# A Matrix Analytic Solution of a Finite Buffer Queue with PH Distributed Customers' Impatience*

Vicent Pla[1], Vicente Casares-Giner[1], Jorge Martínez[1]

Department of Communications. Universidad Politécnica de Valencia (UPV).
ETSIT Camí de Vera s/n, 46022 Valencia, Spain.
{vpla,vcasares,jmartinez}@dcom.upv.es

**Abstract.** We study a multiserver finite buffer queue in which customers have a stochastic deadline of phase-type until the beginning of their service. The following service disciplines are considered: *FCFS* (First-Come First-Served), *LCFS* (Last-Come First-Served) and *SIRO* (Service In Random Order) along with a parameterizable probabilistic push-out mechanism. The analysis of the system is performed using a matrix analytic approach and we obtain performance measures such as probabilities of blocking, expulsion and abandonment as well as the sojourn time distribution in different system conditions.
**Keywords:** impatient customers, service discipline, phase-type, matrix-analytic.

## 1 Introduction

Queuing models in which customers abandon the system if their service has not started by a given deadline have many applications in telecommunications as well as in other disciplines (some examples can be found in [1]).

While this topic has attracted the interest of queuing theorists for a few decades, the existing literature is rather limited (see [2] and references therein). To the best of our knowledge, infinite buffer size is assumed in all queuing models considering the impatience phenomenon, except [3] and [4]. Moreover, in [3] only the relatively simple case of exponentially distributed patience time is contemplated.

Intuition seems to indicate that the service order may have an influence on the number of customers that leave the system without being served. Indeed, in [5] the authors give a characterization of the optimal scheduling discipline that minimizes the number of customers that abandon the system before receiving service, and there are some instances where the optimal policy is not the conventional *FCFS*. Zhao and Alfa [6] consider a system in which impatient customers are served on an *LCFS* basis. The analysis in [6] is approximate and the patience time is assumed to be deterministic. Doshi and Heffes [7] study a model quite related to ours, where customers may "turn bad" after some time although they do not abandon the system, i.e. bad customers are served even though

* This work has been supported by the Spanish Government (PGE, 30%) and the European Commission (FEDER, 70%) through the projects TIC2003-08272 and TEC2004-06437-C05-01.

they do not count as system goodput. The study in [7] considers *FCFS* and *LCFS* service disciplines as well as various customer rejection schemes including blocking and push-out. Notwithstanding the affinities between the model in [7] and the one of this paper, there are significant differences between them, and the analytical solution of the model in [7] cannot be applied to our model.

In this paper we analyze a queuing model that has the following characteristics: the waiting room is finite; the customers' impatience is modeled by a phase-type (PH) distribution; the service discipline can be *FCFS* (First Come First Served), *LCFS* (Last Come First Served) or *SIRO* (Service In Random Order); when a customer arrives and the buffer is full, the buffer management policy can block the newly arrived customer or push-out the head-of-line (HOL) customer in order to allocate the newly arrived one.

The contribution of our work is twofold. First, we solve the $M/M/C/K/FCFS + PH$ model using an alternative approach to that of [4] by deploying a *matrix analytic* solution. The model studied in [4] is, in a sense, more general than ours since it considers a general distribution for the patience time. Nevertheless, the family of PH-distributions is highly versatile and can be fitted to a wide range of experimental data and, from a theoretical point of view, the set of PH-distributions is dense in the set of all probability distributions on $[0, \infty)$ [8]. Secondly, and more importantly, we carry out the analysis for several non-*FCFS* disciplines in the context of customer impatience and a rather general distribution for the patience time, which have not been done before.

The remaining of the paper is structured as follows. In Section 2 the mathematical model of the system is described and its analysis is carried out in Section 3. A numerical example is presented in Section 4. Finally, Section 5 summarizes the paper and draws some conclusions.

## 2 Model Description

We first describe the analytical aspects that are common to all service disciplines and customer rejection schemes, and their specific features are subsequently addressed.

The system has a total of $C$ identical servers and a waiting line limited to $N$ positions. Customers arrive according to a Poisson process of rate $\lambda$ and their service time is exponentially distributed with rate $\mu$. Each customer has a stochastic deadline —starting on arrival— until the beginning of service, if service has not begun by his deadline the customer abandons the system and is lost. Deadlines are considered to be *independent and identically distributed* (iid) phase type (PH) random variables (*rv*) with representation $(\boldsymbol{\beta}, \boldsymbol{T})$; we denote by $m$ the number of phases in the PH distribution (see [9] for more details on PH distributions).

Throughout this paper the following notation is used: $x_i$ denotes the i-th entry of a vector $\boldsymbol{x}$; $M_{ij}$ denotes the entry on the i-th row and the j-th column of a matrix $M$; $\otimes$ denotes the Kronecker product of two matrices (example: $A \otimes B$), see [9, p. 17] for futher details; $\text{diag}\{\cdot\}$ is an operator over a vector that yields a diagonal matrix whose diagonal entries are the elements of the vector; $\boldsymbol{e}$ is a column vector of ones; $\boldsymbol{0}$ is a column vector of zeros and $\boldsymbol{I}$ is the identity matrix.

