theta[1-Q_{00}(t)]+\int_0^t Q_{00}(u)\,du.$$

In particular, from (7.3) the transform of the left side is

$$2\theta\sum_{i=1}^{\infty}\hat{Q}_{i0}(s)=2\theta^2\hat{\gamma}_\rho(s)\sum_{i=1}^{\infty}z_1^i=\hat{\gamma}(s)\hat{P}_{00}(s),$$

while the transform of the right side is

$$\theta[s^{-1} - \hat{Q}_{00}(s)] + s^{-1}\hat{Q}_{00}(s) = s^{-1}\theta\hat{\gamma}_\rho(s)[\hat{\Psi} + 1 - \theta s] = \hat{\gamma}_\rho(s)\theta r_1/s = \hat{\gamma}_\rho(s)\hat{P}_{00}(s).$$

From the basic equation for motion (Chapman–Kolmogorov equations), we have in our time scale

$$Q'_{00}(t) = \theta^{-2}Q_{01}(t) - (1-\theta)\theta^{-2}Q_{00}(t). \tag{7.9}$$

Combining (7.1), (7.8) and (7.9) completes the proof.

We can combine Theorems 7.1 and 7.2 to express $P_{00}(t)$ in terms of $\gamma_\rho(t)$.

Corollary 7.2.1. *$P_{00}(t) = 2\theta + \theta(1-\theta)[2\gamma_\rho(t) - \gamma_{\rho e}(t)] + \theta^3\gamma'_\rho(t)$ where $\gamma_{\rho e}(t)$ is the stationary-excess density associated with $\gamma_\rho(t)$, i.e., $\gamma_{\rho e}(t) = (1-\theta)^{-1} \int_0^t \gamma_\rho(u)\,du$.*

From the conservation law for the first-moment function $m_1(t, i)$ to be established in Theorem 8.1, the first-moment density (starting at the origin) can be expressed as

$$\rho h_1(t) = \theta^{-1}P_{00}(t) - 2, \tag{7.10}$$

so that we obtain an $M/M/1$ generalization of (7.6).

Corollary 7.2.2. *$\rho h_1(t) = (1-\theta)[2\gamma_\rho(t) - \gamma_{\rho e}(t)] + \theta^2\gamma'_\rho(t)$.*

Note that (7.6) is obtained from Corollary 7.2.2 by simply letting $\rho \to 1$.

By Corollary 3.3.1 of AWb, $h_1(t)$ is completely monotone.

Corollary 7.2.3. *$P_{00}(t) - 2\theta$ is completely monotone and thus decreasing and convex with $\int_0^\infty [P_{00}(t) - 2\theta]\,dt = \rho\theta$.*

8. Shape of the moment functions

Our object in this section is to rigorously determine the shape of the moment functions $m_k(t, i)$ when $i > 1$. We apply results in Chapter 9 of van Doorn (1980) to obtain a description paralleling our previous description for RBM in Section 8 of AWa. The shape of $m_1(t, i)$ is essentially the same as for RBM.

Let $m'_k(t, i)$ be the derivative of $m_k(t, i)$ with respect to t and so forth. We begin with a basic conservation law; e.g., (2.7) on p. 178 of Cohen (1982). This conservation law connects $m_1(t, i)$ and its derivatives directly to $P_{i0}(t)$ and its derivatives.

Theorem 8.1 (conservation law). *$m'_1(t, i) = (2\theta^2)^{-1}[P_{i0}(t) - (1-\rho)]$ or, equivalently,*

$$m_1(t, i) = i + (2\theta^2)^{-1}\int_0^t P_{i0}(u)\,du - t\theta^{-1}.$$

First proof. From the Chapman–Kolmogorov equations, $P'_{ij}(t) = P_{i,j+1}(t) - (1+\rho)P_{ij}(t) + \rho P_{i,j-1}(t)$ for $j \geqq 1$ without time scaling, so that elementary algebra yields (still without time scaling)

$$m'_1(t, i) = \sum_{j=1}^{\infty} jP'_{ij}(t) = P_{i0}(t) - (1-\rho).$$

We obtain the second expression by integrating, using $m_1(0, i) = i$.

Second proof. As in the alternate proof of Corollary 5.2.1, $m_1(t, i)$ coincides with the expected virtual waiting time $EW(t)$, but here under the condition that there are initially i customers in the system with unspecified service times. Then, without time scaling,

$$EW(t) = i + \rho t - t + EI(t) = i + \rho t - t + \int_0^t P_{i0}(u)\,du.$$

We apply Lemma 9.4.1(ii) and Theorem 9.4.3(ii) of van Doorn (1980) to determine what the shape $m_1(t, i)$ must be. Theorem 8.1 and (4.3) then allow us to deduce what the shape of $P_{i0}(t)$ and $P_{0i}(t)$ must be as well.

Theorem 8.2 (van Doorn). If $m'_1(t, i) \leqq 0$, then $m''_1(t, i) \geqq 0$.

Corollary 8.2.1 (van Doorn). If $m'_1(t, i) \geqq 0$, then $m'_1(u, i) \geqq 0$ for all $u \geqq t$.

The asymptotic theory (Theorem 6.2) then shows what the shape of $m_1(t, i)$ must be. See Figures 1 and 2 of AWa. (Alternatively we could invoke pp. 63–64 of van Doorn (1980) together with his Theorem 8.2 above.)

Corollary 8.2.2. (a) For $i \geqq (1 + \rho^{-\frac{1}{2}})m_1(\infty)$, $P_{i0}(t)$ and $P_{0i}(t)$ are strictly increasing, and $m_1(t, i)$ is strictly decreasing and convex for all t.

(b) For $1 \leqq i \leqq (1 + \rho^{-\frac{1}{2}})m_1(\infty)$, there is a time t_1 such that $P_{i0}(t_1) = \rho^{-i}P_{0i}(t_1) = 1 - \rho$, $P_{i0}(t)$ and $P_{0i}(t)$ are increasing and $m_1(t, i)$ is decreasing and convex on $(0, t_1)$, and $m_1(t, i)$ is increasing on (t_1, ∞).

