Survey of transient finite server queues *

G. P. Phillips and A. E. Krzesinski
Department of Computer Science
University of Stellenbosch
South Africa
phil@cs.sun.ac.za

Abstract

Computationally efficient models are required for modelling *transient* finite server queueing systems.

A literature survey of transient queueing models is undertaken. The models are classified into two groups, those modelling queues with infinite capacity and those modelling queues with finite capacity. The models are further classified according to whether they are *exact* or *approximate*. Each model is summarized and gives an indication of the model's accuracy and computation.

The survey shows that some papers give solutions for queues without considering their computation while other papers present approximate models that are computationally efficient. If practical models are required then *approximate* models should be selected for further investigation.

*This work was funded by the Telkom Teletraffic Application Development Programme.

1 Introduction

1.1 Problem statement

We are interested in finding computationally efficient models for modelling *transient* finite server queues.

1.2 Scope and limitations

This paper investigates models for transient finite server queues. Two criteria are used to select the models. The first criterion is that the queue has a *finite* number of servers (infinite server queues are not considered). The second criterion is that the model yields an expression for the mean queue length, either explicitly or via the distribution of queue length.

The reason for investigating *transient* models is that communication networks operate under transient conditions caused by fluctuating user demands and changes in the network configuration. Furthermore we believe that there is insufficient evidence that accurate results are obtained when applying steady-state models to transient communication networks.

The accuracy and computation for the models are indicated, however numerical comparisons are outside the scope of this paper.

1.3 Notation

The shorthand queueing notation due largely to Kendall is used [Ken53]. This notation is augmented with the subscript t that denotes a time dependence for the parameters of the particular process. For example, while $M/M/1$ describes a queue with exponential arrival and service processes with constant parameters, $M_t/M/1$ denotes a queue with a time dependent arrival process parameter.

1.4 Overview

The models are classified into two groups, those modelling queues with infinite capacity and those modelling queues with finite capacity. The paper is concluded with a chronological summary of the various models investigated.

2 Queues with infinite capacity

This section surveys models for queues that have infinite capacity. Three exact models are discussed (the $M/M/1$, the $M/E_k/1$ and the non-homogeneous $M_t/M_t/1$ queue) followed by several approximate models. The models are listed according to their queueing structure in table 1.

2.1 Exact models

2.1.1 The M/M/1 queue

A closed-form expression for the transient queue length distribution for the $M/M/1$ queue was first derived by Clarke in an unpublished paper in 1953 [Cla53], while the first *published* results were given by Ledermann & Reuter [LR54]. The claim that Clarke produced the first transient solution is substantiated by a quote from Champernowne [Cha56],

> The simple queue problem with constant parameters, dealt with first by W. Ledermann and G. E. H. Reuter (1954) and by A. B. Clarke (1953) was considered by these writers in relation to far more general problems and the solutions are accordingly less simple than is possible in the special case ...

Ledermann & Reuter used spectral analysis to derive their results for a general birth-death process with state-dependent parameters. They include results for the special case of constant parameters (that is the $M/M/1$ queue) and state that in their opinion the standard method of generating functions cannot be used in this special case ([LR54], section 14). However Bailey shows that the generating function technique can be used to solve the $M/M/1$ queue [Bai54].

Although the queue length distribution for the $M/M/1$ queue is known, numerical procedures have not been given enough emphasis in the literature. Many textbooks present the queue length distribution in terms of modified Bessel functions without any discussion about the evaluation of this distribution. However there are several expressions that require less computation to evaluate.

- Takács derives an expression for queue length using results from the $M/M/1/K$ queue and taking the limit as the size of the waiting-line tends to infinity [Tak62].
- Jones et al. considers the problem of evaluating an infinite sum of Bessel functions [JCJ80]. The infinite sum is eliminated using Marcum's Q-function for which computational procedures are known [Mar60, DJ62, SBS66]. The introduction of this function leads to an expression for queue length in terms of Q-functions and a *finite* number of Bessel functions.
- Cantrell proposes a similar method where the infinite sum of Bessel functions is eliminated using the generalized Q-function [Can86]. The evaluation of the resulting expression is two to eight times faster than the evaluation of the expression given by Takács ([AW89], page 1103).

