AN EXACT FCFS WAITING TIME ANALYSIS FOR A GENERAL CLASS OF G/G/s QUEUEING SYSTEMS

Dimitris BERTSIMAS

Mathematics Department and Operations Research Center, Massachusetts Institute of Technology, Rm. 2-342, Cambridge, Mass. 02139, U.S.A.

Received 18 September 1987
Revised 25 January 1988

Abstract

A closed form expression for the waiting time distribution under FCFS is derived for the queueing system $MGE_k/MGE_m/s$, where MGE_n is the class of mixed generalized Erlang probability density functions (pdfs) of order n, which is a subset of the Coxian pdfs that have rational Laplace transform. Using the calculus of difference equations and based on previous results of the author, it is proved that the waiting time distribution is of the form $1-\sum_{j=1}^{\binom{s+m-1}{s}} L_j e^{-u_j t}$, under the assumption that the roots u_j are distinct, i.e. belongs to the Coxian class of distributions of order $\binom{s+m-1}{s}$. The present approach offers qualitative insight by providing exact and asymptotic expressions, generalizes and unifies the well known theories developed for the G/G/1, G/M/s systems and leads to an $O(k^3\binom{s+m-1}{s}^3)$ algorithm, which is polynomial if only one of the parameters s or m varies, and is exponential if both parameters vary. As an example, numerical results for the waiting time distribution of the $MGE_2/MGE_2/s$ queueing system are presented.

Keywords: Multichannel queues, mixed generalized Erlang pdf, waiting time distribution

1. Introduction

The explicit evaluation, either by analytic or by numerical means, of the waiting time distribution in a general multi-server queueing system is known to present substantial difficulties. In this paper based on previous results of the author (Bertsimas [2]), which are summarized in the end of this section, we derive closed form expressions for the waiting time distribution under FCFS for the $MGE_k/MGE_m/s$ system, where MGE_n is the class of mixed generalized Erlang probability density functions (pdfs) of order n, which is a subset of the pdfs that have rational Laplace transform (R_n). It should be noted, however, that Schassberger [7] showed that a sequence of mixed Erlang distributions can be found which will converge weakly to any arbitrary distribution function. In the sense of pointwise convergence at points of continuity, we can then say that the class of mixed generalized Erlangs, which certainly includes the class of mixed Erlang distributions, is dense in the class of all distribution functions. The denseness of

this family gives an indication of the theoretical comprehensiveness of the mixed generalized Erlang as a practical modeling tool.

Concerning exact calculations for the waiting time distribution of multiserver queueing systems, when we go beyond the exponentially assumption for the service time pdf, which seldom holds in practice, the relevant analytic and computational problems become really challenging, especially if this also happens for the interarrival time pdf. Notable among recent works concerning the derivation of the waiting time distribution are those of Pollaczek [5], for the $G/R_m/s$, Avis [1] for the $M/E_2/2$, $E_k/E_2/2$ and $D/E_2/2$, de Smit [8] for the $G/H_m/s$, Ishikawa [4] for the $G/E_m/s$ and Ramaswami and Lucantoni [6] for the $G/PH/s$. For a thorough account of approximations, purely numerical methods and asymptotic results corresponding to the waiting time distribution of multi-server queues with non-exponential service times, see Tijms [10].

We close this section with a description of the system, an explanation of its structure, the notation used and review of the results for the steady-state probability distribution for the number of customers in the system from [2] that we will use for the derivation of the waiting time distribution. In section 2, we write down the equations that describe the system and then use the calculus of difference equations to derive a closed form expression for the waiting time distribution. Furthermore, we show that the results for two seemingly very different systems $G/G/1$ and $G/M/s$, which are traditionally analysed with very different methods (Wiener-Hopf decomposition and imbedded Markov chain respectively) are unified using the results of the present paper in the case of mixed generalized Erlang distributions. In section 3, we examine the asymptotic behavior of the waiting time distribution and in the final section 4, the established theoretical results, are used to write a computationally efficient algorithm in order to extract numerical results. As an example, the algorithm is applied to the general class of MGE_2 distributions, i.e. for the $MGE_2/MGE_2/s$ system, for which numerical results are reported.

We shall examine, henceforth, an s identical server single waiting line queueing system with interarrival and service time distributions of mixed generalized Erlang type of order k and m respectively, which is a proper subset of the Coxian distributions that have rational Laplace transform. The queue discipline is first-come-first-served (FCFS).

The general Coxian class C_n was introduced in Cox's [3] pioneering paper. The stage representation of the Coxian distribution is presented in fig. 1. It should be noted that this stage representation of the Coxian distribution is purely formal in the sense that the branching probabilities q_i can be negative and the rates μ_i can be complex numbers. The mixed generalized Erlang distribution is a Coxian distribution, where we assume that the probabilities q_i are non-negative and the rates μ_i are reals. As a result, the mixed generalized Erlang distribution has a valid probabilistic interpretation, which is further exploited in this paper.

To analyse the model we conceive of the arrival process as an arrival timing

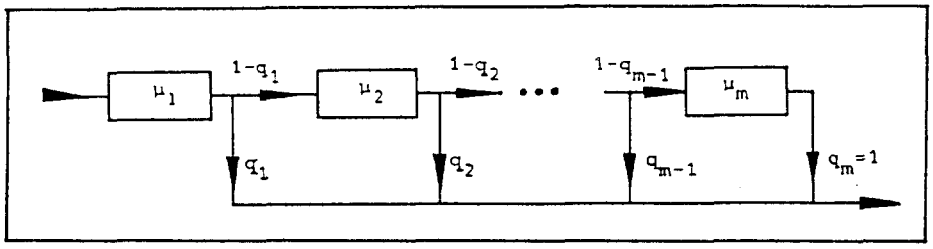

Fig. 1. The C_m class of distributions.

channel (ATC) consisting of k consecutive stages with rates $\lambda_1, \lambda_2, \ldots, \lambda_k$ and with probabilities $p_1, p_2, \ldots, p_k \triangleq 1$ of entering the system after the completion of the 1st, 2nd, ..., kth stage. We remark that as soon as a customer in the ATC enters the system a new customer arrives at stage 1 of the ATC. For the service time distribution we consider as above a service-timing channel (STC) consisting of m consecutive stages with rates $\mu_1, \mu_2, \ldots, \mu_m$ and with probabilities $q_1, q_2, \ldots, q_m \triangleq 1$ of leaving the system.