Under the above assumptions the system model fits within the category of non-homogeneous finite Quasi-Birth-and-Death (QBD) processes. Let $\{X(t) : t > 0\}$ be

the stochastic process for the system with the following two-dimensional state space $S = \big\{(l,k) : 0 \le l \le N;\ 0 \le k \le m^l - 1\big\} \bigcup \{(-1,k) : 0 \le k \le C-1\}$ that can be partitioned into *levels* as $S = \bigcup_{l=-1}^{N} L(l)$, where $L(l_0) = \{(l,k) : l = l_0;\ (l_0,k) \in S\}$. The first coordinate of a state is also referred to as *level* and the second coordinate as *phase*. Level $(-1)$ groups the states in which not all servers are busy, for this level the state phase indicates the number of customers in the system, i.e. being served. Level $l \ge 0$ groups the states in which all servers are busy and there are $l$ customers in the waiting room, In levels $l \ge 0$ the state phase encodes the current phase of the patience time distribution for each of the $l$ waiting customers. The mapping between the state phase $(k)$ and waiting customers phases is as follows. Let the $l$-tuple $(k_1, \ldots, k_l)$, $1 \le k_i \le m$, $1 \le i \le l$ denote the phases of the Markov process associated to the PH distribution of the waiting customers patience time, being $k_i$ the phase of the customer at the $i$-th position in the waiting room. Then, $k = \sum_{1 \le i \le l} k_i m^{l-i}$; in other words, the $l$-tuples are numbered in lexicographical order from $(1, \ldots, 1)$ to $(m, \ldots, m)$.

Let us denote by $\boldsymbol{\pi}$ the stationary probability vector of the process. In the same way as with states, we partition $\boldsymbol{\pi}$ by levels into subvectors $\boldsymbol{\pi}^{(l)}$, $-1 \le l \le N$, where $\boldsymbol{\pi}^{(-1)}$ has $C$ components and $\boldsymbol{\pi}^{(l)}$ $(l \ge 0)$ has $m^l$ components. State transitions are restricted to states in the same level or in two adjacent levels and consequently, the infinitesimal generator $\boldsymbol{Q}$ of the process has a block tridiagonal structure,

$$\boldsymbol{Q} = \left[\begin{array}{c|cc} \boldsymbol{A}_1^{(-1)} & \boldsymbol{A}_0^{(-1)} & \boldsymbol{0} \ \cdots \\ \hline \boldsymbol{A}_2^{(0)} & & \\ \boldsymbol{0} & & \boldsymbol{Q}_p \\ \vdots & & \end{array}\right], \quad \text{where} \quad \boldsymbol{Q}_p = \begin{bmatrix} \boldsymbol{A}_1^{(0)} & \boldsymbol{A}_0^{(0)} & & \\ \boldsymbol{A}_2^{(1)} & \boldsymbol{A}_1^{(1)} & \boldsymbol{A}_0^{(1)} & \\ & \ddots & \ddots & \ddots \\ & & \boldsymbol{A}_2^{(N)} & \boldsymbol{A}_1^{(N)} \end{bmatrix}.$$

Block matrices which are not in $\boldsymbol{Q}_p$ involve the boundary level $(-1)$ and thus they do not conform to the general construction that will be given for the rest of matrices. Matrix entries for these particular cases, which are independent of the service discipline, are as follows:

$$\boldsymbol{A}_1^{(-1)} = \begin{bmatrix} * & \lambda & & & \\ \mu & * & \lambda & & \\ & 2\mu & * & \lambda & \\ & & \ddots & \ddots & \ddots \\ & & & (C-1)\mu & * \end{bmatrix}, \qquad \boldsymbol{A}_0^{(-1)} = \big[0 \cdots 0\ \lambda\big]^t$$

$$\boldsymbol{A}_2^{(0)} = \big[0 \cdots 0\ C\mu\big]$$

The diagonal entries of $\boldsymbol{A}_1^{(-1)}$, which are represented by asterisks for the ease of display, are such that the corresponding rows of $\boldsymbol{Q}$ sum to zero, i.e. $\boldsymbol{A}_1^{(-1)}\boldsymbol{e} + \boldsymbol{A}_0^{(-1)}\boldsymbol{e} = \boldsymbol{0}$.

### 2.1 FCFS Discipline

The matrices $\boldsymbol{A}_0^{(l)}$ correspond to transitions from $L(l)$ to $L(l+1)$, $0 \leq l < N$. These transitions represent the arrival of a customer that will occupy the $(l+1)$-th position of the waiting room. The PH distribution for the patience time of the arriving customer will begin at its $i$-th phase with probability $\beta_i$. It can be easily seen that $\boldsymbol{A}_0^{(l)} = \boldsymbol{I}_{m^l} \otimes \lambda\boldsymbol{\beta}$.

