ON A MULTISERVER FINITE BUFFER QUEUE WITH IMPATIENT CUSTOMERS*

Vicent Pla, Vicente Casares-Giner and Jorge Martínez
Universidad Politécnica de Valencia (UPV)
Department of Communications
ETSIT, Camí de Vera s/n, 46022 Valencia, Spain.
e-mail: (vpla,vcasares,jmartinez)@dcom.upv.es
Telephone: +34 963879733, Fax: +34 963877309

ABSTRACT

In this paper we study a multiserver finite buffer queue in which customers have an 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. Although this model has many applications in the general area of telecommunications as well as in other disciplines, it was primarily motivated by the study of handoff priority schemes in mobile cellular networks. 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.

1. Introduction

Queuing models in which customers leave 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]). In cellular systems the impatience phenomenon is doubly present: in addition to callers' impatience (causing the reneging of waiting new calls), when queuing of handoff attempts is employed as a prioritization scheme [2–5], handoff attempts can stay in the waiting room until the mobile moves out of the handoff area.

While this topic has been attracting the interest of queuing theorists for a few decades, the existing literature is rather moderate (see [6] and references therein). To the best of our knowledge, except [7] and [8], infinite buffer size is assumed in all the queuing models considering the impatience phenomenon, but in [7] only the rather 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 customer that leave the system without being served. Indeed, in [9] the authors give a characterization of the optimal scheduling discipline, that minimizes the number of customer that abandon the system before receiving service. There are some instances where the optimal policy is not the conventional *FCFS*. Zhao and Alfa consider a system in which impatient customers are served on an *LCFS* basis. The analysis in [10] is approximate and the patience time is assumed to be deterministic. Doshi and Heffes [11] studie a model quite related to ours, where customers may "turn bad" after some time although they do not abandon the system, i.e. bad customer are served even though they are not accounted for the system goodput. The paper contemplates *FCFS* and *LCFS* service disciplines as well as various customer rejection schemes including blocking and push-out. Notwithstanding the affinities between the model in [11] and the one of this paper, there are significant differences between them and, the analytical solution of the model in [11] cannot be applied to our model.

While there are numerous papers dealing with the analysis of cellular systems that incorporate impatience in their models (see [2, 12, 13] to name a few) the vast majority of them assume exponentially distributed patience and *FCFS* discipline. By contrast, the results in [14] and [15] suggest a non-exponential distribution for the patience time. In [16–18] the effect of non-exponential patience time has been explored but the assumption of *FCFS* discipline is maintained. Tekinay and Jabbari in [3] study via simulations the performance of non-preemptive priority queueing for handover calls, where, if a channel is released, then the handover call in the queue that has the lowest received signal strength gets served. It is shown that the proposed scheme, which is called measurement-based priority scheme (MBPS), outperforms *FCFS* queueing scheme under all traffic conditions [3]. However, the study in [3] does not take into account the dynamics of user motion. In [19], Ebersman and Tonguz investigate the dynamic queueing of handover calls using a *signal prediction priority queueing* (SPPQ) discipline, where the order of handover calls is

*This work has been supported by the *Spanish Ministry of Science and Technology* under projects TIC2000-1041-C03-02 and TIC2001-0956-C04-04.

not only based on the received signal strength (RSS), but also on the rate of change of RSS. The performance (i.e., new call blocking probability, forced termination probability, etc.) of a cellular system that uses SPPQ scheme is evaluated via extensive computer simulations. Finally, in [20] a queueing discipline with two classes of priority for handoff calls is employed. Two queues, first priority and second priority, are employed for the two priority classes of handoff calls. Arriving handoff requests are classified into first-priority or second-priority, subsequently second-priority can become a first-priority handoff call and join the first-priority handoff queue. Customers in the first-priority queue are served before those in the second-priority queue and within each queue customers are served following a *FCFS* order.

In this paper we analyze a queuing model in which: 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, push-out the head-of-line (HOL) customer in order to allocate the newly arrived one, or use a probabilistic combination of the two.

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 [8] by deploying a *matrix analytic* solution. The model studied in [8] is indeed 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 $[0,\infty)$ [21]. 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. The stationary state probabilities for the model are derived in Section 3 and from these, performance evaluation measures are derived in Section 4. A numerical example is presented in Section 5. Finally, Section 6 summarizes the paper and draws some conclusions.

2. Model Description

2.1. Notation and definitions

C Is the number of identical servers of the system.