For the steady state we introduce the random variables:

1. $N \triangleq$ The number of customers in the system
2. $N^- \triangleq$ The number of customers seen by an arriving customer just before his arrival
3. $R_a \triangleq$ The number of the ATC stage currently occupied by the arriving customer
4. $R_j \triangleq$ The number of customers being served at the jth STC stage ($j = 1, 2, \ldots, m$)
5. $R_j^- \triangleq$ The number of customers being served at the jth STC stage ($j = 1, 2, \ldots, m$), just before the arrival of an entering customer
6. $T_q \triangleq$ The waiting time of an arriving customer
7. $L_q \triangleq$ The length of the queue.

For simplicity of notation we introduce the vectors of random variables

$$\boldsymbol{R} \triangleq (R_1, \ldots, R_m), \quad \boldsymbol{R}^- \triangleq (R_1^-, \ldots, R_m^-)$$

and also we will use the notation:

$$\delta_j \triangleq (0, \ldots, 0, 1, 0, \ldots, 0), \quad a(s, m) \triangleq \binom{s+m-1}{s},$$

where $|\boldsymbol{i}| = s$ means that $\sum_{j=1}^m i_j = s$.

With the above definitions the system can be formulated as a continuous time Markov chain with infinite state space:

$$\left\{(N, R_a, R_1, \ldots, R_m),\ N = 0, 1, \ldots, R_a = 1, 2, \ldots, k,\ \sum_{j=1}^m R_j = \min(N, s)\right\},$$

where the states with $N < s$ (at least one server free) and $N \geq s$ (all servers busy) will be termed "unsaturated" and "saturated" respectively. We now introduce the

following set of probabilities:

$$P_{n,l,\mathbf{i}} \triangleq \Pr\{N=n,\ R_a=l,\ \boldsymbol{R}=\boldsymbol{i}\}, \qquad P^-_{n,i} \triangleq \Pr\{N=n,\ \boldsymbol{R}^-=\boldsymbol{i}\},$$
$$P_n \triangleq \Pr\{N=n\},\ P^-_n \triangleq \Pr\{N^-=n\}.$$

We also define:
$f^*_{T_a}(\theta)$, $f^*_{T_s}(\theta) \triangleq$ The Laplace transform of the interarrival and service time distributions respectively.
$1/\lambda \triangleq$ The mean interarrival time, $1/\mu \triangleq$ The mean service time, $\rho \triangleq \lambda/s\mu =$ The traffic intensity.
$C_a^2 \triangleq$ The squared coefficient of variation of the interarrival distribution.
$C_s^2 \triangleq$ The squared coefficient of variation of the service time distribution.

In Bertsimas [2], after writing the equations for $P_{n,l,i}$ and using separation of variables and a generating function technique, the following closed form expressions were established for $\rho < 1$, under the assumption that the roots w_j are distinct (we did not find any case where the roots w_j are not distinct; in the case of $m=2$ we will prove that the roots w_j are positive and distinct):

$$P_{n,l,i} = \sum_{j=1}^{a(s,m)} B_j \left(\prod_{r=1}^{l-1} \frac{(1-p_r)\lambda_r}{x(w_j)+\lambda_{r+1}} \right) f(\boldsymbol{i},\ w_j) w_j^n$$
$$n \geqslant s,\ l=1,\ldots,k,\ |\boldsymbol{i}| = s, \tag{1}$$

where B_j are computed by solving a linear system of $a(s, m)$ equations with $a(s,m)$ unknowns, $f(\boldsymbol{i},\ w_j)$ are computed from the following equations:

$$f\big((0,0,\ldots,s),\ w_j\big) = 1$$
$$\sum_{r=1}^{m} (1-q_r)\mu_r(i_r+1) f\big(\boldsymbol{i}+\delta_r-\delta_{r+1},\ w_j\big)$$
$$+ w_j \sum_{r=2}^{m} q_r\mu_r(i_r+1) f\big(\boldsymbol{i}+\delta_r-\delta_1,\ w_j\big)$$
$$+ w_j q_1\mu_1 i_1 f(\boldsymbol{i},\ w_j) = f(\boldsymbol{i},\ w_j)\left\{ \sum_{r=1}^{m} i_r\mu_r - x(w_j) \right\},$$
$$|\boldsymbol{i}| = s,\ j=1,\ldots,a(s,\ m). \tag{2}$$

Each of the $a(s, m)$ roots w_j satisfies the following system of nonlinear equations (the subscripts j corresponds to one of the $a(s, m)$ combinations of the vector $\boldsymbol{i}=(i_1, i_2,\ldots, i_m)$, $\Sigma_{r=1}^m i_r = s$ and for simplicity of notation $x(w_j)$ is simply written x):

$$\phi_i(x) \triangleq i_1\theta_1(x) + i_2\theta_2(x) + \ldots + i_m\theta_m(x) = x, \quad i_1+i_2+\ldots+i_m = s, \tag{3}$$

where $\theta_j(x)$ $(j=1,\ldots,m)$ are the m roots of the polynomial equation of degree m:

