OR Spektrum (1995) 17: 19–22

OR Spektrum

Computing transient performance measures for the M/M/1 queue

M. C. T. van de Coevering

Department of Econometrics, Vrije University, 1081 HV Amsterdam, The Netherlands

Received: 31 August 1993/Accepted: 10 August 1994

Abstract. A practically important problem is the computation of transient performance measures for the M/M/1 queue. Trigonometric integral representations are very well suited for that purpose. This paper reviews a number of results that can be found scattered in the literature and also provides the practitioner simple recommendations for calculating routinely the M/M/1 performance measures.

Zusammenfassung. Ein für die Praxis bedeutsames Problem besteht in der numerischen Berechnung nichtstationärer Leistungsmaße für M/M/1-Warteschlangensysteme. Für diesen Zweck sehr gut geeignet sind trigonometrische Integral-Darstellungen. Der vorliegende Aufsatz enthält einen Überblick über eine Anzal in der Literatur zu findenden Ergebnisse und gibt dem Praktiker einfache Empfehlungen zur routinemäßigen Berechnung der M/M/1-Leistungsmaße.

Key words: M/M/1 queue, transient queue length, transient moments, computational methods

Schlüsselwörter: M/M/1-Warteschlangen, nichtstationäre Warteschlangenlängen, nichtstationäre Momente, numerische Methoden

1. Introduction

Analysing transient behaviour of queueing systems is of considerable importance in many practical situations. The steady-state analysis is often not appropriate. In manufacturing and banking situations the system may physically terminate at the end of each day before a steady-state situation is reached. In addition there are many systems for which the input process is time-dependent so that a steady-state situation cannot even be defined. A typical example concerns airport runway operations. The demand profile for runway operations shows considerable variation over time with peaks at certain hours of the day. Steady-state queueing models are not applicable in these situations. Despite the immense literature on queueing systems, even for simple systems few results for transient behaviour are available. Simple Markovian queues are the few queues for which exact closed-form solutions are available, see e.g. Sharma (1990). However, there is a big gap between theoretically derived solutions and practically useful numerical solution methods. Most textbooks give representations of M/M/1 transient quantities in terms of modified Bessel functions, see also Parthasarathy (1987). In the leading text Kleinrock (1976) it was already pointed out that this representation is not practically useful. It is only in recent years that more attention is paid to the numerical calculation of transient performance measures for the M/M/1 queue, see Abate and Whitt (1989), Ackroyd (1982), Cantrell (1986, 1988). An interesting new analytical formula for the transient queue length probabilities in the M/M/1 queue was recently given by Leguesdron et al. (1993). In the special case in which the M/M/1 queue is empty at $t = 0$, a simple and computationally attractive formula for the transient queue length probability has been discussed in the recent paper of Conolly and Langaris (1993). The formulas in the latter two references are computationally more demanding than the trigonometric integral representation to be discussed below. The paper of Abate and Whitt (1989) recognizes how important old integral representations are as starting point for numerical computations. The trigonometric integral representations for the M/M/1 queue go back to Morse (1955) and Takacs (1962). It is important to have a fast and reliable method to compute the transient quantities for the M/M/1 queue. The M/M/1 transient solution is also a building block for the heuristic transient solutions proposed in Roth (1985) for the general G/G/1 queue.

The goal of this paper is to review a number of results that can be found scattered in the literature and to provide the practitioner with simple recommendations for calculating routinely the M/M/1 transient performance measures. In Sect. 2 we first review the trigonometric integral representations including a new one for the variance of the transient queue length. This trigonometric integral representation for the variance is somewhat simpler than the integral representation proposed in Abate and Whitt (1988). Also, Sect. 2 reviews some recent approximations proposed for the mean of the transient number of customers in the system. Section 3 discusses numerical integration methods. It turns out that the simple Gauss-Legendre integration method is excellently suited for practical purposes. Finally, the appendix gives the proof of a new integral representation for the variance of the transient queue length in the M/M/1 queue.

2. Integral representations for the M/M/1 queue

Let us first introduce some notation. In the M/M/1 system customers arrive according to a Poisson process with rate λ and the service time is exponentially distributed with mean $1/\mu$. The traffic intensity is denoted by

$$\varrho = \lambda/\mu .$$

Assuming that there are i customers present at the current time $t = 0$, define for any $t \geq 0$ the random variables

$L(i, t)$ = the number of customers in the system at time t (including any customer in service),
$L_q(i, t)$ = the number of customers waiting in queue at time t (excluding any customer in service).