N Is the buffer size.

λ Is the rate of the arrival Poisson process.

μ^{-1} Is the mean of the exponentially distributed service time.

x_i Is the i-th entry of a vector x.

M_{ij} Is the entry on the i-th row and the j-th column of a matrix M.

$\otimes$ Is the symbol for the Kronecker product of two matrices (example: $A \otimes B$). See [22, p. 17] for futher details.

$\mathrm{diag}\{\cdot\}$ Is an operator over a vector that yields a diagonal matrix whose diagonal entries are the elements of the vector.

e Is a column vector of all ones.

0 Is a column vector of all zeros.

I Is the identity matrix.

If not explicitly indicated e, 0 and I do have the appropriate dimension. Alternatively, in some expressions they might be subscripted to avoid ambiguity or to ease readability.

When the model is applied to a cellular scenario with handoff queuing the arrival rate λ will only refer to handoff requests. New call requests consume cell resources as well and they will have to be considered, This can be done with minor and straightforward modifications to the model. Furthermore, some kind of admission control scheme such as the *guard channel* [2, 23] could also incorporated into the model. The channel holding time, which is the minimum of the cell residence time and the unencumbered call duration, will map to the service time in our model. A handoff request can be waiting until the mobile terminal (MT) leaves the coverage area of base station it was connected to, i.e. until the power received by the MT falls below the *receiver threshold* [4]. In this case the model should also capture the event that a handoff call could finish its call while waiting in the queue. Hence the deadline of queued customers is given by either the instant when the MT abandons the cell or the instant when the call voluntarily finished, whichever happens first. Although these two events are modeled as an abandonment from the queue, their implications in the perceived *Quality of Service* (QoS) by the user are completely different and thus should be accounted separately.

2.2. General Considerations

We first describe those analytical aspects that are common to all service disciplines and customer rejection schemes considered in this paper while their specific features are addressed subsequently.

Each customer has a stochastic deadline starting on arrival and until the beginning of service, if service has not begun by his deadline the customer abandons the system and is lost. In this paper we consider that deadlines are *independent and identically distributed (iid)* phase type (PH) random variables (*rv*). with representation (β, T); we denote by m the number of phases in the PH distribution (see [22] for more details on phase type distributions).

Under the above assumptions the 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

$$S=\Big\{(l,k): 0\le l\le N;\, 0\le k\le m^l-1\Big\} \bigcup\Big\{(-1,k): 0\le k\le C-1\Big\}$$

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 groups the states in which all servers are busy and there are l customers in the waiting room, while for this level the state phase encodes the current phase of the patience time distribution for each of the l waiting customers. The mapping between state phase (k) and waiting customer 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 we denote by π the stationary probability vector of the process. In the same way as with states, we partition π by levels into subvectors $\pi^{(l)}$, $-1\le l\le N$, where $\pi^{(-1)}$ has C components and $\pi^{(l)}$ ($l\ge 0$) has m^l components. Transitions between states are restricted to states in the same level or in the two adjacent levels; that is why it is termed a QBD process. Consequently, infinitesimal generator Q of the process will have a block tridiagonal structure,

$$Q=\left[\begin{array}{c|ccc} A_1^{(-1)} & A_0^{(-1)} & 0 & \cdots \\ \hline A_2^{(0)} & & & \\ 0 & & Q_p & \\ \vdots & & & \end{array}\right]$$

where

$$Q_p=\begin{bmatrix} A_1^{(0)} & A_0^{(0)} & & \\ A_2^{(1)} & A_1^{(1)} & A_0^{(1)} & \\ & \ddots & \ddots & \ddots \\ & & A_2^{(N)} & A_1^{(N)} \end{bmatrix} \tag{1}$$

Block matrices which are not in the lower right part of Q 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 are as follows.

$$A_1^{(-1)}=\begin{bmatrix} * & \lambda & & & \\ \mu & * & \lambda & & \\ & 2\mu & * & \lambda & \\ & & \ddots & \ddots & \ddots \\ & & & (C-1)\mu & * \end{bmatrix}$$

Diagonal elements of $A_1^{(-1)}$, represented by asterisks for the ease of display, are such that the corresponding rows of Q sum to zero, i.e. $A_1^{(-1)}e+A_0^{(-1)}e=0$.

$$A_0^{(-1)}=\begin{bmatrix} 0 & 0 & \cdots & \lambda \end{bmatrix}^t$$

$$A_2^{(0)}=\begin{bmatrix} 0 & 0 & \cdots & C\mu \end{bmatrix}$$

2.3. FCFS Discipline

For this service discipline the remaining blocks are constructed as follows.