2.1.2 The M/E$_k$/1 queue

The Erlang random variable of type k is modelled by the sum of k independent and identically distributed exponential random variables. Therefore if the state of the queue is defined as the number of exponential phases waiting or being processed we have a Markov process (and therefore the queue can be described by Kolmogorov's system of equations).

The $M/E_k/1$ queue was solved for the idle probability $P_0(t)$ by Luchak [Luc56]. The phase probabilities $P_n(t)$ for $n > 0$ were found using $P_0(t)$ and Kolmogorov's system of equations. Note that the $M/E_k/1$ queue includes the $M/M/1$ queue as a special case when $k = 1$.

2.1.3 The non-homogeneous M$_t$/M$_t$/1 queue

The $M_t/M_t/1$ queue was solved by Clarke by converting Kolmogorov's equations into a *single* hyperbolic partial differential equation [Cla56]. This differential equation is solved using a standard method to obtain a solution in terms of a Volterra type integral equation ([Cla56], equations (17) and (18)). The solution is numerically difficult to evaluate since the kernel of the solution tends to infinity at one end of the range of integration ([Cla56], equation (18)). Wragg proposed an alternative method [Wra63]. This method is considered the most favourable out of 8 numerical methods investigated by Lees & Boyd [LB66].

Queue	Model accuracy	Author(s)
$M/M/1$	Exact	Cantrell [Can86], Jones [JCJ80], Takács [Tak62], Luchak [Luc56], Champernowne [Cha56], Morse [Mor55], Bailey [Bai54], Ledermann et al. [LR54], Clarke [Cla56]
	Approx.	Kotiah [Kot78], Rider [Rid76], Rothkopf & Oren [RO79]
$M/E_k/1$	Exact	Luchak [Luc56]
	Approx.	
$BMAP/G/1$	Exact	
	Approx.	Lucantoni, Choudhury & Whitt [LCW93]
$M_t/M/1$	Exact	Lees & Boyd [LB66], Wragg [Wra63]
	Approx.	Filipiak [Fil88], Clark [Cla81], Rothkopf & Oren [RO79], Rider [Rid76]
$M_t/M_t/s$	Exact	Kolmogorov's system of forward DEs
	Approx.	Clark [Cla81], Rothkopf & Oren [RO79]

Table 1: Models for infinite capacity queues.

2.2 Approximate models

2.2.1 Kotiah's model for the M/M/1 queue

The *Laplace transform* for the queue length distribution is given in terms of a complex root inside the unit circle of an auxiliary quadratic equation. Inverting this transform yields a complicated expression for the queue length distribution. Kotiah approximated this transform by rational polynomials that are easy to invert [Kot78]. The model has the notable feature that upper and lower bounds for the approximation error are determined. Unfortunately the model is limited in its current form since results are given only for an initial queue length of zero.

2.2.2 Rider's model for the M$_t$/M$_t$/1 queue

The kernel of Rider's model for the $M_t/M_t/1$ queue is the equation

$$\frac{d}{dt}E[Q(t)] \quad = \quad \lambda(t) - \mu(t)\left(1 - P_0(t)\right)$$

where $E[Q(t)]$ is the mean queue length [Rid76]. (This equation is derived by multiplying Kolmogorov's system of forward DEs by n and summing over n.) Although this equation seems straightforward, an expression for the idle probability $P_0(t)$ is involved even for the simple case of constant $\lambda(t)$ and $\mu(t)$. Instead of evaluating an exact expression for $P_0(t)$, $P_0(t)$ is approximated in terms of known quantities: mean queue length, mean arrival rate and mean service rate. The model is solved by numerical integration.

2.2.3 Rothkopf & Oren's model for the M$_t$/M$_t$/s queue

Rothkopf & Oren's model for the $M_t/M_t/1$ queue [RO79] uses the same kernel as Rider's model. The idle probability $P_0(t)$ however is approximated by the negative binomial distribution. This distribution is chosen for two reasons. Firstly this distribution ensures that the model yields exact steady state results ([RO79], page 525). Secondly the parameters of the negative binomial distribution are uniquely specified by its mean and variance. The significance of this is that by augmenting the kernel with a single equation, the distribution's parameters (and therefore $P_0(t)$) can be determined.

