Available online at www.sciencedirect.com
SCIENCE DIRECT®
Performance Evaluation 63 (2006) 131–145
PERFORMANCE EVALUATION
An International Journal
www.elsevier.com/locate/peva

Solving the ME/ME/1 queue with state–space methods and the matrix sign function

Nail Akar*

Electrical and Electronics Engineering Department, Bilkent University, Bilkent, 06800 Ankara, Turkey

Received 18 September 2003; received in revised form 12 December 2004
Available online 17 March 2005

Abstract

Matrix exponential (ME) distributions not only include the well-known class of phase-type distributions but also can be used to approximate more general distributions (e.g., deterministic, heavy-tailed, etc.). In this paper, a novel mathematical framework and a numerical algorithm are proposed to calculate the matrix exponential representation for the steady-state waiting time in an ME/ME/1 queue. Using state–space algebra, the waiting time calculation problem is shown to reduce to finding the solution of an ordinary differential equation in state–space form with order being the *sum* of the dimensionalities of the inter-arrival and service time distribution representations. A numerically efficient algorithm with quadratic convergence rates based on the matrix sign function iterations is proposed to find the boundary conditions of the differential equation. The overall algorithm does not involve any transform domain calculations such as root finding or polynomial factorization, which are known to have potential numerical stability problems. Numerical examples are provided to demonstrate the effectiveness of the proposed approach.

Keywords: GI/GI/1 queue; Lindley's equation; Matrix exponential distribution; Realization theory; Matrix sign function

1. Introduction

The successive waiting times in a single server queue with a first come first serve (FCFS) service discipline is depicted in Fig. 1. Here, B_n denotes the service time of customer n and A_n denotes the time between the arrival of customers n and $n+1$. The service of the nth customer begins W_n seconds after

* Tel.: +90 312 2902337; fax: +90 312 2664192.
E-mail address: akar@ee.bilkent.edu.tr.

0166-5316/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.peva.2004.12.002

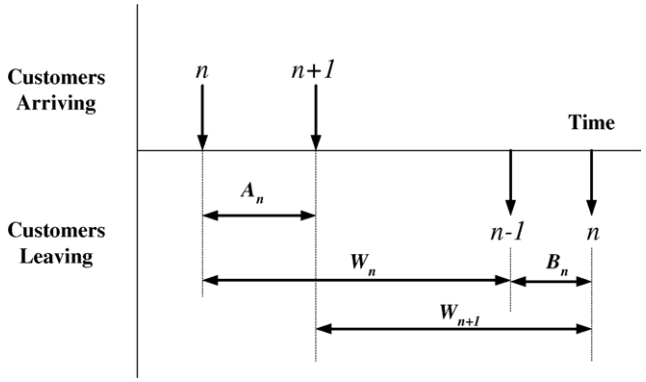

Fig. 1. Successive queue waiting times in a single server queue.

its arrival where W_n denotes the nth customer's waiting time in the queue. We observe from Fig. 1 that the queue waiting times W_{n+1} and W_n of two successive customers in a single server queue are related through the so-called Lindley recurrence relation [21]

$$W_{n+1} = \max(0, W_n + B_n - A_n), \quad n \geq 0. \tag{1}$$

The GI/GI/1 notation (GI stands for general independent) is used to represent this queueing system with one server and infinite waiting room where the successive inter-arrival times (service times) A_n (B_n) are independent and identically distributed. We note that the more general case of non-renewal inter-arrival and/or service times is left outside the scope of the current paper. We also assume in this paper that the inter-arrival and service times possess a matrix exponential (ME) distribution [5,24]. An ME distribution G on $[0, \infty)$ has a density $g(x)$ in the interval $[0, \infty)$ of the form

$$g(x) = \mathbf{v}\mathrm{e}^{\mathbf{T}x}\mathbf{s} + d\delta(x), \tag{2}$$

where $\mathbf{v}$ is a row vector, $\mathbf{T}$ is a square matrix of size m, $\mathbf{s}$ is a column vector, $\delta(x)$ is the unit impulse function, and $d = 1 + \mathbf{v}\mathbf{T}^{-1}\mathbf{s}$ is a scalar indicating the probability mass at zero [20,24]. The density then has the unilateral Laplace transform

$$g^*(s) = L_-[g(x)] = \int_{0^-}^{\infty} \mathrm{e}^{-sx} g(x)\,\mathrm{d}x = \mathbf{v}(s\mathbf{I} - \mathbf{T})^{-1}\mathbf{s} + d, \tag{3}$$

where the unilateral Laplace transform $L_-[\cdot]$ has a lower integration limit that is set to 0^- as in [20]. In short, we use the triple $(\mathbf{v}, \mathbf{T}, \mathbf{s})$ to represent this matrix exponential distribution with dimensionality m. The representation $(\mathbf{v}, \mathbf{T}, \mathbf{s})$ is irreducible if and only if

$$\det(s\mathbf{I} - \mathbf{T}) = k[\text{denominator of } g^*(s)],$$

where k is a constant. By using [24], the ith moment of a matrix exponential distribution for a random variable X with representation $(\mathbf{v}, \mathbf{T}, \mathbf{s})$ can be written as

$$E[X^i] = (-1)^{i+1} i! \mathbf{v}\mathbf{T}^{-(i+1)}\mathbf{s}.$$

The class of ME distributions are also characterized by rational Laplace transforms of their densities. The well-known class PH of phase-type distributions form a sub-case of ME-type distributions; they

have the representation as in (2) with $\mathbf{s} = -\mathbf{Te}$ for a column vector $\mathbf{e}$ of ones and a substochastic matrix $\mathbf{T}$ [26]. Moreover, PH-type distributions have a probabilistic interpretation as the distribution of time till absorption in a continuous-time Markov chain with m transient and one absorbing state for which the matrix $\mathbf{T}$ governs the transitions among the transient states [26].

Matrix exponential distributions are pioneered by Lipsky [24] and queues with matrix exponential inputs have recently been addressed in a number of studies [5,11,25]. The focus of this paper is the ME/ME/1 queue which is an important sub-case of the GI/GI/1 queue and we will study the steady-state distribution of the waiting time $W \triangleq \lim_{n\to\infty} W_n$ in the ME/ME/1 queue, when it exists. There have been successful studies on the approximation of more general (including heavy-tailed) inter-arrival and service times by PH- or ME-type distributions [7,17,30]. Therefore, the approach proposed in this paper can also be viewed as an approximate methodology to find the steady-state waiting times for the more general GI/GI/1 queue.

For classical techniques that rely on root finding and Wiener–Hopf factorizations, we refer the reader to several textbooks on queueing theory [3,18,21]. However, transform domain techniques and particularly root finding may become ill conditioned for large-sized problems [27]. With the introduction of the matrix analytical approach, matrix calculations have taken the role of root finding and such calculations are amenable to algorithms with enhanced numerical features compared to the classical transform domain techniques [29]. Using the matrix analytical approach, Sengupta [29] and Asmussen [4] show that the waiting time has a matrix exponential representation and they present algorithms for the GI/PH/1 queue using a nonlinear matrix equation which can be solved using an iterative technique. One drawback of these algorithms are their linear convergence rates. Recently, Asmussen and Moller studied the more general GI/PH/c and MAP/PH/c queues with arbitrary number of servers again making use of the matrix analytical paradigm [6].