The matrices $\boldsymbol{A}_2^{(l)}$ correspond to transitions from $L(l)$ to $L(l-1)$, $0 < l \leq N$. These transitions represent the departure of a customer from the system which may be due to either a customer abandoning the waiting line (because his deadline has expired) or to a service completion. The former type of transition will be represented by matrix $\boldsymbol{U}_1^{(l)}$ and the latter by matrix $\boldsymbol{U}_2^{(l)}$. Then, $\boldsymbol{A}_2^{(l)} = \boldsymbol{U}_1^{(l)} + \boldsymbol{U}_2^{(l)}$ where

$$\boldsymbol{U}_1^{(l)} = \begin{cases} \boldsymbol{\tau}, & l = 1 \\ \boldsymbol{\tau} \otimes \boldsymbol{I}_{m^{l-1}} + \boldsymbol{I}_m \otimes \boldsymbol{U}_1^{(l-1)}, & 1 < l \leq N \end{cases}; \qquad \boldsymbol{U}_2^{(l)} = C\mu \boldsymbol{e}_m \otimes \boldsymbol{I}_{m^{l-1}}$$

being $\boldsymbol{\tau} = -\boldsymbol{T}\boldsymbol{e}$.

The matrices $\boldsymbol{A}_1^{(l)}$ correspond to transitions between states within $L(l)$. These transitions represent phase changes in the PH processes associated to waiting customers. The expression for this matrix is first given ignoring elements on its main diagonal and next they will be computed using the fact that the rows of $\boldsymbol{Q}$ must sum to zero. For the sake of clarity we introduce the set of matrices $\boldsymbol{D}^{(l)}$ whose entries are equal to the entries of $\boldsymbol{A}_1^{(l)}$, except those on their main diagonal. Now it can be written that

$$\boldsymbol{D}^{(l)} = \begin{cases} \boldsymbol{T}, & l = 1 \\ \boldsymbol{T} \otimes \boldsymbol{I}_{m^{l-1}} + \boldsymbol{I}_m \otimes \boldsymbol{D}^{(l-1)}, & 1 < l \leq N \end{cases} \tag{1}$$

Note that in the expression of $\boldsymbol{D}^{(N)}$ it is assumed that customers arriving while the system is at level $N$, i.e. when the buffer is full, are blocked. Then $\boldsymbol{A}_1^{(l)}$ is given by

$$\boldsymbol{A}_1^{(l)} = \boldsymbol{D}^{(l)} - \operatorname{diag}\left\{\boldsymbol{A}_2^{(l)}\boldsymbol{e} + \boldsymbol{D}^{(l)}\boldsymbol{e} + \boldsymbol{A}_0^{(l)}\boldsymbol{e}\right\}$$

which can be further simplified to

$$\boldsymbol{A}_1^{(l)} = \begin{cases} \boldsymbol{D}^{(l)} - (C\mu + \lambda)\boldsymbol{I}_{m^l}, & l < N \\ \boldsymbol{D}^{(N)} - C\mu\boldsymbol{I}_{m^N}, & l = N \end{cases} \tag{2}$$

by virtue of Proposition 1, that will be proved after the following lemma.

**Lemma 1** $(A \otimes B)\,\boldsymbol{e} = (A\boldsymbol{e}) \otimes (B\boldsymbol{e})$

**Proposition 1** *The following equalities hold for* $1 \leq l \leq N$

$$\operatorname{diag}\left\{\boldsymbol{A}_0^{(l)}\boldsymbol{e}\right\} = \lambda\boldsymbol{I}_{m^l}\ ; \qquad \operatorname{diag}\left\{\boldsymbol{A}_2^{(l)}\boldsymbol{e} + \boldsymbol{D}^{(l)}\boldsymbol{e}\right\} = C\mu\boldsymbol{I}_{m^l}$$

*Proof.* The first equality follows immediately by applying the Lemma to $\boldsymbol{A}_0^{(l)} = \boldsymbol{I}_{m^l} \otimes \lambda\boldsymbol{\beta}$ and noting that $\boldsymbol{\beta e} = 1$ and $\text{diag}\{\boldsymbol{e}\} = \boldsymbol{I}$.

To prove the second equality we first observe that $\boldsymbol{A}_2^{(l)}\boldsymbol{e} + \boldsymbol{D}^{(l)}\boldsymbol{e} = \boldsymbol{U}_1^{(l)}\boldsymbol{e} + \boldsymbol{U}_2^{(l)}\boldsymbol{e} + \boldsymbol{D}^{(l)}\boldsymbol{e}$ and by applying the Lemma to $\boldsymbol{U}_2^{(l)} = C\mu\boldsymbol{e}_m \otimes \boldsymbol{I}_{m^{l-1}}$ it is easily seen that

