detectors are very close to the optimum and more complicated DMF.

REFERENCES

[1] M. Schwartz and L. Shaw, *Signal Processing: Discrete Spectral Analysis, Detection and Estimation.* New York: McGraw-Hill, 1975.

[2] V. Milutinovic, "Suboptimum detection procedure based on the weighting of partial decisions," *Electron Lett.*, vol. 16, pp. 237-238, Mar. 1980.

[3] ——, "Comparison of three suboptimum detection procedures," *Electron Lett.*, vol. 16, pp. 681-683, Aug. 1980.

[4] ——, "Performance comparison of two suboptimum detection procedures in real environment," *Proc. IEE*, part F, vol. 131, pp. 341-344, July 1984.

[5] N. C. Beaulie and C. Leung, "On the performance of three suboptimum detection schemes for binary signaling," *IEEE Trans. Commun.*, vol. COM-33, pp. 241-245, Mar. 1985.

[6] ——, "Optimal detection of hard-limited data signals in different noise environment," *IEEE Trans. Commun.*, vol. COM-34, p. 619, June 1986.

[7] C. C. Lee and J. B. Thomas, "Detectors for multinomial input," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-19, pp. 288-297, Mar. 1983.

Calculating Time-Dependent Performance Measures for the *M/M/1* Queue

JOSEPH ABATE AND WARD WHITT

Abstract—**This correspondence discusses methods for computing transient performance measures for the *M/M/1* queue. These performance measures are often expressed in terms of modified Bessel functions without any discussion about computation. In fact, a common expression for the probability transition function of the *M/M/1* queue length process has an infinite sum of modified Bessel functions. For actually generating numbers, however, it is convenient to use numerical integration with associated integral representations, as was first pointed out by Morse in 1955 [17].**

I. INTRODUCTION

In the IEEE TRANSACTIONS ON COMMUNICATIONS, several papers have proposed numerical procedures for calculating time-dependent performance measures for the $M/M/1$ queue, such as the mean, the variance, and the probability mass function of the queue length at time t for any given initial state. The proposed procedures include finite-state approximations [21], the discrete Fourier transform [4], Q functions [13], and generalized Q functions [7], [8]. These new procedures seem to be effective, but we wish to point out that the need for new procedures is less than it might appear because integral representations for the $M/M/1$ time-dependent performance measures exist that make numerical calculation relatively straightforward by numerical integration, e.g., by Simpson's rule; see [3, Sect. 25.4]. It is also possible to numerically invert Laplace transforms, but numerical integration usually gives much better accuracy (with bounds on the error) with very little effort (using readily available programs). Integral representations for $M/M/1$ transient quantities were first proposed by Ledermann and Reuter [16] and Morse [17] and appear in the 1962 textbooks by Riordan [18] and Takács [22]. New and old integral representations (some remarkably simple) are also discussed in [2]. For example, the busy-period density, the probability that the server is busy at time t starting in zero, the mean queue length at time t starting in zero, and the autocorrelation function of the stationary queue-length process can all be represented as mixtures of exponentials with a probabilistic mixing density that is a simple modification of a beta density.

Paper approved by the Editor for Communication Theory of the IEEE Communications Society. Manuscript received August 4, 1987; revised May 24, 1988.

J. Abate is with AT&T Bell Laboratories, Whippany, NJ 07981.

W. Whitt is with AT&T Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 8930084.

II. GUIDANCE IN THE TEXTBOOKS

Unfortunately, numerical procedures have not been given enough emphasis in the queueing textbooks. Most textbooks present only representations of $M/M/1$ transient quantities in terms of modified Bessel functions, without any discussion about how these expressions are to be used for generating numbers, e.g., Gross and Harris [11, p. 129] and Kleinrock [15, p. 77]. Indeed Kleinrock [15, p. 78] concludes

> "This last expression [the standard representation of the probability transition function in terms of modified Bessel functions] is most disheartening. What it has to say is that an appropriate model for the *simplest interesting* queueing system...leads to an ugly expression for the time-dependent behavior of its state probabilities. As a consequence, we can only hope for greater complexity and obscurity in attempting to find time dependent behavior of more general queueing systems."