Another powerful approach is also a matrix analytical approach, the so-called matrix geometric paradigm pioneered by Neuts, which involves computations for discrete-state structured Markov chains [26,27]. An important sub-case of the GI/GI/1 system, the so-called PH/PH/1 queue, is studied in depth by Neuts using the quasi-birth-and-death (QBD) process framework and the matrix geometric approach [26]. In [26], an iterative algorithm is used with matrices of size being the *product* of the dimensionalities of the inter-arrival and service time distribution representations, to first find the steady-state queue length probabilities in matrix geometric form and then the waiting times. Latouche and Ramaswami extend this analysis by considering the queue length process at the embedded epochs of queue size change [23]. They show that the embedded QBD process can be solved by matrix geometric techniques and the quadratically convergent logarithmic reduction (LR) procedure [22] operating on matrices that have a size of the *sum* of the dimensionalities of the inter-arrival and service time distribution representations, as opposed to their product. This size reduction within the LR iterations is a significant advantage of their proposed algorithm. However, obtaining the matrices required for the LR iterations still require the construction of a matrix with the order of the product of the number of phases in the arrival and service time distributions and further matrix multiplications involving this product-sized matrix [23]. This size dependence on the "product" appears to be a limiting factor for the scalability of the proposed algorithm in [23].

In this paper, we propose state–space methods for reducing the problem of finding the steady-state waiting time distributions in ME/ME/1 queues to the solution of an ordinary differential equation (ODE) with constant coefficients but with unknown initial conditions to be determined. We note that state–space methods are successfully being used as powerful computational tools particularly in the area of control systems and other areas including filtering and signal processing [15,16]. The strength of state–space

methods stems from the availability of a wide variety of advanced linear algebra tools that can be employed in conjunction with the state–space methodology. One such powerful linear algebra tool is the matrix sign function that is used for solving the algebraic Riccati equation that arises in the solution of a number of optimal control and filtering problems [12,28]. The problem of finding the initial conditions of the ODE is shown to be equivalent to an ordinary spectral divide-and-conquer (SDC) problem (see [9]) applied on a certain "coupling matrix" whose size equals the sum of the dimensionalities of the inter-arrival time and service time distribution representations. In our proposed approach, the coupling matrix is very easy to obtain and we do not need the Laplace transforms of the inter-arrival time and service time densities. Instead, we use their natural matrix exponential representations, which makes the overall algorithm numerically stable. We note that the term "coupling matrix" was coined by van de Liefvoort [31] in which the waiting time distribution for the PH/PH/1 queue is obtained by finding the eigenvalues of a certain matrix which is similar to the coupling matrix investigated in this paper. We propose to use the computationally efficient matrix sign function and the corresponding Newton iterations for solving the underlying spectral divide-and-conquer problem [12]. We note that this proposed method does not require the computation of individual eigenvalues and eigenvectors. The reason for choosing the matrix sign method stems from its ease of implementation although we note the existence of other SDC techniques available in the numerical linear algebra literature [8]. We compare our results with the algorithm proposed by Latouche and Ramaswami [23] for the case of the PH/PH/1 queue and also with that of Asmussen and Bladt [5] for the GI/ME/1 queue, in terms of the number of required iterations and accuracy.

The organization of the paper is as follows. In Section 2, the spectral divide-and-conquer problem is described and the matrix sign function is introduced. In Section 3, we present the solution of the ME/ME/1 queue. Section 4 is devoted to numerical examples. We conclude in Section 5.

2. Spectral divide-and-conquer problem and the matrix sign function

We use the following notation in this paper. We denote vectors or matrices by boldface letters to differentiate them from scalars. $\mathbf{I}$ denotes an identity matrix of suitable size. We use the notation * to denote Laplace transforms. All differential equations and functions of x, the indeterminate variable, to be used in this paper are defined in the interval $[0, \infty)$.

For a given $n \times n$ nonsymmetric real matrix $\mathbf{M}$, we are interested in finding an invariant subspace $\mathcal{R}$ (i.e., $\mathbf{M}\mathcal{R} \subseteq \mathcal{R}$) corresponding to the eigenvalues of $\mathbf{M}$ in an a priori specified region $\mathcal{D}$ of the complex plane. Equivalently, we are interested in obtaining an orthogonal matrix $\mathbf{Q} = (\mathbf{Q}_1, \mathbf{Q}_2)$ (i.e., $\mathbf{Q}^{\mathbf{T}}\mathbf{Q} = \mathbf{I}$) with $\mathcal{R} = \text{span}\{\mathbf{Q}_1\}$ such that

$$\mathbf{Q}^{\mathbf{T}}\mathbf{M}\mathbf{Q} = \begin{bmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ 0 & \mathbf{M}_{22} \end{bmatrix}, \tag{4}$$

where the eigenvalues of $\mathbf{M}_{11}$ are the eigenvalues of $\mathbf{M}$ in $\mathcal{D}$. This problem is called the spectral divide-and-conquer problem and is one of the most extensively studied problems of numerical linear algebra [9]. The Schur decomposition-based algorithm [10], inverse-free methods based on the QR decomposition [9], and the matrix sign function-based methods [12,28] are among the most popular methods of numerical linear algebra which address the SDC problem. A comparative study of the existing methods for the SDC

problem in the context of the GI/GI/1 queue is outside the scope of the current paper and we focus only on the matrix sign function-based algorithm due to the ease of its implementation.

Let $\mathcal{C}^-$, $\mathcal{C}^+$, and $\mathcal{C}^0$ denote the open left half plane, open right half plane, and the imaginary axis, respectively. In many applications, the region of interest $\mathcal{D}$ is chosen to be $\mathbf{C}^-$, or equivalently the eigenvalues of the submatrix $\mathbf{M}_{11}$ are the same as those of the matrix $\mathbf{M}$ with negative real parts. The following is based on [28]. The sign function sign($\mathbf{M}$) of a matrix $\mathbf{M}$ with no eigenvalues on the imaginary axis can be defined via the Jordan canonical form of $\mathbf{M}$. Let

$$\mathbf{M} = \mathbf{X}\begin{bmatrix}\mathbf{J}_- & \mathbf{0}\\ \mathbf{0} & \mathbf{J}_+\end{bmatrix}\mathbf{X}^{-1}$$

be the Jordan canonical form of $\mathbf{M}$, where the eigenvalues of $\mathbf{J}_-$ are in the open left half plane, and the eigenvalues of $\mathbf{J}_+$ are in the open right half plane. Then, sign($\mathbf{M}$) is defined as

$$\text{sign}(\mathbf{M}) \triangleq \mathbf{X}\begin{bmatrix}-\mathbf{I} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\end{bmatrix}\mathbf{X}^{-1}. \tag{5}$$