We also apply a result about the first-passage-time density $f(t; 0, i)$ by Keilson (1979) to further describe the shape.

Theorem 8.3. For $1 \leqq i \leqq (1 + \rho^{-\frac{1}{2}})m_1(\infty)$, there is a time $t_2 > t_1$ such that $P_{0i}(t)$ and $P_{i0}(t)$ are increasing and $m_1(t, i)$ is convex on (t_1, t_2), while $P_{0i}(t)$ and $P_{i0}(t)$ are decreasing and $m_1(t, i)$ is concave on (t_2, ∞).

Proof. By first principles, $P_{0i}(t) = \int_0^t f(t-s; 0, i)P_{ii}(s)\,ds$, from which we can deduce that $P_{0i}(t)$ is unimodal. First, from the spectral representation as discussed in van Doorn (1980), $P_{ii}(t)$ is a mixture of exponentials and so decreasing and thus unimodal. Second, by pp. 59, 70 of Keilson (1979), $f(t; 0, i)$ is distributed as the convolution of exponentials, and so is strongly unimodal, i.e., the convolution of it with any unimodal density is again unimodal.

Remark 8.1. The general first-passage-time density $f(t; i, j)$ is known to be unimodal (but not strongly unimodal); Keilson (1981). See Section 10 and Remark 10.1 for further discussion about shape.

Following the first proof of Theorem 8.1, we can also describe the derivatives of the second-moment function.

Theorem 8.4. The derivatives of the second-moment function with respect to t are

$$\begin{aligned}
\text{(a)}\quad m_2'(t, i) &= -2(1-\rho)m_1(t, i) - P_{i0}(t) + 1 + \rho \\
&= -2(1-\rho)m_1(t, i) - m_1'(t, i) + 2\rho \\
&= 2(1-\rho)[m_1(\infty) - m_1(t, i)] - m_1'(t, i) \\
\text{(b)}\quad m_2''(t, i) &= -2(1-\rho)P_{i0}(t) - 2(1-\rho)^2 - P_{i0}'(t).
\end{aligned}$$

From this analysis, we see that $m_k'(t, i)$ can be expressed in terms of $m_{k-1}(t, i)$, $m_{k-2}(t, i), \cdots, m_1(t, i)$, $P_{i0}(t)$, so that the kth derivative of $m_k(t, i)$ can be expressed solely in terms of $P_{i0}(t)$ and its first $k-1$ derivatives. We obtain the following representation result.

Theorem 8.5. For each $k \geqq 1$, $m_k(t, i)$ can be expressed solely (as a polynomial) in terms of $P_{i0}(t)$ and its first $k-1$ derivatives.

Combining Theorem 8.5 and Corollary 4.2.7, we see that $m_k(t, i)$ can be expressed solely in terms of the busy-period c.d.f. $B(t) \equiv F(t; 1, 0)$.

9. Renewal-process operators

An interesting feature of AWa is the way various quantities of interest are related by certain renewal-process operators; see Corollaries 1.5.1 and 1.5.2, Remark 4.5 and Theorem 7.2 of AWa. For example, the two c.d.f.'s $H_2(t)$ and $G_2(t)$ associated with the second-moment function of RBM turn out to be simply the stationary-excess c.d.f.'s of the corresponding c.d.f.'s $H_1(t)$ and $G_1(t)$ associated with the first moment of RBM. The purpose of this section is to present wherever possible $M/M/1$ analogs. (The same result does not hold for $G_2(t)$ in $M/M/1$.) We omit the proofs, which involve relatively elementary algebra.

In terms of transforms, the *stationary-excess operator* SE maps the transform $\hat{f}(s)$ of a density $f(t)$ on $[0, \infty)$ with mean m_1 into the transform of another density according to

$$\text{(9.1)}\qquad \mathrm{SE}(\hat{f})(s) = [1 - \hat{f}(s)]/m_1 s.$$

The normalized *renewal-excess operator* RE maps the same transform $\hat{f}(s)$ according to

$$\text{(9.2)}\qquad \mathrm{RE}\,(\hat{f})(s) = \left(\frac{2}{c^2 - 1}\right)\left[\frac{\hat{f}(s)}{1 - \hat{f}(s)} - \frac{1}{m_1 s}\right].$$

Aside from the normalization, $s[\mathrm{RE}\,(\hat{f})]$ is the transform of $U(t)-t/m_1$ where $U(t)$ is the renewal function determined by renewal intervals with density $f(t)$. The normalization is the steady-state limit. From Brown (1980), (1981), we know that $U(t)-t/m_1$ is increasing in t when $f(t)$ is IMRL (increasing mean residual life). That will be the case for the densities $f(t)$ we consider, so that $\mathrm{RE}\,(\hat{f})$ as well as $\mathrm{SE}\,(\hat{f})$ will be the transform of a bona fide probability density.

Theorem 9.1. The stationary-excess operator SE satisfies:

(a) $\mathrm{SE}\,(\hat{b})=\hat{h}_1=\hat{f}_{\varepsilon 0}\hat{b}$; (b) $\mathrm{SE}\,(\hat{f}_{\varepsilon 0})=(\hat{f}_{\varepsilon 0})^2\hat{b}=\mathrm{SE}\,(\hat{h}_i)$;

(c) $\hat{h}_{(2)}=\hat{b}\,\mathrm{SE}\,(\hat{h}_1)=\hat{h}_1^2$; (d) $\hat{h}_2=\dfrac{1}{(1-\theta)}\mathrm{SE}\,(\hat{h}_1)-\dfrac{\theta}{(1-\theta)}\hat{h}_1$;

(e) $\mathrm{SE}\,(\hat{\gamma}_\rho)=\hat{\gamma}_\rho\hat{f}_{\varepsilon 0}$.

