Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2008, Article ID 183089, 15 pages
doi:10.1155/2008/183089

Research Article

Generalized Truncated Methods for an Efficient Solution of Retrial Systems

Ma Jose Domenech-Benlloch, Jose Manuel Gimenez-Guzman, Vicent Pla, Jorge Martinez-Bauset, and Vicente Casares-Giner

Departamento de Comunicaciones, Universidad Politécnica de Valencia (UPV), Camì de Vera s/n, 46022 Valencia, Spain

Correspondence should be addressed to Jose Manuel Gimenez-Guzman, jogiguz@upvnet.upv.es

Received 2 July 2007; Accepted 5 March 2008

Recommended by Nahum Shimkin

We are concerned with the analytic solution of multiserver retrial queues including the impatience phenomenon. As there are not closed-form solutions to these systems, approximate methods are required. We propose two different generalized truncated methods to effectively solve this type of systems. The methods proposed are based on the homogenization of the state space beyond a given number of users in the retrial orbit. We compare the proposed methods with the most well-known methods appeared in the literature in a wide range of scenarios. We conclude that the proposed methods generally outperform previous proposals in terms of accuracy for the most common performance parameters used in retrial systems with a moderated growth in the computational cost.

1. Introduction

A common assumption when evaluating the performance of communication networks is that users that do not obtain an immediate service leave the system without retrying. However, due to the increasing number of customers and network complexity, the customer behavior, in general, and the retrial phenomenon, in particular, may have a significant impact on the network performance. We can find numerous examples of retrial queues in communication networks. In addition to classical telephony networks [1], the effect of retrials in wireless networks [2, 3] is usually a more delicate matter. We can also find the effect of retrials in data transfers along Internet [4]. However, it must be noted that retrial queues can also be observed in a wider range of systems, such as communication protocols or in many queues in daily life where a customer retries its access to a queue hoping to experience a lower delay in the access to a free server. Regarding the appearance of retrials in communication protocols, we can highlight its common use in the performance evaluation of MAC protocols like CSMA/CD [5]. The retrial phenomenon is also essential in modeling call centers (see [6] for a wealth of references on this issue).

The modeling of repeated attempts has been the subject of numerous investigations. Two functional blocks are typically distinguished in models which consider retrials: a block that accommodates the servers and possibly a waiting queue, and a block where users that retry are accommodated, usually called retrial orbit. More concretely, we consider an $M/M/C$ queueing system (following Kendall's notation $M/M/x$ stands for a queue with exponentially distributed interarrival and service times and an infinite capacity waiting room) with retrials where, following the human behavior, users can abandon the system with certain probability after an unsuccessful retry, what is usually referred to as impatience. Therefore, the resulting model under study will have a nonhomogeneous (because the retrial rate depends on the number of users in the retrial orbit) and infinite state space. It is known that the classical theory [7] is developed for random walks on the semistrip $\{0, \ldots, C\} \times \mathbb{Z}_+$ with infinitesimal transitions subject to conditions of space homogeneity. When the space-homogeneity condition does not hold, that is, in the case described in this paper, the problem of calculating the equilibrium distribution has not been solved beyond approximate methods [8, 9]. Indeed, if we focus on the case of multiserver retrial queues, the absence of closed-form solutions for the main performance characteristics when $C > 2$ [10] can be emphasized. Therefore, it is clear that in this case it is necessary to resort to approximate methods. These methods are usually grouped into three categories: approximations, finite truncated methods, and generalized truncated methods. Although all the mentioned categories are in fact approximations and, therefore, sometimes it is not clear which category a method belongs to, the first category is usually devoted to methods that can be useful in a certain domain of the system parameters or in special extreme cases (low retrial intensity, high blocking probabilities, etc.). Both the finite truncated and generalized truncated methods replace the original infinite state space by a solvable state space, that is, a model where steady state probabilities can be computed. However, while for the finite truncated methods the initial state space is replaced by a finite state space, in generalized truncated methods it is replaced by another infinite but solvable state space.

The main contribution of this paper is the development of two new generalized truncated methods that are able to effectively solve retrial systems with user impatience. These novel methods are inspired in the model proposed by Neuts and Rao [11]. Comparing the proposed methods with the most well-known approaches appeared in the literature, we conclude that the new methods generally outperform the previous proposals in terms of accuracy for the most common performance parameters used in retrial systems and under a wide range of scenarios. Moreover, we show that their computational cost increases moderately compared to the simpler methods. We also show that all the generalized truncated methods clearly outperform the finite truncated methods in terms of accuracy, as it was predicted in [10].

The rest of the paper is structured as follows. Section 2 describes the model of the system under study and Section 3 summarizes previous approaches appeared in the literature. In Section 4 we describe the two proposed methods, and their performance is evaluated and compared to previous approaches in Section 5. Final remarks and a summary of the results are provided in Section 6.

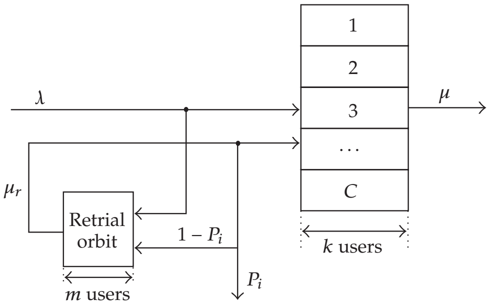


Figure 1: System model.

2. System model

In the system under study (shown in Figure 1), users arrive following a Poisson stream with rate λ and contend for access to a system with C servers, requesting an exponentially distributed service time with rate μ. Furthermore, we define the load of the system by $\rho = \lambda/(C\mu)$. Without loss of generality, we consider that each user occupies one resource unit. When a new request finds all servers occupied, it joins the retrial orbit with probability 1, having considered an infinite capacity for the retrial orbit. After an exponentially distributed time of rate μ_r this user retries, being a successful retrial if it finds a free server. Otherwise, the user leaves the system with probability P_i or returns to the retrial orbit with probability $(1-P_i)$, starting the retrial procedure again.