The simplest scheme to obtain the sign function of $\mathbf{M}$ is the following Newton iteration

$$\mathbf{M}_{j+1} = \frac{1}{2}(\mathbf{M}_j + \mathbf{M}_j^{-1}), \quad j \geq 0, \ \ \mathbf{M}_0 = \mathbf{M}. \tag{6}$$

The above iteration is globally and ultimately quadratically convergent with $\mathbf{M}_\infty \triangleq \lim_{j\to\infty}\mathbf{M}_j = \text{sign}(\mathbf{M})$. Scaling can also be introduced into the iterations (6) as in [2] to speed up the iterations but we will not consider scaling in this paper. Moreover, if an orthogonal matrix $\mathbf{Q}$ is chosen such that its leading columns span the range space of $(\mathbf{I} - \mathbf{M}_\infty)$, then $\mathbf{Q}$ yields the desired decomposition given in (4). This last step can be carried out by a rank revealing QR decomposition [13]. A rank revealing QR decomposition of a matrix $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{Q}\mathbf{R}\mathbf{\Pi},$$

where $\mathbf{R}$ is upper triangular, $\mathbf{Q}$ is orthogonal, and $\mathbf{\Pi}$ is a permutation matrix such that the rank deficiency of $\mathbf{A}$ is exhibited in $\mathbf{R}$ for which the diagonal entries decrease in absolute value with increasing index. If $r = \text{rank}(\mathbf{A})$ is a priori known exactly (as will be the case for the current paper), then the first r columns of $\mathbf{Q}$ are known to span the range space of $\mathbf{A}$.

For the solution of certain structured Markov chains, Akar and Sohraby proposed to use the Newton iterations for a matrix, say $\mathbf{M}$, which turns out to have one single eigenvalue on the imaginary axis (i.e., at the origin) [2]. Let $\mathbf{x}_\text{r}$ and $\mathbf{x}_\text{l}$ be right and left eigenvectors of the matrix $\mathbf{M}$, respectively, corresponding to the eigenvalue at the origin. The following Newton iteration is then suggested in [2]

$$\mathbf{M}_{j+1} = \frac{1}{2}(\mathbf{M}_j + \mathbf{M}_j^{-1}), \quad j \geq 0, \ \ \mathbf{M}_0 = \mathbf{M} + \frac{\mathbf{x}_\text{r}\mathbf{x}_\text{l}}{\mathbf{x}_\text{l}\mathbf{x}_\text{r}}, \tag{7}$$

which converges to $\mathbf{M}_\infty$. A rank revealing QR decomposition of $(\mathbf{I} - \mathbf{M}_\infty)$ yields an orthogonal matrix $\mathbf{Q}$ such that the decomposition (4) holds with the eigenvalues of $\mathbf{M}_{11}$ and $\mathbf{M}_{22}$ being in $\mathcal{C}^-$ and $\mathcal{C}^0 \cup \mathcal{C}^+$, respectively.

3. Waiting times in the ME/ME/1 queue

Consider the ME/ME/1 queue with a matrix exponential inter-arrival distribution A and a matrix exponential service time distribution B with irreducible representations $(\mathbf{v}_a, \mathbf{T}_a, \mathbf{s}_a)$ and $(\mathbf{v}_b, \mathbf{T}_b, \mathbf{s}_b)$, respectively. The matrices $\mathbf{T}_a$ and $\mathbf{T}_b$ are of sizes $m_a \times m_a$ and $m_b \times m_b$, respectively. For the sake of generality, we allow arbitrary probability masses (at zero) $d_a = 1 + \mathbf{v}_a \mathbf{T}_a^{-1} \mathbf{s}_a$ and $d_b = 1 + \mathbf{v}_b \mathbf{T}_b^{-1} \mathbf{s}_b$, $0 \leq d_a, d_b < 1$, for the inter-arrival and service time distributions, respectively. The associated Laplace transforms are rational functions and are expressed as

$$a^*(s) = \frac{p_a(s)}{q_a(s)} = \mathbf{v}_a(s\mathbf{I} - \mathbf{T}_a)^{-1}\mathbf{s}_a + d_a, \tag{8}$$

and

$$b^*(s) = \frac{p_b(s)}{q_b(s)} = \mathbf{v}_b(s\mathbf{I} - \mathbf{T}_b)^{-1}\mathbf{s}_b + d_b. \tag{9}$$

Recall that B_n denotes the service time of customer n, A_n denotes the time between the arrival of customers n and $n+1$, and the waiting time of the nth customer denoted by W_n satisfies the Lindley recurrence relation given in (1). We assume throughout this paper that the load ρ defined by the mean service time divided by the mean inter-arrival time, is strictly less than unity.

Therefore, $W_n \to W$ as $n \to \infty$ in distribution, where W is called the steady-state waiting time distribution and $w(x)$ denotes its density [3]. The Laplace transform of the waiting time density is denoted by $w^*(s)$. We are now ready to give the following theorem that provides an expression for $w^*(s)$. For the proof, we refer the reader to Section 8.2 of [21].

Theorem 1. *Assume that the steady-state waiting time distribution exists. Then, there exists a polynomial* $\phi(s) = \phi_0 + \phi_1 s + \cdots + \phi_{m_a} s^{m_a}$ *such that the following holds*

$$w^*(s) = [1 - a^*(-s)b^*(s)]^{-1} \frac{\phi(s)}{q_a(-s)}. \tag{10}$$

Conversely, the choice of $\phi(s)$ *satisfying* (10) *with* $w^*(s)$ *being analytic in the closed right half of the complex plane and* $w^*(0) = 1$ *gives the steady-state waiting time distribution.*

This theorem can directly be used as a method for finding the waiting times by factorizing the numerator of $[1 - a^*(-s)b^*(s)]$ into two coprime polynomials, one having all its roots in the right half plane, and then choosing $\phi(s)$ appropriately. However, this immediate approach has several disadvantages from a numerical analysis standpoint. Firstly, one has to evaluate the Laplace transform of the inter-arrival and service times. Secondly, root finding in transform domain may become ill conditioned for large-sized problems especially when the roots are close to each other or some roots have multiplicities. Finally, we need to use sophisticated inverse Laplace transform techniques to describe the steady-state waiting times in the time domain. In this paper, we propose an algorithm that avoids ill-conditioned transform domain calculations. Theorem 1 is used as an intermediate tool for proving our results but is not explicitly used in the proposed numerical algorithm which is based on state–space methods and the use of the matrix sign function.

