WAITING TIME ANALYSIS OF MULTI-CLASS QUEUES WITH IMPATIENT CUSTOMERS

Vahid Sarhangian
University of Toronto,
Joseph L. Rotman School of Management
105 St. George Street, Toronto, M5S 3E6, CANADA,
vahid.sarhangian11@rotman.utoronto.ca
Barış Balcıoḡlu
Sabancı University,
Faculty of Engineering and Natural Sciences,
Orhanlı-Tuzla, 34956 Istanbul, Turkey,
balcioglu@sabanciuniv.edu

Abstract

In this paper, we study three delay systems where different classes of impatient customers arrive according to independent Poisson processes. In the first system, a single server receives two classes of customers with general service time requirements, and follows a non-preemptive priority policy in serving them. Both classes of customers abandon the system when their exponentially distributed patience limits expire. The second system comprises parallel and identical servers providing the same type of service for both classes of impatient customers under the non-preemptive priority policy. We assume exponential service times and consider two cases depending on the time-to-abandon distribution being exponentially distributed or deterministic. In either case, we permit different reneging rates or patience limits for each class. Finally, we consider the first-come-first-served policy in single and multi-server settings. In all models, we obtain the Laplace transform of the virtual waiting time for each class by exploiting the level-crossing method. This enables us to compute the steady-state system performance measures.

Short title: Multi-class Queues with Impatient Customers

1 Introduction

Motivated by customer contact centers, health-care systems, and telecommunication networks, in this paper, we analyze multi-class queueing systems with impatient customers. We focus on the non-preemptive priority policy while serving two classes of customers, and also provide some results for the first-come-first-served (FCFS) policy in multi-class systems. Under both policies, once the service for a customer starts, it cannot be preempted in favor of another customer. This appears to be a valid assumption for modeling customer contact centers, where service refers to customers and agents talking on the phone, or health-care systems, where service could be a surgical operation, or telecommunication networks, where service is receiving data or voice packets. In customer contact centers, the Automatic Call Distributor (ACD) can classify callers based on the type of service requested or information provided by the caller revealing if s/he can be considered a premier class customer. If callers, grouped in different classes by the ACD, wait on the line for too long before an agent handles their call, they can hang up and be lost to the system. In health-care systems, the type of health service requested or the severity of the condition of the patient can be used for classification and scheduling resources, yet, patients might either die or choose to go to another health-care facility if they cannot start with medical treatment in a reasonable time. In telecommunication networks, there can be hard deadlines to receive data beyond which the data becomes useless.

In the queueing literature, there is an extensive body of research on single or multi-server systems serving a single class of impatient customers, e.g., [3, 4, 16, 19, 28, 34, 36, 40]. In FCFS multi-class systems with c identical parallel servers, if the exponential service time distribution is the same for all k classes of customers that arrive according to independent Poisson processes, an appropriately constructed single class $M/M/c + H_k$ model (see Baccelli and Hebuterne, [3]) can be used if customers in each class have exponential reneging times. Here H_k denotes a k-stage

Hyperexponential random variable (r.v.) representing the time-to-abandon r.v. of the constructed single class of customers (see the end of Section 4). However, incorporating non-exponential distributions for interarrival, service time and time-to-abandon r.v.s, or considering priority disciplines, especially in systems serving more than two classes of customers, makes it difficult to construct tractable and exact analytical models; instead, one has to resort to simulation studies, e.g., [2]. When there are two classes of customers to consider, we can find some earlier analytical models. Choi, Kim, and Chung [13] study the $M/M/1$ queue where high-priority customers have deterministic impatience time and have preemptive priority over patient low-priority customers. Analyzing the underlying Markov process, they provide the joint distribution of the system size, and the Laplace transform (LT) of the response time of low-priority customers. Brandt and Brandt [8] study the $M/M/1$ system where high-priority customers, for whom general time-to-abandon distribution is assumed, are served under the preemptive-resume priority policy alongside patient low-priority customers. They derive the probability generating function of the joint queue size distribution and the LT of the waiting time of low-priority customers. Iravani and Balcıoğlu [20] study two problems in the $M/GI/1$ setting. In the first problem which considers the preemptive-resume priority discipline, both classes have exponentially distributed times-to-abandon. In the second problem, the non-preemptive priority is worked on, yet, the low-priority class is assumed to be patient. For both problems, they provide the LT of the virtual waiting time for both classes. The non-preemptive priority policy for impatient high-priority customers in multi-server queues has been studied by Brandt and Brandt [7], also by assuming patient low-priority customers. In this model, reneging high-priority customers, whose wait time is below a threshold value, become low-priority customers. They provide the exact waiting time and queue length distributions for high-priority customers as well as approximations for the factorial moments of the number of low-priority customers.

It appears that when non-preemptive priority discipline is taken into consideration, in previous research, impatience behavior is assumed only for high-priority customers. There are few exceptions to this, all in Markovian settings. Jouaini and Dallery [23] analyze a multi-server Markovian system with two classes, yet they assume identically distributed times-to-abandon (as well as identical service time distributions) for both classes. They study the corresponding birth-and-death process and derive the steady-state probabilities of the number of total customers, and high-priority customers in the system. Rozenshmidt [31] considers the same setting, this time with k-priority classes, and obtains the expected waiting time of each class. Wang [38] studies the non-preemptive priority $M/M/1$ queueing system where both classes have identically distributed service requirements, and exponential times-to-abandon with possibly different rates. He analyzes the two-dimensional Markov chain and provides approximations for the performance measures of the system. Stolletz and Helber [35] also construct a Markov chain in modeling a call center with skill-based routing, where a group of flexible servers – in addition to specialized server groups for each class – attends to two impatient classes of customers under the non-preemptive rule.

Even though in some cases prioritization of a specific class of customers can be due to their impatience, in many other settings, such as customer contact centers and health-care systems, different classes of customers can be impatient with different levels of tolerance for waiting in the queue. For instance, patients waiting for special medical treatments are usually categorized into priority groups according to their conditions so that, patients with higher risk of death (i.e., the most impatient customers), receive service first as the highest priority customers. In some other context, prioritization can be even irrelevant of the patience limit of customers. For example, many call centers give exclusive priority to their more valuable customers who are on a contract (e.g., [27, 33]), or are known, according to historical data, to be more profitable. Therefore, there is a need to incorporate impatience for low-priority customers in settings operating under the non-preemptive priority policy.