2.3.1. $A_0^{(l)}$

This matrix corresponds to transitions from $L(l)$ to $L(l+1)$, $0\le l<N$. 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 β_i. Thus it can be seen that

$$A_0^{(l)}=I_{m^l}\otimes\lambda\beta \tag{2}$$

2.3.2. $A_2^{(l)}$

This matrix corresponds to transitions from $L(l)$ to $L(l-1)$, $0<l\le N$. 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 $U_1^{(l)}$ and the latter by matrix $U_2^{(l)}$. Then,

$$A_2^{(l)}=U_1^{(l)}+U_2^{(l)} \tag{3}$$

where

$$U_1^{(l)}=\begin{cases}\tau, & l=1\\ \tau\otimes I_{m^{l-1}}+I_m\otimes U_1^{(l-1)}, & 1<l\le N\end{cases}$$

being $\tau=-Te$, and

$$U_2^{(l)}=C\mu e_m\otimes I_{m^{l-1}} \tag{4}$$

Note that in (4) it has been taken into account that service discipline is *FCFS*.

2.3.3. $A_1^{(l)}$

This matrix corresponds to transitions between states within $L(l)$. These transitions represent phase changes in the PH associated processes 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 Q must sum to zero. For the sake of clarity we introduce the set of matrices $D^{(l)}$ whose entries are equal to the entries of $A_1^{(l)}$, except those on their main diagonal. Now it can be written that

$$D^{(l)} = \begin{cases} T, & l = 1 \\ T \otimes I_{m^{l-1}} + I_m \otimes D^{(l-1)}, & 1 < l \leq N \end{cases} \tag{5}$$

Note that in the expression of $D^{(N)}$ it is assumed that customers arriving while the system is at level N, i.e. when the buffer is full, are blocked. Then $A_1^{(l)}$ is given by

$$A_1^{(l)} = D^{(l)} - \text{diag}\left\{A_2^{(l)}e + D^{(l)}e + A_0^{(l)}e\right\}$$

which can be further simplified to

$$A_1^{(l)} = \begin{cases} D^{(l)} - (C\mu + \lambda)I_{m^l}, & l < N \\ D^{(N)} - C\mu I_{m^N}, & l = N \end{cases} \tag{6}$$

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

Lemma 1

$$(A \otimes B)\, e = (Ae) \otimes (Be) \tag{7}$$

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

$$\text{diag}\left\{A_0^{(l)}e\right\} = \lambda I_{m^l} \tag{8}$$

$$\text{diag}\left\{A_2^{(l)}e + D^{(l)}e\right\} = C\mu I_{m^l} \tag{9}$$

Proof: Equation (8) follows immediately by applying the Lemma to (2) and noting that $\beta e = 1$ and $\text{diag}\{e\} = I$.

To prove (9) we first observe that

$$A_2^{(l)}e + D^{(l)}e = U_1^{(l)}e + U_2^{(l)}e + D^{(l)}e \tag{10}$$

and by applying the Lemma to (4) it is easily seen that

$$U_2^{(l)}e = C\mu e_{m^l} \tag{11}$$

on the other hand

$$\begin{aligned} U_1^{(l)}e_{m^{l-1}} + D^{(l)}e_{m^l} &= (\tau \otimes I_{m^{l-1}} + I_m \otimes U_1^{(l-1)})e \\ &\quad + (T \otimes I_{m^{l-1}} + I_m \otimes D^{(l-1)})e \\ &= (\tau \otimes e_{m^{l-1}} + e_m \otimes (U_1^{(l-1)}e_{m^{l-2}})) \\ &\quad + (Te_m \otimes e_{m^{l-1}} + e_m \otimes (D^{(l-1)}e_{m^{l-1}})) \\ &= e_m \otimes (U_1^{(l-1)}e_{m^{l-2}} + D^{(l-1)}e_{m^{l-1}}) \end{aligned} \tag{12}$$

and by recursive application of (12) we obtain

$$\begin{aligned} U_1^{(l)}e_{m^{l-1}} + D^{(l)}e_{m^l} &= e_m \otimes (U_1^{(l-1)}e_{m^{l-2}} + D^{(l-1)}e_{m^{l-1}}) \\ &= e_{m^2} \otimes (U_1^{(l-2)}e_{m^{l-3}} + D^{(l-2)}e_{m^{l-2}}) \\ &\quad \vdots \\ &= e_{m^{l-1}} \otimes (U_1^{(1)} + D^{(1)}e_m) \\ &= e_{m^{l-1}} \otimes (\tau + Te_m) = e_{m^{l-1}} \otimes 0 = 0 \end{aligned} \tag{13}$$