A linear time-invariant dynamical system with r inputs and t outputs is represented by the following set of ODEs [20]

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{y}(x) = \mathbf{T}\mathbf{y}(x) + \mathbf{s}\mathbf{u}(x), \tag{11}$$

$$\mathbf{w}(x) = \mathbf{v}\mathbf{y}(x) + \mathbf{d}\mathbf{u}(x), \tag{12}$$

where $\mathbf{u}(x) = (u_1(x), \ldots, u_r(x))^{\mathrm{T}}$ and $\mathbf{w}(x) = (w_1(x), \ldots, w_t(x))^{\mathrm{T}}$ denote the input and output vectors, respectively, $\mathbf{y}(x) = (y_1(x), \ldots, y_m(x))^{\mathrm{T}}$ is called the state vector and its components are called the state variables, or simply the states. The matrices $\mathbf{T}$, $\mathbf{v}$, $\mathbf{s}$ and $\mathbf{d}$ in the Eqs. (11) and (12) are real matrices of suitable sizes. Considering zero initial state, the transfer matrix $\mathbf{H}^*(s)$ between the input and output vectors is written as [20]

$$\mathbf{w}^*(s) = \mathbf{H}^*(s)\mathbf{u}^*(s) = (\mathbf{v}(sI - \mathbf{T})^{-1}\mathbf{s} + \mathbf{d})\mathbf{u}^*(s), \tag{13}$$

where $\mathbf{u}^*(s)$ and $\mathbf{w}^*(s)$ are the Laplace transforms of the input and output vectors, respectively. The equations of the form (11) and (12) are said to constitute a state–space description or realization of the given linear time-invariant system with transfer matrix $\mathbf{H}^*(s)$ if (13) holds [20]. The number of states (i.e., m) is referred to as the order or the dimensionality of the state–space representation. Using similarity transformations, one can obtain infinitely many realizations whereas realization theory deals with finding state–space descriptions of linear systems and the properties of these descriptions [14,20]. Methods that use state–space representations (as opposed to transform domain calculations) are called state–space methods and they are frequently used in the fields of control and signal processing [15,16,20].

Our goal now is to find a matrix exponential representation for $w(x)$ in Eq. (10) using the individual representations for the inter-arrival and service times. For this purpose, consider the following linear system (denoted by $\mathcal{S}_\mathrm{a}$) associated with the inter-arrival times in state–space form but with nonzero initial states

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{y}_\mathrm{a}(x) = -\mathbf{T}_\mathrm{a}\mathbf{y}_\mathrm{a}(x) + \mathbf{s}_\mathrm{a}u_\mathrm{a}(x), \quad \mathbf{y}_\mathrm{a}(0^-) = \mathbf{y}_0, \\ &w_\mathrm{a}(x) = -\mathbf{v}_\mathrm{a}\mathbf{y}_\mathrm{a}(x) + d_\mathrm{a}u_\mathrm{a}(x) + d_0\delta(x). \end{aligned} \tag{14}$$

The system $\mathcal{S}_\mathrm{a}$ has two inputs, one being the control input $u_\mathrm{a}(x)$, the other being the unit impulse function $\delta(x)$ feeding in through a amplifier (multiplier) d_0, one output $w_\mathrm{a}(x)$, and a nonzero initial state $\mathbf{y}_0$. The system parameters d_0 and $\mathbf{y}_0$ are not known yet but they are to be determined. Now consider the following linear system (denoted by $\mathcal{S}_\mathrm{b}$) associated with the service times in state–space form

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{y}_\mathrm{b}(x) = \mathbf{T}_\mathrm{b}\mathbf{y}_\mathrm{b}(x) + \mathbf{s}_\mathrm{b}u_\mathrm{b}(x), \quad \mathbf{y}_\mathrm{b}(0^-) = 0, \\ &w_\mathrm{b}(x) = \mathbf{v}_\mathrm{b}\mathbf{y}_\mathrm{b}(x) + d_\mathrm{b}u_\mathrm{b}(x). \end{aligned} \tag{15}$$

The system $\mathcal{S}_\mathrm{b}$ has one control input $u_\mathrm{b}(x)$, one output $w_\mathrm{b}(x)$, and zero initial state. We propose to interconnect the two systems via the following feedback configuration also given in Fig. 2

$$u_\mathrm{a}(x) = w_\mathrm{b}(x) =: u(x), \qquad u_\mathrm{b}(x) = w_\mathrm{a}(x) =: w(x) \tag{16}$$

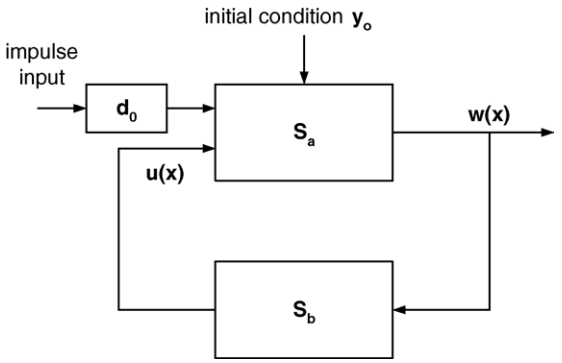

Fig. 2. Feedback interconnection diagram of the two systems $\mathcal{S}_a$ and $\mathcal{S}_b$.

As the next step, we show that the Laplace transform of $w(x)$ defined in (16), namely $w^*(s)$, satisfies the identity (10). In order to prove this, we first show the following by using (14)

$$w^*(s) = \begin{cases} -\mathbf{v}_a \mathbf{y}_a^*(s) + d_a u^*(s) + d_0, \\ -\mathbf{v}_a (sI + \mathbf{T}_a)^{-1}(\mathbf{y}_0 + \mathbf{s}_a u^*(s)) + d_a u^*(s) + d_0, \\ [-\mathbf{v}_a (sI + \mathbf{T}_a)^{-1}\mathbf{y}_0 + d_0] + \underbrace{[-\mathbf{v}_a (sI + \mathbf{T}_a)^{-1}\mathbf{s}_a + d_a]}_{a^*(-s)} u^*(s). \end{cases} \tag{17}$$

Since $u^*(s) = (\mathbf{v}_b(s\mathbf{I} - \mathbf{T}_b)^{-1}\mathbf{s}_b + d_b)w^*(s) = b^*(s)w^*(s)$ from (15), Eq. (17) implies

$$w^*(s) = [1 - a^*(-s)b^*(s)]^{-1}[-\mathbf{v}_a (sI + \mathbf{T}_a)^{-1}\mathbf{y}_0 + d_0], \tag{18}$$

$$w^*(s) = [1 - a^*(-s)b^*(s)]^{-1} \frac{\phi(s)}{q_a(-s)}, \tag{19}$$

where $\phi(s)$ is the numerator polynomial of $[-\mathbf{v}_a(sI + \mathbf{T}_a)^{-1}\mathbf{y}_0 + d_0]$ and the associated denominator polynomial is equal to $q_a(-s)$ up to a constant due to the irreducibility of the inter-arrival time distribution representation. In (19), we obtain the same expression as in Theorem 1 for $w^*(s)$. Therefore, we conclude that the choice of $\mathbf{y}_0$ and the scalar d_0 in the set of ODEs (14) leading to $w^*(s)$ being analytic in the closed right half of the complex plane and $w^*(0) = 1$ gives the steady-state waiting time distribution.