Theorem 9.2. The renewal-excess operator RE satisfies:

(a) $\mathrm{RE}\,(\hat{b})=\hat{h}_1=\hat{f}_{\varepsilon 0}\hat{b}$; (b) $\mathrm{RE}\,(\hat{f}_{\varepsilon 0})=\hat{f}_{\varepsilon 0}$;

(c) $\mathrm{RE}\,(\hat{h}_1)=\hat{h}_1$; (d) $\mathrm{RE}\,(\hat{\gamma}_\rho)=\hat{h}_1$.

Even for the general $GI/G/1$ model, we can study $P_{00}(t)$ by considering the alternating renewal process of successive idle and busy periods; see p. 82 of Cox (1962). Let $\hat{u}(s)$ and $\hat{b}(s)$ be the Laplace–Stieltjes transforms of the idle-period and busy-period distributions. The renewal argument yields

$$\hat{P}_{00}(s)=\frac{1-\hat{u}(s)}{s[1-\hat{u}(s)\hat{b}(s)]}. \tag{9.3}$$

For the special case of the $M/G/1$ model in our time scale an idle time has the distribution of an interarrival time, so that $\hat{u}(s)=\rho/(\rho+2\theta^2 s)$, and

$$\hat{P}_{00}(s)=\left(\frac{2\theta^2}{\rho}\right)\left(\frac{\hat{u}(s)}{1-\hat{u}(s)\hat{b}(s)}\right) \tag{9.4}$$

and the expected cumulative idle time $EI_0(t)$ has transform

$$\hat{I}_0(s)=s^{-1}\hat{P}_{00}(s)=\left(\frac{2\theta^2}{\rho}\right)\left(\frac{\hat{u}(s)}{s[1-\hat{u}(s)\hat{b}(s)]}\right), \tag{9.5}$$

which has the simple interpretation that $EI_0(t)$ equals the mean interarrival time multiplied by the expected number of full idle periods in $[0,t]$ for the alternating renewal process. The $M/G/1$ structure evidently gives this relatively simple form.

It turns out that a more remarkable formula holds for the $M/M/1$ model.

Theorem 9.3.

$$\hat{I}_0(s)=s^{-1}\hat{P}_{00}(s)=\frac{2\theta^2 z_1}{s(1-z_1)}=(2\theta)^2\frac{\hat{b}(s)}{s[1-\hat{b}(s)]}.$$

Theorem 9.3 can be interpreted as $EI_0(t)$ equalling the mean service time multiplied by the mean number of busy periods completed up to time t in an ordinary renewal process with time between renewals having the busy-period distribution. We have yet to develop a direct probabilistic interpretation.

We can apply the renewal function structure in Theorem 9.3 to get bounds and approximations for $EI_0(t)$, e.g., via p. 46 of Cox (1962), Lorden (1970) and Brown (1980), (1981). However, we can do just as well directly via Corollary 4.2.6. Such a direct approach is used by Kumar and Wong (1987).

10. Connections to RBM

The scaling of space and time in (1.6) of AWb makes RBM appear as a special case of the $M/M/1$ queue-length process; i.e., the limit as $\rho\to 1$ appears as a proper limit (Corollary 5.2.2 above), so that RBM appears explicitly as the case $\rho=1$. However, in making the connection to RBM from results for the $M/M/1$ queue here, it is important to remember that in this paper we scaled time but not space. To obtain results for RBM, we thus need to introduce the space scaling by θ and let $\rho\to 1$.

It is even more interesting, though, to establish a connection between RBM and the $M/M/1$ model for $\rho<1$. Such a connection would help us understand how to use RBM to approximate queues; e.g., see Duda (1984) and Whitt (1982). At least in part, the connection between RBM and $M/M/1$ is embodied in the relation between two systems of quadratic equations. First, RBM is characterized by the two functions $r_1\equiv r_1(s)=(1+2s)^{\frac{1}{2}}+1$ and $-r_2\equiv -r_2(s)=(1+s)^{\frac{1}{2}}-1$, which are the roots of the quadratic equation (as functions of s) $r^2-2r-2s=0$, see Section 1.3 of AWa, while the queue-length process in the $M/M/1$ queue with scaled time is characterized by the two functions $z_1=z_1(s)=\rho^{-1}(1-\theta r_1)$ and $z_2\equiv z_2(s)=\rho^{-1}(1+\theta r_2)$, which are the roots of the quadratic equation $\rho z^2-(1+\rho+2\theta^2 s)z+1=0$; see (2.4) and (2.5).

At first glance, there appears to be no simple connection between the RBM system of quadratic equations with its roots $r_1(s)$ and $r_2(s)$ and the $M/M/1$ system of quadratic equations with its roots $z_1(s)$ and $z_2(s)$. Note that r_1 and r_2 for $M/M/1$ in (2.4) converge to r_1 and r_2 for RBM as $\rho\to 1$, but RBM has no counterparts to z_1 and z_2. However, Corollary 3.4.1 suggests a connection. We propose the following *operational calculus*: first, since $M/M/1$ has been time scaled but not space scaled, when the RBM state is x, let the $M/M/1$ state be $n=x/\theta$. (Assume that x is a multiple of θ so that the $M/M/1$ state is an integer.) Next, starting with some RBM quantity based on the functions $r_1(s)$ and $r_2(s)$ in exponential form $\exp(\theta r_1(s))$ and $\exp(-\theta r_2(s))$, obtain the corresponding $M/M/1$ quantity for any given ρ by letting $z_1(s) = \exp(-\theta r_2(s))$ and $z_2(s) = \exp(\theta r_1(s))$ so that $z_1 z_2 = \rho^{-1} = \exp(2\theta) = \exp(1-\rho)$ where $z_1(s)$ and $z_2(s)$ are understood to satisfy the $M/M/1$ quadratic equations for the given ρ, rather than the original RBM quadratic equations. Finally, when the RBM quantities r_1 and r_2 appear alone (not in the exponential form above), simply replace them by their $M/M/1$ counterparts.