Also, define the transient probabilities

$$p_{ij}(t) = P\{L(i,t) = j\} \quad \text{for} \quad t \geq 0 \text{ and } i,j = 0, 1, \ldots .$$

The following integral representation was derived by Morse (1955) and Takacs (1962).

Theorem 1. *For any $t \geq 0$ and $i,j = 0, 1, \ldots,$*

$$p_{ij}(t) = \frac{2}{\pi}\varrho^{(j-i)/2}\int_0^\pi \frac{e^{-\mu t \gamma(y)}}{\gamma(y)} \cdot a_i(y) \cdot a_j(y)\,dy + \begin{cases} (1-\varrho)\varrho^j & \varrho < 1 \\ 0 & \varrho \geq 1, \end{cases} \tag{1}$$

where $\gamma(y)$ and $a_k(y)$ are shorthand notations for

$$\gamma(y) = 1 + \varrho - 2\sqrt{\varrho}\cos(y)$$

and

$$a_k(y) = \sin(k y) - \sqrt{\varrho}\sin((k+1)y) .$$

Next the integral representation for the mean of the transient number in the system is given.

Theorem 2. *For any $t \geq 0$ and $i = 0, 1, \ldots,$*

$$EL(i,t) = \frac{2}{\pi}\varrho^{(1-i)/2}\int_0^\pi \frac{e^{-\mu t \gamma(y)}}{\gamma(y)^2} \cdot a_i(y) \cdot \sin(y)\,dy + \begin{cases} \varrho/(1-\varrho) & \varrho < 1 \\ i + (\lambda - \mu)t + \varrho^{-1}/(\varrho - 1) & \varrho > 1 . \end{cases} \tag{2}$$

This Theorem corrects an error in the original results of Takacs (1962). Professor Takacs (private communication) was so kind to provide the corrected formula and a simple derivation. We have extended this derivation to obtain an trigonometric expression for the second moment of $L(i, t)$.

Theorem 3. *For any $t > 0$ and $i = 0, \ldots,$*

$$EL^2(i,t) = 2(1-\varrho)\frac{2}{\pi}\varrho^{(1-i)/2}\int_0^\pi \frac{e^{-\mu t \gamma(y)}}{\gamma(y)^3} \cdot a_i(y) \cdot \sin(y)\,dy + \begin{cases} 2\varrho(1-\varrho)^{-2} - EL(i,t) & \varrho < 1, \\ 2(\varrho-1)it + (\varrho-1)^2t^2 + 2\varrho^{-i}t + 2\varrho t - EL(i,t) + i + i^2 & \\ -[\varrho + 1 + (2i+1)(\varrho-1)]\varrho^{-i}(\varrho-1)^{-2}, & \varrho \geqq 1 . \end{cases} \tag{3}$$

This explicit expression for the second moment of the queue length is better suited for numerical computations than the alternative expression derived on p. 336 in Abate and Whitt (1988). Using tricky algebra the trigonometric representation in Theorem 3 can be derived from results in the Sects. 7 and 9 of Abate and Whitt (1988). However, a simple and direct proof of Theorem 3 can be given using ideas of Takacs. This proof is included in the Appendix for completeness. Finally, it is noted that the first two moments of the transient queue length can be obtained from

$$EL_q(i,t) = EL(i,t) - (1 - p_{i0}(i,t))$$

and

$$EL_q^2(i,t) = EL^2(i,t) - 2EL(i,t) + 1 - p_{i0}(i,t) .$$

Remark 1. Abate and Whitt (1987) have found an interesting asymptotic expansion for the first moment of $L(0, t)$ when $\varrho < 1$. For t large,

$$EL(0,t) = \varrho/(1-\varrho) - b(\varrho)e^{-t/a(\varrho)}$$

where

$$a(\varrho) = 2(1-\varrho)^{-2}c(\varrho), \quad b(\varrho) = \varrho(1-\varrho)^{-1}(1 + 2c(\varrho))^{-1},$$

and

$$c(\varrho) = 0.25[2 + \varrho + (5 - (1-\varrho)(5+\varrho))^{1/2}] .$$

Our numerical experiments confirm that the above expression is very useful to compute quick engineering estimates for $EL(0, t)$ when $t \geq 2(1-\varrho)^{-2}$. The factor $a(\varrho)$ is a relaxation factor being the dominating factor for the time it takes until the system reaches equilibrium, see also Cohen (1982) and Odoni and Roth (1983).