$$f^*_{T_a}(x) f^*_{T_s}\big(-\theta_j(x)\big) = 1 \tag{4}$$

and

$$w = f_{T_a}^*(x). \tag{5}$$

The generating function of the coefficients $f(\boldsymbol{i}, w_j)$

$$G_j(z) \triangleq \sum_{|\boldsymbol{i}|=s} f(\boldsymbol{i}, w_j) z_1^{i_1} \dots z_m^{i_m}, \quad \boldsymbol{i} = (i_1, \dots, i_m), \quad z = (z_1, \dots, z_m)$$

satisfies the following linear partial differential equation

$$\sum_{r=1}^{m} \frac{\partial G_j(z)}{\partial z_r} \left(\mu_r z_r - w_j z_1 q_r \mu_r - (1 - q_r)\mu_r z_{r+1}\right) = x(w_j) G_j(z) \tag{6}$$

whose solution was found to be (each j corresponding to a vector $\boldsymbol{i}$):

$$G_j(z) = \prod_{r=1}^{m} \left(\frac{b_{1,r}(w_j) z_1 + \dots + b_{m,r}(w_j) z_m}{b_{m,r}(w_j)} \right)^{i_r} \tag{7}$$

where the coefficients $b_{i,r}(w_j)$ are computed from the expansion of a determinant.

The "unsaturated" probabilities $P_{n,l,\boldsymbol{i}}$ are of the form:

$$P_{n,l,\boldsymbol{i}} = \sum_{j=1}^{a(s,m)} B_j g(n, l, \boldsymbol{i}, w_j) \quad n < s, \quad l = 1, \dots, k, \quad |\boldsymbol{i}| = n \tag{8}$$

where the coefficients $g(n, l, \boldsymbol{i}, w_j)$ are computed recursively. Furthermore, the pre-arrival probabilities $P_{n,\boldsymbol{i}}^-$ were found to be:

$$P_{n,\boldsymbol{i}}^- = \frac{1}{\lambda} \sum_{j=1}^{a(s,m)} B_j f(\boldsymbol{i}, w_j)\left(\lambda_1 + x(w_j)\right) w_j^{n+1} \quad n \geqslant s, \ |\boldsymbol{i}| = s. \tag{9}$$

2. Waiting time analysis under FCFS

We denote

$$W(t) \triangleq \Pr\{0 < T_q \leqslant t\}$$

$$F_{T_q}(t) \triangleq \Pr\{T_q \leqslant t\} = W(t) + \Pr\{T_q = 0\}$$

$$F_{n,\boldsymbol{i}}(t) \triangleq \Pr\{0 < T_q \leqslant t \mid N^- = n + s, \ \boldsymbol{R}^- = \boldsymbol{i}\}.$$

In this section we shall derive closed form expressions for $W(t)$ and $F_{T_q}(t)$, the probability distribution for the waiting time under FCFS of an arriving customer.

THEOREM 1

Under the assumption that the roots w_j of the system (3), (4) and (5) are distinct for $\rho < 1$ the waiting time distribution for the $\mathrm{MGE}_k/\mathrm{MGE}_m/s$ is given by:

$$W(t) \triangleq \Pr\{0 < T_q \leqslant t\}$$

$$= \frac{1}{\lambda} \sum_{j=1}^{a(s,m)} B_j G_j(\mathbf{1})\left(\lambda_1 + x(w_j)\right) \frac{w_j^{s+1}}{1 - w_j} \left(1 - \mathrm{e}^{-x(w_j)t}\right).$$

Proof

By conditioning on N^- and $\boldsymbol{R}^-$ we easily find that

$$W(t) = \sum_{|\boldsymbol{i}|=s} \sum_{n=0}^{\infty} F_{n,\boldsymbol{i}}(t) P^-_{n+s,\boldsymbol{i}}.$$

Using (9), we can write this as

$$W(t) = \frac{1}{\lambda} \sum_{j=1}^{a(s,m)} B_j\big(\lambda_1 + x(w_j)\big) w_j^{s+1} \left\{ \sum_{|\boldsymbol{i}|=s} f(\boldsymbol{i}, w_j) \sum_{n=0}^{\infty} F_{n,\boldsymbol{i}}(t) w_j^n \right\}. \tag{10}$$

By conditioning on the next event in the interval $(t, t+\delta t)$ and taking the limit as $\delta t \to 0$ we can write the equations that $F_{n,\boldsymbol{i}}(t)$ satisfy

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t} F_{n,\boldsymbol{i}}(t) + F_{n,\boldsymbol{i}}(t) \sum_{r=1}^{m} i_r \mu_r \\ &\quad = \sum_{r=1}^{m} i_r q_r \mu_r F_{n-1,\boldsymbol{i}-\boldsymbol{\delta}_r+\boldsymbol{\delta}_1}(t) + \sum_{r=1}^{m} (1-q_r)\mu_r i_r F_{n,\boldsymbol{i}-\boldsymbol{\delta}_r+\boldsymbol{\delta}_{r+1}}(t) \\ &\quad n \geqslant 0, \ |\boldsymbol{i}| = s \end{aligned} \tag{11}$$

where $F_{-1,\boldsymbol{i}}(t) \triangleq 1$, by definition. We define the Laplace transforms

$$W^*(\theta) \triangleq \mathscr{L}(W(t)) \text{ and } \Phi^*_{n,\boldsymbol{i}}(\theta) \triangleq \mathscr{L}(F_{n,\boldsymbol{i}}(t))$$

and the quantities

$$A_{\boldsymbol{i},j}(\theta) \triangleq \sum_{n=0}^{\infty} w_j^n \Phi^*_{n,\boldsymbol{i}}(\theta),$$

$$H_j(\theta) \triangleq \sum_{|\boldsymbol{i}|=s} f(\boldsymbol{i}, w_j) A_{\boldsymbol{i},j}(\theta).$$

Then from (10) we obtain

$$W^*(\theta) = \frac{1}{\lambda} \sum_{j=1}^{a(s,m)} B_j\big(\lambda_1 + x(w_j)\big) w_j^{s+1} H_j(\theta). \tag{12}$$