To find the unknowns $\mathbf{y}_0$ and d_0, by using standard matrix arithmetic, we combine the linear dynamical Eqs. (14) and (15) into one linear dynamical equation associated with an autonomous system (i.e., no exogenous inputs) with $m = m_a + m_b$ state variables

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{y}(x) = \begin{bmatrix} \frac{\mathrm{d}}{\mathrm{d}x}\mathbf{y}_a(x) \\ \frac{\mathrm{d}}{\mathrm{d}x}\mathbf{y}_b(x) \end{bmatrix} = \mathbf{C} \begin{bmatrix} \mathbf{y}_a(x) \\ \mathbf{y}_b(x) \end{bmatrix} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{y}_a(x) \\ \mathbf{y}_b(x) \end{bmatrix}, \tag{20}$$

$$\mathbf{y}(0^+) = \begin{bmatrix} \mathbf{y}_a(0^+) \\ \mathbf{y}_b(0^+) \end{bmatrix} = \begin{bmatrix} \mathbf{y}_0 + \mathbf{n}_1 d_0 \\ \mathbf{n}_2 d_0 \end{bmatrix}, \tag{21}$$

$$w(x) = \mathbf{c}\mathbf{y}(x) + n d_0 \delta(x), \tag{22}$$

where

$$n \triangleq (1 - d_a d_b)^{-1}, \qquad \mathbf{n}_1 \triangleq n d_b \mathbf{s}_a, \qquad \mathbf{n}_2 \triangleq n \mathbf{s}_b, \tag{23}$$

$$\mathbf{C}_{11} \triangleq -\mathbf{T}_a - n d_b \mathbf{s}_a \mathbf{v}_a, \qquad \mathbf{C}_{12} \triangleq n \mathbf{s}_a \mathbf{v}_b, \tag{24}$$

$$\mathbf{C}_{21} \triangleq -n \mathbf{s}_b \mathbf{v}_a, \qquad \mathbf{C}_{22} \triangleq \mathbf{T}_b + n d_a \mathbf{s}_b \mathbf{v}_b, \tag{25}$$

and

$$\mathbf{c} \triangleq [\, -n\mathbf{v}_a \quad n d_a \mathbf{v}_b \,]. \tag{26}$$

The ordinary differential Eq. (20) and the output Eq. (22) completely describe the waiting time $w(x)$ if the initial state given in (21) is known. What then remains is the calculation of the unknowns $\mathbf{y_0}$ and d_0 such that the conditions for Theorem 1 are satisfied.

We note from realization theory [20] that the eigenvalues of the so-called coupling matrix $\mathbf{C}$ are exactly the same as the zeros of the rational function $[1 - a^*(-s)b^*(s)]$ and therefore the matrix $\mathbf{C}$, in case of a stable queue (i.e., $\rho < 1$), will have one eigenvalue at the origin, m_b eigenvalues with negative real parts and $m_a - 1$ eigenvalues with positive real parts [21]. Using the identities $a^*(0) = 1$ and $b^*(0) = 1$, one can show through straightforward algebraic manipulations that the vectors defined by

$$\mathbf{x}_l \triangleq [\, \mathbf{v}_a \mathbf{T}_a^{-1} \quad \mathbf{v}_b \mathbf{T}_b^{-1} \,] \tag{27}$$

and

$$\mathbf{x}_r \triangleq \begin{bmatrix} -\mathbf{T}_a^{-1}\mathbf{s}_a \\ \mathbf{T}_b^{-1}\mathbf{s}_b \end{bmatrix} \tag{28}$$

are the left and right eigenvectors, respectively, of the matrix $\mathbf{C}$ associated with the single eigenvalue at the origin. Solving the SDC problem, one can find an orthogonal matrix $\mathbf{U}$ such that

$$\mathbf{U}^T \mathbf{C} \mathbf{U} = \mathbf{R} = \begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0} & \mathbf{R}_{22} \end{bmatrix}, \tag{29}$$

where the $m_b \times m_b$ matrix $\mathbf{R}_{11}$ has all its eigenvalues with negative real parts and the $m_a \times m_a$ matrix $\mathbf{R}_{22}$ has all its eigenvalues with nonnegative real parts (including the eigenvalue at the origin). The transformation matrix $\mathbf{U}$ can be computed by a number of methods given in Section 2. We propose the following Newton iteration outlined in Section 2

$$\mathbf{M}_{k+1} = \frac{1}{2}(\mathbf{M}_k + \mathbf{M}_k^{-1}), \qquad \mathbf{M}_0 = \mathbf{C} + \frac{\mathbf{x}_r \mathbf{x}_l}{\mathbf{x}_l \mathbf{x}_r}, \tag{30}$$

which converges to $\mathbf{M}_\infty$ as $k \to \infty$. Recall that a rank revealing QR factorization of $(\mathbf{I} - \mathbf{M}_\infty)$ yields the desired decomposition (29).

Using the decomposition (29), we introduce a new state vector $\mathbf{z}(x) \triangleq \mathbf{U}^T \mathbf{y}(x)$ so that we rewrite the dynamical Eqs. (20)–(22) in terms of $\mathbf{z}(x)$

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbf{z}(x) = \mathbf{R}\mathbf{z}(x), \tag{31}$$

$$w(x) = \mathbf{c}\mathbf{U}\mathbf{z}(x) + n d_0 \delta(x). \tag{32}$$

We then introduce the following partitions on $\mathbf{z}(x)$, $\mathbf{cU}$, and $\mathbf{U}$

$$\mathbf{z}(x) \triangleq \begin{bmatrix} \mathbf{z}_s(x) \\ \mathbf{z}_u(x) \end{bmatrix}, \qquad \mathbf{cU} \triangleq [\, \mathbf{c}_s \quad \mathbf{c}_u \,], \qquad \mathbf{U} \triangleq \begin{bmatrix} \mathbf{U}_{11} & \mathbf{U}_{12} \\ \mathbf{U}_{21} & \mathbf{U}_{22} \end{bmatrix}, \tag{33}$$

where the sizes of $\mathbf{z}_s(x)$, $\mathbf{z}_u(x)$, $\mathbf{c}_s$, $\mathbf{c}_u$, $\mathbf{U}_{11}$, $\mathbf{U}_{12}$, $\mathbf{U}_{21}$, and $\mathbf{U}_{22}$ are $m_b \times 1$, $m_a \times 1$, $1 \times m_b$, $1 \times m_a$, $m_a \times m_b$, $m_a \times m_a$, $m_b \times m_b$, and $m_b \times m_a$, respectively.