Hence, from (11) and (13) it follows that $A_2^{(l)}e + D^{(l)}e = C\mu e_{m^l}$, and taking the diag$\{\cdot\}$ operator on both sides of this equality yields the desired result. ■

2.4. Other Disciplines

Up to this point we restricted our analysis to the *FCFS* service discipline. Now we explain the differences in the model arising when other service disciplines (*LCFS* and *SIRO*) are considered.

2.4.1. *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 $U_2^{(l)}$ must be modified in the following manner,

$$U_2^{(l)} = I_{m^{l-1}} \otimes C\mu e_m \tag{14}$$

while the rest remains unchanged since, under the new expression for $U_2^{(l)}$, it can be proved that (11) holds in much the same way as done before .

2.4.2. *SIRO* Discipline

By the same reasoning as above only $U_2^{(l)}$ changes and it can be readily shown to be given by

$$\begin{aligned} U_2^{(1)} &= C\mu e_m \\ U_2^{(l)} &= \frac{C\mu}{l} e_m \otimes I_{m^{l-1}} + I_m \otimes \frac{(l-1)}{l} U_2^{(l-1)} \quad 1 < l < N \end{aligned}$$

In this case, (11) can be easily proved by induction on l.

2.5. Buffer management scheme

As for the buffer management or customer rejection scheme, so far we have assumed that arriving customers to a full buffer are blocked.

Here we consider a more general 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

$U \leftarrow A_1^{(N)}$
$R^{(N)} \leftarrow A_0^{(N-1)}(-U)^{-1}$
for $l = N-1, N-2, \ldots, 0, -1$ **do**
  $U \leftarrow A_1^{(l)} + R^{(l+1)} A_2^{(l+1)}$
  $R^{(l)} \leftarrow A_0^{(l-1)}(-U)^{-1}$
end for

solve $\pi^{(-1)}$ **from** $\{\pi^{(-1)}U = 0;\ \pi^{(-1)}e = 1\}$
for $l = 0, 1, \ldots, N$ **do**
  $\pi^{(l)} = \pi^{(l-1)}R^{(l)}$
end for

Figure 1: Algorithm for computing π

within level $L(N)$ and thus only the values entries of matrix $D^{(N)}$ are modified. It is a simple matter to show that the modification of $D^{(N)}$ is as follows

$$D^{(N)} \longleftarrow D^{(N)} + e_m \otimes I_{m^{N-1}} \otimes \lambda\beta$$

Hence, if the push-out and blocking schemes are probabilistically combined together with probabilities q and $(1-q)$, respectively, we have that

$$D^{(l)} = \begin{cases} T, & l = 1 \\ T \otimes I_{m^{l-1}} + I_m \otimes D^{(l-1)}, & 1 < l < N \\ T \otimes I_{m^{N-1}} + I_m \otimes D^{(N-1)} & \\ \qquad + q\,(e_m \otimes I_{m^{N-1}} \otimes \lambda\beta), & l = N \end{cases} \tag{15}$$

Finally, by the same method we used to obtain (6) it follows that

$$A_1^{(l)} = \begin{cases} D^{(l)} - (C\mu + \lambda)I_{m^l}, & l < N \\ D^{(N)} - (C\mu + q\lambda)I_{m^l}, & l = N \end{cases} \tag{16}$$

Note that when $q = 0$, (16) reduces to (6).

3. Model Analysis

In this section we describe the method to calculate the stationary state probabilities (π) for the model. Performance evaluation measures are derived from these probabilities in the next section.

The stationary probabilities (π) are obtained as the solution to the set of simultaneous linear equations $\pi Q = 0$, $\pi e = 1$. Being Q a finite dimension matrix the above system could be solved by standard linear algebra methods. However, as the size of Q may be very large, it is advisable to use more specific algorithms that take advantage of the structure (Q is block-tridiagonal) and the nature of the problem (Q is an infinitesimal generator). We used the *Linear Level Reduction* algorithm [22, 24], which can solve finite level-dependent QBDs. The algorithm is displayed in Figure 1.