To illustrate, suppose that we want to identify the transform $\hat{f}_M(s; n, 0)$ of the first passage time to 0 from n in the $M/M/1$ model with traffic intensity ρ. Of course this transform is z_1^n as given in Theorem 3.1(a), but suppose that we only know about RBM. For RBM the corresponding first-passage-time transform is $\hat{f}_R(s; x, 0) = \exp(-xr_2(s))$; see (1.7) of AWa and (3.5) here. Using the operational calculus specified above, we set

$$\hat{f}_M(s; n, 0) = \hat{f}_R(s; n\theta, 0) = \exp(-n\theta r_2(s)) = z_1(s)^n, \tag{10.1}$$

which produces the correct result. Similarly, the operational calculus produces (3.2)–(3.4) for $M/M/1$ given (3.5)–(3.8) for RBM.

An obvious question is: how do we know when this operational calculus relating $M/M/1$ and RBM will work? From Section 3, we see that it works for the first-passage times both up and down. By Theorem 4.3, we have shown that the $M/M/1$ transition probabilities can be expressed solely in terms of these first-passage times, so that we see it works for the basic Markov transition probabilities too. In particular, suppose that we want to identify the $M/M/1$ complementary c.d.f. $\sum_{j=n}^{\infty} P_{0j}(t)$. By Corollary 4.2.5, we know it is $\rho^n F(t; n, 0)$ with transform $s^{-1}\rho^n z_1^n$. However, suppose that we start with RBM. From (1.8) and Section 1.7 of AWa, we know that the corresponding transform for RBM is $s^{-1}\exp(-r_1 x) = s^{-1}\exp(-(r_2 - 2)x)$. Applying the operational calculus yields $s^{-1}\rho^n z_1^n$ for $n\theta = x$ as desired. The correspondence for the c.d.f. $\sum_{j=0}^{n} P_{ij}(t)$ with general initial condition follows from Theorem 4.3 and the analysis above.

The operational calculus also applies to p.m.f.'s and densities, but with an obvious modification to account for the space scaling. In particular, to obtain the $M/M/1$ p.m.f. $P_{0n}(t)$ from the RBM density $g(x; t, 0)$ we need to multiply the RBM density by θ. To see this, start with $M/M/1$ and apply the operational calculus to obtain

$$\begin{aligned} \hat{P}_{0n}(s) &= \frac{\rho^n z_1^n - \rho^{n+1} z_1^{n+1}}{s} = \frac{\rho^n z_1^n}{s} = \frac{\rho^n z_1^n}{s}(1 - \rho z_1) = \frac{\rho^n z_1^n \theta r_1}{s} \\ &= \theta s^{-1} r_1(s) \exp(-r_1(s)x) = \theta \hat{g}(x; s, 0). \end{aligned} \tag{10.2}$$

Of course, in general the operational calculus for going from RBM to $M/M/1$ must be applied with caution because ρ in $M/M/1$ is replaced by 1 in RBM and thus cannot be identified from RBM. The operational calculus can thus only be a guide.

The limitations are illustrated by the fomulas describing the asymptotic behavior as $t \to \infty$, e.g., Corollary 5.2.3. However, as we noted above, in many cases no extraneous ρ terms arise.

If the operational calculus is valid for $M/M/1$ p.m.f.'s and RBM densities, then we must be able to go from $M/M/1$ to RBM. This limit for $M/M/1$ p.m.f.'s as $\rho \to 1$ does not follow directly from the standard heavy-traffic limit theorems in Iglehart and Whitt (1970); the standard heavy-traffic theorems only yield convergence of the associated c.d.f.'s. However, convergence of the p.m.f.'s (a local limit theorem) can be established, e.g., via the Bessel function representation for $P_{ij}(t)$. Another classic proof would parallel Section 1.3 of Itô and McKean (1965). We present a different proof which is based on the following conjecture, and so is incomplete. We believe that both the conjecture and the proof are of considerable interest, however.

Conjecture 10.1. For each i and j, the p.m.f. $P_{ij}(t)$ is a unimodal function of t.

In Corollary 8.2.2 and Theorem 8.3, we have established Conjecture 10.1 for the cases in which $i = 0$, $j = 0$ and $i = j$. Consequently, the local limit theorem below is established for these cases.

Remark 10.1. Conjecture 10.1 is known to be invalid for general birth-and-death processes, e.g., Rosenlund (1978) and p. 97 of Karlin (1964). A related result for first-passage times is contained in Keilson (1981). Our proof of Theorem 10.1 does not depend on the full strength of Conjecture 10.1

We also employ the following order properties of the p.m.f.'s.

Lemma 10.1. For all positive t, i and j, $P_{i,i+j}(t) \leqq P_{0j}(t) \leqq \rho^j P_{00}(t)$.

Proof. To establish the second inequality, note that the p.m.f.'s starting at 0 increase in the monotone-likelihood-ratio stochastic order as t increases, Theorem 4.5(b) of Keilson and Sumita (1982), so that $P_{0j}(t)/P_{00}(t)$ increases in t. Since $P_{0j}(\infty)/P_{00}(\infty) = \rho^j$, the second inequality holds. To establish the first inequality, apply a coupling argument, as in Section 11 of AWa. Consider two processes, one starting at i and the other at 0, with all potential transitions in both processes generated by a single Poisson process with intensity $(1+\rho)/2\theta^2$ ($\lambda + \mu$ before time scaling). Given an event in the Poisson process, both processes go up with probability $\rho/(1+\rho)$; and both go down with probability $1/(1+\rho)$, with a down transition resulting in no change at the origin. The proof is completed by induction on i, j and the number of transitions in the Poisson process.

It is significant that the refined limit behavior in Theorem 10.1 below is not valid for other $GI/G/1$ systems. (It is not difficult to show this for $M/G/1$ systems.) This confirms that $M/M/1$ is related to RBM in a special way.