Similarly for the $M/M/s$ queue Rothkopf & Oren use the negative binomial distribution to approximate the probabilities $P_0, P_1, \ldots, P_{s-1}$. A correctional term is required to ensure that the model yields the correct steady state results.

2.2.4 Clark's model for the M$_t$/M$_t$/s queue

Clark's model for the $M_t/M_t/s$ queue uses a Polya-Eggenberger distribution [JK69] to reduce Kolmogorov's system of forward DEs to 5 differential equations ([Cla81], equations (15) to (19)). In addition to calculating the mean queue length, the queue length variance and mean waiting time are available from two supplementary equations. This model is more accurate than Rothkopf & Oren's model for $s > 1$.

2.2.5 Filipiak's model for the M$_t$/M/1 queue

Filipiak's model for the $M_t/M/1$ queue has the form [Fil88]

$$\frac{d}{dt}E[Q(t)] \quad = \quad -\mu\frac{E[Q(t)]}{E[Q(t)]+1} + \lambda(t)$$

If $\lambda(t)$ is constant this equation is solved by separation of variables [Agn75]. However if $\lambda(t)$ is time dependent then numerical integration is used to solve the model.

Filipiak's model approximates mean queue length and mean carried traffic. The model yields exact steady state results if the queue is homogeneous (that is if $\lambda(t)$ is constant), however an estimate of the model's approximation error under transient conditions is not given. The model is computationally efficient and is suited for modelling transient queueing *networks*.

2.2.6 Lucantoni's model for the BMAP/G/1 queue

The $BMAP/G/1$ queue is modelled by Lucantoni et al. [LCW93]. This queue has a batch Markovian arrival process that is able to model a general class of Markovian arrival processes. (The arrival process includes the phase-type renewal process and the Markov modulated Poisson process as special cases and allows correlated inter-arrival times and batch sizes.) The service times are independent and identically distributed random variables.

Numerical results for the queue length distribution are obtained by numerically inverting its Laplace transform.

3 Queues with finite capacity

This section surveys models for queues that have a finite capacity. Exact models are covered first followed by several approximate models. The models are listed in table 2 according to their queueing structure.

3.1 Exact models

The first solution for the queue length distribution for the $M/M/1/K$ queue was given by Morse [Mor58][1]. Takács derives an equivalent result using techniques from the theory of Linear Algebra [Tak62].

The $M/M/1/K$ queue with balking is solved by Abou-el-ata et al. [AAK93]. The solution for the queue length distribution is intricate and includes integrals of Bessel functions. No consideration is given to the evaluation of the solution.

[1] This author has not found a solution for the $M/M/1/K$ queue prior to the Morse's paper in 1958.

Queue	Model accuracy	Author(s)
$M/M/1/K$	Exact	Abou-el-ata et al. [AAK93], Takács [Tak62], Morse [Mor58]
	Approx.	Stern [Ste79]
$M_t/M/1/K$	Exact	Kolmogorov's system of forward DEs
	Approx.	Filipiak [Fil88], Moore [Moo75]
$M_t^X/E_Y/1/K$	Exact	
	Approx.	Moore [Moo75]

Table 2: Models for finite capacity queues.

3.2 Approximate models

3.2.1 Stern's model for the M/M/1/K queue

Stern's model for the $M/M/1/K$ queue takes advantage of the form of the queue length distribution for the exact model [Ste79]. The distribution is expressed as a sum of exponential terms and having the form

$$E[Q(t)] \;=\; L_0 + \sum_{i=1}^{K} L_i e^{-\alpha_i t}$$

where $\alpha_i > 0$. The computation required to evaluate $E[Q(t)]$ is reduced by removing insignificant terms from the summation. The terms considered insignificant are those terms $L_i e^{-\alpha_i t}$ with large values of α_i since these terms decay rapidly as t increases. The approximation error of the model is easily determined since the largest error occurs when t is zero.

3.2.2 Filipiak's model for the M$_t$/M/1/K queue

Filipiak's model for the $M_t/M/1/K$ queue is similar to Filipiak's model for the $M/M/1$ queue [Fil88]

$$\frac{d}{dt}E[Q(t)] \;=\; -\mu\, G(E[Q(t)]) + \lambda(t)$$

The functional form of $G(\cdot)$ is determined such that the model yields a small error over a wide range of conditions.