We now transform (11) and obtain

$$\begin{aligned} &\left(\theta + \sum_{r=1}^{m} i_r \mu_r\right) \Phi^*_{n,\boldsymbol{i}}(\theta) \\ &\quad = \sum_{r=1}^{m} i_r q_r \mu_r \Phi^*_{n-1,\boldsymbol{i}-\boldsymbol{\delta}_r+\boldsymbol{\delta}_1}(\theta) + \sum_{r=1}^{m} (1-q_r)\mu_r i_r \Phi^*_{n,\boldsymbol{i}-\boldsymbol{\delta}_r+\boldsymbol{\delta}_{r+1}}(\theta) \\ &\quad n \geqslant 0, \ |\boldsymbol{i}| = s \end{aligned} \tag{13}$$

where

$$\Phi^*_{-1,\boldsymbol{i}}(\theta) \triangleq \mathscr{L}\{F_{-1,\boldsymbol{i}}(t)\} = \mathscr{L}\{1\} = 1/\theta.$$

The strategy for obtaining a closed form expression for $W^*(\theta)$ is the following:

1. we obtain a difference equation for $A_{i,j}(\theta)$, and
2. we multiply the equation for $A_{i,j}(\theta)$ by the coefficients $f(\boldsymbol{i}, w_j)$, add with respect to $\boldsymbol{i}$ and using the equations (2), we are able to solve for $H_j(\theta)$.

We multiply (13) with w_j^n and add with respect to n to find

$$\left(\theta + \sum_{r=1}^{m} i_r\mu_r\right)A_{i,j}(\theta) = \frac{1}{\theta}\sum_{r=1}^{m} i_r q_r \mu_r + w_j \sum_{r=1}^{m} i_r q_r \mu_r A_{i-\delta_r+\delta_1,j}(\theta) + \sum_{r=1}^{m} (1-q_r)\mu_r i_r A_{i-\delta_r+\delta_{r+1},j}(\theta). \tag{14}$$

Performing now the second step we take

$$\theta H_j(\theta) + \sum_{|\boldsymbol{i}|=s} f(\boldsymbol{i}, w_j)A_{i,j}(\theta)\sum_{r=1}^{m} i_r\mu_r = \frac{1}{\theta}\sum_{r=1}^{m} q_r\mu_r \sum_{|\boldsymbol{i}|=s} i_r f(\boldsymbol{i}, w_j) + \sum_{|\boldsymbol{i}|=s} A_{i,j}(\theta)$$
$$\left\{ w_j q_1 \mu_1 i_1 f(\boldsymbol{i}, w_j) + w_j \sum_{r=2}^{m} q_r\mu_r(i_r+1)f(\boldsymbol{i}+\delta_r-\delta_1, w_j) + \sum_{r=1}^{m}(1-q_r)\mu_r(i_r+1)f(\boldsymbol{i}+\delta_r-\delta_{r+1}, w_j)\right\}. \tag{15}$$

Substituting (2) into the right side of (15) we find

$$\theta H_j(\theta) + \sum_{|\boldsymbol{i}|=s} f(\boldsymbol{i}, w_j)A_{i,j}(\theta)\sum_{r=1}^{m} i_r\mu_r = \frac{1}{\theta}\sum_{r=1}^{m} q_r\mu_r \sum_{|\boldsymbol{i}|=s} i_r f(\boldsymbol{i}, w_j) + \sum_{|\boldsymbol{i}|=s} A_{i,j}(\theta) f(\boldsymbol{i}, w_j)\left(\sum_{r=1}^{m} i_r\mu_r - x(w_j)\right). \tag{16}$$

Since in (16) the term

$$\sum_{|\boldsymbol{i}|=s} f(\boldsymbol{i}, w_j)A_{i,j}(\theta)\sum_{r=1}^{m} i_r\mu_r$$

cancels from both sides we get

$$\theta H_j(\theta) = \frac{1}{\theta}\sum_{r=1}^{m} q_r\mu_r \sum_{|\boldsymbol{i}|=s} i_r f(\boldsymbol{i}, w_j) - x(w_j)H_j(\theta),$$

which gives

$$H_j(\theta) = \frac{\sum_{r=1}^{m} q_r\mu_r \sum_{|\boldsymbol{i}|=s} i_r f(\boldsymbol{i}, w_j)}{\theta(\theta + x(w_j))}. \tag{17}$$

From (6) for $z = \mathbf{1} \triangleq (1, 1, \ldots, 1)$ we find that

$$\sum_{r=1}^{m} q_r \mu_r \sum_{|i|=s} i_r f(i, w_j) = \sum_{r=1}^{m} q_r \mu_r \frac{\partial G_j(z)}{\partial z_r}\bigg|_{z=\mathbf{1}} = \frac{x(w_j)}{1-w_j} G_j(\mathbf{1}).$$

Therefore

$$H_j(\theta) = \frac{x(w_j) G_j(\mathbf{1})}{1-w_j} \frac{1}{\theta(\theta + x(w_j))}. \tag{18}$$

Substituting (18) into (12) we obtain

$$W^*(\theta) = \frac{1}{\lambda} \sum_{j=1}^{a(s,m)} B_j G_j(\mathbf{1})(\lambda_1 + x(w_j)) \frac{w_j^{s+1}}{1-w_j} \frac{x(w_j)}{\theta(\theta + x(w_j))}.$$

Using partial fractions we find

$$W^*(\theta) = \frac{1}{\lambda} \sum_{j=1}^{a(s,m)} B_j G_j(\mathbf{1})(\lambda_1 + x(w_j)) \frac{w_j^{s+1}}{1-w_j} \left(\frac{1}{\theta} - \frac{1}{\theta + x(w_j)} \right). \tag{19}$$