We are now ready to state the conditions on $\mathbf{y}_0$ and d_0 such that the conditions of Theorem 1 are satisfied. Firstly, in order for $w^*(s)$ to be analytic in the closed right half plane, either $\mathbf{c}_u$ is zero or $\mathbf{z}_u(0^+)$ should be the zero vector, since otherwise the eigenvalues of $\mathbf{R}_{22}$ would appear as the poles of $w^*(s)$. The former cannot be true since it would then lead to an infinite number of solutions for the single server queue which violates the uniqueness of the solution when it exists [21]. On the other hand, the latter condition is mathematically equivalent to

$$\mathbf{z}_u(0^+) = [\, \mathbf{U}_{12}^T \quad \mathbf{U}_{12}^T\mathbf{n}_1 + \mathbf{U}_{22}^T\mathbf{n}_2 \,] \begin{bmatrix} \mathbf{y}_0 \\ \mathbf{d}_0 \end{bmatrix} = 0. \tag{34}$$

When $\mathbf{z}_u(0^+) = 0$, we can rewrite $w(x)$, using (32) as follows

$$w(x) = \begin{cases} \mathbf{cUz}(x) + nd_0\delta(x), \\ \mathbf{c}_s e^{\mathbf{R}_{11}x}\mathbf{z}_s(0) + nd_0\delta(x), \\ \mathbf{c}_s e^{\mathbf{R}_{11}x}[\, \mathbf{U}_{11}^T \quad \mathbf{U}_{11}^T\mathbf{n}_1 + \mathbf{U}_{21}^T\mathbf{n}_2 \,] \begin{bmatrix} \mathbf{y}_0 \\ d_0 \end{bmatrix} + nd_0\delta(x). \end{cases} \tag{35}$$

Using (35) and by the requirement $w^*(0) = 1$, we obtain one other equation for the unknowns $\mathbf{y_0}$ and d_0:

$$[\, -\mathbf{c}_s\mathbf{R}_{11}^{-1}\mathbf{U}_{11}^T \quad -\mathbf{c}_s\mathbf{R}_{11}^{-1}(\mathbf{U}_{11}^T\mathbf{n}_1 + \mathbf{U}_{21}^T\mathbf{n}_2) + n \,] \begin{bmatrix} \mathbf{y}_0 \\ d_0 \end{bmatrix} = 1. \tag{36}$$

Eqs. (34) and (36) form $m_a + 1$ linear equations with $m_a + 1$ unknowns. This linear system should be nonsingular since otherwise one would violate the existence and uniqueness of a solution for the stable GI/GI/1 queue. Solving for the unknowns $\mathbf{y}_0$ and d_0 from these two equations, the steady-state waiting time has an ME distribution

$$w(x) = \mathbf{v}e^{\mathbf{T}x}\mathbf{s} + d\delta(x), \tag{37}$$

where

$$\mathbf{v} \triangleq \mathbf{c}_s, \qquad \mathbf{T} \triangleq \mathbf{R}_{11}, \qquad \mathbf{s} \triangleq \mathbf{U}_{11}^T\mathbf{y}_0 + (\mathbf{U}_{11}^T\mathbf{n}_1 + \mathbf{U}_{21}^T\mathbf{n}_2)d_0, \qquad d \triangleq nd_0.$$

The technique described in this section for the calculation of the steady-state waiting time distribution of the ME/ME/1 queue is based on the solution of an SDC problem. Since we opted to use the matrix sign function (MSF) iterations as the numerical engine in such computations, the whole algorithm will in short be referred to as the MSF algorithm.

4. Numerical examples

We study two numerical examples in this section. The first example is a PH/PH/1 queue for which the waiting times can efficiently be obtained by the particular matrix geometric technique proposed in [23] that uses the logarithmic reduction algorithm as its computational engine [22]. For this example, we compare the MSF algorithm proposed in this paper with that of [23] in terms of accuracy and convergence speeds. The quadratically convergent logarithmic reduction-based algorithm and its variants are known to result in the most efficient computational algorithms to date for PH/PH/1 queues, therefore such a comparison is crucial to determine the effectiveness of the proposed algorithm of this paper. The second example we study is a *D*/ME/1 queue from the existing literature [5]. For this example, the deterministic inter-arrival time distribution is not of ME-type but can be approximated by ME distributions. We note the possibility of extending the computationally efficient logarithmic reduction algorithm for solving the more general ME/ME/1 queue for comparison purposes but such a possible extension is left outside the scope of this paper. Instead, for the *D*/ME/1 example, we compare the MSF algorithm with that of [5] in terms of accuracy and convergence speeds. We note that the latter algorithm is known to have linear convergence rates unlike the LR iterations.

4.1. Example 1: PH/PH/1 queue

As a PH/PH/1 system, we study an IPP/E_k/1 queue, where the IPP (interrupted Poisson process) is a PH-type process with two phases, namely the OFF and ON phases, and E_k denotes the Erlangian distribution with k stages [19]. The mean service rate is set to 100 in this numerical example. In an IPP, the arrivals are Poisson with rate λ in the ON phase and there are no arrivals in the OFF phase; the IPP has the following ME representation ($\mathbf{v}_a$, $\mathbf{T}_a$, $\mathbf{s}_a$) given in [19]

$$\mathbf{v}_a = \begin{bmatrix} 0 & 1 \end{bmatrix}, \qquad \mathbf{T}_a = \begin{bmatrix} -\gamma_{01} & \gamma_{01} \\ \gamma_{10} & -(\gamma_{10}+\lambda) \end{bmatrix}, \qquad \mathbf{s}_a = \begin{bmatrix} 0 \\ \lambda \end{bmatrix}.$$

E_k distributions have natural ME representations given in [26]. The burstiness b of an IPP is defined as the ratio between the arrival rate in a burst and the overall average arrival rate. In this numerical example, we fix $\gamma_{01} = 10$ and choose γ_{10} so as to fix the burstiness $b = 4$. The rate parameter λ is then chosen so as to attain a desired load ρ on the queueing system.

The algorithm of [23] uses the LR iterative procedure for the PH/PH/1 queue. The LR procedure was first introduced in [22]. The advantage of the algorithm [23] stems from the reduced size of the matrices that are used within the LR procedure; the order of the matrices are the sum of the phases (i.e., $m = m_a + m_b$) in the arrival and service time distributions in [23]. This is in contrast with matrices of size being their product (i.e., $m_p = m_a m_b$) in the original matrix geometric algorithm given in [26]. This order reduction brings a considerable computational advantage. However, calculation of the input matrices to the LR procedure still require the construction of a matrix with the order of the product of the number of phases in the arrival and service time distributions and further matrix multiplications involving this product-sized matrix [23]. Therefore, the overall algorithm of [23] requires large computation times and storage space when the product m_p is large. We note that this algorithm employs one matrix inversion and eight matrix multiplications (involving matrices of size m and less) per iteration of the loop. On the other hand, the MSF algorithm proposed in this paper does not use matrices of size m_p in any step of the algorithm and

Table 1
The number of iterations required for LR and MSF iterations for the IPP/E_k/1 queue as a function of ρ and the number of stages of the Erlangian service time distribution

ρ	k	Number of iterations		Δ
		LR	MSF	
0.6	4	7	13	2.8×10^{-11}
	16	7	15	1.8×10^{-11}
	64	7	17	1.6×10^{-11}
	256	7	19	1.9×10^{-11}
0.9	4	10	13	1.4×10^{-12}
	16	10	15	4.4×10^{-13}
	64	10	17	2.7×10^{-12}
	256	10	19	7.3×10^{-11}
0.9999	4	19	13	5.2×10^{-6}
	16	19	14	4.2×10^{-6}
	64	19	16	4.2×10^{-6}
	256	19	21	5.9×10^{-6}

the matrix sign iterations require only one matrix inversion of size m. We use both the algorithms to solve for the IPP/E_k/1 queue as a function of the number of stages k of the Erlangian distribution and also as a function of the utilization ρ of the system. The stopping criterion we use is $\varepsilon = 10^{-8}$ where ε is the normalized difference between the 1-norms of the successive iterated matrices for both algorithms.

