Stationary Delays for a Two-Class Priority Queue with Impatient Customers

Oualid Jouini
Ecole Centrale Paris
Laboratoire Génie Industriel
Grande Voie des Vignes
92295 Châtenay-Malabry Cedex, France
walid.jouini@ecp.fr

Yves Dallery
Ecole Centrale Paris
Laboratoire Génie Industriel
Grande Voie des Vignes
92295 Châtenay-Malabry Cedex, France
yves.dallery@ecp.fr

ABSTRACT

We consider a Markovian multiserver queue with two types of impatient customers, high and low priority ones. The first type of customers has a non-preemptive strict priority over the other type. After entering the queue, a customer will wait a random length of time for service to begin. If service has not begun by this time he will renege and be lost. We focus on deriving performance measures in terms of the sojourn times of customers in queue, either before starting service, or before reneging. We provide an exact analysis for systems where customers within each type are served under the FCFS discipline of service.

Keywords multiserver queues, queueing delays, reneging, non-preemptive priority.

1. INTRODUCTION

In this paper, we consider a Markovian multiserver queueing system with two types of impatient customers and a non-preemptive priority schema. Temporal limitation or reneging (or also abandonment) is an important feature in a wide variety of situations that may be encountered in real-time computing, manufacturing systems of perishable goods, telecommunication systems, call centers, etc. Models incorporating reneging are therefore closer to reality, and necessary to obtain more accurate analysis. The authors in [13] define the impatience through three different forms. The first is balking, that is, the reluctance of a customer to join a queue upon arrival. The second is reneging, which means the reluctance to remain in queue after joining and waiting. Finally, the third is jockeying between separate queues. Jockeying means that one customer has the possibility to change to one queue while he is waiting in another. In this paper, we only consider the second form of impatience. In [21], the author refines the definition of the second form of impatience by distinguishing two models. In the first model, a customer keeps his deadline only until the beginning of his service, and will remain in system while being served until he completes all service requirements. In the second model, a customer retains his deadline until the end of service, so he may interrupt his service because he has missed his deadline. In this paper, we are dealing with the former model of customer behavior, i.e., once customers get access to a server they are no longer impatient. To underline the importance of the abandonment modeling in the call center field, the authors in [11] and in [20] give some numerical examples that point out the effect of abandonment on performances.

The literature on queueing models with reneging focuses especially on performance evaluation. We refer the reader to [2], [12], and references therein for simple models assuming exponential reneging times. In [12], the authors study the subject of Markovian abandonments. They suggest an asymptotic analysis of their model under the heavy-traffic regime. Their main result is to characterize the relation between the number of servers, the offered load and system performances such as the probability of delay and the probability to abandon. This can be seen as an extension of the results of [14] by adding reneging. The author in [8] derived the limiting distribution of the virtual waiting time in a $GI/G/1$ queue. A simpler form of the latter distribution is obtained by [10] for an $M/G/1$ queue. A number of approximations for the probability to abandon are developed in [4]. The author have considered a simple Markovian multiserver queue but with generally distributed impatience times. Other works have treated the impatience phenomenon under various assumptions. Related studies include those by [1], [3], [5], [26], and references therein.

The second central feature of the model under consideration in this paper is the priority schema. Priority mechanisms are a useful scheduling method that allows different customer types to receive differentiated performance levels. Priority queueing comes up in many applications such as communication networks with differentiated services, call centers with VIP and less important customers, and more. Priority schemes are in addition known for their ease of implementation which explain their prevalence in practice. Much of queueing literature is devoted to analyzing priority queues. Most papers are restricted to two priority classes. There are two possible refinements in priority situations, namely preemption and non-preemption. In the preemptive case, a customer with high priority is allowed to enter service immediately even if another one with lower priority is already present in service at his arrival epoch. On the other hand, a priority discipline is said to be non-preemptive if there is no interruption. A customer with higher priority just

Valuetools '07, October 23-25, 2007, Nantes,France
Copyright 2007 ICST 978-963-9799-00-4.

goes to the head of the queue and waits for his turn. In this work, we are dealing with non-preemptive priority policies. In the following, we mention some of research works on priority queues. We refer the reader to [18] and [9] for a simple Markovian non-preemptive queue where all classes have the same mean service time. The author in [25] considers multiserver non-preemptive priority systems with a Markovian arrival process, service times having phase type distributions and both finite and infinite queueing space. Other references considering more complicated models include those by [17], [22], and [24]. As for preemption schemes, we refer the reader to [15], and [23] references therein. In [23], the authors derived approximations for a wide range of relevant performance characteristics, such as the moments of the number of customers of a certain type, in a Markovian queue where customers have different mean values of service times. The work in [15] introduces a new technique to reduce the Markov chain dimensionality of an $M/PH/s$ model with an arbitrary number of preemptive-resume priority classes.

Although reneging and priority systems have each received attention separately, few papers have addressed both of them. We refer the reader to [7], where the authors derived several performance measures for an $M/M/1$ queue with two classes of impatient customers in which class 1 customers have impatience of constant duration, and class 2 customers have no impatience and low priority level. An extension of the latter model is done in [6] for general distributed impatience times.

In this paper, we consider a Markovian multiserver queue with two classes of impatient customers, high and low priority ones. We assume common exponential distributions for service times as well as times before reneging for both customer types. We derive various performance measures related to queueing delays. To the best of our knowledge, the derived formulas for low priority customers are new in the literature.

The remainder of this paper is structured as follows. In Section 2, we describe the queueing model under consideration. In Section 3, we give the definitions of the performance measures of interest and develop some preliminary results that would help us in the rest of the analysis. In Sections 4 and 5, we present the results of performance evaluation when high and low priority customers are served under the FCFS basis, respectively. In Section 6, we give some concluding remarks.