$$\boldsymbol{U}_2^{(l)}\boldsymbol{e} = C\mu\boldsymbol{e}_{m^l}. \tag{3}$$

On the other hand

$$\begin{aligned}\boldsymbol{U}_1^{(l)}\boldsymbol{e}_{m^{l-1}} + \boldsymbol{D}^{(l)}\boldsymbol{e}_{m^l} &= (\boldsymbol{\tau} \otimes \boldsymbol{I}_{m^{l-1}} + \boldsymbol{I}_m \otimes \boldsymbol{U}_1^{(l-1)})\boldsymbol{e} + (\boldsymbol{T} \otimes \boldsymbol{I}_{m^{l-1}} + \boldsymbol{I}_m \otimes \boldsymbol{D}^{(l-1)})\boldsymbol{e} \\ &= (\boldsymbol{\tau} \otimes \boldsymbol{e}_{m^{l-1}} + \boldsymbol{e}_m \otimes (\boldsymbol{U}_1^{(l-1)}\boldsymbol{e}_{m^{l-2}})) + (\boldsymbol{T}\boldsymbol{e}_m \otimes \boldsymbol{e}_{m^{l-1}} + \boldsymbol{e}_m \otimes (\boldsymbol{D}^{(l-1)}\boldsymbol{e}_{m^{l-1}})) \\ &= \boldsymbol{e}_m \otimes (\boldsymbol{U}_1^{(l-1)}\boldsymbol{e}_{m^{l-2}} + \boldsymbol{D}^{(l-1)}\boldsymbol{e}_{m^{l-1}})\end{aligned} \tag{4}$$

and by recursive application of (4) we obtain

$$\begin{aligned}\boldsymbol{U}_1^{(l)}\boldsymbol{e}_{m^{l-1}} + \boldsymbol{D}^{(l)}\boldsymbol{e}_{m^l} &= \boldsymbol{e}_m \otimes (\boldsymbol{U}_1^{(l-1)}\boldsymbol{e}_{m^{l-2}} + \boldsymbol{D}^{(l-1)}\boldsymbol{e}_{m^{l-1}}) \\ &\vdots \\ &= \boldsymbol{e}_{m^{l-1}} \otimes (\boldsymbol{\tau} + \boldsymbol{T}\boldsymbol{e}_m) = \boldsymbol{e}_{m^{l-1}} \otimes \boldsymbol{0} = \boldsymbol{0}\end{aligned} \tag{5}$$

Hence, from (3) and (5) it follows that $\boldsymbol{A}_2^{(l)}\boldsymbol{e} + \boldsymbol{D}^{(l)}\boldsymbol{e} = C\mu\boldsymbol{e}_{m^l}$, and taking the $\text{diag}\{\cdot\}$ operator on both sides of this equality yields the desired result.

### 2.2 *LCFS* Discipline

Using the *LCFS* discipline only affects the selection of the queued customer that will start service, and therefore will be removed from the queue, after a service completion. Thus the expression for $\boldsymbol{U}_2^{(l)}$ must be modified in the following manner: $\boldsymbol{U}_2^{(l)} = \boldsymbol{I}_{m^{l-1}} \otimes C\mu\boldsymbol{e}_m$. The rest remains unchanged since under the new expression for $\boldsymbol{U}_2^{(l)}$ it can be proved, in much the same way as before, that (3) holds .

### 2.3 *SIRO* Discipline

By the same reasoning as in *LCFS* discipline, only $\boldsymbol{U}_2^{(l)}$ changes and it can be readily shown to be given by

$$\boldsymbol{U}_2^{(1)} = C\mu\boldsymbol{e}_m \quad \text{and} \qquad \boldsymbol{U}_2^{(l)} = \frac{C\mu}{l}\boldsymbol{e}_m \otimes \boldsymbol{I}_{m^{l-1}} + \boldsymbol{I}_m \otimes \frac{(l-1)}{l}\boldsymbol{U}_2^{(l-1)} \quad 1 < l < N$$

In this case Eq. (3) can be easily proved by induction on $l$.

### 2.4 Buffer management scheme

So far we have assumed that arriving customers to a full buffer are blocked. Here we consider a more general buffer management scheme in which a customer arriving when

the buffer is full either is blocked or it pushes-out the HOL waiting customer. The choice between these two options is done randomly and independently for each customer. The random component of the algorithm is tuned by parameter $q$ which represents the probability that a customer that finds a full queue upon arrival will push-out the HOL customer; therefore, a customer finding a full queue will be blocked with probability $(1-q)$.

The arrival of a customer that finds a full buffer and pushes-out the HOL customer is modeled as a transition within level $L(N)$ and thus only the values entries of matrix $\boldsymbol{D}^{(N)}$ are modified as follows: $\boldsymbol{D}^{(N)} \longleftarrow \boldsymbol{D}^{(N)} + \boldsymbol{e}_m \otimes \boldsymbol{I}_{m^{N-1}} \otimes \lambda\boldsymbol{\beta}$. Hence, if the push-out and blocking schemes are probabilistically combined together with probabilities $q$ and $(1-q)$, respectively, we have that

$$\boldsymbol{D}^{(l)} = \begin{cases} \boldsymbol{T}, & l = 1 \\ \boldsymbol{T} \otimes \boldsymbol{I}_{m^{l-1}} + \boldsymbol{I}_m \otimes \boldsymbol{D}^{(l-1)}, & 1 < l < N \\ \boldsymbol{T} \otimes \boldsymbol{I}_{m^{N-1}} + \boldsymbol{I}_m \otimes \boldsymbol{D}^{(N-1)} + q\,(\boldsymbol{e}_m \otimes \boldsymbol{I}_{m^{N-1}} \otimes \lambda\boldsymbol{\beta}), & l = N \end{cases} \tag{6}$$

Finally, by the same method we have used to obtain (2) it follows that

$$\boldsymbol{A}_1^{(l)} = \begin{cases} \boldsymbol{D}^{(l)} - (C\mu + \lambda)\boldsymbol{I}_{m^l}, & l < N \\ \boldsymbol{D}^{(N)} - (C\mu + q\lambda)\boldsymbol{I}_{m^N}, & l = N \end{cases} \tag{7}$$

Note that when $q = 0$, (7) reduces to (2).