4. Performance Evaluation

In this section we derive useful expressions for performance measures of the system. These are given in terms of the systems state probabilities, which can be obtained as explained in Section 3., and some of the system model matrices and parameters, which are defined in Section 2..

4.1. Distribution of the Number of Customers

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

$$p_k = \begin{cases} \pi_k^{(-1)}, & k < C \\ \pi^{(0)}, & k = C \\ \pi^{(k-C)}e, & C < k \le C+N \end{cases} \tag{17}$$

Likewise the distribution of the number of customers in the waiting room is given by

$$q_k = \begin{cases} \sum_{i=0}^{C} p_k, & k = 0 \\ p_{k+C}, & 0 < k \le N \end{cases} \tag{18}$$

4.2. Probabilities of Blocking, Expulsion and Reneging

Since arrivals are Poisson, by *PASTA* property [25] we have that the probability that an arriving customer sees the buffer full is p_{C+N}. Therefore, the probabilities of being blocked (P_b) and expulsion (P_e) are respectively given by

$$P_b = (1-q)p_{C+N} = (1-q)\pi^{(N)}e \tag{19}$$

$$P_e = qp_{C+N} = q\pi^{(N)}e \tag{20}$$

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,

$$\frac{\sum_{l=1}^{N} \pi^{(l)} U_1^{(l)} t_0 e + o(t_0)}{\lambda t_0 + o(t_0)}$$

and letting $t_0 \to \infty$, we get that

$$P_r = \frac{1}{\lambda} \sum_{l=1}^{N} \pi^{(l)} U_1^{(l)} e \tag{21}$$

4.3. 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 $\bigcup_{l=0}^{N} L(l)$, and let us denote by T_c the the sojourn time *rv* in such condition. Similarly, let us define the blocking condition as the state in which the number of costumer 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 *rv* in the blocking condition. Next we obtain the distribution of these *rv* and derive their mean values.

A congestion condition period starts when 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 $PH(\beta^{(c)}, T^{(c)})$ where

$$\beta^{(c)} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \end{bmatrix} \quad \text{and} \quad T^{(c)} = Q_p$$

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 $T^{(c)}$. 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 $PH(\beta^{(b)}, T^{(b)})$ where

$$\beta^{(b)} = \frac{1}{\pi^{(N-1)} A_0^{(N-1)} e} \pi^{(N-1)} A_0^{(N-1)}$$
$$T^{(b)} = A_1^{(N)}$$

From this it follows that [22, Eq. (2.13)]

$$\overline{T}_b = \beta^{(b)} \left(-T^{(b)}\right)^{-1} e \qquad (22)$$

and noting that $A_0^{(N-1)} e = \lambda e$ and

$$\pi^{(N-1)} A_0^{(N-1)} + \pi^{(N)} A_1^{(N)} = 0$$

we can rewrite (22) as

$$\overline{T}_b = \frac{\pi^{(N)} e}{\lambda \pi^{(N-1)} e}$$

5. Numerical Example

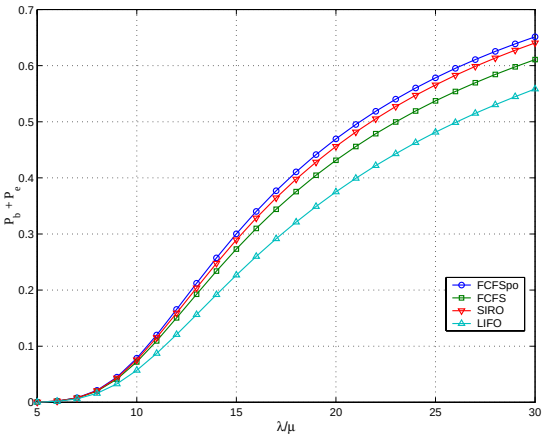
In this section we present a numerical example to illustrate the analysis carried out in the previous sections. In this example the system parameters are: $C = 10$ $N = 5$. Four combinations of service discipline and buffer management schemes are considered: *FCFS* (*FCFS* and $q = 0$), *FCFSpo* (*FCFS* and $q = 1$), *SIRO* (*SIRO* and $q = 0.5$) and *LCFS* (*LCFS* and $q = 0$). Arrival rate (λ) and transition rates of the PH distribution (T) are normalized with respect to μ. Two different instances of the patience time are examined, one whose hazard rate is increasing (e.g. Erlang) and the other decreasing (e.g. Hyper exponential). Their PH representations are, respectively,