2. BASIC MODEL

Consider a queueing model with two classes (types) of customers: important customers type A, and less important ones type B. The model consists of two infinite queues type A and B, and a set of s parallel, identical servers. All servers are able to handle all types of customers. The system is work-conserving, i.e., a server is never forced to be idle with customers waiting. So upon arrival, a customer is addressed by one of the available servers, if any. If not, the customer must join one of the queues. The scheduling policy of service assigns customers A and B to queues A and B, respectively. Customers A (waiting in queue A) have priority over customers B (waiting in queue B) in the sense that agents are providing assistance to customers A first. The priority rule is non-preemptive, which simply means that a server currently serving a B customer, while a new type A arrival enters system, will complete this service before turning to queue A customer. Within each queue, customers are served in order of their arrivals, i.e., under the FCFS manner. Arrival processes of type A and B customers follow a Poisson process with rates λ^A and λ^B, respectively. Let λ^T be the total arrival rate, $\lambda^T = \lambda^A + \lambda^B$. Successive service times are assumed to be i.i.d., and follow a common exponential distribution with rate μ for both types of customers.

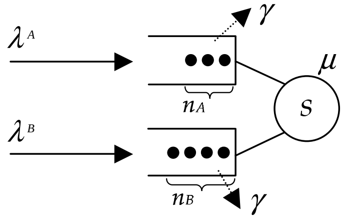

Figure 1: Basic Model

In addition, we let customers be impatient. After entering the queue, a customer will wait a random length of time for service to begin. If service has not begun by this time he will renege (leaves the queue). Times before reneging, for both customer types, are assumed to be i.i.d. and exponentially distributed with a common rate, say γ. Finally, retrials are ignored, and reneging is not allowed once one customer starts his service. Following similar arguments, the behavior of the system can be viewed as a two-class $M/M/s + M$ queueing system. The symbol M after the + is to indicate the Markovian assumption for times before reneging. The resulting model is referred to as the Basic Model, and is shown in Figure 1. Note that owing to abandonments (reneging), the system is unconditionally ergodic.

3. NOTATIONS AND PRELIMINARIES

In this section, we first present notations and definitions of the performances we are interested in. The performance measures are defined in terms of the waiting time in queue. Second, we present some preliminary derivations that we will need along the way.

3.1 Notations

We denote by m the type of one customer, $m \in \{A, B\}$. In a distant future, we define the following random variables.

- Q^m is the stationary mean number of type m customers in queue m.
- X is the unconditional stationary queueing delay of a customer (regardless of his type).
- X^m is the stationary queueing delay of type m customers.
- X_s^m is the conditional stationary queueing delay of a type m customer, given that he will enter service. We denote by P_s^m the stationary probability to enter service for type m new arrivals.
- X_r^m is the conditional stationary queueing delay of a type m customer, given that he will renege in queue.

Also, let P_r^m be the stationary probability to abandon while waiting in queue for type m customers.

- X_d^m is the conditional stationary queueing delay of a type m customer, given that he has to wait (all servers are busy). The probability that a new arrival has to wait is type of the customer independent and is referred to as the probability of delay, say P_d.
- $X_{s,d}^m$ is the conditional stationary queueing delay of a type m customer, given that he will enter service and that he was queueing. We denote by $P_{s,d}^m$ the probability that a type m waiting customer in queue will enter service.

In this paper, we compute the moments of the distributions of X_s^m and X_r^m. We also derive the expressions of the probabilities P_d and P_r^m. By doing so, one may easily deduce the analysis for all remaining random variables we have defined. For $k \geq 0$, we denote by $E(Y^k)$ the kth order moment of a given random variable Y.

Since the arrival processes are Poisson, the probability that a new arrival is of type m is $\frac{\lambda^m}{\lambda^A+\lambda^B}$. So,

$$E(X^k) = \frac{\lambda^A}{\lambda^A+\lambda^B} \cdot E(X^{A,k}) + \frac{\lambda^B}{\lambda^A+\lambda^B} \cdot E(X^{B,k}), \ k \geq 0. \tag{1}$$

A customer who does not renege will necessarily enter service, then $P_s^m + P_r^m = 1$. For $m \in \{A, B\}$, one may write

$$E(X^{m,k}) = P_s^m \cdot E(X_s^{m,k}) + P_r^m \cdot E(X_r^{m,k}), \ k \geq 0. \tag{2}$$

Upon arrival, a customer is immediately addressed by one of the available servers, if any. If not, he has to wait and joins one of the queues (with probability P_d). Thus,

$$E(X_d^{m,k}) = \frac{E(X^{m,k})}{P_d}, \ k \geq 0. \tag{3}$$

For a customer who joins the queue, there are two possibilities: either he reneges while waiting in queue, or he gets service. So, $P_d = P_r^m + P_{s,d}^m$. Also

$$E(X_d^{m,k}) = P_{s,d}^m \cdot E(X_{s,d}^{m,k}) + P_r^m \cdot E(X_r^{m,k}), \ k \geq 0, \tag{4}$$

which allows to determine $E(X_{s,d}^{m,k})$, for $k \geq 0$.

3.2 Preliminaries

We start by computing the stationary probabilities of system states. We denote by n_A, n_B and n_T the numbers in queues of customers type A, type B and the total one $n_T = n_A + n_B$, respectively. Computing the stationary probabilities for n_B or the couple (n_A, n_B) is a hard task. We only consider the processes $\{n_T(t), t > 0\}$ and $\{n_A(t), t > 0\}$ and compute their corresponding stationary probabilities. Recall that all stationary probabilities exist due to the ergodicity condition (which holds for any $\gamma > 0$). For the rest of the paper, an empty sum is being interpreted as zero, and an empty product is being interpreted as one.