## 3 Model Analysis

In this section we describe the method to calculate the stationary state probabilities ($\boldsymbol{\pi}$) of the model. From these, performance evaluation measures are derived.

The stationary probabilities ($\boldsymbol{\pi}$) are obtained as the solution to the set of simultaneous linear equations $\boldsymbol{\pi Q} = \boldsymbol{0}$, $\boldsymbol{\pi e} = 1$. Being $\boldsymbol{Q}$ a finite dimension matrix the above system could be solved by standard linear algebra methods. However, as the size of $\boldsymbol{Q}$ may be very large, it is advisable to use more specific algorithms that take advantage of the structure ($\boldsymbol{Q}$ is block-tridiagonal) and the nature of the problem ($\boldsymbol{Q}$ is an infinitesimal generator). We used the *Linear Level Reduction* algorithm [9, 10], which can solve level-dependent finite QBDs.

1: $U \leftarrow \boldsymbol{A}_1^{(N)}$
2: $R^{(N)} \leftarrow \boldsymbol{A}_0^{(N-1)}\,(-U)^{-1}$
3: **for** $l = N-1, N-2, \ldots, 0, -1$ **do**
4: $\quad U \leftarrow \boldsymbol{A}_1^{(l)} + R^{(l+1)}\boldsymbol{A}_2^{(l+1)}$
5: $\quad R^{(l)} \leftarrow \boldsymbol{A}_0^{(l-1)}\,(-U)^{-1}$
6: **end for**
7: **solve** $\boldsymbol{\pi}^{(-1)}$ **from** $\{\boldsymbol{\pi}^{(-1)}U = \boldsymbol{0};\ \boldsymbol{\pi}^{(-1)}\boldsymbol{e} = 1\}$
8: **for** $l = 0, 1, \ldots, N$ **do**
9: $\quad \boldsymbol{\pi}^{(l)} = \boldsymbol{\pi}^{(l-1)}R^{(l)}$
10: **end for**

### 3.1 Distribution of the Number of Customers

Let $p_k (0 \le k \le N + C)$ denote the probability that $k$ customers are in the system, then we have that

$$p_k = \begin{cases} \pi_k^{-1}, & k = 0, \ldots, C-1 \\ \boldsymbol{\pi}^{(k-C)}\boldsymbol{e}, & k = C, \ldots, C+N \end{cases}$$

### 3.2 Probabilities of Blocking, Expulsion and Reneging

Since arrivals are Poisson, by *PASTA* property [11] we have that the probability that an arriving customer sees the buffer full is $p_{C+N}$. Therefore, the blocking probability is given by $P_b = (1-q)p_{C+N} = (1-q)\boldsymbol{\pi}^{(N)}\boldsymbol{e}$ and the expulsion probability is given by $P_e = qp_{C+N} = q\boldsymbol{\pi}^{(N)}\boldsymbol{e}$. The probability of reneging ($P_r$) is measured by taking the average number of customers who renege divided by the average number of customers that arrived over a sufficiently long period, say $t_0$, $(\sum_{l=1}^{N} \boldsymbol{\pi}^{(l)}\boldsymbol{U}_1^{(l)} t_0\boldsymbol{e} + o(t_0))/(\lambda t_0 + o(t_0))$ and letting $t_0 \to \infty$ we obtain that $P_r = \sum_{l=1}^{N} \boldsymbol{\pi}^{(l)}\boldsymbol{U}_1^{(l)}\boldsymbol{e}/\lambda$.

### 3.3 Sojourn Time in Congestion Condition and Blocking Condition

We say that the system is congested if an arriving customer has to wait, i.e. the system is in one of the states in $\bigcup_{l=0}^{N} L(l)$. Let $T_c$ denote the sojourn time in the congestion condition *rv* . Similarly, let us define the blocking condition as the state in which the number of costumers in the system is at its maximum $(C+N)$, so that if a new customer arrives it will be blocked or a waiting customer will be pushed-out. Let $T_b$ be the sojourn time in the blocking condition *rv*. Next we obtain the distribution of these *rv* and derive their mean values.

A congestion condition period starts when the system enters level $L(0)$ and lasts until its first visit to level $L(-1)$. During this period the system will visit states in $\bigcup_{l=0}^{N} L(l)$ whose residence times are all exponential. Thus it may be concluded that the distribution of $T_c$ is phase-type and it is easy to check that its representation is $PH(\boldsymbol{\beta}^{(c)}, \boldsymbol{T}^{(c)})$ where $\boldsymbol{\beta}^{(c)} = \begin{bmatrix} 1\,0\,0\ldots 0 \end{bmatrix}$ and $\boldsymbol{T}^{(c)} = \boldsymbol{Q}_p$.