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

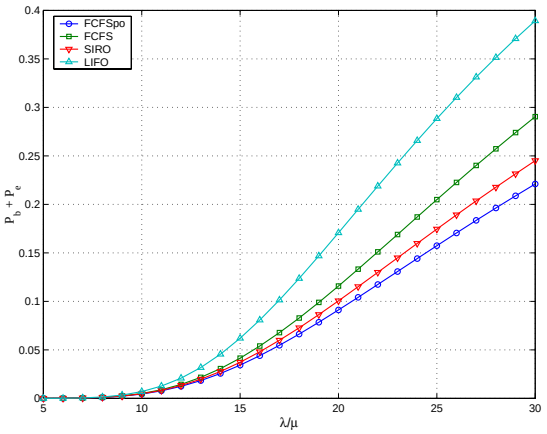
and

$$\beta = \tfrac{1}{3} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}, \quad T = \tfrac{-56}{150} \operatorname{diag}\left\{ \begin{bmatrix} 50 & 10 & 1 \end{bmatrix} \right\}$$

Figure 2 represents the sum of the probabilities of blocking and expulsion as a function of the offered traffic, and Fig. 3 does the same for the probability of reneging.



(a) Erlang



(b) Hyper exponential

Figure 2: Blocking probability plus expulsion probability

6. Conclusion

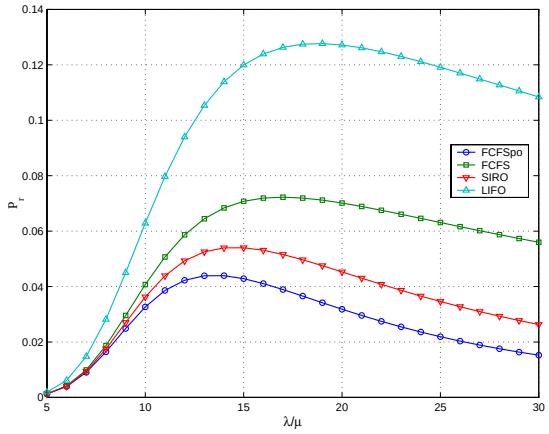
In this paper, we developed a stochastic model for a multiserver finite buffer queue with impatient customers where the patience time is modeled by a *rv* of phase-type. Furthermore, the model considers different service disciplines (*FCFS*, *LCFS* and *SIRO*) along with a probabilistically weighted buffer management scheme that combines two modes of operation: customers who arrive when system is full are blocked and customers who arrive when system is full push-out the HOL customer. The model is constructed and analyzed using matrix analytic methods. As a result of the analysis, expressions for performance evaluation are derived.

A future step in our research would be to devise a numerical procedure for computing stationary state probabilities (π) especially tailored for this model and therefore more efficient.

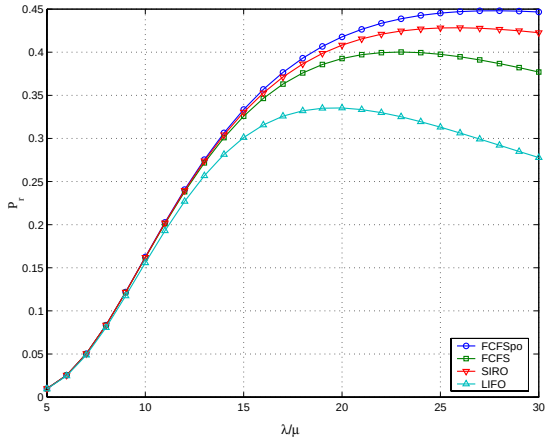
REFERENCES

[1] O. Boxma and P. de Waal, "Multiserver queues with impatient customers," in *Proceedings of ITC 14*. Elsevier Science, 1994, pp. 743–756.

[2] D. Hong and S. S. Rappaport, "Traffic model and performance analysis for cellular mobile radio tele-



(a) Erlang



(b) Hyper exponential

Figure 3: Reneging probability

phone systems with prioritized and nonprioritized handoff procedures," *IEEE Transactions on Vehicular Technology*, vol. VT-35, no. 3, pp. 77–92, Aug. 1986, see also: CEAS Technical Report No. 773, June 1, 1999, College of Engineering and Applied Sciences, State University of New York, Stony Brook, NY 11794, USA.