Let us start by considering the process $\{n_T(t),\ t > 0\}$. Service times as well as times before reneging are memoryless and common for both types of customers. Thereafter due to the work-conserving property of our system, the total number of customers in queue does not depend on the discipline of service. With regard to the stationary probabilities of n_T, our system is equivalent to a single multiserver queue with a single class of customers. The arrival process is Poisson with intensity $\lambda^T = \lambda^A + \lambda^B$. Taking the associated birth-death process as shown in Figure 2, one may obtain in the long run a set of infinite recursive relations relating the steady state probabilities. Proceeding to solve by iteration and using the normalization condition, we get the solutions $p(n_T) = q(n_T + s)$, where

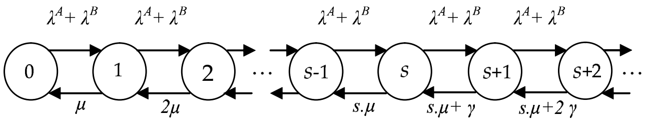

Figure 2: Birth-death process of the total number of customers in system

$$q(i) = \frac{\lambda^i}{i!\,\mu^i} q(0), \text{ for } 0 \leq i \leq s, \tag{5}$$

$$q(i) = \frac{\lambda^i}{s!\mu^s \prod_{j=1}^{i-s}(s\,\mu + j\,\gamma)} q(0), \text{ for } i > s,$$

and $q(0)$ is the stationary probability to have no customers in system (in service and in queue). It is given by

$$q(0) = \left(\sum_{i=0}^{s} \frac{\lambda^i}{i!\,\mu^i} + \frac{1}{s!\,\mu^s} \sum_{i=s+1}^{\infty} \frac{\lambda^i}{\prod_{j=1}^{i-s}(s\,\mu + j\,\gamma)} \right)^{-1}. \tag{6}$$

So, the probability of delay P_d can be determined by

$$P_d = 1 - \sum_{k=0}^{s-1} q(k). \tag{7}$$

To compute the stationary probabilities for $\{n_A(t), t > 0\}$, we consider a special two-dimension Markov chain as shown in Figure 3. The state of the system is defined by the total number of customers in system (regardless of their type) if less than s customers are in system (i.e., all customers are in service), and by the couple (n_A, n_B) if s customers or more are in system (i.e., all servers are busy). During the stationary regime, we have from the Markov chain presented in Figure 3

$$p(n_A = k) = \frac{(\lambda^A)^k}{\prod_{j=1}^{k} \ (s\mu + j\gamma)} p(n_A = 0), \text{ for } k \geq 0, \tag{8}$$

where $p(n_A = 0)$ is the probability to have all servers busy and no type A customers waiting in queue.

To get $p(n_A = 0)$, we come back to the modeling of our system using the birth-death process of Figure 2. It is clear that the probability to be in state i, $0 \leq i \leq s-1$, in the Markov chain of Figure 3 is equivalent to that already derived in Equation (5) for the birth-death process of Figure 2, $q(i)$. The normalization condition for our Markov chain may be written as

$$\sum_{i=0}^{s-1} p(i) + \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} p(n_A = k,\ n_B = i) = 1. \tag{9}$$

Recall that $p(n_A = k) = \sum_{i=0}^{\infty} p(n_A = k,\ n_B = i)$, for $k \geq 0$.

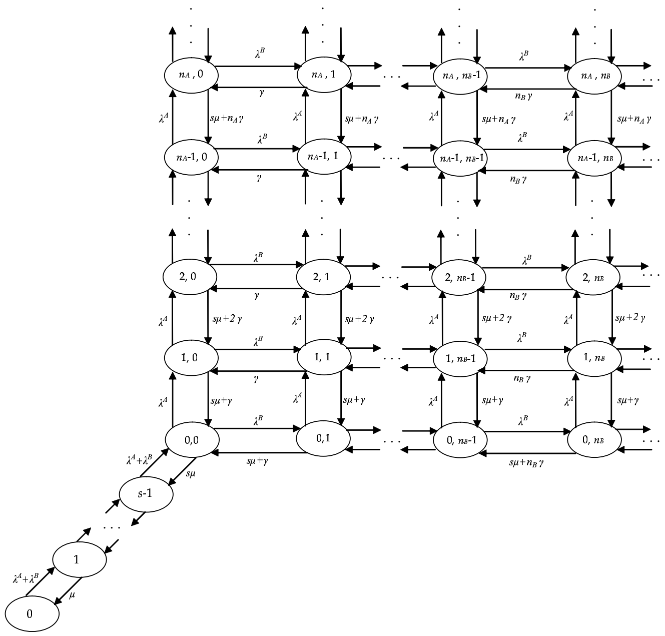

Figure 3: Markov chain for the number of customers in queue

Combining thereafter Equations (8) and (9) leads to

$$\sum_{k=0}^{s-1} q(k) + \left(\sum_{k=0}^{\infty} \frac{(\lambda^A)^k}{\prod_{j=1}^{k}(s\mu + j\gamma)} \right) \cdot p(n_A \;=\; 0) = 1, \quad (10)$$

or equivalently

$$p(n_A \;=\; 0) = \left(1 - \sum_{k=0}^{s-1} q(k)\right) \cdot \left(\sum_{k=0}^{\infty} \frac{(\lambda^A)^k}{\prod_{j=1}^{k}(s\mu + j\gamma)} \right)^{-1}. \quad (11)$$

Having in hand $p(n_A = k)$ and $p(n_T = k)$ for $k \geq 0$, the stationary mean number in queue of type A customers, Q^A, and that of both customer types, Q^T, are therefore given by

$$Q^A = \sum_{k=1}^{\infty} k \cdot p(n_A = k), \text{ and } Q^T = \sum_{k=1}^{\infty} k \cdot p(n_T = k). \quad (12)$$

As a consequence, the stationary mean number in queue of type B customers, Q^B, may be deduced by $Q^B = Q^T - Q^A$.