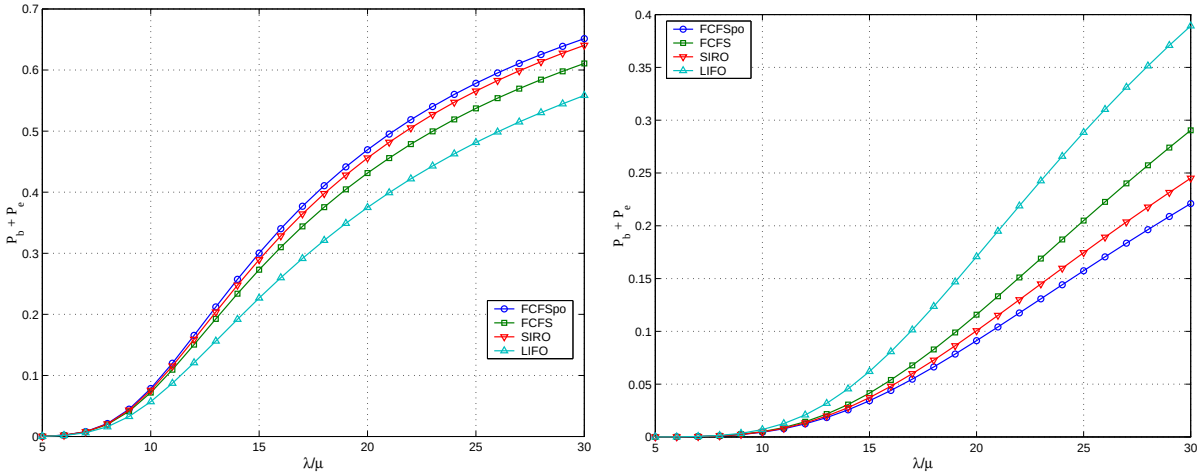
In order to obtain the mean value of $T_c$ we will use a probabilistic argument instead of using its distribution which would entail inverting matrix $\boldsymbol{T}^{(c)}$. This reasoning is based on the observation that for an infinitely large time period, say $t_0$, it holds that, the mean sojourn time per visit to a set of states equals the total sojourn time in that set of states divided by the number of visits. Hence, we can write

$$\overline{T}_c = \lim_{t_0\to\infty} \frac{\left(\sum_{k=C}^{C+N} p_k\right) t_0 + o(t_0)}{\lambda p_{C-1} t_0 + o(t_0)} = \frac{1}{\lambda p_{C-1}} \sum_{k=C}^{C+N} p_k.$$

In the same manner we can see that $T_b$ is phase-type and its representation is $PH(\boldsymbol{\beta}^{(b)}, \boldsymbol{T}^{(b)})$ where $\boldsymbol{\beta}^{(b)} = \boldsymbol{\pi}^{(N-1)}\boldsymbol{A}_0^{(N-1)}/(\boldsymbol{\pi}^{(N-1)}\boldsymbol{A}_0^{(N-1)}\boldsymbol{e})$ and $\boldsymbol{T}^{(b)} = \boldsymbol{A}_1^{(N)}$. Hence [9, Eq. (2.13)] $\overline{T}_b = \boldsymbol{\beta}^{(b)}\left(-\boldsymbol{T}^{(b)}\right)^{-1}\boldsymbol{e}$, which noting that $\boldsymbol{A}_0^{(N-1)}\boldsymbol{e} = \lambda\boldsymbol{e}$ and $\boldsymbol{\pi}^{(N-1)}\boldsymbol{A}_0^{(N-1)} + \boldsymbol{\pi}^{(N)}\boldsymbol{A}_1^{(N)} = \boldsymbol{0}$ we can rewrite as $\overline{T}_b = \boldsymbol{\pi}^{(N)}\boldsymbol{e}/(\lambda\boldsymbol{\pi}^{(N-1)}\boldsymbol{e})$.

## 4 Numerical Example

In this section we present a numerical example to illustrate the analysis carried out in the previous sections. In this example the system parameters are: $C = 10$, $N = 5$. Four combinations of service discipline and buffer management schemes are considered: *FCFS* (*FCFS* and $q = 0$), *FCFSpo* (*FCFS* and $q = 1$), *SIRO* (*SIRO* and $q = 0.5$) and *LCFS*




(a) Erlang (b) Hyper exponential

**Fig. 1.** Blocking probability plus expulsion probability.

(*LCFS* and $q = 0$). Arrival rate ($\lambda$) and transition rates of the PH distribution ($\boldsymbol{T}$) are normalized with respect to $\mu$. Two different instances of the patience time are examined, one whose hazard rate function[1] is increasing (e.g. Erlang) and the other decreasing (e.g. Hyper exponential). Their PH representations are, respectively,

$$\boldsymbol{\beta} = \begin{bmatrix} 1\,0\,0 \end{bmatrix}, \boldsymbol{T} = \begin{bmatrix} -3 & 3 & 0 \\ 0 & -3 & 3 \\ 0 & 0 & -3 \end{bmatrix}$$

and $\boldsymbol{\beta} = 1/3 \begin{bmatrix} 1\,1\,1 \end{bmatrix}$, $\boldsymbol{T} = -56/150\,\mathrm{diag}\left\{ \begin{bmatrix} 50\,10\,1 \end{bmatrix} \right\}$.