Now the inversion of (19) is an easy task. Thus, under the assumption that the roots w_j are distinct, so that equation (9) holds, we have

$$W(t) \triangleq \Pr\{0 < T_q \leqslant t\}$$

$$= \frac{1}{\lambda} \sum_{j=1}^{a(s,m)} B_j G_j(\mathbf{1})(\lambda_1 + x(w_j)) \frac{w_j^{s+1}}{1-w_j} (1 - e^{-x(w_j)t}). \quad \square \tag{20}$$

As a check on the algebra we verify that

$$\lim_{t \to \infty} W(t) = \Pr\{T_q > 0\} = \sum_{n=s}^{\infty} P_n^- = \sum_{n=s}^{\infty} \sum_{|i|=s} P_{n,i}^-.$$

Also

$$F_{T_q}(t) = 1 - \frac{1}{\lambda} \sum_{j=1}^{a(s,m)} B_j G_j(\mathbf{1}) \frac{w_j^{s+1}}{1-w_j} (\lambda_1 + x(w_j))\, e^{-x(w_j)t} \tag{21}$$

where from (5) $w_j = f_{T_a}^*(x(w_j))$. It is remarkable that the waiting time pdf has also rational Laplace transform, i.e. it belongs to the Coxian class of distributions of order $\binom{s+m-1}{s}$. From (21) it is easy to find the following compact expression for the rth moment of T_q:

$$E\{T_q^r\} = \frac{r!}{\lambda} \sum_{j=1}^{a(s,m)} B_j G_j(\mathbf{1}) \frac{w_j^{s+1}}{1-w_j} \frac{(\lambda_1 + x(w_j))}{[x(w_j)]^r}.$$

As an additional check on the algebra we calculate the factorial moments of L_q, namely

$$E\{L_q(L_q-1)\dots(L_q-r+1)\} \triangleq \sum_{n=s}^{\infty}(n-s)(n-s-1)\dots(n-s-r+1)P_n.$$

Since for $n \geqslant s$

$$P_n = \sum_{l=1}^{k}\sum_{|i|=s} P_{n,l,i}$$

we find from (1)

$$P_n = \sum_{j=1}^{a(s,m)} B_j G_j(\mathbf{1})\frac{\lambda_1 + x(w_j)}{x(w_j)} w_j^n(1-w_j), \quad n \geqslant s$$

from which, after algebraic manipulations and using the identity

$$\sum_{n=0}^{\infty} n(n-1)\dots(n-r+1)a^{n-r} = \frac{r!}{(1-a)^{r+1}}$$

we find that

$$E\{L_q(L_q-1)\dots(L_q-r+1)\} = r!\sum_{j=1}^{a(s,m)} B_j G_j(\mathbf{1})\frac{w_j^{s+r}}{(1-w_j)^r}\frac{(\lambda_1 + x(w_j))}{x(w_j)}.$$

For $r=1$ we verify Little's formula $E\{L_q\} = \lambda E\{T_q\}$. For $k=1$ the model becomes M/MGE$_m$/s and we obtain the well known result which holds for M/G/s:

$$E\{L_q(L_q-1)\dots(L_q-r+1)\} = \lambda^r E\{T_q^r\},$$

where we used the fact that in this case

$$w_j = f_{T_a}^*(x(w_j)) = \frac{\lambda}{\lambda + x(w_j)} \quad (\lambda_1 = \lambda).$$

We conclude this section with some examples, which show that the present approach generalizes and unifies the well known results for the G/G/1, G/M/s systems, when the distributions involved are mixed generalized Erlangs.

1. MGE$_k$/MGE$_m$/1

Since the only combination of $\boldsymbol{i}$ for $s=1$ are of the type $\boldsymbol{i} = (0,\dots,0,1,0,\dots,0)$ ($a(1, m) = m$) we verify a well-known result from G/G/1 theory that $F_{T_q}(t)$ is a

sum of m exponentials of the form of (21) where in this case $x(w_j)$, $j = 1, \ldots, m$ are the m roots of the equation

$$f_{T_a}^*(x) f_{T_s}^*(-x) = 1$$

subject to the constraint $|f_{T_a}^*(x)| < 1$.

2. $\mathrm{MGE}_k/\mathrm{M}/s$

For $m = 1$, (4) becomes

$$f_{T_a}^*(x) f_{T_s}^*\left(-\frac{x}{s}\right) = 1 \Rightarrow x = s\mu\left(1 - f_{T_a}^*(x)\right).$$

Since $a(s, 1) = 1$ we find the well known result for $\mathrm{G}/\mathrm{M}/s$

$$1 - F_{T_q}(t) = K_1 \, \mathrm{e}^{-xt},$$

x being the unique root of the above equation and K_1 a constant.

3. $\mathrm{MGE}_k/\mathrm{MGE}_2/s$

For $m = 1, 2, 3, 4$ we can solve (4) in the closed form, since it is a polynomial equation of degree m. For $m = 2$ we find that

$$\theta_1(x),\ \theta_2(x) = \tfrac{1}{2}\left\{\mu_1 + \mu_2 - q_1\mu_1 f_{T_a}^*(x) \pm \sqrt{\Delta\left(f_{T_a}^*(x)\right)}\right\},$$

where $\theta_1(x)$ corresponds to the $+$sign and $\theta_2(x)$ corresponds to the $-$sign and

$$\Delta(y) = (\mu_1 + \mu_2 - yq_1\mu_1)^2 - 4\mu_1\mu_2(1 - y).$$

Therefore (3) becomes

$$(s - 2i)\sqrt{\Delta\left(f_{T_a}^*(x)\right)} - s\left(\mu_1 + \mu_2 - q_1\mu_1 f_{T_a}^*(x)\right) + 2x = 0, \quad i = 0, 1, \ldots, s. \tag{22}$$

In this case we can prove that the roots $x(w_j)$ are positive and distinct.