The model approximates mean queue length, mean rejected traffic and mean carried traffic. The model is computationally efficient and is suited for modelling transient queueing *networks*.

3.2.3 Moore's model for the M$_t^X$/E$_Y$/1/K queue

Moore's model approximates the $M_t^X/E_Y/1/K$ queue [Moo75]. This queue has an arrival process that is modelled by a Poisson distribution with arrival groups of random size X, and service times that are modelled by a weighted-sum Erlang distribution. The model approximates the non-homogeneous arrival process by dividing the time axis into intervals such that the arrival process in each interval is considered homogeneous. The imbedded Markov chain method is used to calculate the queue length distribution at the departure instants.

4 Chronological summary

The following table provides a chronological summary of the queueing models investigated in the previous sections. (The most recent paper is given first.)

The column labeled E shows whether the model is *exact* or approximate, the column labeled NR shows whether the paper includes *numerical results* for queue length while the column labeled QD shows whether the model calculates the *queue length distribution*. The last column shows some of the references made by the article.

Author(s)	Queue	E	NR	QD	Method	Importance of article	Refs made
Abou-el-ata et al. [AAK93]	$M/M/1/K$	✓	×	✓	Uses a generating function and Bessel functions.	Includes balking.	[Par87]
Lucantoni et al. [LCW93]	$BMAP/G/1$	×	✓	✓	Numerically inverts a Laplace transform.	Yields numerical results for a general Markov arrival process.	[AW92] [Ç67]
Filipiak [Fil88]	$M_t/M/1$ $M_t/M/1/K$ $M_t/M/c/c$	×	✓	×	Numerically integrates a single non-linear DE.	Yields a computationally efficient approximate solution.	[Rid76] [Moo75] [LB66]
Cantrell [Can86]	$M/M/1$	✓	✓	✓	The generalized Q-function replaces an infinite sum of Bessel functions.	Yields a computationally efficient solution.	[Ack82] [Ste79]
Jones et al. [JCJ80]	$M/M/1$	✓	✓	✓	Marcum's Q-function replaces an infinite sum of Bessel functions.	Avoids difficulties in the evaluation of an infinite sum of Bessel functions.	[Ste79]
Clark [Cla81]	$M_t/M_t/s$	×	✓	×	Approximates $P_s(t)$ and $P_{s+1}(t)$ using a Polya distribution. Integrates 5 DEs.	Improves accuracy of Rothkopf & Oren's model for $s > 1$.	[RO79]
Rothkopf & Oren [RO79]	$M_t/M_t/s$	×	✓	×	Approximates $P_0(t)$ using a negative binomial distribution. Integrates 2 DEs.	Improves accuracy of Rider's model.	[Rid76]
Stern [Ste79]	$M/M/1/K$	×	✓	×	Approximates the solution by the slowest transient component.	Yields a computationally efficient approximate solution.	[CS61]
Kotiah [Kot78]	$M/M/1$	×	✓	✓	Approximates a complex root by a rational polynomial.	Gives an error-bounds analysis for $P_n(t)$ and $E[Q(t)]$.	[Gav66]
Rider [Rid76]	$M_t/M/1$	×	✓	×	Approximates $P_0(t)$. Numerically integrates a single non-linear DE.	Yields a computationally efficient approximate solution.	[Cla56] [Koo72]