Theorem 10.1. For any $t>0$,

$$\begin{aligned}\lim_{\rho\to 1}\theta^{-1}P_{[x\theta^{-1}],[y\theta^{-1}]}(t)&=g(y;t,x)\\&=t^{-\frac{1}{2}}\left[\phi\left(\frac{-y+x-t}{\sqrt{t}}\right)+\exp(-2y)\phi\left(\frac{-y-x+t}{\sqrt{t}}\right)\right]\\&\quad+2\exp(-2y)\Phi\left(\frac{-y-x+t}{\sqrt{t}}\right),\end{aligned}$$

where $g(y;t,x)$ is the density of RBM at time t starting at x.

Proof based on Conjecture 10.1 (*and thus complete when* $x=0$, $y=0$ *or* $x=y$). First, note that $\theta^{-1}P_{[x\theta^{-1}],[y\theta^{-1}]}(t)$ can be regarded as a probability density function in y (in the histogram form) for each θ and t. By Stone (1963) or Iglehart and Whitt (1970), for each t the associated c.d.f.'s converge to the RBM c.d.f. $G(y;t,x)$ associated with the density $g(y;t,x)$ for all x, y and t. This serves to identify the limit of any convergent subsequence of the sequence of normalized p.m.f.'s. (This identification can also be established directly with the transforms: the time transform of $\sum_{j=n}^{\infty}P_{ij}(t)$ is $\rho^n\hat{F}(s;n,0)/\hat{f}(s;0,i)$ by Theorem 4.3(a). Convergence of the space-scaled version then follows as in the proof of Corollary 3.4.1.)

The proof is completed by a compactness argument. It suffices to show for any $\varepsilon>0$ and ρ_0 with $0<\rho_0<1$ that the set $\{\theta^{-1}P_{[x\theta^{-1}],[y\theta^{-1}]}(t):\rho_0\leqq\rho<1\}$ of real-valued functions on the interval $[\varepsilon,\infty)$ is compact in a topology inducing pointwise convergence, i.e., that every subsequence has a further subsubsequence converging pointwise to an integrable limit. Since the p.m.f.'s are unimodal in y for each θ and t, see Keilson and Kester (1978), so is any limit as $\theta\to 0$, which implies that any limit function is integrable. As in Theorem 3.5, we can apply the Lebesgue dominated convergence theorem (with Lemma 10.1) to get convergence of the associated c.d.f.'s from convergence of the densities. Hence, the limit of any convergent subsequence must have c.d.f. $G(y;t,x)$ and thus must be $g(y;t,x)$.

We established the desired compactness by applying Conjecture 10.1. The unimodality implies that $P_{in}(t)$ is a function of bounded variation, so that it can be expressed as the difference of two non-decreasing functions, say, $P_{in}(t)=A_{in}(t)-B_{in}(t)$ where $B_{in}(0)=0$. We can thus apply the Helly selection theorem with each monotone component separately; p. 227 of Billingsley (1968). Moreover, the unimodality implies that

$$\sup_{t\geqq\varepsilon}\max\{A_{in}(t),B_{in}(t)\}\leqq\sup_{t\geqq\varepsilon}P_{in}(t), \tag{10.3}$$

so that to establish compactness it suffices to demonstrate boundedness for the normalized form of $P_{in}(t)$, i.e., to show that

$$\sup_{\substack{t\geqq\varepsilon\\ \rho_0\leqq\rho<1}}\theta^{-1}P_{[x\theta^{-1}],[y\theta^{-1}]}(t)<\infty. \tag{10.4}$$

By Lemma 10.1, to establish (10.4) it suffices to consider only the case $x = y = 0$. However, this special case is covered by Theorem 3.5 and Corollary 4.2.3:

$$\begin{aligned}\lim_{\rho\to 1}\theta^{-1}P_{00}(t) &= \lim_{\rho\to 1}\frac{2\rho[1-B(t)]}{1-\rho}+2\\ &= 2t^{-\frac{1}{2}}\phi(t^{\frac{1}{2}})-2[1-\Phi(t^{\frac{1}{2}})]+2 \qquad (10.5)\\ &= 2t^{-\frac{1}{2}}\phi(t^{\frac{1}{2}})+2\Phi(t^{\frac{1}{2}})=g(0;t,0).\end{aligned}$$

Note that (10.5), Corollary 4.2.3 and Theorem 10.1 explain the form of the heavy-traffic limit for the busy-period c.d.f. in Theorem 3.5.

11. Summary

In this paper we have developed some new ways to analyze the transient behavior of the $M/M/1$ queue. Perhaps the main idea is that some transient results of interest can be obtained quite easily without deriving or applying the complete expression for $P_{ij}(t)$. First, the transform $\hat{P}_{i0}(s)$ in (2.6) emerges early in the analysis. Then the transform $\hat{f}(s, i, 0)$ for the first-passage time down can easily be obtained from it, as indicated in the proof of Theorem 3.1(a). Next, by various means (Section 4), we can obtain interesting expressions for $P_{0j}(t)$, one of which is solely in terms of the first-passage-time c.d.f.'s $F(t; i, 0)$. Although this procedure does not seem to yield new ways to get $P_{ij}(t)$ when neither $i = 0$ nor $j = 0$, from Theorem 4.3 we see that $P_{ij}(t)$ can also be expressed solely in terms of the first-passage-time distributions (up as well as down). The connection to the first-passage times can be explained in various ways: via the Laplace transform relations here, via the probabilistic proofs in AWa, b and via duality; see Siegmund (1976), Chapter 3 of van Doorn (1980) and Clifford and Sudbury (1985).

A special role is played by the initial condition starting at the origin. Not only do we develop nice expressions for $P_{0j}(t)$ in Section 4, but we obtain nice expressions for the factorial moments $m_{(k)}(t, 0)$ starting at the origin in Section 5 (and in AWb). Focusing on the zero initial condition yields the new factorization in Section 2, which is exploited in Sections 5 and 6 to produce nice characterizations of the moment function $m_1(t, i)$ for general initial state i. The asymptotic behavior of the first moment $m_1(t, i)$ as $t\to\infty$ in Theorem 6.2 is especially easy to derive this way.