The model considered can be represented as a bidimensional continuous-time Markov chain (CTMC) whose state space is defined by

$$\mathcal{S} := \{s = (k, m) \,:\, k \leq C; m \in \mathbb{Z}_+\}, \tag{2.1}$$

being k the number of users being served and m the number of users in the retrial orbit, constituting a level dependent Quasi Birth and Death Process (QBD) [7] whose transition diagram is depicted in Figure 2. In QBD related literature, the term *level* refers to a set of states with the same second coordinate. The infinitesimal generator of this process has the following infinite block tridiagonal structure:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{A}_1^{(0)} & \mathbf{A}_0^{(0)} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{A}_2^{(1)} & \mathbf{A}_1^{(1)} & \mathbf{A}_0^{(1)} & \mathbf{0} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{A}_2^{(2)} & \mathbf{A}_1^{(2)} & \mathbf{A}_0^{(2)} & \mathbf{0} & \cdots \\ \mathbf{0} & \mathbf{0} & \ddots & \ddots & \ddots & \cdots \end{bmatrix}. \tag{2.2}$$

The main characteristics of this model are its infinite state space ($\{0, \ldots, C\} \times \mathbb{Z}_+$) and also its space heterogeneity produced by the fact that the retrial rate depends on the number of customers in the retrial orbit.

2.1. Computation of performance parameters

The most common performance parameter used in retrial systems is the blocking probability (P_b), which is defined as the probability of the system being in a state where arrivals are not

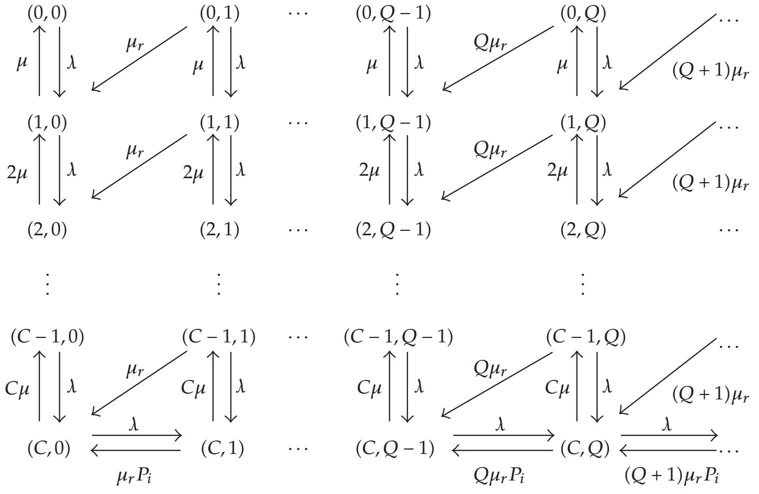


Figure 2: Retrial model under study.

accepted. Notwithstanding, other performance parameters can describe the behavior of retrial systems more accurately. Those performance parameters are the mean number of users in the retrial orbit (N_{ret}), the immediate service probability (P_{is}), the delayed service probability (P_{ds}), and the nonservice probability (P_{ns}). P_{is} is defined as the probability of a user being served in its first attempt, P_{ds} as the probability of obtaining service but not in the first attempt, and P_{ns} as the probability of leaving the system due to impatience without having been served. Obviously, it must be met that $P_{\mathrm{is}} + P_{\mathrm{ds}} + P_{\mathrm{ns}} = 1$.

3. Overview of approximate methods

As stated in Section 1, it is known [10] that neither a closed-form analytical solution nor a direct algorithmic computation of the steady state probabilities has been obtained for the model under study, so we must resort to approximate methods. Those approximations can be classified into three categories [10].

(i) Approximations: this category includes those solutions in which the original intractable model is simplified making some assumptions that usually suppose a good approximation only in a certain domain of the system parameters or in special extreme cases.

(ii) Finite truncated methods: these methods consist in replacing the original infinite state space by a finite one, so the resulting model is solvable.

(iii) Generalized truncated methods: in this case, the original infinite state space is replaced by another infinite but simplified and solvable state space.

It must be noted that the difference between the category *approximations* and the rest is sometimes not clear because all the methods are in fact approximations, and usually the accuracy of a method is better in a certain domain of the problem.

3.1. Approximations

An intuitive first approach to solve retrial systems is to consider those retrials as new arrivals requesting the access to a server. With this approximation, known as *loss method* [12, Section 2.8], the system becomes a simple loss model whose solution can be obtained using Erlang's B formula with an arrival rate that is the sum of new users and retrials. Although this solution is very simple, intuitively its main drawback is the expected inaccuracy when the retrial rate is high.

On the other side, for high values of the retrial rate the system can be approximated by an $M/M/C$ queueing system, so that the blocking probability can be computed using Erlang's C formula. As the loss method and the $M/M/C$ approach are useful for low and high retrial rates, respectively, in [12, Section 2.8], an interpolation of the above two methods is proposed to obtain a more accurate solution. Hereafter, we will refer to this method as *Int*.

In [13], Greenberg and Wolff propose another approximation (denoted by GW) based on the assumption that the returning customers see time averages. Obviously, the approximation made in GW is expected to work more accurately when the retrial rate decreases, because when a customer is blocked and reattempts quickly, the probability of finding the system busy is high.

3.2. Finite truncated methods

The first finite truncated method was proposed by Wilkinson in [14]. This method, denoted by *Wil*, is based on the truncation of the state space of the QBD beyond a level Q, that is, the method restricts the maximum number of users in the retrial orbit to Q. Obviously, the method is expected to be more accurate as we use larger values for the truncation level Q, but its computational cost will also increase. Other authors present more efficient truncated methods, for example, in [15] the truncation is based on the exclusion of those states with negligible stationary probabilities.

There also exist other finite truncated methods that modify the truncated state space to introduce, in some sense, the effect of retrials. The model presented by Fredericks and Reisner [16], and called hereafter FR, reduces the dimensionality of the model to a one dimensional state space, eliminating the dimension corresponding to the number of users in the retrial orbit and introducing its effect as a new arrival rate that depends on the state of the system.