We are now ready to compute the stationary probability to renege and that to enter service for a type m new arrival. The quantity P_r^m can be viewed as the proportion of customers who renege, i.e., the fraction of the stationary mean rate of abandoned customers over that of arrivals. Thus, it is calculated as

$$P_r^m = \frac{\gamma \cdot Q^m}{\lambda^A}. \quad (13)$$

The probability to enter service is only the complementary probability (no possible events of blocking or balking). A customer who does not indeed renege, will necessarily enter service. So,

$$P_s^m = 1 - P_r^m. \quad (14)$$

4. ANALYSIS FOR HIGH PRIORITY CUSTOMERS

In this section, we tackle the quantitative analysis for high priority customers, namely type A customers. In [27], the author has derived the first and second moments of the distributions of the random variables X_s^A and X_r^A in the case of a finite single queue with a single class of customers. Owing to his higher priority, the quantitative analysis of X_s^A and X_r^A for our two-class model is not far from that analysis.

Our approach is based on system state probabilities seen by a randomly chosen new arrival A. From the PASTA property (Poisson Arrivals See Time Averages), these probabilities coincide with those seen by an outside random observer, i.e., simply the probabilities that the system is in a given state at a random instant. The PASTA property is based on the memoryless property of the Poisson process, which allows to generate a sequence of arrivals that take a random look at the system. We refer the reader to [19] for further explanation, and [28] for a rigorous proof.

Consider a new arrival A who finds all servers busy and n_A waiting customers ahead of him in queue A, $n_A \geq 0$. It goes without saying, for the remaining cases (at least one server is idle), that our customer will get service immediately. Because of their lower priority, type B customers already waiting in queue B, as well as those who will arrive after our arrival A of interest, will not affect the sojourn time in queue of the latter. Conditioning on the system state, the new arrival will start his sojourn at position $n_A + 1$ in queue A. Two cases are possible. The first is that the customer of interest does not renege until starting service with a given conditional probability, say Ψ_{n_A+1}. The second case is that he reneges at one of the positions he occupies during his sojourn in queue A, with probability $(1 - \Psi_{n_A+1})$. Such a situation may be analyzed by a pure death process with state-dependent death rates, see Figure 4. We do not consider birth rates because all future type A arrivals have no priority over the customer of interest (queue A is working under the FCFS discipline of service).

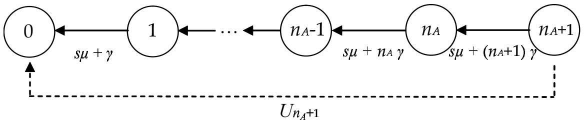

Figure 4: A new type A arrival who finds n_A customers in queue A

The process moves from state i to state $i-1$, $1 \leq i \leq n_A + 1$, further to a departure event, i.e., further to either a service completion with rate $s\mu$, or an abandonment with a rate equal to the number of waiting customers times the reneging rate, $i\gamma$. The memoryless property of service times as well as times before reneging allows us to state the following claim. When being in state i, the probability that the process moves down due to the event of reneging of our customer of interest is given by $\frac{\gamma}{s\mu + i\gamma}$. Next, the conditional probability, Ψ_{n_A+1}, that our customer does not renege while waiting in queue, given the system state he sees at his arrival epoch, may be written as

$$\Psi_{n_A+1} = \prod_{i=1}^{n_A+1} \left(1 - \frac{\gamma}{s\mu + i\gamma}\right). \qquad (15)$$

In other words, the latter event means that our customer does not renege in all possible queue positions he may occupy, starting from position n_A+1 until position 1 and enters service afterwards.

Let us now define the conditional random variable U_{n_A+1} denoting the time it takes to empty queue A of n_A+1 waiting customers (without considering eventual future arrivals). Our customer will thereafter enter service with probability Ψ_{n_A+1} in U_{n_A+1} units of times. Conditioning on a state seen by a new arrival A and averaging thereafter over all possibilities, the k-th moment of the sojourn time in queue A and being served afterwards is given by $\sum_{n_A=0}^{\infty} p(n_A) \cdot \Psi_{n_A+1} \cdot E(U_{n_A+1}^k)$, where $E(U_{n_A+1}^k)$ denotes the k-th order moment of U_{n_A+1}. Thus, the k-th order moment of the conditional random variable $X_s^{A,k}$ given service is

$$E(X_s^{A,k}) = \frac{\sum_{n_A=0}^{\infty} p(n_A) \cdot \Psi_{n_A+1} \cdot E(U_{n_A+1}^k)}{P_s^A}. \qquad (16)$$

It remains for us to derive the expression of $E(U_i^k)$, $i \geq 1$. The random variable U_i can be viewed as the first passage time at state 0 starting from state i in the pure death process of Figure 4. Then, the distribution of U_i is the convolution of i independent exponential distributions with parameters $s\mu+\gamma$, $s\mu+2\gamma$, ..., and $s\mu+i\gamma$, which is an hypoexponential distribution. So, all moments of U_i may be derived in a closed form. We only give here its mean and variance. They are $\sum_{j=1}^{i} \frac{1}{s\mu+j\gamma}$ and $\sum_{j=1}^{i} \frac{1}{(s\mu+j\gamma)^2}$, respectively.