Remark 2. Consider the M/G/1 queue in which customers arrive according to a Poisson process with rate λ and the service time S of a customer has a general probability distribution. An exact algorithm to compute the transient M/G/1 quantities has been derived in Van de Coevering (1992) for Erlangian distributed services and hyperexponentially distributed services. This algorithm is time-consuming and is not suitable for quick calculations for practical purposes. However, we used the algorithm to test extensively an interesting heuristic for the transient queue length that was proposed in Roth (1985). For the

M/G/1 queue starting with an empty system and having an offered load $\varrho = \lambda E(S) < 1$, this heuristic is

$$E_{\text{app}}[L_q(0,t)] = 1/2(1 + C_S^2) E_{\text{exp}}[L_q(0, 2t(1 + C_S^2)^{-1})], \quad t \geq 0,$$

where $E_{\text{exp}}[L_q(0,t)]$ is the transient queue length in the corresponding M/M/1 queue and C_S^2 is the ratio of the squared mean and the variance of the service time S. Our numerical experiments indicate that this is a practically useful approximation provided that C_S^2 is not too large (say, $C_S^2 \leq 2$). Details can be found in Van de Coevering (1992).

3. Numerical integration

The choice of the numerical integration method to evaluate the integral representation is important. The method should be simple and reliable. How good a particular method works depends much on the properties of the integrand. The difficulty of computing the integrals depends much on the values of the initial state and the offered load. Abate and Whitt (1989) proposed the basic trapezoidal integration rule. This method works indeed quite well, but a considerably faster method is provided by Gauss-Legendre integration, particularly when ϱ is close to 1. In case many applications of the integration routine are required, it is important to have such a fast method. The Gauss-Legendre integration method is relatively unknown, though it has some popularity in physics. The idea behind Gauss-Legendre is simple. For some appropriate choice of n, we compute the integral through

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \approx \sum_{i=1}^{n} w_i f(x_i),$$

where the weights w_i and the abscissas x_i are given numbers that do not depend on the integrand f. The w_i and the x_i are determined in such a way that the integral is exact for polynomials up to degree $2n - 1$. The values of the w_i and the x_i have been tabulated for various values of n, see Abramowitz and Stegun (1965). Alternatively, the practitioner can find ready-to-use Pascal or Fortran computer codes in Press et al. (1986). Once you know the w_i and the x_i, the integral can be cheaply evaluated. This is particularly important when the integral must be repeatedly computed. Two remarks are in order. First, a higher value of n does not necessarily imply a higher accuracy. Second, the integration procedure also applies to any finite integral $\int_a^b f(x)\,\mathrm{d}x$ by writing $\int_a^b f(x)\,\mathrm{d}x = \frac{1}{2}(b-a)\int_{-1}^{1} g(u)\,\mathrm{d}u$ where $g(u) = f(\frac{1}{2}(a+b) + \frac{1}{2}(b-a)u)$.

A drawback of Gauss-Legendre integration is that the appropriate value of n depends on the form of the integrand f and has to be determined experimentally. It turned out that for our integral representations an appropriate value of n depends strongly on the value of the initial state i and the offered load ϱ. In particular, it turned out that the integral representations cannot be evaluated for ϱ very close to 1. Below we will recommend a choice for n that is "uniform" in the parameter values. Therefore we have restricted the values of the initial state i and the offered load ϱ to $0 \leq i \leq 100$ and $0 \leq \varrho \leq 10$ with $\varrho \notin (0.99, 1.01)$. Also for fixed ϱ, the value of the initial state i has been further restricted to $i \leq r(\varrho)$ with $r(\varrho) = -16\ln(10)/\ln(\varrho)$. This is not a severe restriction, since $r(\varrho) \geq 100$ for $\varrho \geq 0.7$ and since for ϱ small the values of the relevant states are small too.

Experimentally we found that for $0 \leq i \leq \min(100, r(\varrho))$, $j \leq 100$, and $0 \leq \varrho \leq 10$ with $\varrho \notin (0.99, 1.01)$, the following holds:

- Using $n = 200$ (or $n = 275$) gives for $p_{ij}(t)$ an accuracy of at least seven decimals regardless of the value of t.
- Using $n = 200$ (or $n = 275$) gives for $EL(i,t)$ an accuracy of at least six decimals regardless of the value of t.
- Using $n = 275$ gives for the standard deviation of $L(i,t)$ an accuracy of at least five decimals regardless of the value of t.

This accuracy is more than sufficient for practical purposes. The Gauss-Legendre integration method has the great advantage that the computing time of the integrals is not dependent on the parameter values. Using $n = 275$ support points, the three integrals for $p_{ij}(t)$, $EL(i,t)$ and $EL^2(i,t)$ are together computed within a second on an Olivetti M380 computer with a mathematical coprocessor.