In this paper, we first analyze the two-class non-preemptive priority $M/GI/1+M$ queue in Section 2. We allow general and different service time distributions for each class as well as different reneging rates. Then, we extend the problem in Section 3 to two-class multi-server queues. We assume identical exponential service time distributions for both classes yet retain different rates for exponentially distributed patience limits in Section 3.1, and different deterministic patience limits in Section 3.2. Same service time distribution is a reasonable assumption, as exemplified by Milner and Olsen [27], in call centers where the service is the same for all classes but prioritization differentiates contract customers from those without a contract. In Section 3.1, we also visit the case with patient high-priority customers and low-priority customers that can abandon the system if their patience expires. Examples for this case are service firms that offer a guarantee to their special or VIP customers on the service delivery time [18, 33]; if this guarantee is not met, high-priority customers are not charged, which can lead high-priority customers to wait patiently. The non-VIP customers, who have to pay for service no matter how long they wait, can be deterred if they wait for too long and can renege from the system. Finally, in Section 4, we consider the FCFS policy when serving k classes of customers. Under the FCFS policy, we model the $M/GI/1$ queue with class specific service time distributions, yet, for tractability, we have to assume that the reneging rates are the same across all customer classes. In the extension to the multi-server case, we resort to identical service time distributions for each class. When the customer arrival rates are so high that the servers are busy almost all the time, the analysis can be carried out via fluid approximations. For instance, Talreja and Whitt [37] consider multiple groups of servers attending to different classes of customers on an FCFS basis. Assuming general service time distributions depending on the customer class and the server group, they develop fluid queueing models which accurately capture the steady-state routing flow rates which can be used as the input of a single class fluid model to approximate the performance measures of the system.

In all the problems we study, we obtain the LT of the virtual waiting time for all classes by employing the level-crossing technique due to to Brill [9, 11] (see also Brill [12]). This enables us to write down the integral equations for the density functions of the virtual waiting time for classes. Eventually, we obtain their LT's. Then, by relating the virtual waiting time to the actual waiting time, we are able to compute the steady-state performance measures of the queueing systems analyzed, such as the waiting time distributions with their mean values, and proportion of reneging customers for each class.

The rest of the paper is organized as follows. In Section 2, we analyze the $M/GI/1+M$ queue serving two classes of customers according to the non-preemptive priority rule. In Section 3, the analyses of the $M/M/c+M$ and $M/M/c+D$ queues operating under the non-preemptive priority policy are presented. Finally, in Section 4, we consider the FCFS policy in single and multi-server queues.

2 The Single Server Priority Queue with Impatient Customers

In this section, we model a queueing system in which a single server attends to two classes of impatient customers under the non-preemptive priority rule. High-priority (class 1) customers have to wait for the completion of the service time of the low-priority (class 2) customer they might see under service upon their arrival at the system. Upon completion of a service, only if there are no high-priority customers in the system, the next customer to serve is the first low-priority customer in line (provided that there are any low-priority customers waiting). For both classes, we assume that once their service starts, they are patient until their service is over. That is, reneging behavior is observed only among customers during their waiting time in the queue.

In this setting, class i customers arrive at the queueing system according to an independent Poisson process with rate λ_i, $i = 1, 2$. The independent and identically distributed (i.i.d.) service-time r.v. for class i is denoted by B_i, and it follows a general distribution function $B_i(x)$ (with $\overline{B}_i(x) = 1 - B_i(x)$ denoting the complementary service time distribution) with mean $1/\mu_i$. Furthermore, the Laplace-Stieltjes transform (LST) of the service-time distribution function is denoted by $\widetilde{b}_i(s)$. We assume that the time-to-abandon distribution for class i customers is exponential with rate ω_i. The service-time distributions and reneging rates of each class could be different from one another. Additionally, service-time and time-to-abandon r.v.s are independent of each other and the Poisson arrival processes.

Let $V_i(t)$ denote the virtual waiting time for class i customers at time $t \geq 0$ (the amount of time a class i customer arriving at time t would have to wait if its patience were infinite). A class i customer with patience R arriving at the system at time t must wait for the $\min\{V_i(t), R\}$, at the end of which it either reaches the server and its service commences if $V_i(t-) \leq R$, or abandons the queue otherwise, with its patience expired. We consider the virtual waiting time for class i when the system is in steady-state. Denote the steady-state probability density function of $V_i(t)$ by $f_i(x)$ with $x > 0$. We eventually obtain the LT of $f_1(x)$ and $f_2(x)$ denoted by $\tilde{f}_1(s)$ and $\tilde{f}_2(s)$, respectively, from which the steady-state waiting time distribution for each class can be computed. To do this, we employ the level-crossing theorem due to Brill [9, 11], Brill and Posner [10], Cohen [14], Cohen and Rubinovitch [15], and Shanthikumar [32]. This method briefly asserts that in steady-state, for every level x of the virtual waiting time, the value of the density function at x is equal to the rate of downcrossing level x, which is in turn equal to the rate of upcrossing level x.

We first carry out the analysis for high-priority customers. The following lemma provides the density function of the virtual waiting time for class 1 customers.