The expression displayed in [15, p.77, eq. (2.163)] *is* somewhat daunting. With λ denoting the arrival rate, μ the service rate, and $\rho \equiv \lambda/\mu$ the traffic intensity, the expression is

$$P_{ij}(t) = e^{-(\lambda+\mu)t}\left[\rho^{(j-i)/2} I_{j-i}(2\sqrt{\lambda\mu}\,t) + \rho^{(j-i-1)/2} I_{j+i+1}(2\sqrt{\lambda\mu}\,t) + (1-\rho)\rho^{j}\sum_{k=j+i+2}^{\infty}\rho^{-k/2} I_k(2\sqrt{\lambda\mu}\,t)\right] \quad (1)$$

where

$$I_k(x) = \sum_{m=0}^{\infty}\frac{(x/2)^{k+2m}}{(k+m)!\,m!} \quad (2)$$

is the modified Bessel function of the first kind of order k; see [3, Sect. 9.6]. Note that (1) not only involves the modified Bessel functions in (2), but an *infinite sum* of these modified Bessel functions. Obviously, there is a gap between (1) and numerical results.

III. FINITE—STATE APPROXIMATIONS

Consistent with Kleinrock's conclusion, Stern motivates his development of approximations for $M/M/1$ transient behavior in [21] with the remark

> "it is well known that even for the simplest case, the $M/M/1$ queue, the exact expression for $\bar{N}(t)$ [the mean queue length at time t] involves an infinite sum of Bessel functions. Clearly approximations are necessary."

What Stern does then in [21] is approximate the $M/M/1$ model by a truncated model having a finite waiting room and

analyze this finite-state model exactly. This procedure works very well because, first, the time-dependent behavior of a finite-state birth-and-death process has a relatively simple exponential from (e.g., see Keilson [14, Sect. 3.2]) and second, the time-dependent behavior in the two models differs very little when the waiting room is sufficiently large. Indeed, for a sufficiently large waiting room, the difference between the two models is invariably negligible compared to the quality of the model fit in applications.

Stern's approximation is somewhat disconcerting, though because we usually introduce models involving infinite quantities in order to obtain mathematical simplification, and now we are proceeding in the reverse direction. Indeed, from most perspectives, the standard $M/M/1$ model has a more elementary description than its finite-waiting room counterpart. In fact, it is really not so easy to see what the transient behavior is like from the formulas with the finite waiting room. Recognizing this, Stern looks for additional structure and approximations.

In their seminal paper, Ledermann and Reuter [16] previously pursued this line of reasoning (for general birth-and-death processes). An improved version of this approach for the $M/M/1$ queue is also presented by Takács [22]. In [22, ch. 1, Sect. 1], Takács derives the time-dependent probability mass function for $M/M/1$ queue with finite waiting room. In [22, ch. 1, Sect. 2], Takács then uses this result to *derive* the corresponding result for the $M/M/1$ system with unlimited waiting room by taking the limit as the size of the waiting room tends to infinity. The finite-waiting-room expression involving a sum of trigonometric terms is a Riemann sum approximating the limiting integral. Thus, the finite-waiting-room approximation turns out to be equivalent to a numerical integration of an integral. In this way, Takács obtains a new proof of the trigonometric integral representation for the probability transition function due to Morse [17]. The trigonometric integrand may look unsightly to the human eye, but the computer is pleased. Indeed, contrary to Stern's remark above, the mean queue length (number in system) at time t starting in state i with $\rho < 1$ has a relatively tractable integral representation without any Bessel functions, namely,

$$m(t, i) \equiv \sum_{j=1}^{\infty} jP_{ij}(t) = \frac{\rho}{1-\rho} - \frac{2\rho^{-i/2}}{\pi} \int_0^{\pi} \frac{e^{-\gamma(y)\mu t}}{\gamma(y)^2} \cdot \sin y \,(\sin (i+1)y - \rho^{-1/2} \sin iy)\, dy \quad (3)$$

for $\gamma(y) = 1 + \rho - 2\sqrt{\rho}\cos y$; see [22, p.27]. The integrand in (3) and its derivatives can easily be bounded to produce bounds on the error in numerical integration. Greater accuracy for a given number of points can be obtained by allocating more points to the region where $\gamma(y)$ is small. These integrals are relatively well behaved, e.g., a numerical evaluation of (3) in the case $i = 0$ using a simple trapezoidal rule with only 200 points produces an absolute error of about 10^{-7} for $\rho \leq 0.85$. For $\rho > 0.85$, you need to use more points near the minimum of $\gamma(y)$.