A major goal in this paper has been to express many transient descriptions in terms of basic building blocks. We have just reviewed how we can go from $P_{i0}(t)$ to $F(t; i, 0)$ to $P_{0n}(t)$. Formula (4.3) shows that we can go back and forth between $P_{i0}(t)$ and $P_{0i}(t)$. By Corollary 4.2.3, we can in turn express the busy-period c.d.f. $B(t) \equiv F(t; 1, 0)$ directly in terms of $P_{00}(t)$. (More generally, for the $M/G/1$ queue the connection between $F(t; 1, 0)$ and $P_{00}(t)$ follows from the theory of regenerative phenomena by regarding $P_{00}(t)$ as the p-function; see Kingman (1966), (1972), e.g., p. 432 of the former.) Section 8 shows that we can also express the moment functions $m_k(t, i)$ for all k and i in terms of $P_{i0}(t)$ (and its derivatives), which can in turn be expressed in terms of $P_{00}(t)$. Theorem 7.2 shows that we can further express $P_{00}(t)$ solely in terms of the corresponding transition probability function in the unrestricted process, $Q_{00}(t)$, or equivalently in terms of the density function $\gamma_\rho(t) = \theta^{-1}Q_{00}(t)$ in (7.7), a principal ingredient being the modified Bessel function $I_0(v)$. It is rather remarkable that all these functions of time can be expressed in terms of $\gamma_\rho(t)$. Figure 1 depicts the logical connections.

The full transition probability function $P_{in}(t)$ can be expressed in terms of $Q_{0n}(t)$, $-\infty < n < \infty$, (Champernowne (1956)) or $F(t; n, 0)$ and $F(t; 0, n)$ (Theorem 4.3). We remark that we can obtain $Q_{in}(t)$ from $P_{in}(t)$ via $Q_{in}(t) = \lim_{m\to\infty} P_{m,m+n-i}(t)$.

$$\begin{tikzcd}
e^{-v}I_0(v) \arrow[d, leftrightarrow] & & P_{i0}(t) \arrow[dl] \arrow[d, leftrightarrow] \arrow[dr] & \\
Q_{00}(t) \arrow[r] \arrow[d, leftrightarrow] & P_{00}(t) \arrow[r] & P_{0n}(t) \arrow[d, leftrightarrow] & m_k(t, i) \\
\gamma_\rho(t) & & F(t; 1, 0) \arrow[ul, leftrightarrow] &
\end{tikzcd}$$

Figure 1. Logical relations among formulas for the $M/M/1$ transient behavior. ($A \to B$ means that B can be expressed directly in terms of A.)

The last three sections of this paper present supplementary results. In Section 8 we apply Section 9 of van Doorn (1980) to show that the shape of the $M/M/1$ moment functions is indeed essentially the same as for RBM; cf. Section 8 of AWa. We also show that $m_k(t, i)$ can be expressed in terms of $P_{00}(t)$ there. In Section 9 we establish some curious connections to renewal processes. Finally, in Section 10 we propose an operational calculus for obtaining $M/M/1$ formulas directly from RBM formulas. Overall, we have succeeded in producing $M/M/1$ analogs for almost all the RBM formulas derived in AWa. (We have not obtained a nice characterization of the second-moment difference c.d.f. $G_2(t)$; it does not satisfy the nice stationary-excess relation in (7.9) of AWa.) It is well known that the $M/M/1$ queue-length process is the discrete analog of RBM; e.g., see Stone (1963) and references there. With the proper scaling, it is not difficult to see the close connection in the descriptive characteristics.

In conclusion, the $M/M/1$ model can be viewed from quite a few different perspectives. The structure here should also be available from other approaches; e.g., Bessel functions, the spectral representation and martingales. A detailed analysis of the $M/M/1$ model is important in part because it is an elementary special case of so many models.

Appendix: Discussion of the time scaling

In this appendix we discuss the time scaling in (2.1). Our time scale measures time in units of $1/(2\theta^2) = 2/(1-\rho)^2$ mean service times. Thus a time t in the original $M/M/1$ model with arrival rate λ and service rate μ is transformed into $2\theta^2\mu t$ in the new scale with arrival rate $\rho/(2\theta^2)$ and service rate $1/(2\theta^2)$. As indicated in Section 2.2 of AWb, this scaling is designed to reveal the close connections to RBM as $\rho \to 1$. As $\rho \to 1$, the new parameter θ in (2.4) satisfies $\theta = (1-\rho)/2 \to 0$.

A simple prescription for converting to this time scale is to transform the arrival rate λ into $\rho/(2\theta^2)$ and the service rate μ to $1/(2\theta^2)$ in any equation involving λ and μ. To go back to the original (λ, μ)-time scale, replace θ by $\sqrt{1/2\mu}$ and time t by $t/(2\theta^2\mu)$. In any time scale, $\rho = (\text{arrival rate})/(\text{service rate})$.