Lemma 1. *The density function of the virtual waiting time for high-priority (class*

1) customers in the $M/GI/1+M$ system satisfies the integral equation

$$\begin{aligned} f_1(x) &= \lambda_1\overline{B}_1(x)P_0 + \lambda_2\overline{B}_2(x)P_0 + \lambda_1\int_0^x \overline{B}_1(x-y)e^{-\omega_1 y}f_1(y)dy \\ &\quad +\lambda_2\overline{B}_2(x)\int_0^\infty e^{-\omega_2 y}f_2(y)dy, \qquad (1) \end{aligned}$$

where P_0 is the steady-state probability of finding the system empty satisfying the normalizing equation

$$P_0 + \int_0^\infty f_1(x)dx = 1.$$

Proof. According to the level-crossing theory, the rate of downcrossing level x on a sample path of the virtual waiting time of high-priority customers is equal to $f_1(x)$. The rate of upcrossing level x, as given on the right hand side (RHS), has four parts: The first two terms are the upcrossing rates of x due to high- and low-priority customers arriving at an empty system. The third term is the upcrossing rate caused by high-priority customers that must wait a positive time $0 < y < x$. That is, a "tagged" high-priority customer arriving at a busy system can cause an upcrossing if its patience is more than the amount of virtual waiting time y, and its service requirement is in excess of $x - y$. The last term on the RHS is the upcrossing rate caused by low-priority customers arriving at a busy system. These customers do not contribute to the virtual waiting time of high-priority customers until it is their turn to seize the server, i.e., when the virtual waiting time of class 1 customers hits level 0. Since the queue is stable due to reneging, all these customers reach the server in a finite amount of time if they do not abandon the queue. Therefore, a tagged low-priority customer that must wait for a positive time $y > 0$ causes an upcrossing, if it is patient enough to survive this wait, and if its service requirement exceeds x. Note that the virtual waiting time of class 2 customers with density function $f_2(x)$ comprises the amount of work due to patient high- and low-priority customers that the tagged low-priority customer finds in the system upon its arrival, plus, the additional work that patient high-priority customers bring in during the queue time of the tagged low-priority customer. □

To obtain the density function $f_2(x)$, we need the distribution of the busy period generated by both high- and low-priority customers arriving at an empty system. The busy period generated by a high-priority customer arriving at an idle system has a distribution function of $L_0(x)$ (with $\overline{L}_0(x) = 1 - L_0(x)$), the LST of which is given in Eq. (1) of Iravani and Balcıoğlu [20] who adapt the result which Rao [30] provides for an $M/GI/1 + M$ queue with a single class of impatient customers

We denote the r.v. of the busy period initiated by a low-priority customer arriving at an empty system by H. Let $H(x)$ and $\widetilde{h}(s)$ denote the distribution function and the LT of H, respectively. We refer the reader to Eq. (25) of [20] for $\widetilde{h}(s)$, which is used to numerically compute $H(x)$ (with $\overline{H}(x) = 1 - H(x)$).

Given this, the following lemma provides the density function of the virtual waiting time for class 2 customers.

Lemma 2. *The density function of the virtual waiting time for low-priority (class 2) customers in the $M/GI/1 + M$ system satisfies the integral equation*

$$f_2(x) = \lambda_1 \overline{L}_0(x) P_0 + \lambda_2 \overline{H}(x) P_0 + \lambda_2 \int_0^x \overline{H}(x-y) e^{-\omega_2 y} f_2(y) dy, \tag{2}$$

with the normalizing equation

$$P_0 + \int_0^\infty f_2(x) dx = 1.$$

Proof. We again employ the level crossing theorem. In Eq. (2), $f_2(x)$ on the LHS is the rate of downcrossing level x on a sample path of the virtual waiting time of low-priority customers. The rate of upcrossing level x, as given on the RHS, has three parts: The terms $\lambda_1 \overline{L}_0(x) P_0$ and $\lambda_2 \overline{H}(x) P_0$ give the upcrossing rates of x due to high- and low-priority customers arriving at an empty system, respectively. Such customers increase the virtual waiting time for low-priority customers by the busy period they generate. The last term is the upcrossing rate due to low-priority customers that must wait a positive time $0 < y < x$. Such customers also increase the virtual waiting time by a busy period if they can survive the offered waiting time

y. Note that the contribution of high-priority customers arriving at a busy system in the virtual waiting time of low-priority customers is already included in these three components presented on the LHS of Eq. (2). □

When we take the LT of both sides of Eqs. (1) and (2), we have

$$\begin{aligned} \widetilde{f}_1(s+\omega_1) - \frac{\widetilde{f}_1(s)}{\lambda_1\widetilde{\beta}_1(s)} &= -P_0 - \frac{\lambda_2\widetilde{\beta}_2(s)}{\lambda_1\widetilde{\beta}_1(s)}(P_0 + \widetilde{f}_2(\omega_2)), && (3) \\ \widetilde{f}_2(s+\omega_2) - \frac{\widetilde{f}_2(s)}{\lambda_2\widetilde{\kappa}(s)} &= -P_0 - \frac{\lambda_1\widetilde{g}_0(s)}{\lambda_2\widetilde{\kappa}(s)}P_0, && (4) \end{aligned}$$

where $\widetilde{\beta}_i(s)$ is the LT of the complementary service time distribution for class i, $\widetilde{\kappa}(s)$ is the LT of $\overline{H}(x)$, and $\widetilde{g}_0(s)$ is that of the $\overline{L}_0(x)$. The equations are inhomogeneous difference equations of the form

$$u(s+\omega) - a(s)u(s) = b(s), \quad \omega > 0,$$

studied in e.g., Jagerman [22] (p. 115). Given that the series is absolutely convergent, the general solution is given by (Jagerman [22] (p. 116))

$$u(s) = cv(s) - \sum_{j=0}^{\infty} \frac{b(s+j\omega)}{a(s)a(s+\omega)...a(s+j\omega)}, \tag{5}$$

where c is a constant and $v(s)$ is the solution of the corresponding homogeneous equation $v(s+\omega) - a(s)u(s) = 0$. In Appendix A, we prove the convergence of the series for Eqs. (3) and (4), and also show that $c = 0$ for both equations, which we later use in proposition 1.

Let $E[L_0]$ and $E[H]$ denote the expected length of busy periods initiated by high- and low-priority customers, respectively, which can be computed by using their corresponding LT's. We are now ready to present the LT of the virtual waiting time of each class in the following proposition.