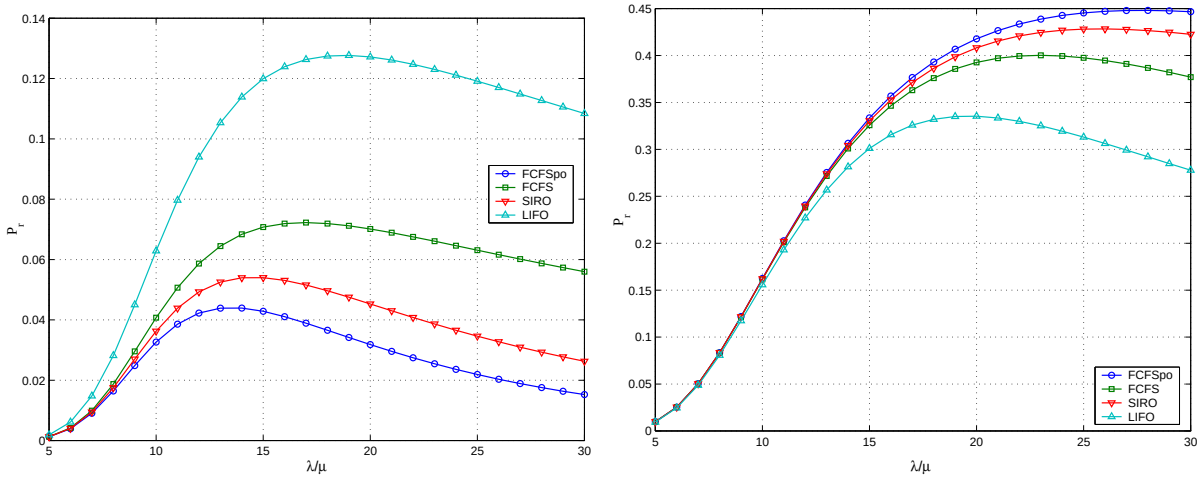
Figure 1 represents the sum of the probabilities of blocking and expulsion as a function of the offered traffic, and Fig. 2 does the same for the probability of reneging.

The total computational cost incurred in order to find the state probabilities is dominated by the cost of inverting the matrix $\boldsymbol{A}_1^{(N)}$, which is of size $m^N$. Obviously, increasing the buffer size $N$ and the number of phases $m$ can make the numerical solution simply unfeasible, which of course is a limitation of our model. Nevertheless, in such situation our model can still provide a better approximation than the simple first order exponential approximation. An example comparing these approximations is shown in figures 3 and 4. In these two examples we consider a 5-phase PH distribution for which the exact performance measures are computed: in Fig. 3 we used a hyper exponential distribution and in Fig. 4 an Erlang distribution . Then, the original PH distribution is approximated (using moment matching) by a 1-phase PH (exponential) and a 2-phase PH. As expected, the 2-phase approximation offers better accuracy than the exponential one.

## 5 Conclusion

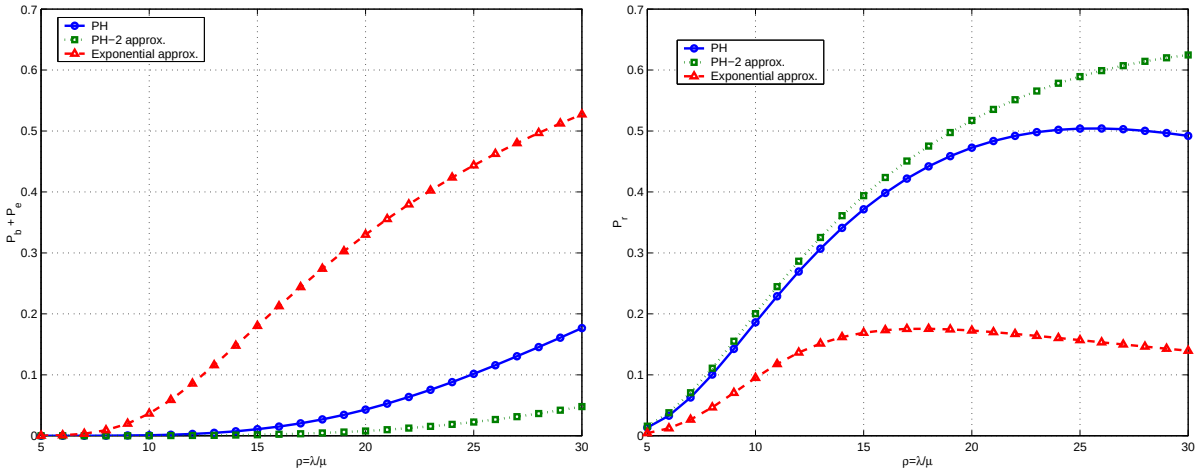
In this paper, we developed a stochastic model for a multiserver finite buffer queue with impatient customers where the patience time is modeled by a *rv* of phase-type. Further-

[1] The hazard rate function is also known as the failure rate function. Let $f(x)$ and $F(x)$ be the *pdf* and the *CDF* of a *rv* $X$, the hazard rate function $h(x)$ of $X$ is defined as $h(x) = f(x)/(1 - F(x))$.