In Table 1, we report the number of iterations required for the LR and the MSF procedures. Moreover, we introduce a parameter Δ that is indicative of the accuracy of the proposed algorithm. The parameter Δ is calculated as the maximum of three normalized absolute differences with respect to the results of [23]; differences being in the probability mass at the origin, in the mean, and in the variance, of the steady-state waiting time using the MSF algorithm. Table 1 demonstrates that both the LR and MSF algorithms have rapid convergence rates (i.e., quadratic). It generally took fewer iterations for the convergence of the LR algorithm whereas for very heavy loads we have observed cases for which the MSF algorithm required fewer iterations. Increasing the load also increased the number of iteration steps. For this particular queueing problem, the number of iterations for LR did not depend on the system size, whereas the system size was shown to have a slight effect on the convergence of MSF; increasing k also increased the required iteration steps for MSF. In all cases, we have obtained very close results for both algorithms whereas the normalized difference between the results of the two algorithms is shown to increase for critically loaded systems, as would be expected.

4.2. Example 2: D/ME/1 queue

We consider a D/ME/1 queue studied in [5] where the service time distribution has the ME representation $(\mathbf{v}_b, \mathbf{T}_b, \mathbf{s}_b)$

$$\mathbf{v}_b = [\,1 + 4\pi^2 \quad 0 \quad 0\,], \qquad \mathbf{T}_b = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1-4\pi^2 & -3-4\pi^2 & -3 \end{bmatrix}, \qquad \mathbf{s}_b = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

and the arrival process is assumed to be deterministic (i.e., denoted by D) and is therefore degenerate at the point $a = \mathbf{v}_b \mathbf{T}_b^{-2} \mathbf{s}_b / \rho$ for a given load ρ. The matrix analytical approach of [5] involves the iteration

$$\mathbf{v}_+^{(k+1)} = \mathbf{v}_b e^{a(\mathbf{T}_b + \mathbf{s}_b \mathbf{v}_+^{(k)})}, \qquad \mathbf{v}_+^{(0)} = 0, \tag{38}$$

which converges to $\mathbf{v}_+$ as $k \to \infty$. It is shown in [5] that the waiting time has the ME representation $(\mathbf{v}_+, \mathbf{T}_b + \mathbf{s}_b \mathbf{v}_+, (1 + \mathbf{v}_+ \mathbf{T}_b^{-1} \mathbf{s}_b) \mathbf{s}_b)$. A numerical algorithm using squaring and scaling is proposed in [5] for matrix exponentiation at each step of the iteration (38). As an alternative, we suggest to use the approach of this paper for the solution of the D/ME/1 queue. However, the deterministic arrival distribution is not of ME-type and therefore we propose to make use of Pade or Erlangian approximations for the deterministic inter-arrival time [1]. We note that Erlang distribution is of PH-type and therefore it is also ME. A Pade approximation has a matrix exponential representation but it is not guaranteed to be associated with a probability density unless the degree of the approximation is sufficiently high. A Pade(l) or Erlang(l) notation denotes a rational approximation to the irrational Laplace transform e^{-sa} with numerator and polynomial degrees being at most l. The goal of this numerical example is to show if there may be any potential computational benefit of the proposed algorithm in the numerical solution of GI/GI/1 type queues by reporting the number of required iterations. We note that the size of the matrices in the Newton iteration (30) are $m_a + m_b \times m_a + m_b$ as opposed to $m_b \times m_b$ matrices of the iteration (38) which is an advantage of the matrix analytical approach of [5]. However, note that each iteration of [5] requires matrix exponentiation which is computationally more intensive than one matrix inversion required for MSF.

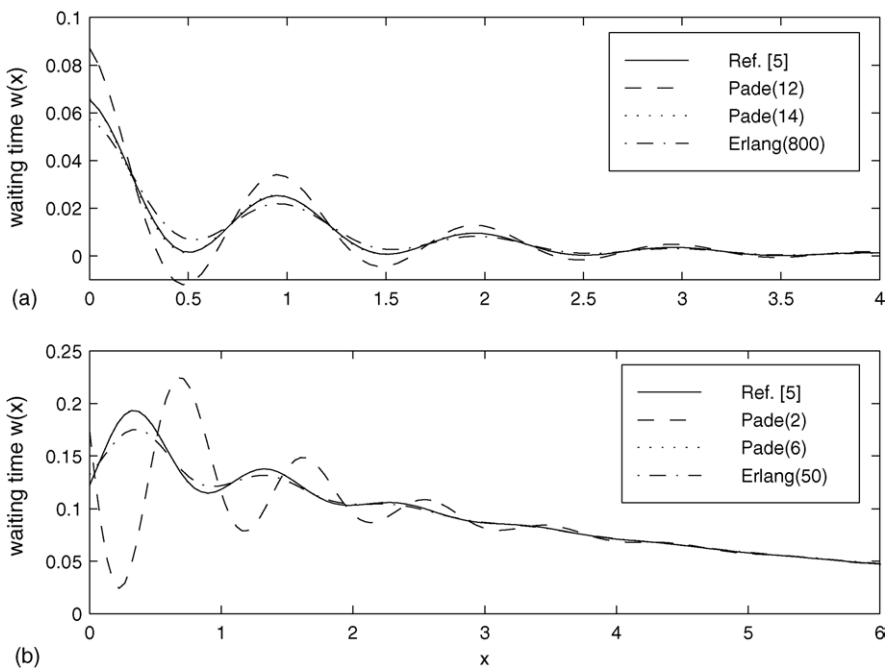

Fig. 3. Steady-state waiting time distribution computed via different methods for the two cases: (a) $\rho = 0.3$; (b) $\rho = 0.9$. The probability masses at the origin are not depicted in the plots.