In a similar way, Marsan et al. [17] propose a method (denoted by *Mar*) that reduces the infinite state space to a finite one grouping all the levels of the QBD where there are users in the retrial orbit into a single level. In [18] we propose a generalized version of Mar method, called FM. Comparing FM and Mar, FM shows a substantial improvement in the accuracy at the expense of only a marginal increase in the computational cost. These two methods include the effect of the eliminated states by modifying the transition rates of the last level of the resulting QBD. The difference between both methods lies in the number of levels taken into account in the finite truncated model. While Mar groups all the states with retrials into a single level, FM defines a value Q so the aggregation is performed beyond level Q. Mar can be seen as a particular case of FM, where $Q = 1$. Note also that by increasing the value of Q both the accuracy and the computational cost will also increase. Due to that aggregation two new parameters are introduced: Θ and p. The parameter Θ denotes the mean number of users in the retrial orbit conditioned to those states where there are at least Q users in the orbit, that is,

$\Theta = E(m \mid m \geq Q)$. The probability that after a successful retrial the number of users in the retrial orbit does not drop below Q is represented by p. The expressions needed to compute those parameters are [18]

$$p = \frac{\pi(C,Q)}{\pi(C,Q-1) + \pi(C,Q)},$$
$$\Theta = \frac{\lambda\left(\pi(C,Q-1) + \pi(C,Q)\right)}{\mu_r\left[\sum_{k=0}^{k=C-1}\pi(k,Q) + P_i\pi(C,Q)\right]}. \tag{3.1}$$

As there is a mutual dependence between Θ and p and the steady state probabilities $(\pi(k,m))$, to find these values we followed a fixed-point iterative procedure starting with $p = 0$ and $\Theta = Q$.

3.3. Generalized truncated methods

In this section, we discuss the use of generalized truncated methods which approximate the $M/M/C$ retrial queue by some infinite although solvable system. We comment separately on the method proposed by Neuts and Rao in [11] as the methods proposed in this paper are based on it.

The method introduced by Falin in [19] (denoted by *Fal*) is based on the assumption that the retrial rate becomes infinite when the number of customers in orbit exceeds a certain level Q. Therefore, the system is approximated beyond level Q by a standard $M/M/1$ queue, and the retrial rate will depend on the number of users in the retrial orbit as follows:

$$\mu_r(m) = \begin{cases} m\mu_r & \text{if } m < Q, \\ \infty & \text{if } m \geq Q. \end{cases} \tag{3.2}$$

Artalejo and Pozo [10] proposed a method (denoted by AP) based on the $M/M/2$ queue, instead of the $M/M/1$ queue. Therefore, the retrial rate will depend not only on the number of users in the retrial orbit (m) but also on the number of busy servers (k) as follows:

$$\mu_r(m) = \begin{cases} \infty & \text{if } k < C-1,\ m \geq Q, \\ m\mu_r & \text{otherwise.} \end{cases} \tag{3.3}$$

3.3.1. NR method

The method proposed by Neuts and Rao in [11], and called hereafter NR, which was proved to converge to the original model in [20], is based on the homogenization of the model beyond a given level Q of the QBD, which supposes to restrict the maximum retrial rate, that is,

$$\mu_r(m) = \begin{cases} m\mu_r & \text{if } m < Q, \\ Q\mu_r & \text{if } m \geq Q. \end{cases} \tag{3.4}$$

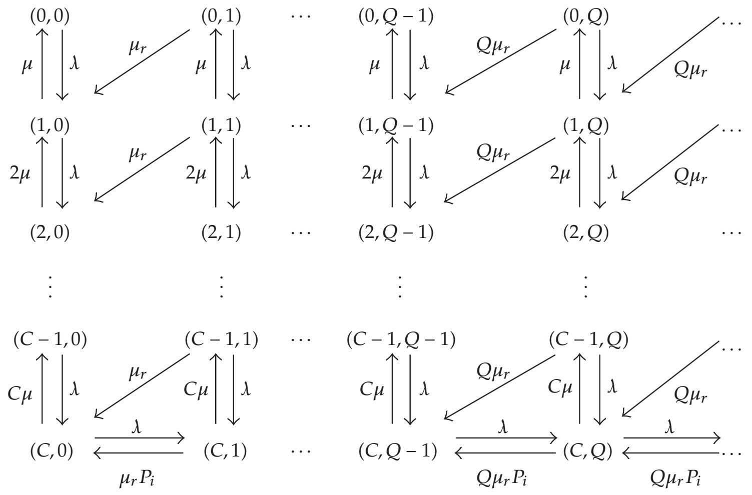


Figure 3: NR model.

Figure 3 depicts the system model used in NR method. The resulting approximated model can be solved by the matrix-geometric approach [7]. Let

$$\widehat{\mathbf{Q}} = \begin{bmatrix} \mathbf{A}_1^{(0)} & \mathbf{A}_0^{(0)} & 0 & \cdots & 0 & 0 \\ \mathbf{A}_2^{(1)} & \mathbf{A}_1^{(1)} & \mathbf{A}_0^{(1)} & \cdots & 0 & \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{A}_1^{(Q-1)} & \mathbf{A}_0^{(Q-1)} \\ 0 & 0 & 0 & \cdots & \mathbf{A}_2 & \mathbf{A}_1 + \mathbf{R}\mathbf{A}_2 \end{bmatrix}, \tag{3.5}$$

where $\mathbf{R}$ is the *rate matrix* defined as the unique minimal nonnegative solution of the quadratic matrix equation $\mathbf{R}^2\mathbf{A}_2 + \mathbf{R}\mathbf{A}_1 + \mathbf{A}_0 = 0$. In order to compute matrix $\mathbf{R}$, we have used the logarithmic reduction algorithm as proposed in [9, Section 8.4], using a precision of 10^{-8} for the iterative procedure. The steady state probabilities for states where $m \leq Q$ are obtained by solving $[\boldsymbol{\pi}^{(0)} \cdots \boldsymbol{\pi}^{(Q)}]\widehat{\mathbf{Q}} = \mathbf{0}$ combined with the normalization condition

$$\sum_{l=0}^{Q-1} \boldsymbol{\pi}^{(l)}\mathbf{e} + \boldsymbol{\pi}^{(Q)}(\mathbf{I} - \mathbf{R})^{-1}\mathbf{e} = 1. \tag{3.6}$$

As $\widehat{\mathbf{Q}}$ is a finite matrix, this system can be solved by any of the standard methods defined in the classical linear algebra. Finally, the steady state probabilities of the states in the homogeneous part of the model are easily computed by $\boldsymbol{\pi}^{(Q+n)} = \boldsymbol{\pi}^{(Q)}\mathbf{R}^n$.

4. Proposed methods

4.1. HM1 method

When the number of users retrying is higher than Q, the NR method considers the retrial rate to be $Q\mu_r$. Obviously, this is a rough approximation that can be inaccurate in many cases.