LEMMA 2

The roots w_j are positive and distinct and satisfy

$$0 < x(w_0) < x(w_1) < \ldots < x(w_s), \quad 0 < w_s < w_{s-1} < \ldots < w_0 < 1. \tag{23}$$

Proof

Denote by $\phi_i(x)$ the expression in the left side of (22). We observe that $\phi_i(x)$ is a continuous function on the set of real numbers $\mathscr{R}$ with the properties:

1. $\lim_{x \to \infty} \phi_i(x) = \infty$;
2. $\phi_i(0) = -2i(\mu_2 + \mu_1(1 - q_1)) < 0$ for $i = 1, 2, \ldots, s$; and
3. $\phi_0(x) = -2x(1 - \rho)/\rho + o(x)$ as $x \to 0$. (This result is established by using Taylor expansion of $\phi_0(x)$.)

From the above properties it is clear that the root $x(w_i)$ of the equation $\phi_i(x) = 0$ are positive and thus $w_i = f^*_{T_a}(x(w_i))$ are positive too. Furthermore, $\phi_i(x) = a(x) - ib(x)$, with $b(x)$ being positive for all x. Therefore for a fixed value of x we have that $\phi_0(x) > \phi_1(x) > \ldots > \phi_s(x)$, which gives $0 < x(w_0) < x(w_1) < \ldots < x(w_s)$. Since the function $f^*_{T_a}(x)$ is decreasing for $x > 0$, it follows that

$$0 < w_s < w_{s-1} < \ldots < w_0 < 1. \quad \square$$

A surprising consequence of the above lemma is the fact that we can analyse this model using real arithmetic. This is not true for any m.

3. Asymptotic results

In this section we verify and extend some asymptotic results for the $\mathrm{MGE}_k/\mathrm{MGE}_m/s$ queue. Takahashi [9] proved that in a $\mathrm{PH}/\mathrm{PH}/s$ system the stationary probability Π_m that there are more than m customers waiting in the queue behaves asympototically as

$$\Pi_m \sim K_2 \eta^m \quad (m \to \infty),$$

where the constant K_2 was not computed and $\eta = f^*_{T_a}(sy)$, y being the unique positive root of the equation

$$f^*_{T_a}(sy) f^*_{T_s}(-y) = 1.$$

Furthermore, he proved that as $t \to \infty$

$$1 - F_{T_q}(t) \sim K_3 \, \mathrm{e}^{-syt}, \quad \frac{K_2}{K_3} = \frac{\lambda(1-\eta)}{sy}.$$

In order to see the connection between these results and the results of the present paper, we observe that

$$\Pi_m = \sum_{n=m+s+1}^{\infty} P_n = \sum_{j=1}^{a(s,m)} B_j G_j(\mathbf{1}) \frac{\lambda_1 + x(w_j)}{x(w_j)} w_j^{s+1} w_j^m.$$

Thus asymptotically as $m \to \infty$

$$\Pi_m \sim B_1 G_1(\mathbf{1}) \frac{\lambda_1 + x(w_1)}{x(w_1)} w_1^{s+1} w_1^m,$$

where w_1 is the root corresponding to the combination of $\boldsymbol{i} = (0, \ldots, s)$. Specializing (3), (4) and (5) we find that $w_1 = f^*_{T_a}(x(w_1))$ where $x(w_1)$ is the unique positive root of the equation

$$f^*_{T_a}(x) f^*_{T_s}\left(-\frac{x}{s}\right) = 1.$$

Letting

$$y = \frac{x(w_1)}{s}, \quad \eta = w_1$$

we see that the two results are identical. We also observe from (21) that as $t \to \infty$

$$1 - F_{T_q}(t) \to B_1 G_1(\mathbf{1})(\lambda_1 + x(w_1)) \frac{w_1^{s+1}}{1 - w_1} e^{-x(w_1)t}.$$

Again the two results are identical and we can also easily verify that

$$\frac{K_2}{K_3} = \frac{\lambda(1 - w_1)}{x(w_1)}.$$

It should also be stressed that in our expressions we are able to compute explicitly the constants K_2, K_3.

4. Computational and complexity considerations

In order to extract numerical results from the formulae presented in the section 3 the author in [2] has proposed an algorithm with complexity $O(k^3(\binom{s+m-1}{s})^3)$, which is polynomial if only one of the parameters s or m varies, but is exponential if both parameters vary. In other words, for an arbitrary interarrival distribution and a given service time distribution the problem of determining the waiting time distribution under FCFS can be solved in time polynomial in the number of servers. This algorithm is summarized as follows:

1. determination of the $\binom{s+m-1}{s}$ roots w_j of the system of equations (3)–(5);
2. recursive determination of the coefficients $f(\boldsymbol{i}, w_j)$ from (2);
3. recursive determination of the coefficients $g(n, l, \boldsymbol{i}, w_j)$ in (8) from the equations that $P_{n,l,\boldsymbol{i}}$ satisfy for $n < s$; and
4. determination of the $\binom{s+m-1}{s}$ unknowns B_j as a solution of a nonhomogeneous linear system with $\binom{s+m-1}{s}$ equations.

We are currently investigating a parallel implementation of the above algorithm, based on the fact that the equations (3)–(5) are separable.

To fully gauge the performance of the proposed algorithm we programmed it in FORTRAN 77 on a SUN 3 for the general class of the MGE_2 distributions, i.e. for the $MGE_2/MGE_2/s$ system. The reasons for choosing this model are that it is representative of the general behavior of the algorithm for more general models, is rather unexpectedly (as shown in lemma 2) in real arithmetic, its complexity is $O(s^3)$ and allows the determination of exact results when the coefficients of variation of the interarrival and of the service time pdf are both bigger than 1/2. In the implementation of the above algorithm in this case we used the Newton-Raphson method to determine the $s + 1$ roots w_j. Exploiting

the established ordering of the roots (equation (23)), as a starting point for the next root we used the previous one. In this way we observed that the method converged very fast. Steps (2), (3) and (4) were implemented in a straightforward way. From a computational point of view the heaviest part of the algorithm is the third step, because there is a large number of unknowns ($O(s^3)$ in this case).