The densities of the waiting time are plotted using various methods for $\rho = 0.3$ and 0.9 in Fig. 3. For both approaches, the iterations (30) and (38) are stopped when the norm of the difference among the successive values of the iterated variables becomes less than 10^{-10}. For the low load case of $\rho = 0.3$, the matrix analytical approach of [5] required few iterations (i.e., 12 iterations) for convergence and to reach the same level of accuracy, we had to use a Pade approximation with a dimensionality of 14. The convergence of the solution with the Erlangian approximation $E(l)$ to the actual solution as $l \to \infty$ is observed to be very slow in this case since there is a notable difference between the two densities even for the $E(800)$ case. The two Pade approximations (i.e., Pade(12) and Pade(14)) and the Erlang(800) approximation required 11 and 13 Newton iterations, respectively.

We observe a substantial advantage of the approach proposed in this paper for the heavy load case $\rho = 0.9$. In this case, the matrix analytical approach of [5] required 192 iterations for convergence and a dimensionality of 6 for the Pade approximation is shown to be sufficient for the same level of accuracy. Furthermore, we observe that the number of iterations required for the Newton iterations did not increase with load, i.e., the two Pade approximations Pade(2) and Pade(6), and the Erlang(50) approximation, required 9 and 11 Newton iterations, respectively. We are led to believe that it is the reduced system size that resulted in less number of iterations for the heavy load case.

5. Conclusions

A novel approach for the numerical computation of the steady-state waiting times in ME/ME/1 type queues is presented in the current paper. This approach is based on a state–space description of a feedback interconnection system, which is easily derivable from the matrix exponential representations of the individual inter-arrival and service times. Using this state–space description, we formulate the waiting time calculation problem as an SDC problem and we propose to use the Newton iterations for the underlying SDC. The advantages of the proposed approach are the quadratic convergence rates of the proposed iterations, the lack of need for any transform domain calculations, and the fact that the "sum", rather than the "product", of the dimensionalities of the representations for the inter-arrival and service times, determines the computational complexity. Validation of the proposed approach is done by comparing the results with two existing algorithms. We are currently working on extending the results of this paper to Markov renewal queueing systems that arise frequently in the performance analysis of computer and communication systems and networks.

Acknowledgement

This work is supported by a grant from the Science and Technical Research Council of Turkey (project EEEAG-101E025).

References

[1] N. Akar, E. Arıkan, A numerically efficient algorithm for the MAP/D/1/K queue via rational approximations, Queueing Syst. 22 (1996) 97–120.

[2] N. Akar, K. Sohraby, An invariant subspace approach in *M*/*G*/1 and *G*/*M*/1 type Markov chains, Commun. Stat. Stochastic Models 13 (3) (1997) 381–416.
[3] S. Asmussen, Applied Probability and Queues, Wiley, New York, 1987.
[4] S. Asmussen, Phase-type representations in random walk and queueing problems, Ann. Prob. 20 (1992) 772–789.
[5] S. Asmussen, M. Bladt, Renewal theory and queueing algorithms for matrix exponential distributions, in: S.R. Chakravarthy, S. Alfa (Eds.), Matrix Analytic Methods in Stochastic Models, Marcel Dekker, New York, 1997, pp. 313–342.
[6] S. Asmussen, J.R. Moller, Calculation of the steady-state waiting time distribution in GI/PH/*c* and MAP/PH/*c* queues, Queueing Syst. 37 (2001) 9–29.
[7] S. Asmussen, O. Nerman, Fitting phase-type distributions via the EM algorithm, in: Symposium i Anvendt Statistik, 1991, pp. 333–345.
[8] Z. Bai, J. Demmel, Design of a parallel nonsymmetric eigenroutine toolbox. Part I, in: Sixth SIAM Conference on Parallel Processing for Scientific Computing, 1993.
[9] Z. Bai, J. Demmel, M. Gu, Inverse free parallel spectral divide and conquer algorithms for nonsymmetric eigenproblems, Numer. Math. 76 (1997) 389–396.
[10] Z. Bai, J.W. Demmel, On swapping diagonal blocks in real Schur form, Linear Algebra Appl. 186 (1993) 73–95.
[11] M. Bladt, M.F. Neuts, Matrix-exponential distributions: calculus and interpretation via flows, Stochastic Models 19 (1) (2003) 113–124.
[12] R. Byers, Solving the algebraic Riccati equation with the matrix sign function, Linear Algebra Appl. 85 (1987) 267–279.
[13] T.F. Chan, Rank revealing QR factorizations, Linear Algebra Appl. 88–89 (1987) 67–82.
[14] C.T. Chen, Linear System Theory and Design, second ed., Holt, Rinehart, and Winston, New York, 1984.
[15] P.M. Van Dooren, Numerical linear algebra for signals, systems, and control, Draft notes prepared for the Graduate School in Systems and Control, 2003.
[16] P.M. Van Dooren, R.V. Patel, A.J. Laub, Numerical Linear Algebra Techniques for Systems and Control, IEEE Press, Piscataway, NJ, 1994.
[17] A. Feldmann, W. Whitt, Fitting mixtures of exponentials to long-tail distributions to analyze network performance, Perform. Eval. 31 (1998) 245–279.
[18] D. Gross, C.M. Harris, Fundamentals of Queueing Theory, third ed., Wiley, New York, 1998.
[19] B.R. Haverkort, Performance of Computer Communication Systems: A Model-Based Approach, Wiley, New York, 1998.
[20] T. Kailath, Linear Systems, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[21] L. Kleinrock, Queuing Systems. Theory, vol. 1, Wiley, New York, 1975.
[22] G. Latouche, V. Ramaswami, A logarithmic reduction algorithm for quasi-birth–death processes, J. Appl. Prob. 30 (1993) 650–674.
[23] G. Latouche, V. Ramaswami, The PH/PH/1 queue at epochs of queue size change, Queueing Syst. 25 (1997) 97–114.
[24] L.R. Lipsky, Queueing Theory: A Linear Algebraic Approach, Macmillan, New York, 1992.
[25] K. Mitchell, A. van de Liefvoort, Approximation models of feed-forward *G*/*G*/1/*N* queueing networks with correlated arrivals, Perform. Eval. 51 (2003) 137–152.
[26] M.F. Neuts, Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach, Johns Hopkins University Press, Baltimore, MD, 1981.
[27] M.F. Neuts, Structured Stochastic Matrices of *M*/*G*/1 Type and Their Applications, Marcel Dekker, New York, 1989.
[28] J.D. Roberts, Linear model reduction and solution of the algebraic Riccati equation by the use of the sign function, Int. J. Control 32 (1980) 677–687.
[29] B. Sengupta, Markov processes whose steady state distribution is matrix-exponential with an application to the GI/PH/1 queue, Adv. Appl. Prob. 21 (1) (1989) 159–180.
[30] D. Starobinski, M. Sidi, Modeling and analysis of heavy-tailed distributions via classical teletraffic methods, Queueing Syst. 26 (2000) 243–267.
[31] A. van de Liefvoort, The waiting time distribution and its moments of the PH/PH/1 queue, OR Lett. 9 (1990) 261–269.