It is also worth noting that there are other integral representations besides the trigonometric integral representations; e.g., two others are given for $m(t, i)$ in [2, Theorem 4.1 and Corollary 7.2]. The integrands in these other integral representations are somewhat easier to understand (for someone who rarely works with trigonometric functions), but we found the trigonometric integral representations to be superior for obtaining numerical results because the intervals of integration tend to be shorter and the integrands tend to be much more nearly uniform. In a numerical experiment using the trapezoidal rule with 25 equally spaced subintervals (obviously not aiming for high precision), we found that the trigonometric integral representation in (3) to give an order of magnitude better accuracy for the case $i = 0$ than the other two integral representations in [2].

Unfortunately, however, Ledermann and Reuter [16] and Takács [22] do not discuss the importance of the integral representations for computation, so that readers may not realize how useful the results are. The importance of the integral representations for computation *is* emphasized and *demonstrated* by Morse [17], though. Morse's early contribution is especially impressive in retrospect.

IV. Avoiding Infinite Sums of Bessel Functions

It appears that (1) was the first expression found for $P_{ij}(t)$, derived by Clarke in an unpublished 1953 report (see [9], [10, p. 659], and [18, p. 215], but soon there were also other expressions which do not involve infinite sums of Bessel functions. The first published results were Ledermann and Reuter's [16] and Bailey's [5] expression for the derivative $P'_{ij}(t)$ in terms of six modified Bessel functions. An alternate closed-form expression for $P_{ij}(t)$ was found by Bailey [6, eq. (12), p. 328], which is given in (4.31) of Cohen [10, p. 82]. (See Syski [20, p. 335–340] for a good historical account.) However, the best form from the computational point of view seems to be the trigonometric integral representation derived by Morse [17]. (The alternate finite-form representation in [10, p. 82] involves integrals of Bessel functions, so that it reduces to a *double* integral.) The trigonometric form for $P_{ij}(t)$ is also given by Riordan [18, p. 45] and Takács [22, p. 23].

Some of the recent papers in these Transactions seem to have been written without the authors being aware that there are better representations for $P_{ij}(t)$ than (1). Motivated by Stern [21], the problem involving the infinite sum of Bessel functions was directly attacked by Jones *et al.* [13], who observe that the term involving the infinite sum of Bessel functions can be expressed in terms of a function

$$Q(a, b) = \int_b^{\infty} \exp\left(-\frac{[a^2+x^2]}{2}\right) I_0(ax)x\, dx \quad (4)$$

called the *circular coverage function* of *Marcum's Q function,* for which computational procedures are known; see Helstrom [12, Appendix F] and Schwartz *et al.* [19, Appendix A].

More recently, Cantrell [7] and Cantrell and Beall [8] have proposed another numerical procedure based on *generalized Q functions* $Q_m(\alpha, \beta)$ which have expansions

$$1 - Q_m(\alpha, \beta) = \exp\left(-\frac{\alpha^2-\beta^2}{2}\right) \sum_{k=m}^{\infty} \left(\frac{\beta}{\alpha}\right)^k I_k(\alpha\beta) \quad (5)$$

the case $m = 1$ being equivalent to Marcum's Q function. The algorithms for calculating these Q functions involve power series or Neumann series, the latter being used in [7] and [8]. In fact, these numerical procedures based on Q functions seem very effective, yielding very good accuracy using little computer time. [Indeed, in a personal communication, Cantrell reports comparisons in which the generalized Q function approach is two–eight times faster than (3).] However, the integral representations seem to be attractive alternatives, yielding good accuracy using little *learning* or *programming* time (because standard methods and standard programs can be applied). We believe that for most applications, human time (programming and analysis) is far more important than computer time, given that the difference in required computer time is not inordinately great, so that most people will prefer numerical integration with the integral representations. (This might not apply to those already familiar with Q functions.) At any rate, new procedures should certainly be compared to numerical integration based on integral representations instead

of numerical integration based on the Chapman-Kolmogorov differential equations, as is done in [7], [8]. It seems clearly better to use integral representations than the Chapman–Kolmogorov equations.