Therefore, we propose to approximate the retrial rate beyond level Q by an estimation of its mean value, that is, $M = E[m \mid m \geq Q]$, keeping the idea of the homogenization:

$$\mu_r(m) = \begin{cases} m\mu_r & \text{if } m < Q, \\ M\mu_r & \text{if } m \geq Q. \end{cases} \tag{4.1}$$

The physical meaning of the proposed approximation is founded on assigning to the retrial rate beyond level Q a value that is the mean number of users retrying, when there are at least Q users retrying. This value can be computed by using

$$M = E[m \mid m \geq Q] = \frac{\sum_{r \geq Q} r\boldsymbol{\pi}^{(r)}\mathbf{e}}{\sum_{r \geq Q} \boldsymbol{\pi}^{(r)}\mathbf{e}} = \frac{\boldsymbol{\pi}^{(Q)}\left[\mathbf{R}(\mathbf{I}-\mathbf{R})^{-1} + Q\mathbf{I}\right](\mathbf{I}-\mathbf{R})^{-1}\mathbf{e}}{\boldsymbol{\pi}^{(Q)}(\mathbf{I}-\mathbf{R})^{-1}\mathbf{e}}. \tag{4.2}$$

However, it must be noticed that M depends on the value of Q, which is a configuration parameter, and more importantly, on the steady state probabilities. Therefore, we obtain a set of nonlinear equations relating M, $\mathbf{R}$, and the probability vectors $\boldsymbol{\pi}^{(0)}, \ldots, \boldsymbol{\pi}^{(Q)}$.

To solve this system, we use an approximation based on the assumption that when the number of customers in the orbit is sufficiently high then it is very likely that all servers are being used. Hence, $\boldsymbol{\pi}^{(Q)} \approx \pi_{(C,Q)}\boldsymbol{\psi}$, where $\boldsymbol{\psi} = [0\ 0 \cdots 0\ 1]^t$, and thus M and $\mathbf{R}$ no longer depend on $\boldsymbol{\pi}^{(Q)}$:

$$M \approx \frac{\boldsymbol{\psi}\left[\mathbf{R}(\mathbf{I}-\mathbf{R})^{-1} + Q\mathbf{I}\right](\mathbf{I}-\mathbf{R})^{-1}\mathbf{e}}{\boldsymbol{\psi}(\mathbf{I}-\mathbf{R})^{-1}\mathbf{e}}. \tag{4.3}$$

However, as the relationship between $\mathbf{R}$ and M still holds, we deploy an iterative procedure as follows: (1) set $M_0 = Q$; (2) given M_n compute $\mathbf{R}$ using the logarithmic reduction algorithm, then compute M_{n+1} using (4.3); (3) if $|M_{n+1} - M_n|/M_n \geq \varepsilon = 10^{-3}$ go to step 2; (4) with the final values of M and $\mathbf{R}$ compute the steady state probabilities solving the finite QBD.

4.2. HM2 method

We must notice that the approximation that performs HM1 ($\boldsymbol{\pi}^{(Q)} \approx \pi_{(C,Q)}\boldsymbol{\psi}$) may not be very accurate when the blocking probabilty is low, although it is expected to outperform NR. For that reason, we propose the HM2 method, which computes the value of M using (4.2). More specifically, HM2 method uses the iterative procedure sketched next: (1) set $M_0 = Q$; (2) given M_n compute $\mathbf{R}$ and the steady state probabilities solving the finite QBD, then compute M_{n+1} using (4.2); (3) if $|M_{n+1} - M_n|/M_n \geq \varepsilon = 10^{-3}$ go to step 2.

Obviously, the performance of the HM2 method is expected to be higher than that of the HM1 method, but the computational cost of HM2 is also expected to be higher due to the fact that in HM1 the values of $\boldsymbol{\pi}^{(0)}, \ldots, \boldsymbol{\pi}^{(Q)}$ are computed only once, and this is not true in HM2.

4.3. Computation of performance parameters

For the methods HM1 and HM2, we have used the next expressions to compute the performance parameters described in Section 2.1.

$$\begin{aligned}
P_b &= \sum_{m=0}^{Q-1} \boldsymbol{\pi}^{(m)}\mathbf{z} + \boldsymbol{\pi}^{(Q)}(\mathbf{I}-\mathbf{R})^{-1}\mathbf{z} \quad \text{with} \quad \mathbf{z} = [0,0,\ldots,0,1], \\
P_{\text{is}} &= 1 - P_b, \\
P_{\text{ds}} &= \lambda^{-1}\mu_r\left[\sum_{m=0}^{Q-1} m\boldsymbol{\pi}^{(m)}\mathbf{o} + M\boldsymbol{\pi}^{(Q)}(\mathbf{I}-\mathbf{R})^{-1}\mathbf{o}\right] \quad \text{with} \quad \mathbf{o} = [1,1,\ldots,1,0], \\
P_{\text{ns}} &= \lambda^{-1}P_i\mu_r\left[\sum_{m=0}^{Q-1} m\boldsymbol{\pi}^{(m)}\mathbf{z} + M\boldsymbol{\pi}^{(Q)}(\mathbf{I}-\mathbf{R})^{-1}\mathbf{z}\right] \quad \text{with} \quad \mathbf{z} = [0,0,\ldots,0,1], \\
N_{\text{ret}} &= \sum_{m=0}^{Q-1} m\boldsymbol{\pi}^{(m)}\mathbf{e} + \boldsymbol{\pi}^{(Q)}\left(\mathbf{R}(\mathbf{I}-\mathbf{R})^{-1} + Q\mathbf{I}\right)(\mathbf{I}-\mathbf{R})^{-1}\mathbf{e} \quad \text{with} \quad \mathbf{e} = [1,1,\ldots,1].
\end{aligned} \tag{4.4}$$