Proposition 1. *In the $M/GI/1+M$ system, the LT of the virtual waiting time of high-priority (class 1) customers, $\widetilde{f}_1(s)$, and that of low-priority (class 2) customers,*

$\widetilde{f}_2(s)$, *are given, respectively, by*

$$\widetilde{f}_1(s) = \sum_{j=0}^{\infty}\left(P_0 + \frac{\lambda_2\widetilde{\beta}_2(s+j\omega_1)}{\lambda_1\widetilde{\beta}_1(s+j\omega_1)}\left(P_0+\widetilde{f}_2(\omega_2)\right)\right)\prod_{m=0}^{j}\lambda_1\widetilde{\beta}_1(s+m\omega_1). \quad (6)$$

$$\widetilde{f}_2(s) = \sum_{j=0}^{\infty}\left(P_0 + \frac{\lambda_1\widetilde{g}_0(s+j\omega_2)}{\lambda_2\widetilde{\kappa}(s+j\omega_2)}P_0\right)\prod_{m=0}^{j}\lambda_2\widetilde{\kappa}(s+m\omega_2), \quad (7)$$

where

$$P_0 = \left(\sum_{j=0}^{\infty}\left(1+\frac{\lambda_1\widetilde{g}_0(j\omega_2)}{\lambda_2\widetilde{\kappa}(j\omega_2)}\right)\prod_{m=0}^{j}\lambda_2\widetilde{\kappa}(m\omega_2)\right)^{-1}, \quad (8)$$

$$\widetilde{f}_2(\omega_2) = \frac{(1-P_0)-(\lambda_1E[L_0]+\lambda_2E[H])P_0}{\lambda_2E[H]}. \quad (9)$$

Proof. Eqs. (6) and (7) are simply obtained using the solution form in Eq. (A.44). Eq. (7) helps to find P_0, the probability of having an idle system: by letting $s \to 0$ in both sides of Eq. (7), and using the normalizing equation $P_0+\int_0^\infty f_2(x)dx = 1$, which implies that $\widetilde{f}_2(0) = 1-P_0$, we get Eq. (8). Finally, in order to obtain $\widetilde{f}_2(\omega_2)$, which appears on the RHS of Eq. (6), we let $s \to 0$ in Eq. (4), which gives Eq. (9), □

Given $\widetilde{f}_i(s)$ in Eqs. (7) and (6), by numerically inverting $\widetilde{f}_i(s)/s$ using techniques such as the ones due to Abate and Whitt [1] and Jagerman [21], one can compute $F_i(x) = P_0 + \int_0^x f_i(y)dy$, that is the virtual waiting time distribution for class i customers. Let W_i denote the steady-state waiting time of a class i customer in the queue. Also let S denote the event that a customer is successfully served, and A the event that a customer abandons the queue. We denote the probability that a type i customer is served (reneges) by $P_i(S)$ ($P_i(A)$). In the remainder of this section, we summarize how certain steady-state performance measures can be found (see, Stanford [34]).

The conditional distribution of waiting time given that a class i customer is eventually served is given by

$$P(W_i \le x|S) = \frac{P_0 + \int_0^x e^{-\omega_i t}f_i(t)dt}{P_i(S)}, \quad (10)$$

that has a mean of $E[W_i|S]$, which is

$$E[W_i|S] = \frac{\int_0^\infty xe^{-\omega_i x} f_i(x)dx}{P_i(S)}. \tag{11}$$

where

$$P_i(S) = 1 - P_i(A) = P_0 + \int_0^\infty e^{-\omega_i y} f_i(y)dy = P_0 + \widetilde{f}_i(\omega_i). \tag{12}$$

Note that in Eq. (10), $\int_0^x e^{-\omega_i t} f_i(t)dt$ can be computed by numerically inverting its LT, $\widetilde{f}_i(\omega_i + s)/s$. Having only the LT of $f_i(x)$ in hand, the numerical evaluation of the integral in Eq. (11) can be computationally demanding. In cases where $f_i(x)$ is directly available (see e.g., Eq. 24), this computation can be carried out efficiently.

Next, the waiting time distribution of an arbitrary (served or reneging) type i customer in the queue is

$$P(W_i \leq x) \;\; = \;\; 1 - e^{-\omega_i x} + e^{-\omega_i x} F_i(x), \tag{13}$$

and for its expected value we have

$$\begin{aligned} E[W_i] &= \int_0^\infty P(W_i > x)dx = \int_0^\infty e^{-\omega_i x} \overline{F}_i(x)dx \\ &= \frac{1 - \widetilde{f}_i(\omega_i) - P_0}{\omega_i} = \frac{P_i(A)}{\omega_i}. \end{aligned} \tag{14}$$

Finally, the conditional distribution of waiting time given that a class i customer eventually reneges, i.e., $P(W_i \leq x|A)$, can be obtained from

$$P(W_i \leq x) \;\; = \;\; P_i(S)P(W_i \leq x|S) + P_i(A)P(W_i \leq x|A), \tag{15}$$

and its mean $E[W_i|A]$ from

$$E[W_i] = P_i(S)E[W_i|S] + P_i(A)E[W_i|A]. \tag{16}$$

We close this section by demonstrating a nice relationship between P_0 and $P_i(S)$. If we integrate both sides of Eq. (1) on $(0, \infty)$, we have

$$\begin{aligned} 1 - P_0 &= \rho_1 P_0 + \rho_2 P_0 + \rho_1 \int_0^\infty e^{-\omega_1 y} f_1(y)dy + \rho_2 \int_0^\infty e^{-\omega_2 y} f_2(y)dy, \\ P_0 &= 1 - \rho_1 \left(P_0 + \int_0^\infty e^{-\omega_1 y} f_1(y)dy \right) - \rho_2 \left(P_0 + \int_0^\infty e^{-\omega_2 y} f_2(y)dy \right), \\ &= 1 - \rho_1 P_1(S) - \rho_2 P_2(S), \end{aligned}$$

where $\rho_1 = \lambda_1/\mu_1$ and $\rho_2 = \lambda_2/\mu_2$. This expression for P_0 is similar to $P_0 = 1-\rho_1-\rho_2$, the probability of finding an idle server in a two-class $M/GI/1$ queue with patient customers where all customers are served ($P_1(S) = P_2(S) = 1$) provided that $\rho_1+\rho_2 < 1$.