We conclude by noting that integral representations are also available for related quantities. For infinite sums of Bessel functions and Q functions, see [1, eq. (7.8)], [3, p. 376, Eq. (9.6.33), [12, p. 453, eq. (F.19), and [19, p. 588, eq. (A-8-6)]. Integral representations for all moments of the queue length (explicit for the first two) are given in [2, Sect. 7]. Thus, [2] offers a simple alternative to the variance calculation recently proposed in [8].

ACKNOWLEDGMENT

We thank P. Cantrell for commenting on this correspondence and sharing his numerical experience.

REFERENCE

[1] J. Abate and W. Whitt, "Transient behavior of the $M/M/1$ queue via Laplace transforms," *Adv. Appl. Prob.*, vol. 20, pp. 145–178, 1988.

[2] ——, "Simple spectral representations for the $M/M/1$ queue," *Queueing Syst.*, vol. 3, pp. 321–346, 1988.

[3] M. Abramowitz, and I. A. Stegun, Ed., *Handbook of Mathematical Functions.* New York: Dover, 1972.

[4] M. H. Ackroyd, "$M/M/1$ transient state occupancy probabilities via the discrete Fourier transform," *IEEE Trans. Commun.*, vol. COM-30, pp. 557–559, 1982.

[5] N. T. J. Bailey, "A continuous-time treatment of a simple queue using generating functions," *J. Roy. Statist. Soc.*, ser. B, vol. 16, pp. 288–291, 1954.

[6] ——, "Some further results in the non-equilibrium theory of a simple queue," *J. Roy. Statist. Soc.*, ser. B, vol. 19, pp. 326–333, 1957.

[7] P. E. Cantrell, "Computation of the transient $M/M/1$ queue cdf, pdf, and mean with generalized Q-functions," *IEEE Trans. Commun.*, vol. COM-34, pp. 814–817, 1986.

[8] P. E. Cantrell and G. R. Beall, "Transient $M/M/1$ queue variance computations using generalized Q-functions," *IEEE Trans. Commun.*, to be published.

[9] A. B. Clarke, "A waiting line process of Markov type," *Ann. Math. Statist.*, vol. 27, pp. 452–459, 1956.

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

[11] D. Gross and C. M. Harris, *Fundamentals of Queueing Theory,* 2nd ed. New York: Wiley, 1985.

[12] C. W. Helstrom, *Statistical Theory of Signal Detection,* 2nd ed. Oxford: Pergamon, 1968.

[13] S. K. Jones, R. K. Cavin, III, and D. A. Johnston, "An efficient computation procedure for the evaluation of $M/M/1$ transient state occupancy probabilities," *IEEE Trans. Commun.*, vol. COM-28, pp. 2019–2020, 1980.

[14] J. Keilson, *Markov Chain Models—Rarity and Exponentiality.* New York: Springer-Verlag, 1979.

[15] L. Kleinrock, *Queueing Systems, Vol. I.* New York: Wiley, 1975.

[16] W. Ledermann and G. E. H. Reuter, "Spectral theory for the differential equations of simple birth-and-death processes," *Phil. Trans. Roy. Soc. London.*, ser. A, vol. 246, pp. 321–369, 1954.

[17] P. M. Morse, "Stochastic properties of waiting lines," *Oper. Res.*, vol. 3, pp. 255–261, 1955.

[18] J. Riordan, *Stochastic Service Systems.* New York: Wiley, 1962.

[19] M. Schwartz, W. R. Bennet, and S. Stein, *Communication systems and Techniques.* New York: McGraw-Hill, 1966.

[20] R. Syski, *Introduction to Congestion Theory in Telephone Systems.* London: Oliver and Boyd, 1960.

[21] T. E. Stern, "Approximations of queue dynamics and their application to adaptive routing in communications networks," *IEEE Trans. Commun.*, vol. COM-27, pp. 1331–1335, 1979.

[22] L. Takács, *Introduction to the Theory of Queues.* New York: Oxford Univ. Press, 1962.