5. Results

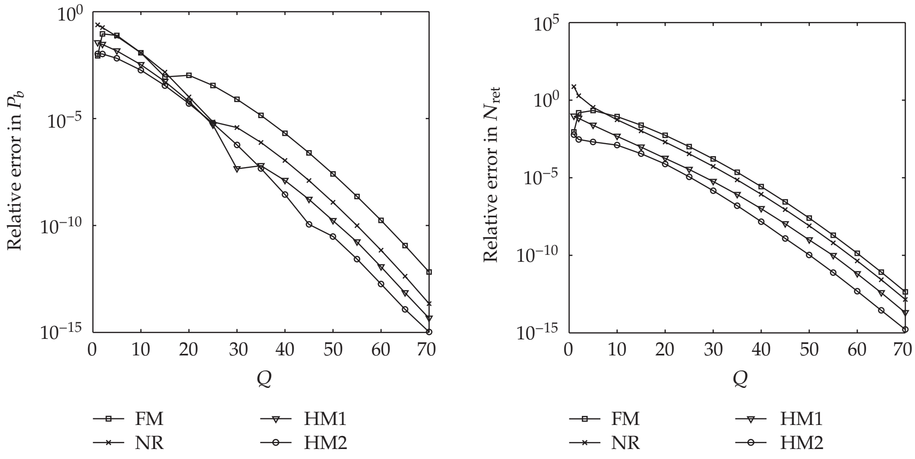
This section is aimed to compare the proposed methods HM1 and HM2 with the most well-known approaches appeared in the literature so far. In order to evaluate and compare the proposed algorithms with previous approaches, we have studied a wide range of scenarios. Letting $\rho = \lambda/(C\mu)$, we have studied different system loads by modifying λ and keeping $C = 50$ and $\mu^{-1} = 180\,s$ fixed. It must be noted that, due to the introduction of the impatience phenomenon, we will be able to consider values of $\rho > 1$. We have also studied different configurations of the retrial behavior by varying μ_r using in all cases $P_i = 0.2$. Note that it is not necessary to modify P_i to have different loads in the retrial orbit, as we use μ_r for that purpose.

The comparison of HM1 and HM2 is done with FM and NR methods, which can be considered representative examples of a finite truncated method and a generalized truncated method, respectively. It must be noted that the resulting QBD has been solved using the algorithm proposed in [21] in all cases, although other algorithms have been proposed in the literature [22, 23]. For the iterative procedures in FM, HM1, and HM2 the accuracy has been chosen in all cases to be $\varepsilon = 10^{-3}$.

For comparing the results, we have used the relative error of different performance parameters, defined as $|\Psi^{\text{approx}} - \Psi^{\text{exact}}|/\Psi^{\text{exact}}$ for a generic performance parameter Ψ. In order to obtain an accurate enough estimate of Ψ which can be used as Ψ^{exact}, we ran all methods with increasing and sufficiently high values of Q so that the value of Ψ had stabilized up to the 14th decimal digit. As an example and for the particular case $\rho = 0.8$ and $\mu_r = 0.01$, Figure 4 depicts that the general trend is that the higher the value of Q the lower the relative error is. This is due to the fact that the system under consideration becomes more similar to the exact model as Q increases. As expected all methods converged to the same value in the performance parameters under study, $\Psi \in \{P_b, P_{\text{ds}}, P_{\text{ns}}, N_{\text{ret}}\}$. In Table 1 we show the values of P_b and N_{ret} that are going to be considered as the exact values, obtained as described above. Obviously, to obtain those values we need higher values of Q as P_b and N_{ret} increase. In general, when we use values

Table 1: Estimations of P_b and N_{ret}.

ρ	$\mu_r = 0.001$	$\mu_r = 0.01$	$\mu_r = 0.1$	$\mu_r = 1.0$
0.4	$P_b = 7.664\cdot10^{-9}$ $N_{\text{ret}} = 8.558\cdot10^{-7}$	$P_b = 7.951\cdot10^{-9}$ $N_{\text{ret}} = 9.276\cdot10^{-8}$	$P_b = 9.299\cdot10^{-9}$ $N_{\text{ret}} = 1.487\cdot10^{-8}$	$P_b = 9.360\cdot10^{-9}$ $N_{\text{ret}} = 3.017\cdot10^{-9}$
0.6	$P_b = 2.262\cdot10^{-4}$ $N_{\text{ret}} = 3.800\cdot10^{-2}$	$P_b = 2.566\cdot10^{-4}$ $N_{\text{ret}} = 4.627\cdot10^{-3}$	$P_b = 3.378\cdot10^{-4}$ $N_{\text{ret}} = 8.947\cdot10^{-4}$	$P_b = 3.050\cdot10^{-4}$ $N_{\text{ret}} = 1.531\cdot10^{-4}$
0.8	$P_b = 2.548\cdot10^{-2}$ $N_{\text{ret}} = 5.883$	$P_b = 3.318\cdot10^{-2}$ $N_{\text{ret}} = 0.876$	$P_b = 3.994\cdot10^{-2}$ $N_{\text{ret}} = 0.160$	$P_b = 2.897\cdot10^{-2}$ $N_{\text{ret}} = 2.009\cdot10^{-2}$
1.0	$P_b = 0.347$ $N_{\text{ret}} = 1.359\cdot10^{2}$	$P_b = 0.333$ $N_{\text{ret}} = 14.332$	$P_b = 0.273$ $N_{\text{ret}} = 1.573$	$P_b = 0.172$ $N_{\text{ret}} = 0.154$
1.2	$P_b = 0.6482$ $N_{\text{ret}} = 450.19$	$P_b = 0.6387$ $N_{\text{ret}} = 44.74$	$P_b = 0.5436$ $N_{\text{ret}} = 4.300$	$P_b = 0.3541$ $N_{\text{ret}} = 0.3928$
1.4	$P_b = 0.7563$ $N_{\text{ret}} = 745.68$	$P_b = 0.7524$ $N_{\text{ret}} = 74.44$	$P_b = 0.7028$ $N_{\text{ret}} = 7.290$	$P_b = 0.5032$ $N_{\text{ret}} = 0.6700$
2.0	$P_b = 0.8713$ $N_{\text{ret}} = 1598.46$	$P_b = 0.8706$ $N_{\text{ret}} = 159.81$	$P_b = 0.8619$ $N_{\text{ret}} = 15.935$	$P_b = 0.7527$ $N_{\text{ret}} = 1.538$
5.0	$P_b = 0.9613$ $N_{\text{ret}} = 5780.45$	$P_b = 0.9612$ $N_{\text{ret}} = 578.04$	$P_b = 0.9607$ $N_{\text{ret}} = 57.80$	$P_b = 0.9539$ $N_{\text{ret}} = 5.770$
10.0	$P_b = 0.9821$ $N_{\text{ret}} = 12728.44$	$P_b = 0.9821$ $N_{\text{ret}} = 1272.84$	$P_b = 0.9820$ $N_{\text{ret}} = 127.28$	$P_b = 0.9809$ $N_{\text{ret}} = 12.725$



(a) P_b

(b) N_{ret}

Figure 4: Evolution of the relative error for the different methods deployed.

of $\rho > 1$, the value of Q needed to obtain a tight estimation of Ψ increases significantly, so the computational complexity makes the resolution to be unfeasible from a practical point of view when $\rho > 10$. For that reason and since blocking probabilities between 50% and 75% are unacceptable for virtually any practical application, the comparison among different methods is done in a range of $0.4 \le \rho \le 1.4$.