For example, in the original time scale the double transform $\bar{P}_i(z, s)$ in (2.7) is

$$\text{(A.1)} \qquad \bar{P}_i(z, s) = \frac{z^{i+1} - \mu(1-z)\hat{P}_{i0}(s)}{\lambda(z-\xi)(\eta - z)}$$

where

$$\text{(A.2)} \qquad \xi = \frac{(\lambda+\mu+s) - \sqrt{(\lambda+\mu+s)^2 - 4\lambda\mu}}{2\lambda}$$

and $\eta = \mu/\lambda\xi$; see (1.29) on p. 9 of Prabhu (1965). We obtain (2.7) and the quadratic equation $\rho z^2 - (1 + \rho + 2\theta^2 s)z + 1 = 0$ by simply substituting $\rho/(2\theta^2)$ for λ and $1/(2\theta^2)$ for μ in (A.1)

This prescription also works for the Chapman–Kolmogorov equations of motion: without time scaling we have

$$\text{(A.3)} \qquad P_n'(t) = \mu P_{n+1}(t) - (\lambda+\mu)P_n(t) + \lambda P_{n-1}(t);$$

after time scaling, we have

$$\text{(A.4)} \qquad 2\theta^2 P_n'(t) = P_{n+1}(t) - (\rho+1)P_n(t) + \rho P_{n-1}(t).$$

To illustrate how we can go back and forth this way, note that before time scaling the conservation law in Theorem 8.1 is

$$\text{(A.5)} \qquad m_1'(t, i) = \lambda - \mu[1 - P_{i0}(t)]$$

and the busy period density in Theorem 3.1(b) is

$$\text{(A.6)} \qquad b(t) = \frac{\mu}{\sqrt{\lambda\mu}t}\exp(-(\lambda+\mu)t)I_1(2\sqrt{\lambda\mu}t);$$

p. 16 of Prabhu (1965). After time scaling, we have

$$\text{(A.7)} \qquad m_1'(t, i) = \frac{\rho}{2\theta^2} - \frac{1}{2\theta^2}[1 - P_{i0}(t)] = \frac{1}{2\theta^2}[P_{i0}(t) - (1-\rho)]$$

and

$$\text{(A.8)} \quad b(t) = \frac{1}{\sqrt{\rho}t}\exp(-(1+\rho)t/2\theta^2)I_1(\sqrt{\rho}t/\theta^2) = \frac{1}{\sqrt{\rho}t}\exp(-t/\tau)\exp(-v)I_1(v)$$

where $v = \sqrt{\rho}t/\theta^2$ and $\tau = (1+\sqrt{\rho})^2/2$ as defined in Section 3.

Since the derivative $(g'(t))$ of a function $g(t) = f(ct)$ satisfies $g'(t) = cf'(ct)$ for a scalar c, the operator (d/dt) is mapped into $[1/(2\mu\theta^2)](d/dt)$ when t is mapped into $t/2\theta^2$; i.e., $f'(t)$ in the original scale is mapped into $(2\mu\theta^2)^{-1}f'(t)$ in the new scale. To illustrate, $m_1'(t, i)$ appears without time scaling in (A.5). To get to the time-scaled version in (A.7), it suffices to divide by $2\mu\theta^2$.

Acknowledgment

We are grateful to D. R. Smith for Remark 4.3 and to E. van Doorn for Remarks 4.4 and 10.1. We thank the referee for a careful reading.

References

ABATE, J. AND WHITT, W. (1987a) Transient behavior of regulated Brownian motion, I and II. *Adv. Appl. Prob.* **19,** 560–598, 599–631.

ABATE, J. AND WHITT, W. (1987b) Transient behavior of the $M/M/1$ queue: starting at the origin. *Queueing Systems* **2,** 41–65.

ABATE, J. AND WHITT, W. (1987c) The correlation functions of RBM and $M/M/1$. Submitted for publication.

ABRAMOWITZ, M. AND STEGUN, I. A. (eds). (1972) *Handbook of Mathematical Functions.* Dover, New York.

ASMUSSEN, S. AND THORISSON, H. (1986) Large deviation results for time-dependent queue length distributions. Unpublished paper, Institute of Mathematical Statistics, Copenhagen.

BAILEY, N. T. J. (1954) A continuous time treatment of a simple queue using generating functions. *J. R. Statist. Soc.* **B16,** 288–291.

BAILEY, N. T. J. (1957) Some further results in the non-equilibrium theory of a simple queue. *J. R. Statist. Soc.* **B19,** 326–333.

BAILEY, N. T. J. (1964) *The Elements of Stochastic Processes with Applications to the Natural Sciences.* Wiley, New York.

BENEŠ, V. E. (1963) *General Stochastic Processes in the Theory of Queues.* Addison-Wesley, Reading, Mass.

BILLINGSLEY, P. (1968) *Convergence of Probability Measures.* Wiley, New York.

BROWN, M. (1980) Bounds, inequalities and monotonicity properties for some specialized renewal processes. *Ann. Prob.* **8,** 227–240.

BROWN, M. (1981) Further monotonicity properties for specialized renewal processes. *Ann. Prob.* **9,** 891–895.

CHAMPERNOWNE, D. G. (1956) An elementary method of solution of the queueing problem with a single server and constant parameters. *J. R. Statist. Soc.* **B18,** 125–128.

CLIFFORD, P. AND SUDBURY, A. (1985) A sample path proof of the duality for stochastically monotone Markov processes. *Ann. Prob.* **13,** 558–565.

COHEN, J. W. (1982) *The Single Server Queue,* 2nd edn. North-Holland, Amsterdam.

CONOLLY, B. (1975) *Lecture Notes on Queueing Systems.* Wiley, New York.

COOPER, R. B. (1972) *Introduction to Queueing Theory.* Macmillan, New York.

COX, D. R. (1962) *Renewal Theory.* Methuen, London.

COX, D. R. AND ISHAM, V. (1986) The virtual waiting-time and related processes. *Adv. Appl. Prob.* **18,** 558–573.

DARLING, D. A. AND SIEGERT, A. J. F. (1953) The first passage problem for a continuous Markov process. *Ann. Math. Statist.* **24,** 624–639.

DELBROUCK, L. E. N. (1976) Approximations for certain congestion functions in single server queueing systems. *Proc. 8th International Teletraffic Congress,* Melbourne, 233, 1–5.

DOETSCH, G. (1974) *Introduction to the Theory and Application of the Laplace Transformation,* translated by W. Nader. Springer-Verlag, New York.

DUDA, A. (1984) Transient diffusion approximation for some queueing systems. *Performance Evaluation Rev.* **12,** 118–128.

ERDÉLYI, A. (1956) *Asymptotic Expansions.* Dover, New York.