Let us now focus on deriving $E(X_r^{A,k})$. Assume that our customer of interest will renege while waiting in queue A, and let V_{n_A+1} denote the random variable measuring his sojourn time in queue before reneging. Again, conditioning on a state seen by a new arrival A and averaging thereafter over all possibilities, the k-th moment of the sojourn time in queue A and not being served afterwards is given by $\sum_{n_A=0}^{\infty} p(n_A) \cdot E(V_{n_A+1}^k)$, where $E(V_{n_A+1}^k)$ denotes the k-th order moment of V_{n_A+1}. Thus, the k-th order of the conditional random variable $X_r^{A,k}$ given reneging is

$$E(X_r^{A,k}) = \frac{\sum_{n_A=0}^{\infty} p(n_A) \cdot E(V_{n_A+1}^k)}{P_r^A}. \qquad (17)$$

Note that computing the moments of V_i involves hypoexponential distributions, and are also easy to derive. One may see that the probability to abandon at position j, $1 \leq j \leq i$, is $(1 - \frac{\gamma}{s\mu+i\gamma}) \cdot (1 - \frac{\gamma}{s\mu+(i-1)\gamma}) ... (1 - \frac{\gamma}{s\mu+(j+1)\gamma}) \cdot \frac{\gamma}{s\mu+j\gamma}$. Knowing that our customer will renege at position j, the time to abandon, say $V_i(j)$, is the sum of $i-j+1$ independent exponential random variables with parameters $s\mu+i\gamma$, $s\mu+(i-1)\gamma$, ..., and $s\mu+j\gamma$, which has an hypoexponential distribution. Averaging on all possibilities, we get

$$E(V_i^k) = \sum_{j=1}^{i} \left(\prod_{k=j+1}^{i} (1 - \frac{\gamma}{s\mu+k\gamma}) \right) \cdot \frac{\gamma}{s\mu+j\gamma} \cdot E(V_i^k(j)). \qquad (18)$$

Up to now, we computed the k-th order moment of the random variables X_s^A and X_r^A. As for the stationary unconditional queueing delay of a customer A, $E(X^{A,k})$, it is given using the relation $E(X^{A,k}) = P_s^A \cdot E(X_s^{A,k}) + P_r^A \cdot E(X_r^{A,k})$. To close the analysis for type A customers, we give the expression of the k-th order moment of the conditional stationary queueing delay of a customer A, given all servers are busy. It is simply computed as $E(X_d^{A,k}) = \frac{E(X^{A,k})}{P_d}$.

5. ANALYSIS FOR LOW PRIORITY CUSTOMERS

In this section, we focus on evaluating the performance measures for type B customers, which is to the best of our knowledge new. We specifically address the quantitative analysis of X_s^B and X_r^B. In the following, we start by deriving the expression of the kth order moment of X_s^B.

Knowing that all servers are busy, let n_A and n_B be the number of types A and B waiting customers seen by a new type B arrival in queues A and B, respectively. In our analysis, we ignore all future type B arrivals because the discipline of service within queue B is FCFS. However, all future type A arrivals have to be considered because of their higher priority over the customer of interest. We note that the sojourn time in queue of this customer does not depend on the couple (n_A, n_B) but on the total number of customers ahead of him, $n_T = n_A + n_B$ (common distribution of times before reneging for both customer types). Our customer of interest will start his sojourn in queue B at a total position $n_T + 1 = n_A + n_B + 1$ for both queues. Before leaving the queue, he will occupy queue positions ahead of position $n_T + 1$, but also he may occupy those behind his initial position $n_T + 1$. The associated birth-death process is shown in Figure 5. The state 0 in the latter corresponds to have no waiting customers in queues A and B. We ignore the birth rate at state 0 because we are interested on the first passage time at that state, starting at state $n_T + 1$. This duration corresponds indeed to the epoch at which the queue becomes empty and our customer will therefore start service (in case he does not renege in between).

Let us now define the random variable R_{n_T+1} as the time it takes to empty the queue of the $n_T + 1$ already waiting customers, as well as all future type A customers (who arrive in between). Recall that a customer leaves the queue by either starting service, or by reneging.

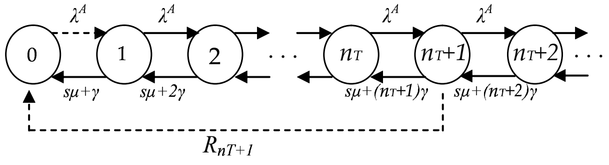

Figure 5: A new type B arrival who finds n_T customers in queues A and B

By considering a general birth-death process, the authors in [16] gave closed-form expressions for any moment of order $k \geq 1$ of several random variables related to first passage times. We use their results in our context here to get the moments of R_{n_T+1}. To simplify the presentation, we define the potential coefficients, say π_i, of the birth-death process presented in Figure 5, as follows.

$$\pi_0 = 1, \text{ and } \pi_i = \frac{(\lambda^A)^i}{\prod_{j=1}^{i}(s\mu + j\gamma)}, \text{ for } i \geq 1. \qquad (19)$$

From [16], the mean, $E(R_{n_T+1})$, and variance, $Var(R_{n_T+1})$, of the random variable R_{n_T+1} are given by

$$E(R_{n_T+1}) = \frac{1}{\lambda^A}\sum_{i=1}^{n_T+1}\frac{1}{\pi_{i-1}}\sum_{j=i}^{\infty}\pi_j, \tag{20}$$

$$\begin{aligned} Var(R_{n_T+1}) = &\sum_{i=1}^{n_T+1}\left(\frac{2}{\lambda^A\pi_{i-1}}\sum_{j=i+1}^{\infty}\frac{1}{\lambda^A\pi_{j-1}}\left(\sum_{k=j}^{\infty}\pi_k\right)^2\right) \\ &+\sum_{i=1}^{n_T+1}\left(\frac{1}{(\lambda^A)^2\pi_{i-1}^2}\left(\sum_{k=i}^{\infty}\pi_k\right)^2\right). \end{aligned} \tag{21}$$

Note that one may derive all higher order moments of R_{n_T+1}, which allows us to derive their full distributions. However, the expressions would be cumbersome. For presentation issues, we content ourself with only the mean and variance.