[3] S. Tekinay and B. Jabbari, "A measurement-based prioritization scheme for handovers in mobile cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 8, pp. 1343–1350, Oct. 1992.

[4] I. Katzela and M. Naghshineh, "Channel assignment schemes for cellular mobile telecommunication systems: A comprehensive survey," *IEEE Personal Communications*, pp. 10–31, June 1996.

[5] N. D. Tripathi, J. H. Reed, and H. F. VanLandingham, "Handoff in cellular systems," *IEEE Personal Communications*, pp. 26–37, 1998.

[6] A. Brandt and M. Brandt, "Asymptotic results and a markovian approximation for the $M(n)/M(n)/s+GI$ system," *Queueing Systems*, vol. 41, pp. 73–94, 2002.

[7] C. J. Ancker and A. Gafarian, "Some queueing problems with balking and reneging," *Operations Research*, vol. 11, pp. 88–100, 1963.

[8] A. Movaghar, "On queueing with customer impatience until the beginning of service," *Queuing Systems*, vol. 29, no. 2, pp. 337–350, 1998.

[9] D. Towsley and S. Panwar, "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.

[10] Y. Q. Zhao and A. S. Alfa, "Performance analysis of a telephone system with both patient and impatient customers," *Telecommunication Systems*, vol. 4, pp. 201–215, 1995.

[11] B. T. Doshi and H. Heffes, "Overload performance of several processor queueing disciplines for the $M/M/1$ queue," *IEEE Transactions on Communications*, vol. COM-34, no. 6, pp. 538–546, jun 1986.

[12] C.-J. Chang, T.-T. Su, and Y.-Y. Chiang, "Analysis of a cutoff priority cellular radio system with finite queueing and renegingdropping," *IEEE/ACM Transactions on Networking*, vol. 2, no. 2, pp. 166–175, Apr. 1994.

[13] V. Pla and V. Casares, "Delay-loss analysis of channel assignment schemes in mobile cellular with handoff priority and hysteresis control," in *Proceedings of 14th ITC Specialist Seminar on Access Networks and Systems*, 2001, pp. 221–230.

[14] V. Pla and V. Casares-Giner, "Effect of the handoff area sojourn time distribution on the performance of cellular networks," in *Proceedings of IEEE MWCN*, Sept. 2002.

[15] M. Ruggieri, F. Graziosi, and F. Santucci, "Modeling of the handover dwell time in cellular mobile communications systems," *IEEE Transactions on Vehicular Technology*, vol. 47, no. 2, pp. 489–498, May 1998.

[16] V. Phan-Van and S. Glisic, "Performance analysis of queueing schemes for priority handoff and call admission control," in *Procceding of IEEE International Conference on Communications (ICC)*, vol. 7, June 2001, pp. 2291–2295.

[17] V. Phan-Van and S. Glisic, "An analytical modeling for a class of queueing priority handoff and call admission control schemes in micro-cellular PCN's," in *Proceedings of IEEE VTS 53rd Vehicular Technology Conference*, vol. 2, May 2001, pp. 901–905.

[18] V. Pla and V. Casares-Giner, "Analytical-numerical study of the handoff area sojourn time," in *Proceedings of IEEE GLOBECOM*, Nov. 2002.

[19] H. G. Ebersman and O. K. Tonguz, "Handoff ordering using signal prediction priority queuing in personal communication systems," *IEEE Transactions on Vehicular Technology*, vol. 48, no. 1, pp. 20–35, Jan. 1999.

[20] A. E. Xhafa and O. K. Tonguz, "Dynamic priority queueing of handover calls in wireless networks: An analytical framework," *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 5, pp. 904 – 916, June 2004.

[21] M. Neuts, *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*. The Johns Hopkins University Press, 1981.

[22] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM, 1999.

[23] E. C. Posner and R. Guérin, "Traffic policies in cellular radio that minimize blocking of handoff calls," in *Proceedings of ITC 11*, 1985.

[24] D. Gaver, P. Jacobs, and G. Latouche, "Finite birth-and-death models in randomly changing environments," *Advances in Applied Probability*, vol. 16, pp. 715–731, 1984.

[25] R. W. Wolff, "Poisson arrivals see time averages," *Operation Research*, vol. 30, no. 2, pp. 223–231, 1982.