The different methods are compared by using a metric based on the minimum truncation level needed to obtain a certain relative error, which has been widely used in the literature

Table 2: Minimum value of Q to obtain relative errors lower than 10^{-4}.

		$\mu_r = 0.001$				$\mu_r = 0.01$				$\mu_r = 0.1$				$\mu_r = 1.0$			
ρ		P_b	N_{ret}	P_{ds}	P_{ns}	P_b	N_{ret}	P_{ds}	P_{ns}	P_b	N_{ret}	P_{ds}	P_{ns}	P_b	N_{ret}	P_{ds}	P_{ns}
0.4	FM	6	8	6	14	8	10	7	13	5	8	8	9	4	4	5	**4**
	NR	5	9	4	11	7	9	6	11	**4**	7	6	8	4	5	**4**	**4**
	HM1	5	9	4	11	7	8	9	10	5	6	7	8	**3**	**3**	**4**	5
	HM2	**1**	**2**	**2**	**6**	**4**	**5**	**3**	**8**	**4**	**5**	**4**	**6**	**3**	**3**	**4**	**4**
0.6	FM	17	18	16	27	15	18	13	22	10	11	11	12	5	5	5	5
	NR	11	16	11	20	13	14	12	18	8	10	9	10	**4**	5	5	5
	HM1	10	15	17	20	11	13	15	18	**5**	8	10	11	**4**	**4**	5	6
	HM2	**4**	**4**	**2**	**13**	**6**	**8**	**4**	**12**	6	**7**	**5**	**8**	**4**	**4**	**4**	**4**
0.8	FM	56	61	53	76	30	32	32	37	13	12	14	14	6	**4**	6	5
	NR	43	49	42	58	21	29	24	32	12	13	12	12	5	6	**5**	5
	HM1	33	39	50	57	20	22	28	32	10	11	13	14	**4**	5	6	6
	HM2	**16**	**23**	**18**	**39**	**19**	**20**	**19**	**22**	**9**	**9**	**9**	**10**	**4**	**4**	**5**	**4**
1.0	FM	264	250	266	266	58	54	59	58	17	15	17	16	6	5	7	5
	NR	226	254	230	236	50	56	51	50	15	17	15	14	6	7	6	5
	HM1	**137**	211	249	256	41	46	54	57	13	13	16	17	**5**	5	6	7
	HM2	144	**190**	**182**	**197**	**38**	**35**	**40**	**40**	**12**	**10**	**12**	**11**	**5**	4	**5**	**4**
1.2	FM	566	552	571	535	87	83	88	77	20	18	21	17	7	6	7	5
	NR	516	564	523	**487**	76	86	78	68	18	20	18	15	6	7	**6**	5
	HM1	**453**	**498**	560	562	63	73	84	86	15	16	19	20	**5**	5	7	7
	HM2	454	500	**499**	500	**62**	**62**	**63**	**64**	**14**	**11**	**15**	**13**	**5**	**4**	**6**	**3**
1.4	FM	856	837	864	791	115	110	118	99	23	21	24	18	7	6	8	5
	NR	792	861	803	**750**	103	117	105	**87**	20	24	21	16	6	8	7	5
	HM1	**748**	**792**	856	858	84	101	115	117	**17**	19	23	24	6	6	7	8
	HM2	**748**	795	**794**	795	**83**	**91**	**92**	91	**17**	**14**	**18**	**15**	**5**	**5**	**6**	**4**

[10, 11]. Table 2 shows the minimum value of Q needed to obtain a relative error lower than 10^{-4} for P_b, P_{ds}, P_{ns} and N_{ret}. Note that a wide range of operation points for the occupancy of the servers and the retrial orbit has been chosen. The number in bold indicates the lowest value of Q for all the models studied. Finally, it is important to note that immediate service probability has been omitted as it is the complementary value of P_b in the system under study.

Results show that the best performance is obtained by HM2, followed by HM1, NR, and FM. This trend is obtained in almost every scenario studied and for all the performance parameters. It is not unexpected that HM2 is more accurate than NR and HM1 as these can be considered as approximations of HM2. As it could be expected generalized truncated methods outperform finite truncated methods, as predicted in [12]. More specifically, the NR method improves FM, as it requires a lower value of Q for obtaining a concrete relative error. HM1 method obtains better accuracy than NR method for the same value of Q for P_b and N_{ret}, but this is not true for P_{ds} and P_{ns}, in which the inverse behavior is obtained. Finally, HM2 method clearly outperforms all methods, as it needs much lower values of Q to obtain the desired accuracy for all the performance parameters under study.

Nonetheless, as we are dealing with numerical methods some attention must be paid to the computational cost. Note that the same methodology for solving the QBD and the same precision for the iterative procedures have been used in all cases. In Figure 5 we show the

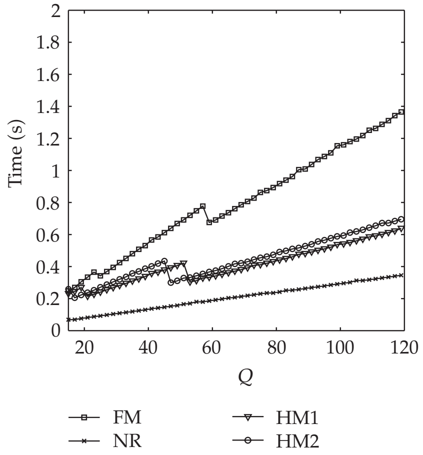


Figure 5: Computation time for the different methods deployed.

time needed to solve the system when $\rho = 0.8$ and $\mu_r = 10^{-2}$ for different values of Q. Results have been obtained running Matlab in an Intel Pentium Core 2. As observed, the generalized truncated methods are not only more accurate than FM, but also they are computationally more efficient. However, we must note that times are negligible from a human point of view in all the generalized truncated methods. If we compare the results with those obtained in Figure 4, we conclude that we can obtain high accuracies with small computation times, specially for the generalized truncated methods.