Conditioning again on a state seen by a new arrival B and averaging thereafter over all possibilities, the k-th moment of the sojourn time in queue B and being served afterwards is given by $\sum_{n_T=0}^{\infty} p(n_T)\cdot\Upsilon_{n_T+1}\cdot E(R^k_{n_T+1})$, where Υ_{n_T+1} is the probability that a new arrival B who finds n_T waiting customers in queues A and B does not renege until starting service. Hence, the k-th order moment of the conditional random variable $X_s^{B,k}$ given service is

$$E(X_s^{B,k}) = \frac{\sum_{n_T=0}^{\infty} p(n_T)\cdot\Upsilon_{n_T+1}\cdot E(R^k_{n_T+1})}{P_s^B}, \tag{22}$$

where the quantities $p(n_T)$ are given using Equations (5) and (6). To explicitly get $E(X_s^{B,k})$, it remains for us to compute Υ_{n_T+1}. Roughly speaking, the quantity Υ_{n_T+1} is the probability that the customer of interest does not renege in all the positions he occupies up to starting service. Deriving Υ_{n_T+1} is more complicated than that for type A customers. The complexity comes from uncertain future type A customers who arrive and get the priority over our customer B of interest. During his sojourn in queue, the latter will occupy positions n_T+1, n_T, ..., and position 1 at least for one time. In addition, he may occupy or not any position (n_T+2, n_T+3,...). In other words, his position may take any strictly non-negative integer value, see Figure 5. If he occupies one of those positions, the number of times he did that is random. To be rigorous, let r_i be the discrete random variable denoting the number of times the customer of interest occupies a total position i in queues, $i\geq 1$. Saying that our customer does not renege until being served means that he does not renege in any position i he may occupy every time he visits that position, i.e., r_i times. By averaging on all possibilities (number of visits j at a given position i), the probability that our customer does not renege at a position i is $\sum_{j=1}^{\infty} p(r_i=j)\cdot(1-\frac{\gamma}{s\mu+i\gamma})^j$, for $i\geq 1$. So, the probability Υ_{n_T+1} may be written as

$$\Upsilon_{n_T+1} = \prod_{i=1}^{\infty}\left(\sum_{j=1}^{\infty} p(r_i=j)\cdot(1-\frac{\gamma}{s\mu+i\gamma})^j\right). \tag{23}$$

In what follows, we go on to compute the probabilities $p(r_i=j)$ using the notion of ruin probabilities. We first start with some preliminaries that will help us to derive these quantities.

Consider again the birth-death process presented in Figure 5. Let ${}^k\eta_{i\,j}$ be the ruin probability that the particle, starting at i, reaches j first before k, $1\leq j<i<k$. Moreover, let ${}^k\nu_{i\,j}$ be the ruin probability that the process, starting at i, reaches j first before k, $0\leq k<i<j$.

One may then deduce the quantities ${}^k\eta_{i\,j}$, for $1\leq j<i<k$, and ${}^k\nu_{i\,j}$, for $0\leq k<i<j$, as follows.

$${}^k\eta_{i\,j} = \prod_{m=j+1}^{i}{}^k\eta_{m\,m-1}, \text{ and } {}^k\nu_{i\,j} = \prod_{m=i}^{j-1}{}^k\nu_{m\,m+1} \tag{24}$$

We derive ${}^k\eta_{i\,i-1}$, $1\leq i<k$, and ${}^k\nu_{i-1\,i}$, $0\leq k<i$, using recursive relations. The quantity ${}^k\eta_{i\,i-1}$ is the ruin probability that the particle, starting at i, reaches $i-1$ first before k, $1\leq i<k$. We denote by μ_i the death rate associated to a state i in the birth-death process of Figure 5, $i\geq 1$. It is clear that the ruin probability ${}^k\eta_{k-1\ k-2}$ to reach $k-2$ starting at $k-1$, without visiting k, is given by $\frac{\mu_{k-1}}{\lambda^A+\mu_{k-1}}$. For a given i, $1\leq i<k-1$, we define the event ${}^kE_{i\,i-1}$ that the particle reaches first $i-1$ starting from i, without visiting k. Let us calculate now the probability that ${}^kE_{i\,i-1}$ occurs, namely ${}^k\eta_{i\,i-1}$. Starting at state i, two events may occur: either the process goes down to $i-1$, say event ${}^kF_{i\,i-1}$, or the process goes up to $i+1$ which is the complementary event of ${}^kF_{i\,i-1}$, say ${}^k\bar{F}_{i\,i-1}$. Hence, we can write

$$\begin{aligned} Pr({}^kE_{i\,i-1}) =& Pr({}^kE_{i\,i-1}\mid {}^kF_{i\,i-1})\cdot Pr({}^kF_{i\,i-1}) \\ &+ Pr({}^kE_{i\,i-1}\mid {}^k\bar{F}_{i\,i-1})\cdot Pr({}^k\bar{F}_{i\,i-1}). \end{aligned} \tag{25}$$