3 The multi-server non-preemptive priority queue with impatient customers

In this Section, we extend the model studied in Section 2 by assuming c identical and parallel servers. When service times are non-exponential r.v.s, an exact analysis of the $M/GI/c$ queue even with a single class of patient customers is not possible. Thus, we consider only exponential service times. Furthermore, we assume that the service time distribution with rate μ is the same for both classes. As in Section 2, high-priority (class 1) and low-priority (class 2) customers arrive in accordance with independent Poisson processes with rate λ_i, and renege from the system if they are not served before their patience expires. The service of a low-priority customer cannot be preempted. In order for a waiting low-priority customer to reach a server, there should not be any high-priority customers waiting in the queue. We study this system in two sections where models differ due to the time-to-abandon distributions assumed. In Section 3.1, we analyze the case in which both classes have exponential time-to-abandon distributions, and in Section 3.2, we study the "time-out problem" where customers have deterministic patience limits.

3.1 The two-priority class $M/M/c+M$ queue

In this Section, we assume that class i customers have exponentially distributed times-to-abandon with rate ω_i, $i = 1, 2$. This model can apply to call centers that can distinguish between customer classes with the help of the ACD such as the one analyzed via simulation by Saltzman and Mehrotra [33].

As in Section 2, we need the LT's of the virtual waiting time density functions for both classes. Before proceeding further, let P_j be the probability of having $j \leq c-1$ servers busy in steady-state. Since the number of busy servers is a birth-and-death process, P_j can be expressed as $P_0\rho^j/j!$, where $\rho = \lambda/\mu$, $\lambda = \lambda_1 + \lambda_2$, and P_0 is the steady-state probability of having all servers idle. Using this, we can express the probability of no wait for a customer (the probability that $c-1$ or fewer servers are busy) as

$$P(W=0) = \sum_{j=0}^{c-1} \frac{\rho^j}{j!} P_0,$$

and we have the following normalizing equation

$$P(W=0) + \int_0^\infty f_i(y)dy = 1, \quad i = 1, 2. \tag{17}$$

Next, employing the level-crossing theorem for the multi-server model, for high-priority customers, we have the following lemma.

Lemma 3. *The density function of the virtual waiting time for high-priority (class 1) customers in the $M/M/c+M$ system satisfies the integral equation*

$$\begin{aligned} f_1(x) &= \lambda P_{c-1} e^{-c\mu x} + \lambda_1 \int_0^x e^{-c\mu(x-y)} e^{-\omega_1 y} f_1(y)dy \\ &\quad + \lambda_2 e^{-c\mu x} \int_0^\infty e^{-\omega_2 y} f_2(y)dy. \end{aligned} \tag{18}$$

Proof. Note that Eq. (18) is very similar in spirit to Eq. (1). The differences are that arrivals that find fewer than $c-1$ servers do not contribute to the virtual waiting time, and the amount of contribution of those who find all servers busy and survive the offered waiting time is exponentially distributed with rate $c\mu$, which is the time until the next departure when all servers are busy. □

To obtain $f_2(x)$, we need the distribution of the busy period initiated by high- and low-priority customers. The busy-period starts when all servers become busy and ends when no high-priority customers are left in the system. Since service time distributions are the same for both classes, the distributions of busy periods generated by high- and low-priority customers are the same, which we denote by $L^c(x)$.

Observe that $L^c(x)$ is the same distribution as the distribution of the busy period in an $M/M/c{+}M$ queue receiving a single class of impatient customers according to a Poisson process with rate λ_1, where each server has a service rate of μ and customers a reneging rate of ω_1. In this single-class $M/M/c+M$ queue (similar to the $M/M/c$ queue with patient customers, see Daley and Servi, [17]), we define the busy period as the time starting from the instant when all servers become busy until we have $c-1$ servers busy again. Since all c servers of the single-class $M/M/c+M$ queue are busy during the busy period, clearing customers at a rate of $c\mu$, the distribution of its busy period is the same as that of an $M/M/1+M$ queue with λ_1 Poisson arrival rate, ω_1 reneging rate, and $c\mu$ service rate. Thus, we conclude that $L^c(x)$ is identical in distribution to the busy period of the $M/M/1+M$ queue and the LST of $L^c(x)$, which we denote by $\widetilde{l}^c(s)$, is found from Eq. (1) of [20] by substituting $\widetilde{b}_1(s) = c\mu/(c\mu+s)$. Now we employ the level-crossing theorem for the low-priority customers in the following lemma where Eq. (19) is obtained similar to Eq. (2), and thus, the proof is omitted to avoid repetitions.

Lemma 4. *The density function of the virtual waiting time for low-priority (class 2) customers in the $M/M/c+M$ system satisfies the integral equation*

$$f_2(x) \quad = \quad \lambda P_{c-1}\overline{L}^c(x) + \lambda_2 \int_0^x \overline{L}^c(x-y)e^{-\omega_2 y} f_2(y)dy. \tag{19}$$

If we take the LT of Eqs. (18) and (19), we get

$$\widetilde{f}_1(s+\omega_1) - \frac{c\mu+s}{\lambda_1}\widetilde{f}_1(s) \quad = \quad -\frac{\lambda}{\lambda_1}P_{c-1} - \frac{\lambda_2}{\lambda_1}\widetilde{f}_2(\omega_2), \tag{20}$$

$$\widetilde{f}_2(s+\omega_2) - \frac{\widetilde{f}_2(s)}{\lambda_2\widetilde{g}_0(s)} \quad = \quad -\frac{\lambda}{\lambda_2}P_{c-1}, \tag{21}$$

where $\widetilde{g}_0(s)$ is LT the $\overline{L}^c(x)$.

Let $E[L^c]$ denote the expected length of a busy period. From Perry and Asmussen [29] and Boxma el al. [6], we have

$$E[L^c] \quad = \quad \sum_{k=0}^{\infty} \frac{\lambda_1^k}{\prod_{j=0}^{k}(c\mu + j\omega_1)}. \tag{22}$$

The following proposition gives the density function of the virtual waiting time of class 1 customers, the LT of that of class 2 customers, and the probability of finding all servers idle.