Ramaswami and Lucantoni [6] extended in an elegant way the potential of the matrix-geometric method developed by M. Neuts. They used the randomization technique to obtain an algorithm for computing the waiting time distribution for

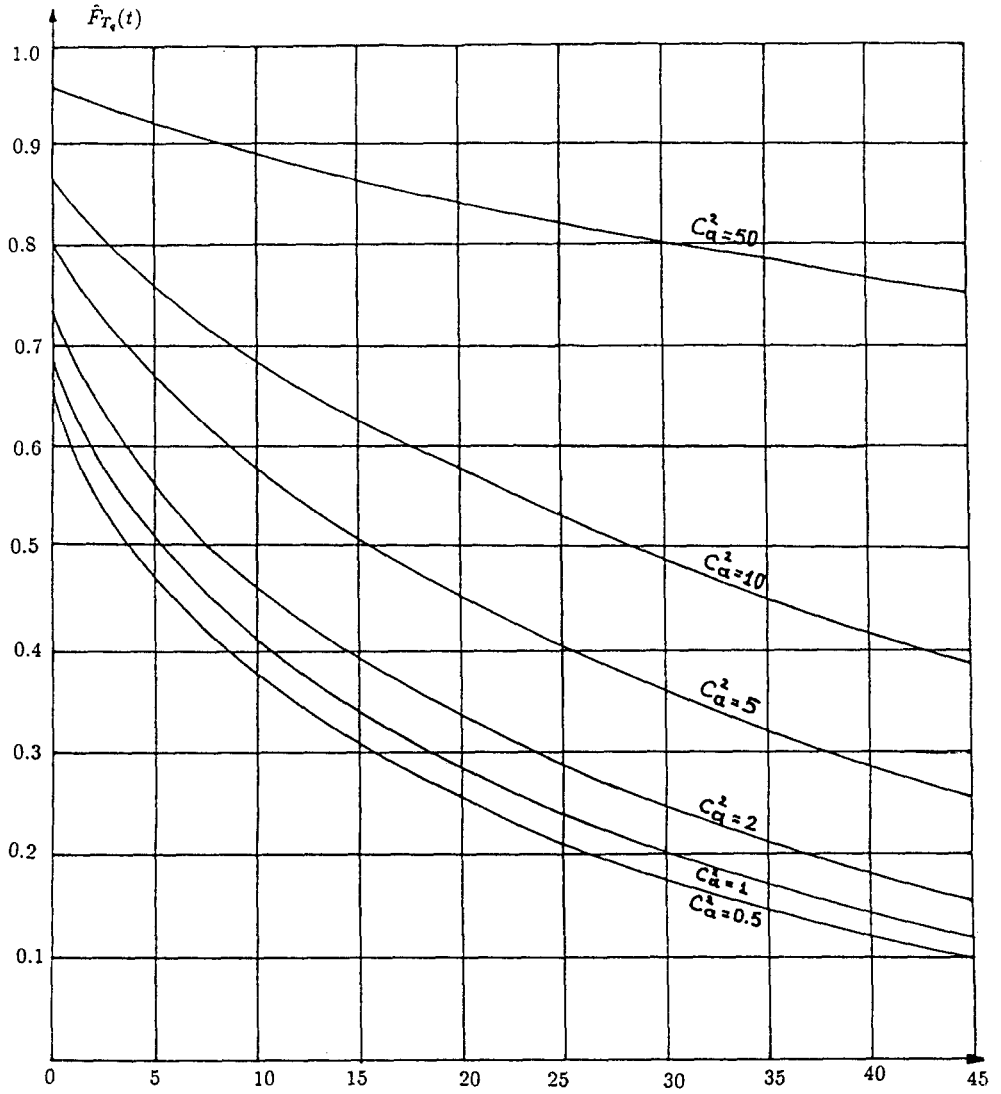

Fig. 2. $\hat{F}_{T_q}(t)$ as a function of C_a^2 for the $MGE_2/MGE_2/10$ system ($\rho = 0.9$, $C_s^2 = 5.0$).

the G/PH/s queue. Their formula for the complementary distribution function of the stationary waiting time is of the form

$$\sum_{j=0}^{\infty} d_j \, \mathrm{e}^{-\theta x} \frac{(\theta x)^j}{j!}.$$

Since this expression involves an infinite sum, truncation should be used in order to extract numerical results. The method has good stability properties, because