The event ${}^kE_{i\,i-1}\mid {}^kF_{i\,i-1}$ is to reach $i-1$ when being at $i-1$ without visiting k, which obviously occurs with probability 1 since the process is already in state $i-1$. The event ${}^kE_{i\,i-1}\mid {}^k\bar{F}_{i\,i-1}$ is to reach $i-1$ first before k when being at $i+1$, which is equivalent to the following: starting at $i+1$, the process reaches i without visiting k, then starting at i, it reaches $i-1$ without visiting k. So, $Pr({}^kE_{i\,i-1}\mid {}^k\bar{F}_{i\,i-1}) = {}^k\eta_{i+1\,i}\,{}^k\eta_{i\,i-1}$. Furthermore, the event ${}^kF_{i\,i-1}$ occurs with probability $\frac{\mu_i}{\lambda^A+\mu_i}$, and the event ${}^k\bar{F}_{i\,i-1}$ with probability $\frac{\lambda^A}{\lambda^A+\mu_i}$. These arguments lead to the following recursive relation

$${}^k\eta_{i\,i-1} = \frac{\mu_i}{\lambda^A+\mu_i}+\frac{\lambda^A}{\lambda^A+\mu_i}{}^k\eta_{i+1\ i}\,{}^k\eta_{i\,i-1}, \text{ for } 1\leq i<k-1, \tag{26}$$

or equivalently

$${}^k\eta_{i\,i-1} = \frac{\mu_i}{\mu_i+\lambda^A(1-^k\eta_{i+1\,i})}, \text{ for } 1\leq i<k-1, \tag{27}$$

$$\text{starting with } {}^k\eta_{k-1\,k-2} = \frac{\mu_{k-1}}{\mu_{k-1}+\lambda^A}.$$

With a similar approach as described above, we give the following recursive relation for the ruin probability ${}^k\nu_{i-1\,i}$.

$${}^k\nu_{i-1\,i} = \frac{\lambda^A}{\lambda^A+\mu_{i-1}(1-^k\nu_{i-2\,i-1})}, \text{ for } i>k+2, \tag{28}$$

$$\text{starting with } {}^k\nu_{k+1\,k+2} = \frac{\lambda^A}{\lambda^A+\mu_{k+1}}.$$

Let us come back to computing the probability distribution of r_i, i.e., $p(r_i=j)$ for $i,j\geq 1$. To do so, we further define the quantity $\alpha_i = \frac{\mu_i}{\lambda^A+\mu_i}$. Consider now the initial state at which the customer of interest B starts his sojourn, namely state n_T+1. Saying that our customer visits that state only one time is equivalent to say that the process in

Figure 5 moves down to state n_T, then it reaches state 0 (starting from state n_T) first before n_T+1. The probability of this event is given by $p(r_{n_T+1}=1) = \alpha_{n_T+1} \cdot {}^{n_T+1}\eta_{n_T\;0}$. The event in which our customer visits state n_T+1 exactly two time is equivalent to one of the following events: The first is that the process moves down to n_T starting at n_T+1, then it visits again n_T+1 first before 0 starting at n_T, then it moves down to n_T starting at n_T+1, it finally reaches state 0 first before n_T+1. The second event is that, it goes up to n_T+2 starting at n_T+1 (with probability $1-\alpha_{n_T+1}$), then it moves down again to state n_T+1 (with probability 1 after a finite time), then it moves down to n_T starting at n_T+1, finally it reaches state 0 first before n_T+1. So, the probability of visiting state n_T+1 exactly two times is

$$p(r_{n_T+1}=2) = \alpha_{n_T+1} \cdot {}^{0}\nu_{n_T\;n_T+1} \cdot \alpha_{n_T+1} \cdot {}^{n_T+1}\eta_{n_T\;0} \qquad (29)$$
$$+ (1-\alpha_{n_T+1}) \cdot 1 \cdot \alpha_{n_T+1} \cdot {}^{n_T+1}\eta_{n_T\;0}.$$

Continuing with the same reasoning, we get for a state i such that $2 \leq i \leq n_T+1$,

$$p(r_i=j) = \alpha_i \cdot {}^{i}\eta_{i-1\;0} \cdot (\alpha_i \cdot {}^{0}\nu_{i-1\;i} + 1 - \alpha_i)^{j-1}, j \geq 1. \qquad (30)$$

For state 1, the expression is simpler. We have $p(r_1=j) = \alpha_1 \cdot (1-\alpha_1)^{j-1}$, for $j \geq 1$.

As for the remaining states, $i \geq n_T+2$, the quantities $p(r_i=j)$ for $j \geq 1$ has a generic expression slightly different from that given in Equation (30). For example, the event to be in position n_T+2 exactly one time is equivalent to say that the process visits n_T+2 starting at n_T+1 first before 0 (with probability ${}^{0}\nu_{n_T+1\;n_T+2}$), then starting at n_T+2 it again moves down to state n_T+1 (with probability α_{n_T+2}), finally starting at n_T+1 it reaches 0 first before n_T+2 (with probability ${}^{n_T+2}\eta_{n_T+1\;0}$). So, the probability of the latter event is $p(r_{n_T+2}=1) = {}^{0}\nu_{n_T+1\;n_T+2} \cdot \alpha_{n_T+2} \cdot {}^{n_T+2}\eta_{n_T+1\;0}$. In the same way, we get for $i \geq n_T+2$

$$p(r_i=j) = {}^{0}\nu_{n_T+1\;i} \cdot \alpha_i \cdot {}^{i}\eta_{i-1\;0} \cdot (\alpha_i \cdot {}^{0}\nu_{i-1\;i} + 1 - \alpha_i)^{j-1}, j \geq 1. \qquad (31)$$

The probability Υ_{n_T+1} is now determined using Equations (30) and (31). Finally, it suffices to come back to Equation (22) in order to explicitly get any kth order moment of the distribution of the conditional random variable X_s^B, given service.

Let us now move on to address the analysis of the conditional queueing delay of a type B customer given that he reneges, namely X_r^B. This is not quite so simple, because we have to fully characterize all possible sample paths of a particle in the birth-death process of Figure 5. In the following, we only give the mean value of X_r^B.