FELLER, W. (1971) *An Introduction to Probability Theory and Its Applications,* Vol. II, 2nd edn. Wiley, New York.

GAVER, D. P., JR. (1968) Diffusion approximations and models for certain congestion problems. *J. Appl. Prob.* **5,** 607–623.

GAVER, D. P., JR. AND JACOBS, P. A. (1986) On inference and transient response for $M/G/1$ models. Naval Postgraduate School, Monterey.

HEYMAN, D. P. AND SOBEL, M. J. (1982) *Stochastic Models in Operations Research,* I. McGraw-Hill, New York.

IGLEHART, D. L. AND WHITT, W. (1970) Multiple channel queues in heavy traffic, II: sequences, networks and batches. *Adv. Appl. Prob.* **2,** 355–369.

ITÔ, K. AND MCKEAN, H. P. JR. (1965) *Diffusion Processes and Their Sample Paths.* Springer-Verlag, New York.

KARLIN, S. (1964) Total positivity, absorption probabilities and applications. *Trans. Amer. Math. Soc.* **111,** 34–107.

KARLIN S. AND MCGREGOR, J. (1958) Many server queueing processes with Poisson input and exponential service times. *Pacific J. Math.* **8,** 87–118.

KEILSON, J. (1979) *Markov Chain Models—Rarity and Exponentiality.* Springer-Verlag, New York.

KEILSON, J. (1981) On the unimodality of passage-time densities in birth–death processes. *Statist. Neerlandica* **35,** 49–55.

KEILSON, J. AND KESTER, A. (1978) Unimodality preservation in Markov chains. *Stoch. Proc. Appl.* **7,** 179–190.

KEILSON, J. AND SUMITA, U. (1982) Uniform stochastic ordering and related inequalities. *Can. J. Statist.* **10,** 181–198.

KELLY, F. P. (1979) *Reversibility and Stochastic Networks.* Wiley, Chichester.

KELTON, W. D. (1985) Transient exponential-Erlang queues and steady state simulation. *Comm. ACM* **28,** 741–749.

KELTON, W. D. AND LAW, A. M. (1985) The transient behavior of the $M/M/s$ queue, with implications for steady-state simulation. *Operat. Res.* **33,** 378–396.

KINGMAN, J. F. C. (1966) An approach to the study of Markov processes. *J. R. Statist. Soc.* **B28,** 417–438.

KINGMAN, J. F. C. (1972) *Regenerative Phenomena.* Wiley, New York.

KOSTEN, L. (1973) *Stochastic Theory of Service Systems.* Pergamon Press, Oxford.

KUMAR, A. AND WONG, W. S. (1987) Some mean value formulas for the transient $M/M/1$ queue. Unpublished paper, AT & T Bell Laboratories, Holmdel, NJ.

LEDERMANN, W. AND REUTER, G. E. (1954) Spectral theory for the differential equations of simple birth and death process. *Phil. Trans. R. Soc. London* A **246,** 321–369.

LEE, I. (1985) *Stationary Markovian Queueing Systems: An Approximation for the Transient Expected Queue Length.* M.S. dissertation, Department of Electrical Engineering and Computer Science, MIT, Cambridge.

LEE, I. AND ROTH, E. (1986) Stationary Markovian queueing systems: an approximation for the transient expected queue length. Unpublished paper.

LORDEN, G. (1970) On excess over the boundary. *Ann. Math. Statist.* **41,** 520–527.

MIDDLETON, M. R. (1979) Transient Effects in $M/G/1$ Queues. Ph.D. dissertation, Stanford University.

ODONI, A. R. AND ROTH, A. (1983) An empirical investigation of the transient behavior of stationary queueing systems. *Operat. Res.* **31,** 432–455.

OTT, T. J. (1977a) The covariance function of the virtual waiting-time process in an $M/G/1$ queue. *Adv. Appl. Prob.* **9,** 158–168.

OTT, T. J. (1977b) The stable $M/G/1$ queue in heavy traffic and its covariance function. *Adv. Appl. Prob.* **9,** 169–186.

PEDGEN, C. D. AND ROSENSHINE, M. (1982) Some new results for the $M/M/1$ queue. *Management Sci.* **28,** 821–828.

PRABHU, N. U. (1965) *Queues and Inventories.* Wiley, New York.

PRABHU, N. U. (1980) *Stochastic Storage Processes.* Springer-Verlag, New York.

RIORDAN, J. (1961) Delays for last-come first-served service and the busy period. *Bell System Tech. J.* **40,** 785–793.

RIORDAN, J. (1962) *Stochastic Service Systems.* Wiley, New York.

ROSENLUND, S. I. (1978) Transition probabilities for a truncated birth–death process. *Scand. J. Statist.* **5,** 119–122.

ROTH, E. (1981) An Investigation of the Transient Behavior of Stationary Queueing Systems. Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge.

SIEGMUND, D. (1976) The equivalence of absorbing and reflecting barrier problems for stochastically monotone Markov processes. *Ann. Prob.* **4,** 914–924.

SIEGMUND, D. (1985) *Sequential Analysis.* Springer-Verlag, New York.

STONE, C. (1963) Limit theorems for random walks, birth and death processes, and diffusion processes. *Illinois J. Math.* **7,** 638–660.

STOYAN, D. (1983) *Comparison Methods for Queues and Other Stochastic Models,* ed. D. J. Daley. Wiley, Chichester.

TAKACS, L. (1967) *Combinatorial Methods in the Theory of Stochastic Processes.* Wiley, New York.

VAN DOORN, E. (1980) *Stochastic Monotonicity and Queueing Applications of Birth–Death Processes.* Lecture Notes in Statistics **4,** Springer-Verlag, New York.

WHITT, W. (1980) Some useful functions for functional limit theorems. *Math. Operat. Res.* **5,** 67–85.

WHITT, W. (1982) Refining diffusion approximations for queues. *Operat. Res. Letters* **1,** 165–169.