5.1. Scenario with no impatience

Although the impatience phenomenon is part of the human behavior, some of the most well-known methods to solve retrial systems do not consider it. For that reason, in this section we consider the particular case of the proposed model in which $P_i = 0$, that is, we do not take into account user impatience. In this way we will be able to make a global comparison of all methods. Note that in this scenario, the load offered to the system (ρ) has to be less than one. Moreover, we have decreased the number of servers to $C = 10$, because the computational complexity of AP makes it infeasible to solve the system when $C \geq 50$ servers. We have focused only in the blocking probability because when we do not consider the impatience phenomenon $P_{ns} = 0$ and, therefore, $P_{ds} = P_b$. We can classify all the methods described in Section 3 into two categories: those in which the truncation level Q is a configurable parameter and those that offer a unique solution. In the former the precision can be adjusted through the parameter Q while in the latter precision is an intrinsic feature of the method and hence it is fixed. In the first group we find Wil, FM, Fal, NR, AP, HM1, and HM2 while in the latter we find FR, GW, Loss, Int and Mar. In a first part, we study the performance of the latter, showing P_b and the relative error for each method in Table 3. As we can see, errors are usually lower when we increase the time between retrials but, in any case, errors are unacceptable in most cases. As Mar is the best method in all of them and it is a particular case of FM in the sequel only the FM method will be considered as it yields an upper bound for the accuracy of fixed precision methods.

Table 3: P_b and its relative error obtained with fixed precision methods.

ρ		$\mu_r = 0.001$	$\mu_r = 0.01$	$\mu_r = 0.1$
0.6	FR	$1.3987{\cdot}10^{-2}$	$5.4617{\cdot}10^{-2}$	$4.6031{\cdot}10^{-2}$
	GW	$4.3889{\cdot}10^{-2}$	0.2187	0.4022
	Loss	$3.4677{\cdot}10^{-2}$	0.2112	0.3964
	Int	$3.4484{\cdot}10^{-2}$	0.2096	0.3846
	Mar	$8.6043{\cdot}10^{-4}$	$8.385{\cdot}10^{-3}$	$2.066{\cdot}10^{-2}$
		$P_b = 5.683{\cdot}10^{-2}$	$P_b = 6.954{\cdot}10^{-2}$	$P_b = 9.089{\cdot}10^{-2}$
0.8	FR	$2.8113{\cdot}10^{-2}$	0.1022	$7.9129{\cdot}10^{-2}$
	GW	$4.3607{\cdot}10^{-2}$	0.2079	0.3702
	Loss	$4.1717{\cdot}10^{-2}$	0.2064	0.3690
	Int	$4.1451{\cdot}10^{-2}$	0.2042	0.3521
	Mar	$1.1167{\cdot}10^{-2}$	$2.3249{\cdot}10^{-2}$	$2.3980{\cdot}10^{-2}$
		P_b=0.2469	P_b=0.2982	P_b=0.3750

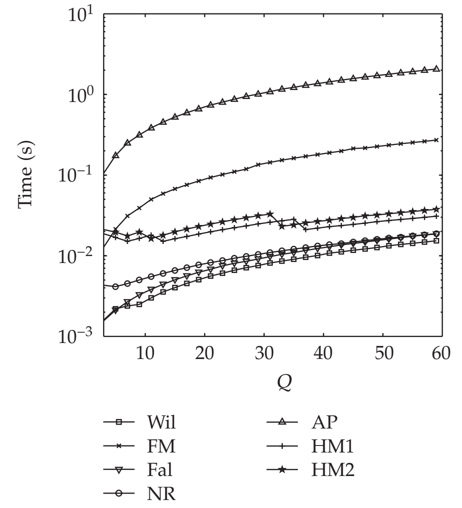


Figure 6: Computation time for the different methods deployed. $P_i = 0$.

Focusing on the configurable methods, Table 4 shows the minimum value of Q needed to obtain a relative error in the blocking probability lower than 10^{-4}. Note that the number in bold represents the best choice for each scenario. As we can see, in scenarios without impatience, generalized truncated methods (Fal, NR, AP, HM1, and HM2) also outperform finite truncated methods (Wil and FM). Among the generalized truncated methods, the best performance is obtained by HM2 and AP. While the first is specially useful for low values of μ_r, the latter is the best for high values of μ_r ($\mu_r/\mu > 10$).

In Figure 6 we show the computational cost associated to all the presented methods when $\rho = 0.8$ and $\mu_r = 0.01$ for different values of Q. Results show that the lowest computation costs are for the simplest methods (Wil, Fal, NR) followed by HM1 and HM2. Finally, and computationally speaking, the worst methods are FM and AP.

Table 4: Minimum value of Q to obtain relative errors for P_b lower than 10^{-4} when $P_i = 0$.

ρ		$\mu_r = 0.001$	$\mu_r = 0.01$	$\mu_r = 0.1$	$\mu_r = 1.0$
0.4	Wil	9	10	10	10
	FM	9	9	7	4
	Fal	11	9	7	4
	NR	7	7	5	4
	AP	10	7	**4**	**1**
	HM1	7	6	5	3
	HM2	**3**	**5**	**4**	3
		$P_b = 5.633 \cdot 10^{-3}$	$P_b = 6.442 \cdot 10^{-3}$	$P_b = 7.998 \cdot 10^{-3}$	$P_b = 8.696 \cdot 10^{-3}$
0.6	Wil	24	19	18	17
	FM	23	16	11	5
	Fal	25	16	11	7
	NR	18	13	9	6
	AP	22	13	**6**	**1**
	HM1	15	10	7	4
	HM2	**5**	**9**	7	4
		$P_b = 5.683 \cdot 10^{-2}$	$P_b = 6.954 \cdot 10^{-2}$	$P_b = 9.089 \cdot 10^{-2}$	$P_b = 9.979 \cdot 10^{-2}$
0.8	Wil	74	45	40	39
	FM	68	34	17	21
	Fal	70	35	23	14
	NR	53	27	17	10
	AP	57	24	**9**	**1**
	HM1	41	21	13	7
	HM2	**36**	**20**	13	7
		$P_b = 0.2469$	$P_b = 0.2982$	$P_b = 0.3750$	$P_b = 0.4043$