Table 3: A summary of models

Author(s)	Queue	E	N R	Q D	Method	Importance of article	Refs made
Moore [Moo75]	$M_t^X/E_Y/1/K$	$\times$	$\checkmark$	$\checkmark$	Uses the imbedded Markov chain method.	Yields numerical results for a non-homogeneous bulk arrival system.	[Koo72] [GW58]
Lees & Boyd [LB66]	$M_t/M_t/1$	$\checkmark$	$\checkmark$	$\times$	Uses Wragg's method of solution.	Compares 8 numerical methods. Yields numerical results for Wragg's method.	[Wra63]
Wragg [Wra63]	$M_t/M/1$	$\checkmark$	$\times$	$\checkmark$	Gives solution in terms of a Volterra integral equation.	Avoids numerical difficulties found in [Cla56].	[Cla56]
Takács [Tak62]	$M/M/1$ $M/M/1/K$	$\checkmark$	$\times$	$\checkmark$	Solves Kolmogorov's system of DEs using results from Linear Algebra.	Derives the $M/M/1$ solution from the $M/M/1/K$ solution.	[Mor58]
Morse [Mor58]	$M/M/1/K$	$\checkmark$	$\times$	$\checkmark$	Solves Kolmogorov's system of DEs directly.	Yields queue length probabilities for the $M/M/1/K$ queue.	
Luchak [Luc56]	$M/E_k/1$	$\checkmark$	$\checkmark$	$\checkmark$	System state represents the number of service phases to be completed.	Generalizes the $M/M/1$ queue to include the Erlang service distribution.	[Bai54] [LR54] [Mor55]
Clarke [Cla56]	$M_t/M_t/1$	$\checkmark$	$\times$	$\checkmark$	Uses a generating function to reduce system of DEs to a partial DE.	Yields results for the non-homogeneous $M/M/1$ queue.	[Bai54] [LR54]
Champernowne [Cha56]	$M/M/1$	$\checkmark$	$\times$	$\checkmark$	Uses advanced combinatorial methods.	Gives an alternate method of solution for the $M/M/1$ queue.	[LR54] [Cla53]
Morse [Mor55]	$M/M/1$	$\checkmark$	$\times$	$\checkmark$	Solves Kolmogorov's system of DEs directly.	Yields a solution in terms of integrals of trigonometrical terms.	
Bailey [Bai54]	$M/M/1$	$\checkmark$	$\times$	$\checkmark$	Uses the Laplace transform-generating function technique.	Yields the results of [LR54] using a generating function.	[LR54] [Ken53]
Ledermann & Reuter [LR54]	State dependent birth-death process.	$\checkmark$	$\times$	$\checkmark$	Uses spectral analysis.	Yields results for constant parameters λ and μ as a special case.	[Ken51]
Clarke [Cla53]	$M/M/1$	$\checkmark$	$\times$	$\checkmark$		First expression for queue length probabilities for the $M/M/1$ queue.	

Table 3: A summary of models (cont.)

References

[AAK93] M. O. Abou-el-ata, R. O. Al-seedy, and K. A. M. Kotb. Transient solution of the state-dependent queue: M/M/1/N with balking and reflecting barrier. *Microelectronics & Reliability*, 33(5):681–688, 1993.

[Ack82] Martin H. Ackroyd. M/M/1 transition state occupancy probabilities via the discrete fourier transform. *IEEE Transactions on Communications*, COM 30(3):557–559, 1982.

[Agn75] Carson E. Agnew. Dynamic modelling and control of congestion-prone systems. *Operations Research*, 24(3):400–419, 1975.

[AW89] Joseph Abate and Ward Whitt. Calculating time-dependent performance measures of the M/M/1 queue. *Queueing Systems*, 37(10):1102–1104, October 1989.

[AW92] Joseph Abate and Ward Whitt. The fourier-series method for inverting transforms of probability distributions. *Queueing Systems*, 10:5–88, 1992.

[Bai54] Norman T. J. Bailey. A continuous time treatment of a simple queue using generating functions. *Journal of the Royal Statistical Society (Series B)*, 16:288–291, 1954.

[Can86] Pierce E. Cantrell. Computation of the transient M/M/1 queue cdf, pdf, and mean generalized Q-Functions. *IEEE Transactions on Communications*, COM-34(8):814–817, August 1986.

[Ç67] E. Çinlar. Queues with semi-markovian arrivals. *Journal of Applied Probability*, 4:365–379, 1967.

[Cha56] D. G. Champernowne. An elementary method of solution of the queueing problem with a single server and constant parameters. *Journal of the Royal Statistical Society (Series B)*, 18(1):125–128, 1956.

[Cla53] A. B. Clarke. The time-dependent waiting line problem. Technical report, University of Michigan Engineering Research Institute, 1953.

[Cla56] A. B. Clarke. A waiting line process of Markov type. *The Annals of Mathematical Statistics*, 27:452–459, 1956.

[Cla81] Gordon M. Clark. Use of Polya distributions in approximate solutions to nonstationary M/M/s queues. *Communication of the ACM*, 24(4):206–217, April 1981.