(a) Erlang (b) Hyper exponential

**Fig. 2.** Reneging probability.

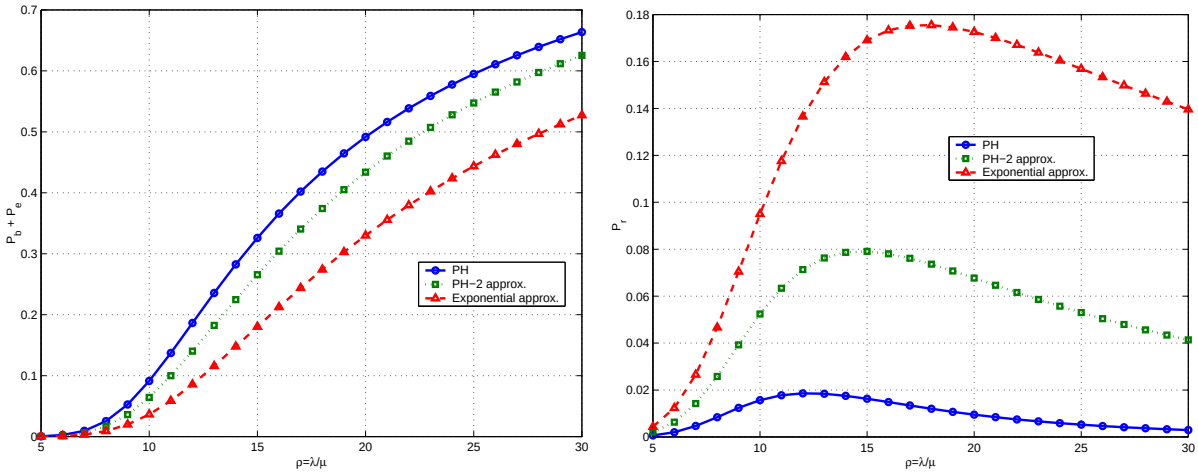



(a) Blocking probability plus expulsion probability (b) Reneging probability

**Fig. 3.** Approximation of a hyper exponential distribution. LIFO discipline.

more, the model considers different service disciplines (*FCFS*, *LCFS* and *SIRO*) along with a probabilistically weighted buffer management scheme that combines two modes of operation: customers who arrive when system is full are blocked or push-out the HOL customer. The most significant achievent of this paper is to develop a model for the performance analysis of a queue with impatient customers under non-*FCFS* service disciplines. Secondly, our model considers the finite buffer case and a fairly general distribution of the customer patience.

The analytical model is constructed and analyzed using matrix analytic methods and we obtain expressions for the performance evaluation of the system. The complexity of the computations in the analysis of our model is mainly dependent on buffer size and the number of phases of the customer patience distribution. While this complexity can make the numerical analysis unfeasible in some cases, in these cases our model can still




(a) Blocking probability plus expulsion probability (b) Reneging probability

**Fig. 4.** Approximation of an Erlang distribution. FIFOpo discipline.

provide a better approximation than the one obtained using an exponentially distributed customer patience.

## References


1. Boxma, O., de Waal, P.: Multiserver queues with impatient customers. In: Proceedings of ITC 14, Elsevier Science (1994) 743–756
2. Brandt, A., Brandt, M.: Asymptotic results and a markovian approximation for the $M(n)/M(n)/s + GI$ system. Queueing Systems **41** (2002) 73–94
3. Ancker, C.J., Gafarian, A.: Some queueing problems with balking and reneging. Operations Research **11** (1963) 88–100
4. Movaghar, A.: On queueing with customer impatience until the beginning of service. Queuing Systems **29** (1998) 337–350.
5. Towsley, D., Panwar, S.: Optimality of the stochastic earliest deadline policy for the $G/M/c$ queue serving customers with deadlines. In: Proceedins of the 2nd ORSA Telecommunications Conference. (1992).
6. Zhao, Y.Q., Alfa, A.S.: Performance analysis of a telephone system with both patient and impatient customers. Telecommunication Systems **4** (1995) 201–215.
7. Doshi, B.T., Heffes, H.: Overload performance of several processor queueing disciplines for the $M/M/1$ queue. IEEE Transactions on Communications **COM-34** (1986) 538–546.
8. Neuts, M.: Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach. The Johns Hopkins University Press (1981)
9. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling. ASA-SIAM (1999)
10. Gaver, D., Jacobs, P., Latouche, G.: Finite birth-and-death models in randomly changing environments. Advances in Applied Probability **16** (1984) 715–731
11. Wolff, R.W.: Poisson arrivals see time averages. Operation Research **30** (1982) 223–231