Proposition 2. *In the $M/M/c+M$ system, the LT of the virtual waiting time for low-priority (class 2) customers is given by*

$$\widetilde{f_2}(s) = \frac{\lambda P_{c-1}}{\lambda_2} \sum_{j=0}^{\infty} \prod_{k=0}^{j} \lambda_2 \widetilde{g}_0(s + k\omega_2). \tag{23}$$

Also, the density function of the virtual waiting time of high-priority (class 1) customers is given by

$$f_1(x) = (\lambda P_{c-1} + \lambda_2 \widetilde{f_2}(\omega_2)) e^{\{-c\mu x + \lambda_1 (1 - e^{-\omega_1 x})/\omega_1\}}, \tag{24}$$

where

$$\widetilde{f_2}(\omega_2) = \frac{(1 - P(W=0)) - \lambda P_{c-1} E[L^c]}{\lambda_2 E[L^c]}, \tag{25}$$

and the probability of finding all servers idle is

$$P_0 \quad = \quad \left(\sum_{i=0}^{c-1} \frac{\rho^i}{i!} + \frac{\lambda}{\lambda_2} \frac{\rho^{c-1}}{(c-1)!} \sum_{j=0}^{\infty} \prod_{k=0}^{j} \lambda_2 \widetilde{g}_0(k\omega_2) \right)^{-1}. \tag{26}$$

Proof. As in Section 2, the solution of Eq. (21) can be found by using Eq. (A.44), which is presented in Eq. (23). We can write the solution for Eq. (20) with the help of Eq. (A.44) as

$$\widetilde{f_1}(s) = \frac{\lambda P_{c-1} + \lambda_2 \widetilde{f_2}(\omega_2)}{\lambda_1} \sum_{j=0}^{\infty} \prod_{k=0}^{j} \frac{\lambda_1}{c\mu + s + k\omega_1}. \tag{27}$$

Let

$$m(x) = \frac{e^{-c\mu x}}{j!} \left(\frac{1 - e^{-\omega_1 x}}{\omega_1} \right)^j,$$

which has the LT (Jagerman, [22], p. 122)

$$\widetilde{m}(s) = \prod_{k=0}^{j} \frac{1}{c\mu + s + k\omega_1}.$$

Then, using $m(x)$, we can explicitly invert the LT $\widetilde{f}_1(s)$ as

$$\begin{aligned} f_1(x) &= \frac{\lambda P_{c-1} + \lambda_2 \widetilde{f}_2(\omega_2)}{\lambda_1} \lambda_1 e^{-c\mu x} \sum_{j=0}^{\infty} \frac{1}{j!} \left(\frac{\lambda_1 (1 - e^{-\omega_1 x})}{\omega_1} \right)^j \\ &= (\lambda P_{c-1} + \lambda_2 \widetilde{f}_2(\omega_2)) e^{\{-c\mu x + \lambda_1 (1 - e^{-\omega_1 x})/\omega_1\}}. \end{aligned}$$

In order to obtain $\widetilde{f}_2(\omega_2)$, which appears on the RHS of Eq. (20), we let $s \to 0$ in Eq. (21), which gives Eq. (25).

The probability of finding all servers idle, P_0, is found by letting $s \to 0$ in Eq. (23), and using Eq. (17) with $P_{c-1} = \rho^{c-1} P_0/(c-1)!$, as

$$1 - P(W = 0) = \frac{\lambda}{\lambda_2} \frac{\rho^{c-1}}{(c-1)!} P_0 \sum_{j=0}^{\infty} \prod_{k=0}^{j} \lambda_2 \widetilde{g}_0(k\omega_2),$$

from which Eq. (26) is obtained.

□

We can use Eqs. (10), (12), and (14) by replacing P_0 with $P(W = 0)$ to compute $P(W_i \leq x|S)$, $P_i(A) = 1 - P_i(S)$, and $E[W_i]$. The other distribution functions and mean waiting times can be found from Eqs. (11), (13), (15), and (16).

We close this section by considering a special case in which high-priority customers are assumed to be patient. This model is relevant to service industry that offers a guarantee to their VIP customers on service delivery time: if this guarantee is not met, the service will be free for VIP customers (see Ho and Zheng,[18], and Saltzman and Mehrotra [33]). In this scenario, it is reasonable to assume that high-priority (VIP) customers are patient, because they know that, in the worst case, if they wait too long, they will not be charged any service fees. Since high-priority customers are patient, Eq. (18) becomes (the difference is the second term on the RHS)

$$f_1(x) = \lambda P_{c-1} e^{-c\mu x} + \lambda_1 \int_0^x e^{-c\mu(x-y)} f_1(y) dy + \lambda_2 e^{-c\mu x} \int_0^\infty e^{-\omega_2 y} f_2(y) dy,$$

that has a solution of

$$f_1(x) = (\lambda P_{c-1} + \lambda_2 \widetilde{f}_2(\gamma_2))e^{-(c\mu-\lambda_1)x}.$$

The busy period distribution required for $f_2(x)$ in Eq. (19) is identical to the busy period distribution in an ordinary $M/M/1$ queue with the same arrival rate and $c\mu$ as the service rate, which has the following LST (e.g., Kleinrock [24], p. 215)

$$\widetilde{l}^c(s) = \frac{\lambda_1 + c\mu + s - \sqrt{(\lambda_1 + c\mu + s)^2 - 4\lambda_1 c\mu}}{2\lambda_1},$$

from which its mean can be found as

$$E[L^c] \;=\; \frac{1}{c\mu - \lambda_1}.$$

3.2 The two-priority class $M/M/c + D$ queue

In this Section, we analyze the multi-server queueing system where the patience limit of customers are constants: if a class i customer does not start receiving service in τ_i time units after its arrival, it abandons the system without being served. The single-class version of this problem was studied by Boots and Tijms [5], Xiong, Jagerman and Altiok [39] and Liu and Kulkarni [25, 26]. Such models apply to telecommunication systems where data become useless if not received within a hard deadline.

To include two priority classes, we again employ the level-crossing theorem, and for high-priority customers we have the following lemma.

Lemma 5. *In the $M/Mc+D$ system, the density function of the virtual waiting time for high-priority (class 1) customers satisfies the following integral equation:*

$$\begin{aligned} f_1(x) \;&=\; \lambda P_{c-1}e^{-c\mu x} + \lambda_1 \int_0^{x\wedge\tau_1} e^{-c\mu(x-y)} f_1(y)dy \\ &\quad +\lambda_2 e^{-c\mu x}\int_0^{\tau_2} f_2(y)dy, \end{aligned} \tag{28}$$

where $a \wedge b = \min(a, b)$.