4. Appendix

This appendix gives the proof of Theorem 3. The starting point of the proof is provided by the relations

$$EL(i,t) = (\lambda - \mu)t + \mu\int_0^t p_{i0}(x)\,\mathrm{d}x + i \tag{4}$$

and

$$EL^2(i,t) = 2(\lambda - \mu)\int_0^t EL(i,x)\,\mathrm{d}x + (\lambda + \mu)t + \mu\int_0^t p_{i0}(x)\,\mathrm{d}x + i^2 \tag{5}$$

for $i = 0, 1, \ldots$ and $t \geq 0$. These relations are obtained by substituting Kolmogorov's forward differential equations (see e.g. Cohen (1982))

$$p'_{ij}(t) = \mu p_{i,j+1}(t) + \lambda p_{i,j-1}(t) - (\lambda + \mu)p_{ij}(t), \quad t \geq 0,$$

(with $p_{i,-1}(t) = 0$) in $\mathrm{d}EL(i,t)/\mathrm{d}t = \sum j p'_{ij}(t)$ and $\mathrm{d}EL^2(i,t)/\mathrm{d}t = \sum j^2 p'_{ij}(t)$ and next integrating with respect to t. The following shorthand notation is helpful,

$$A_i(t) = \frac{2}{\pi}\varrho^{(1-i)/2}\int_0^\pi \frac{\mathrm{e}^{-\mu\gamma(y)t}}{\gamma(y)^2}\sin(iy)\sin(y)\,\mathrm{d}y \tag{6}$$

and

$$B_i(t) = \frac{2}{\pi}\varrho^{(1-i)/2}\int_0^\pi \frac{\mathrm{e}^{-\mu\gamma(y)t}}{\gamma(y)^3}\sin(iy)\sin(y)\,\mathrm{d}y \tag{7}$$

for $i = 0, 1, \ldots$ and $t \geq 0$, where $\gamma(y)$ is given in Theorem 1. Distinguish now between two cases.

Case $\lambda < \mu$. Substituting the expression for $p_{i0}(t)$ from Theorem 1 into the relation (4) yields

$$EL(i,t) = A_i(t) - \varrho A_{i+1}(t) - A_i(0) + \varrho A_{i+1}(0) + i \tag{8}$$

for all $i \geq 0$ and $t \geq 0$. Using that $A_i(t) \to 0$ as $t \to \infty$ and using the well-known result $EL(i,t) \to \varrho/(1-\varrho)$ for all i (see e.g. Cohen (1982)), it follows from (8) that

$$\varrho A_{i+1}(0) - A_i(0) = \varrho/(1-\varrho) - i. \tag{9}$$

Substituting (9) into (8) yields

$$EL(i,t) = A_i(t) - \varrho A_{i+1}(t) + \varrho/(1-\varrho) \tag{10}$$

for all $i \geq 0$ and $t \geq 0$. From this relation, we obtain by integration

$$\int_0^t EL(i,x)\,dx = \frac{1}{\mu}[-B_i(t) + \varrho B_{i+1}(t) + B_i(0) - \varrho B_{i+1}(0)] + \varrho t/(1-\varrho). \tag{11}$$

Together (4), (10), and (11) yield

$$EL^2(i,t) = 2(\varrho-1)[-B_i(t) + \varrho B_{i+1}(t) + B_i(0) - \varrho B_{i+1}(0)] - EL(i,t) + i + i^2. \tag{12}$$

Letting $t \to \infty$ and using that $EL^2(i,t) \to (\lambda\mu + \lambda^2)/(\mu-\lambda)^2$ (see e.g. Cohen (1982)), it follows that

$$2(\varrho-1)[B_i(0) - \varrho B_{i+1}(0)] = \frac{2\lambda\mu}{(\mu-\lambda)^2} - i - i^2. \tag{13}$$

Substituting (13) into (12) yields

$$EL^2(i,t) = 2(\varrho-1)[-B_i(t) + \varrho B_{i+1}(t)] - EL(i,t) + 2\varrho(1-\varrho)^{-2}.$$

This verifies Theorem 2 for the case of $\lambda < \mu$. Before we turn to the other case, we state that for the case of $\lambda < \mu$,

$$A_i(0) = \frac{i}{1-\varrho} \quad \text{and} \quad B_i(0) = \frac{(1-\varrho)i^2 + (1+\varrho)i}{2(1-\varrho)^3} \tag{14}$$

for all $i \geq 0$. This follows by induction from (9) and (13) after some algebra.