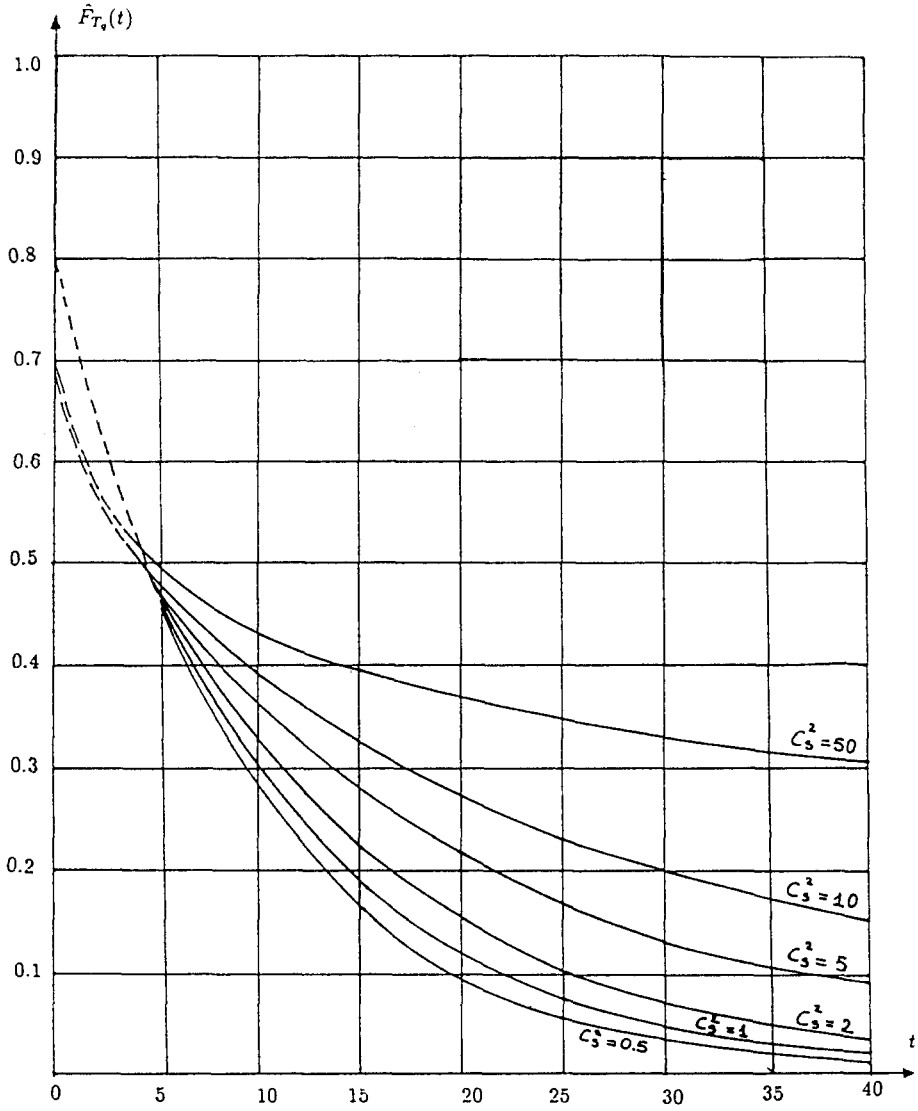

Fig. 3. $\hat{F}_{T_q}(t)$ as a function of C_s^2 for the $\mathrm{MGE}_2/\mathrm{MGE}_2/5$ system ($\rho = 0.8$, $C_a^2 = 5.0$).

the summation involves only positive terms and the error introduced from the truncation can be computed and bounded.

In comparison with this approach, the present approach offers qualitative insight by providing closed form expressions for the stationary waiting time distribution as a finite sum of $\binom{s+m-1}{s}$ exponentials (equation (20)), the exponents of which are computed from equations that involve only the Laplace transforms of the interarrival and service time distributions (equations (3), (4)). In the matrix-geometric method one has to compute a rate matrix R as a solution to the nonlinear matrix equation $R = \sum_{n=0}^{\infty} R^n A_n$, where the dimensions of the matrices involved are $\binom{s+m-1}{s}$; in the present approach one has to compute $\binom{s+m-1}{s}$ roots $x(w_j)$, which satisfy the separable equations (3), (4). Thus one can compute the $x(w_j)$ independently of each other and so the implementation of this part of the algorithm (determination of $x(w_j)$) is less complex than the determination of the matrix R and can be done in parallel.

As an illustration of the stability and accuracy of the present algorithm, we present in fig. 2 some results for the waiting time complementary distribution $\hat{F}_{T_q}(t) \triangleq 1 - F_{T_q}(t)$ for the $\mathrm{MGE}_2/\mathrm{MGE}_2/s$ system as C_a^2 varies ($s = 10$, $\rho = 0.9$, $C_s^2 = 5.0$). In fig. 3 we show the dependence of $\hat{F}_{T_q}(t)$ for the $\mathrm{MGE}_2/\mathrm{MGE}_2/s$ system on C_s^2 ($s = 5$, $\rho = 0.8$, $C_a^2 = 5.0$).

Acknowledgment

I would like to thank the referee for several constructive comments, which improved the paper significantly.

References

[1] D. Avis, Computing waiting times in $\mathrm{GI}/\mathrm{E}_k/s$ queueing system, TIMS Studies in Management Science 7 (1977) 215–232.

[2] D. Bertsimas, An analytic approach to a general class of queueing systems, Working paper, Operations Research Center, MIT, OR 156-87, 1987 (submitted to Operations Research).

[3] D.R. Cox, A use of complex probabilities in the theory of stochastic processes, Proc. Camb. Phil. Soc. 51 (1955) 313–319.

[4] A. Ishikawa, Stationary waiting time distribution in a $\mathrm{GI}/\mathrm{E}_k/m$ queue, Oper. Res. Soc. Japan 27 (1984) 130–149.

[5] F. Pollaczek, Concerning an analytic method for the treatment of queueing problems, 1-42 in: *Congested Theory*, eds. W.L. Smith and R.I. Wilkinson (Univ. of North Carolina Press, 1964).

[6] V. Ramaswami and D.M. Lucantoni, Stationary waiting time distribution in queues with phase type service and in quasi-birth-and-death-processes, Stochastic Models 1 (1985) 125–136.

[7] R. Schassberger, On the waiting time in the queueing system GI/G/1, Ann. Math. Statist. 41 (1970) 182–187.

[8] J.H.A. de Smit, The queue GI/M/s with customers of different types or the queue GI/H_m/s. Adv. Appl. Prob. 15 (1983) 392–419.
[9] Y. Takahashi, Asymptotic exponentially of the tail of the waiting time distribution in a PH/PH/c queue, Adv. Appl. Prob. 13 (1981) 619–630.
[10] H. Tijms, *Stochastic Modelling and Analysis; a Computational Approach* (John Wiley, New York, 1986).