Proof. Eq. (28) is similar to Eq. (18) with the following differences: The second term on the RHS is the upcrossing rate caused by high-priority customers arriving at the system when all servers are busy. In this case, if $x > \tau_1$, a tagged high-priority customer can cause an upcrossing only if it arrives when the virtual waiting time is less than its patience, i.e., $0 < y < \tau_1$ (otherwise, it abandons the system). If $0 < x < \tau_1$, it suffices to arrive at a system with a virtual waiting time less than x, i.e., $0 < y < x$, because in this case, the customer will not abandon and receive service. The third term is the upcrossing rate caused by low-priority customers that must wait for a positive amount of time y. A tagged low-priority customer causes an upcrossing if it reaches the server and if the time until next departure after it reaches the server is in excess of x. □

To write the equation for $f_2(x)$, we need the busy period distribution $L^c(x)$ as in Section 3.1, this time considering deterministic patience limits. In the following lemma, we provide the LT of the busy period in this $M/M/c+D$ queue. First, let

$$\begin{aligned}
\alpha_1 &= \frac{(c\mu - s - \lambda_1) + \sqrt{(s+\lambda_1-c\mu)^2 + 4c\mu s}}{2}, \\
\alpha_2 &= \frac{(c\mu - s - \lambda_1) - \sqrt{(s+\lambda_1-c\mu)^2 + 4c\mu s}}{2}, \\
\gamma_i &= (c\mu - \alpha_i - \frac{\lambda c\mu}{s+c\mu})e^{-\alpha_i\tau_1}, \quad i = 1, 2.
\end{aligned}$$

Lemma 6. *The LT of the busy period in the $M/M/c+D$ queue is*

$$\widetilde{l}^c(s) = \frac{c\mu}{\gamma_1-\gamma_2}\left[\frac{\gamma_2\left(1-e^{-\tau_1(c\mu-\alpha_2)}\right)}{\alpha_2-c\mu} - \frac{\gamma_1\left(1-e^{-\tau_1(c\mu-\alpha_1)}\right)}{\alpha_1-c\mu} + \frac{(\gamma_1 e^{\alpha_1\tau_1} - \gamma_2 e^{\alpha_2\tau_1})e^{-c\mu\tau_1}}{c\mu+s}\right]. \tag{29}$$

Proof. Similar arguments made in Section 3.1 apply here and $L^c(x)$ is identically distributed as the busy period in a single-class $M/M/1+D$ queue with λ_1 as the arrival rate, $c\mu$ as the service rate, and τ_1 as the deterministic patience limit. We resort to Perry and Asmussen [29] and Liu and Kulkarni [25] who provide $\widetilde{l}_0(s,\xi)$, which is the conditional LT of the busy period in a single-class $M/M/1+D$ queue, given that the service time initiating the busy period is a constant ξ:

$$\widetilde{l}_0(s,\xi) = \begin{cases} \frac{\alpha_1 e^{\alpha_1 \xi} - \alpha_2 e^{\alpha_2 \xi}}{\gamma_1 - \gamma_2} & 0 \le \xi \le \tau_1, \\ e^{-s(\xi-\tau_1)} \widetilde{l}_0(s,\tau_1) & \xi > \tau_1. \end{cases}$$

For our problem, we need to remove the condition on the first service time:

$$\widetilde{l}^c(s) = \int_0^{\tau_1} \frac{\gamma_1 e^{\alpha_1 \xi} - \gamma_2 e^{\alpha_2 \xi}}{\gamma_1 - \gamma_2} c\mu e^{-c\mu\xi} d\xi + \int_{\tau_1}^{\infty} e^{-s(\xi-\tau_1)} \widetilde{l}_0(s,\tau_1) c\mu e^{-c\mu\xi} d\xi,$$

which, after some simplification, reduces to Eq. (29). □

Recalling that $\overline{L}^c(x)$ denotes the complementary distribution of the busy period, similar to Eq. (19), for low-priority customers we have the following result, which is presented without a proof.

Lemma 7. *The density function of the virtual waiting time for low-priority (class 2) customers in the $M/Mc+D$ queue satisfies the integral equation*

$$f_2(x) \;=\; \lambda P_{c-1} \overline{L}^c(x) + \lambda_2 \int_0^{x \wedge \tau_2} \overline{L}^c(x-y) f_2(y) dy. \tag{30}$$

Note that similar to Eq. (28), the second term on the RHS in Eq. (30) points out that the upcrossing rate caused by low-priority customers arriving at the system when all servers are busy depends on the level x.

Similar to Liu and Kulkarni [25], we introduce

$$k_1(x) \;=\; \overline{L}^c(x) + \lambda_2 \int_0^{x} \overline{L}^c(x-y) k_1(y) dy, \tag{31}$$

$$k_2(x) \;=\; \overline{L}^c(x) + \lambda_2 \int_0^{\tau_2} \overline{L}^c(x-y) k_1(y) dy, \tag{32}$$

which are used in the following proposition that gives the density functions of the virtual waiting times. We denote the probability that a low-priority customer arriving at the system when all servers are busy is eventually served by $P_2^c(S)$ (noting that $P(W=0) + P_2^c(S) = P_2(S)$), which also appears in proposition 3. By definition,

$$P_2^c(S) \;=\; \int_0^{\tau_2} f_2(y) dy = \lambda P_{c-1} \int_0^{\tau_2} k_1(x) dx. \tag{33}$$

We will elaborate on how to compute $P_2^c(S)$ in the proof of proposition 3.