[CS61] D. R. Cox and Walter L. Smith. *Queues*. Spottiswoode, Ballantyne & Co. Ltd., 1961.

[DJ62] A. R. DiDonata and M. P. Jarnagin. A method for computing the circular coverage function. *Mathematics of Computation*, 16:347–355, July 1962.

[Fil88] Janusz Filipiak. *Modelling and Control of Dynamic Flows in Communication Networks*. Springer-Verlag, 1988. ISBN 3-540-18292-6.

[Gav66] D. P. Gaver. Observing stochastic processes, and approximate transform inversion. *Operations Research*, 14:444–459, 1966.

[GW58] Herbert P. Galliher and R. Clyde Wheeler. Nonstationary queueing probabilities for landing congestion of aircraft. *Operations Research*, 6:607–623, 1958.

[JCJ80] Stephen K. Jones, R. K. Cavin, and D. A. Johnston. An efficient computational procedure for the evaluation of the $M/M/1$ transient state occupancy probabilities. *IEEE Transactions on Communications*, COM-28(12):2019–2020, December 1980.

[JK69] Norman L. Johnson and Samuel Kotz. *Distributions in statistics: Discrete distributions.* Houghton Mifflin Company (Boston, MA), 1969.

[Ken51] David G. Kendall. Some problems in the theory of queues. *Journal of the Royal Statistical Society (Series B)*, 13:151–185, 1951.

[Ken53] David G. Kendall. Stochastic processes occuring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, 24:338–354, 1953.

[Koo72] Bernard O. Koopman. Air-terminal queues under time-dependent conditions. *Operations Research*, 20:1089–1114, 1972.

[Kot78] T. C. T. Kotiah. Approximate transient analysis of some queueing systems. *Operations Research*, 26(2):333–346, March-April 1978.

[LB66] Eric L. Leese and D. W. Boyd. Numerical methods of determining the transient behaviour of queues. *Journal of the Operational Research Society*, 4:1–13, 1966.

[LCW93] David M. Lucantoni, Gagan L. Choudhury, and Ward Whitt. Computing transient distributions in general single-server queues. In *IEEE Global Telecommunications Conference (Houston, Texas)*, pages 1045–1050, 1993.

[LR54] W. Ledermann and G. E. H. Reuter. Spectral theory for the differential equations of simple birth and death processes. *Royal Society of London. Philosophical transactions A*, 246:321–369, 1954.

[Luc56] George Luchak. The solution of the single-channel queueing equations characterized by a time-dependent Poisson-distributed arrival rate and a general class of holding times. *Operations Research*, 4:711–732, 1956.

[Mar60] J. I. Marcum. A statistical theory of target detection by pulsed radar. *IRE Transactions on Information Theory*, IT-6:259–267, April 1960.

[Moo75] S. C. Moore. Approximating the behavior of nonstationary single server queues. *Operations Research*, 23(5):1011–1023, September-October 1975.

[Mor55] P. M. Morse. Stochastic properties of waiting lines. *Operations Research*, 3:255–261, 1955.

[Mor58] P. M. Morse. *Queues, Inventories and Maintenance.* John Wiley & Sons, Inc., 1958.

[Par87] P. R. Parthasarathy. A transient solution to a M/M1 queue: A simple approach. *Advances in Applied Probability*, 19:997–998, 1987.

[Rid76] K. L. Rider. A simple approximation to the average queue size in the time-dependent queue. *Journal of the ACM*, 23(2):361–367, 1976.

[RO79] Michael H. Rothkopf and Shmuel S. Oren. A closure approximation for the nonstationary M/M/s queue. *Management Science*, pages 522–534, June 1979.

[SBS66] M. Schwartz, W. R. Bennett, and S. Stein. *Communication Systems and Techniques*. Mc Graw-Hill, 1966.

[Ste79] Thomas E. Stern. Approximations of queue dynamics and their application to adaptive routing in computer communication networks. *IEEE Transactions on Communications*, 27(9):1331–1335, September 1979.

[Tak62] Lajos Takács. *Introduction to the Theory of Queues*. Oxford University Press, 1962.

[Wra63] A. Wragg. The solution of an infinite set of differential-difference equations occuring in polymerization and queueing problems. *Proceedings of the Cambridge Philosophical Society*, 59:117–124, 1963.