Case $\lambda > \mu$. The relations (8) and (12) now become

$$EL(i,t) = A_i(t) - \varrho A_{i+1}(t) - A_i(0) + \varrho A_{i+1}(0) + i + (\lambda-\mu)t \tag{15}$$

and

$$EL^2(i,t) = 2(\varrho-1)[-B_i(t) + \varrho B_{i+1}(t) + B_i(0) - \varrho B_{i+1}(0)] - EL(i,t) + i + i^2 + 2(\varrho-1)it + (\varrho-1)^2 t^2 + 2\varrho^{-i} t + 2\varrho t. \tag{16}$$

The relations (9) and (13) are no longer true for the case of $\lambda > \mu$, since there exists no steady-state limit when $\varrho > 1$. We now have to proceed in another way. The following trick is used. Define $\bar{A}_i(t)$ and $\bar{B}_i(t)$ as $A_i(t)$ and $B_i(t)$ except that the roles of λ and μ are interchanged in the definitions (6) and (7). It is easy to verify that for all $i \geq 0$ and $t \geq 0$

$$\bar{A}_i(t) = \varrho^{i+1} A_i(t) \quad \text{and} \quad \bar{B}_i(t) = \varrho^{i+2} B_i(t). \tag{17}$$

The quantities $\bar{A}_i(0)$ and $\bar{B}_i(0)$ are given by (14) in which the roles of λ and μ are interchanged, that is, $\varrho = \lambda/\mu$ is replaced by $\varrho^{-1} = \mu/\lambda$. Using the formulas (17) with $t = 0$, it next follows after some algebra that for all $i \geq 0$

$$A_i(0) = \frac{i}{\varrho^i(\varrho-1)} \quad \text{and} \quad B_i(0) = \frac{(\varrho-1)i^2 + (1+\varrho)i}{2\varrho^i(\varrho-1)^3}. \tag{18}$$

Together the formulas (15), (16), and (18) yield after some algebra the desired expression for $EL^2(i,t)$ for the case of $\lambda > \mu$.

Acknowledgement. The author would like to thank Professor Dr. H. C. Tijms for his help with the research and the writing of this article.

References

Abate J, Whitt W (1987) Transient behaviour of the M/M/1 queue starting at the origin. Queueing Syst 2:41–65

Abate J, Whitt W (1988) Simple spectral representations for the M/M/1 queue. Queueing Syst 3:321–346

Abate J, Whitt W (1989), Calculating time-dependent performance measures for the M/M/1 queue. IEEE Trans Commun 37:1102–1104

Abramowitz M, Stegun I (1965) Handbook of Mathematical Functions. Dover, New York

Ackroyd MH (1982) M/M/1 transient state occupancy probabilities via the discrete Fourier transform. IEEE Trans Commun 30: 357–559

Cantrell PE (1986) Computation of the transient M/M/1 queue cdf, pdf, and mean with generalized Q-functions. IEEE Trans Commun 34:814–817

Cantrell PE, Beall GR (1988) Transient M/M/1 queue variance computation using generalized Q-functions. IEEE Trans Commun 36:756–758

Cohen JW (1982) The Single Server Queue, 2nd edn. North Holland, Amsterdam

Coevering MCT van de (1992) The transient behavior of the M/G/1 queue (in Dutch), Master Thesis, Dept of Econometrics, Vrije University, Amsterdam

Conolly BW, Langaris Ch (1993) On a new formula for the transient state probabilities for M/M/1 queues and computational implications. J Appl Probab 30:237–246

Kleinrock L (1976) Queueing Syst Vol I. Wiley, New York

Leguesdron P, Pellaumail J, Rubino G, Sericola B (1993) Transient analysis of the M/M/1 queue. Adv Appl Probab 25:702–713

Morse PM (1955) Stochastic properties of waiting lines. Oper Res 3:225–261

Odoni AR, Roth E (1983) An empirical investigation of the transient behaviour of stationary queueing systems. Oper Res 31:432–455

Parthasarathy PR (1987) A transient solution to an M/M/1 queue: a simple approach. Adv Appl Probab 19:997–998

Press WH et al (1986) Numerical recipes. Cambridge University Press, Cambridge

Roth E (1985), A heuristic technique for the transient behaviour of Markovian queueing systems. Oper Res Lett 3:(6)301–305

Sharma OP (1990) Markovian queues. Ellis Horwood, New York

Takacs (1962) Introduction to the theory of queues. Oxford University Press, New York