6. Conclusions

We proposed and compared two novel methods that effectively obtain the value of typical performance parameters in retrial systems with user impatience. As there are not closed-form solutions to these systems when there are more than two servers, approximate methods are required to solve such systems. The proposed methods are improvements of the method proposed by Neuts and Rao in [11] being also based on the homogenization of the state space beyond a certain level. The results show the better performance of generalized truncated methods compared to finite truncated methods in terms of accuracy. Comparing the generalized truncated methods among them, we conclude that the proposed HM2 method outperforms, in almost all cases, previous proposals in terms of accuracy for a wide range of scenarios and for the performance parameters studied, with moderated computation cost growths compared to the simpler methods.

Acknowledgments

This work was supported by the Spanish Government (30% PGE) and the European Commission (70% FEDER) through Projects TSI2005-07520-C03-03 and TSI2007-66869-C02-02 and by *Cátedra Telefónica de Internet y Banda Ancha* (e-BA) from the Universidad Politécnica de Valencia. Besides, M. Jose Domenech-Benlloch was supported by the Spanish Ministry of Education and Science under Contract AP-2004-3332.

References

[1] G. Jonin and J. Sedol, “Telephone systems with repeated calls,” in *Proceedings of the 6th International Teletraffic Congress (ITC ′70)*, pp. 435.1–435.5, Munich, Germany, September 1970.

[2] P. Tran-Gia and M. Mandjes, “Modeling of customer retrial phenomenon in cellular mobile networks,” *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 8, pp. 1406–1414, 1997.

[3] E. Onur, H. Deliç, C. Ersoy, and M. U. Çağlayan, “Measurement-based replanning of cell capacities in GSM networks,” *Computer Networks*, vol. 39, no. 6, pp. 749–767, 2002.

[4] T. Bonald and J. W. Roberts, “Congestion at flow level and the impact of user behaviour,” *Computer Networks*, vol. 42, no. 4, pp. 521–536, 2003.

[5] B. D. Choi, Y. W. Shin, and W. C. Ahn, “Retrial queues with collision arising from unslotted *CMSA/CD* protocol,” *Queueing Systems*, vol. 11, no. 4, pp. 335–356, 1992.

[6] A. Mandelbaum, “Call centers (centres), research bibliography with abstracts,” Tech. Rep., Faculty of Industrial Engineering and Management Technion—Israel Institute of Technology, Technion City, Israel, 2004. http://ie.technion.ac.il/serveng/.

[7] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*, vol. 2 of *Johns Hopkins Series in the Mathematical Sciences*, Johns Hopkins University, Baltimore, Md, USA, 1981.

[8] L. Bright and P. G. Taylor, “Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes,” *Communications in Statistics. Stochastic Models*, vol. 11, no. 3, pp. 497–525, 1995.

[9] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, ASA-SIAM Series on Statistics and Applied Probability 5, SIAM, Philadelphia, Pa, USA, 1999.

[10] J. R. Artalejo and M. Pozo, “Numerical calculation of the stationary distribution of the main multiserver retrial queue,” *Annals of Operations Research*, vol. 116, no. 1–4, pp. 41–56, 2002.

[11] M. F. Neuts and B. M. Rao, “Numerical investigation of a multiserver retrial model,” *Queueing Systems*, vol. 7, no. 2, pp. 169–189, 1990.

[12] G. Falin and J. Templeton, *Retrial Queues*, Chapman & Hall/CRC, Boca Raton, Fla, USA, 1997.

[13] B. S. Greenberg and R. W. Wolff, “An upper bound on the performance of queues with returning customers,” *Journal of Applied Probability*, vol. 24, no. 2, pp. 466–475, 1987.

[14] R. I. Wilkinson, “Theories for toll traffic engineering in the USA,” *The Bell System Technical Journal*, vol. 35, no. 2, pp. 421–514, 1956.

[15] S. N. Stepanov, “Markov models with retrials: the calculation of stationary performance measures based on the concept of truncation,” *Mathematical and Computer Modelling*, vol. 30, no. 3-4, pp. 207–228, 1999.

[16] A. A. Fredericks and G. A. Reisner, “Approximations to stochastic service systems, with an application to a retrial model,” *The Bell System Technical Journal*, vol. 58, no. 3, pp. 557–576, 1979.

[17] M. A. Marsan, G. de Carolis, E. Leonardi, R. Lo Cigno, and M. Meo, “Efficient estimation of call blocking probabilities in cellular mobile telephony networks with customer retrials,” *IEEE Journal on Selected Areas in Communications*, vol. 19, no. 2, pp. 332–346, 2001.

[18] M. J. Doménech-Benlloch, J. M. Giménez-Guzmán, J. Martínez-Bauset, and V. Casares-Giner, “Efficient and accurate methodology for solving multiserver retrial systems,” *Electronics Letters*, vol. 41, no. 17, pp. 967–969, 2005.

[19] G. I. Falin, “Calculation of probability characteristics of a multiline system with repeat calls,” *Moscow University Computational Mathematics and Cybernetics*, no. 1, pp. 43–49, 1983.

[20] V. V. Anisimov and J. R. Artalejo, “Approximation of multiserver retrial queues by means of generalized truncated models,” *Top*, vol. 10, no. 1, pp. 51–66, 2002.

[21] D. P. Gaver, P. A. Jacobs, and G. Latouche, “Finite birth-and-death models in randomly changing environments,” *Advances in Applied Probability*, vol. 16, no. 4, pp. 715–731, 1984.

[22] L. D. Servi, “Algorithmic solutions to two-dimensional birth-death processes with application to capacity planning,” *Telecommunication Systems*, vol. 21, no. 2–4, pp. 205–212, 2002.

[23] J. Ye and S.-Q. Li, “Folding algorithm: a computational method for finite QBD processes with level-dependent transitions,” *IEEE Transactions on Communications*, vol. 42, no. 234, part 1, pp. 625–639, 1994.