On the one hand, the unconditional stationary mean queueing delay (before reneging or starting service) for type B customers is related to the quantities $E(X_s^B)$ and $E(X_r^B)$ via the relation $E(X^B) = P_s^B \cdot E(X_s^B) + P_r^B \cdot E(X_r^B)$. On the other hand, applying the Little law on queue B leads to $\lambda^B \cdot E(X^B) = Q^B$. Recall that the stationary mean number of customer B waiting in queue is given using $Q^B = Q^T - Q^A$ and Equation (12). Combining the last two relations implies

$$E(X_r^B) = \frac{Q^B}{\lambda^B\, P_r^B} - \frac{P_s^B}{P_r^B} E(X_s^B). \qquad (32)$$

This closes our discussions about the analysis of type B customers.

6. CONCLUSIONS

In this paper, we considered a non-preemptive priority queueing system in which customers wait for service for a limited time only and leave system if service has not begun within that time. Practical examples of queueing systems with customer impatience include real-time telecommunication systems, inventory systems with perishable items, and more. We derived several closed-form expressions of useful performance measures related to queueing delays of high and low priority customers.

In a future study, it would be interesting to investigate approximations or numerical methods for computing performance measures. This would be helpful to avoid numerical instabilities given that the closed-form expressions of interest are somewhat cumbersome. We also want to relax some assumptions: arbitrary number of customer types, different mean values for customer types service times and times before reneging, etc.

7. REFERENCES

[1] E. Altman and A. Borovkov. On the Stability of Retrial Queues. *Queueing Systems*, 26:343–363, 1997.

[2] C. J. Ancker and A. Gafarian. Queueing with Impatient Customers Who Leave at Random. *Journal of Industrial Engineering*, 13:84–90, 1962.

[3] F. Baccelli and G. Hebuterne. On Queues With Impatient Customers. *Performance'81 North-Holland Publishing Company*, pages 159–179, 1981.

[4] O. Boxma, G. Koole, and Z. Liu. Queueing-Theoretic Solution Methods for Models of Parallel and Distributed Systems. *Performance Evaluation of Parallel and Distributed Systems - Solution Methods, CWI Tract 105 & 106, Amsterdam*, 1994.

[5] A. Brandt and M. Brandt. Assymptotic Results and a Markovian Approximation for the M(n)/M(n)/C + GI System. *Queueing Systems: Theory and Applications (QUESTA)*, 41:73–94, 2002.

[6] A. Brandt and M. Brandt. On the Two-class M/M/1 System under Preemptive Resume and Impatience of the Prioritized Customers. *Queueing Systems*, 47:147–168, 2005.

[7] B. Choi, B. Kim, and J. Chung. M/M/1 Queue with Impatient Customers of Higher Priority. *Queueing Systems*, 38:49–66, 2001.

[8] D. Daley. General Customer Impatience in the Queue GI/G/1. *Journal of Applied Probability*, 2:186–205, 1965.

[9] R. Davis. Waiting-Time Distribution of a Multi-Server Priority Queuing System. *Operations Research*, 14:133–136, 1965.

[10] A. de Kok and T. H.C. A Queueing System with Impatient Customers. *Journal of Applied Probability*, 22:688–696, 1965.

[11] N. Gans, G. Koole, and A. Mandelbaum. Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management*, 5:73–141, 2003.

[12] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a Call Center with Impatient Customers. *Manufacturing & Service Operations Management*, 4:208–227, 2002.

[13] D. Gross and C. Harris. *Fundamentals of Queueing Theory*. Wiley series in probability and mathematical statistics, 1998. 3rd Edition.

[14] S. Halfin and W. Whitt. Heavy-Traffic Limits for Queues with Many Exponential Servers. *Operations Research*, 29:567–588, 1981.

[15] M. Harchol-Balter, A. Scheller-Wolf, and A. Wierman. Multi-server Queueing System with Multiple Priority Classes. *Queueing Systems*, 51:331–360, 2005.

[16] O. Jouini and Y. Dallery. Moments of First Passage Times in General Birth-Death Processes. *Mathematical Methods of Operational Research*, 2007. To appear.

[17] E. Kao and S. Wilson. Analysis of Non-preemptive Priority Queues with Multiple Servers and two Priority Classes. *European Journal of Operational Research*, 118:181–193, 1999.

[18] O. Kella and U. Yechiali. Waiting Times in the Non-Preemptive Priority M/M/c Queue. *Stochastic Models*, 1:257–262, 1985.

[19] L. Kleinrock. *Queueing Systems, Theory*, volume I. A Wiley-Interscience Publication, 1975.

[20] A. Mandelbaum and S. Zeltyn. Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers. 2006. Working paper, Thechnion, Haifa, Israel.

[21] A. Movaghar. On Queueing with Customer Impatience until the Beginning of Service. *Queueing Systems*, 29:337–350, 1998.

[22] A. Sleptchenko. Multi-class, Multi-server Queues with Non-preemptive Priorities. 2003. Technical Report 2003-016, EURANDOM, Eindhoven University of Technology.

[23] A. Sleptchenko and M. van der Heijden. An Exact Solution for the State Probabilities of the Multi-class, Multi-server Queue with Preemptive Priorities. *Queueing Systems*, 50:81–107, 2005.

[24] T. Takine. The Non-preemptive Priority MAP/G/1 Queue. *Operations Research*, 47:917–927, 1999.

[25] D. Wagner. Analysis of Mean Values of a Multi-server Model with Non-preemptive Priorities and Non-renewal Inputs. *Communications in Statistics - Stochastic Models*, 13:67–84, 1997.

[26] A. Ward and P. Glynn. A Diffusion Approximation for a Markovian Queue with Reneging. *Queueing Systems: Theory and Applications (QUESTA)*, 43:103–128, 2003.

[27] W. Whitt. Improving Service by Informing Customers about Anticipated Delays. *Management Science*, 45:192–207, 1999.

[28] R. Wolff. Poisson Arrivals See Time Averages. *Operations Research*, 30:223–231, 1982.