Proposition 3. *In the $M/M/c+D$ queue, the density function of the virtual waiting time for high-priority (class 1) customers, $f_1(x)$, and that of low-priority customers, $f_2(x)$, are given, respectively, by*

$$f_1(x) = \begin{cases} \left(\lambda_2 P_2^c(S) + \lambda P_{c-1}\right) e^{-(c\mu-\lambda_1)x}, & x < \tau_1, \\ \left(\lambda_2 P_2^c(S) + \lambda P_{c-1}\right) e^{\lambda_1\tau_1} e^{-c\mu x}, & x \geq \tau_1, \end{cases} \tag{34}$$

$$f_2(x) = \begin{cases} \lambda P_{c-1} k_1(x), & x < \tau_2, \\ \lambda P_{c-1} k_2(x), & x \geq \tau_2, \end{cases} \tag{35}$$

where $P_{c-1} = \rho^{c-1}P_0/(c-1)!$, in which

$$P_0 \;=\; \frac{1 - \lambda_2 P_2^c(S)\left[\frac{1}{c\mu-\lambda_1}\left(1 - e^{-(c\mu-\lambda_1)\tau_1}\right) + \frac{1}{c\mu}e^{-(c\mu-\lambda_1)\tau_1}\right]}{\sum_{i=0}^{c-1}\frac{\rho^i}{i!} + \frac{\lambda\rho^{c-1}}{(c-1)!}\left[\frac{1}{c\mu-\lambda_1}\left(1 - e^{-(c\mu-\lambda_1)\tau_1}\right) + \frac{1}{c\mu}e^{-(c\mu-\lambda_1)\tau_1}\right]}. \tag{36}$$

Proof. We start with Eq. (30), the solution of which depends on the value of x through $x \wedge \tau_2$, and with which we can express $f_2(x)$ as in Eq. (35).

Using the boundary equation $P(W = 0) + \int_0^\infty f_2(x)dx = 0$, and the fact that $P_{c-1} = \rho^{c-1}P_0/(c-1)!$, we write

$$\begin{aligned} 1 - P(W = 0) &= \frac{\lambda\rho^{c-1}}{(c-1)!}P_0\left(\int_0^{\tau_2} k_1(x)dx + \int_{\tau_2}^\infty k_2(x)dx\right), \\ P_0 &= \left(\sum_{i=0}^{c-1}\frac{\rho^i}{i!} + \frac{\lambda\rho^{c-1}}{(c-1)!}\left(\int_0^{\tau_2} k_1(x)dx + \int_{\tau_2}^\infty k_2(x)dx\right)\right)^{-1}. \end{aligned}$$

In order to solve for $k_1(x)$, we take the LT of both sides of Eq. (31). Letting $\widetilde{k}_1(s)$ denote the LT of $k_1(x)$, and as in previous sections, $\widetilde{g}_0(s)$ the LT of $\overline{L}^c(x)$, we have

$$\widetilde{k}_1(s) = \frac{\widetilde{g}_0(s)}{1 - \lambda_2\widetilde{g}_0(s)}, \tag{37}$$

which can be used first to calculate $k_2(x)$, and then $f_2(x)$. Using Eq. (37), we can also obtain $P_2^c(S)$ in Eq. (33) where $\int_0^{\tau_2} k_1(x)dx$ can be computed by numerically inverting $\widetilde{k}_1(s)/s$. However, we need P_{c-1} that requires P_0, which in return calls for evaluating $\int_{\tau_2}^{\infty} k_2(x)dx$. As will be demonstrated shortly, this can be by-passed easily.

The same solution approach taken for $f_2(x)$ can be applied to solve Eq. (28), which yields Eq. (34).

To bypass computing $k_2(x)$, which requires inverting both $\widetilde{k}_1(s)$ and $\widetilde{g}_0(s)$ numerically, we first note that it is more convenient to use Eq. (17) for $f_1(x)$, i.e.,

$$\begin{aligned} P(W=0) + (\lambda_2 P_2^c(S) + \lambda P_{c-1}) \left[\int_0^{\tau_1} e^{-(c\mu-\lambda_1)x} dx + e^{\lambda_1 \tau_1} \int_{\tau_1}^{\infty} e^{-c\mu x} dx \right] &= 1, \\ P(W=0) + \left(\lambda_2 P_2^c(S) + \frac{\lambda \rho^{c-1}}{(c-1)!} P_0 \right) \left[\frac{1}{c\mu - \lambda_1} \left(1 - e^{-(c\mu-\lambda_1)\tau_1} \right) + \frac{1}{c\mu} e^{-(c\mu-\lambda_1)\tau_1} \right] &= 1, \end{aligned}$$

which, after simplification, gives us Eq. (36). Also from Eq. (33), we have

$$P_0 = \frac{P_2^c(S)}{\frac{\lambda \rho^{c-1}}{(c-1)!} \int_0^{\tau_2} k_1(x)dx}. \tag{38}$$

Equating Eqs. (36) and (38), we can obtain $P_2^c(S)$ and using either of these equations gives P_0. □

We close this section by relating the virtual and actual waiting time distributions, and explaining how to calculate the expected waiting times. The virtual waiting time distribution for class i is

$$F_i(x) = P(W=0) + \int_0^x f_i(y)dy.$$

Thus,

$$P(W_i \le x) = \begin{cases} F_i(x), & x < \tau_i, \\ 1, & x \ge \tau_i, \end{cases}$$

and for served customers

$$P(W_i \le x | S) = \begin{cases} F_i(x)/P_i(S) & x < \tau_i, \\ 1, & x \ge \tau_i, \end{cases}$$

where

$$P_i(S) = P(W=0) + \int_0^{\tau_i} f_i(x)dx.$$

One can see that waiting time distributions can be calculated by only numerically inverting $\widetilde{k}_1(s)/s$, without computing $k_2(x)$ in Eq. (32).

To compute the expected waiting times, we first note that $E[W_i|A] = \tau_i$. The conditional expected waiting time given that a type i customer is served is given by

$$E[W_i|S] = \frac{\int_0^{\tau_i} x f_i(x)dx}{P_i(S)}.$$

For low-priority customers using Eq. (35) this becomes

$$E[W_2|S] = \frac{\lambda P_{c-1} \int_0^{\tau_1} x k_1(x)dx}{P_2(S)},$$

which can be computed by inverting $\widetilde{k}_1(s)$ and numerically evaluating the integral. For high-priority customers, however, the computation can be carried out more easily using Eq. (34), and we have

$$\begin{aligned} E[W_1|S] &= \frac{(\lambda_2 P_2^c(S) + \lambda P_{c-1}) \int_0^{\tau_2} x e^{-(c\mu-\lambda_1)x} dx}{P_1(S)} \\ &= \frac{(\lambda_2 P_2^c(S) + \lambda P_{c-1}) \left(1 - e^{-\tau_2(c\mu-\lambda_1)}(1 + \tau_2(c\mu - \lambda_1)\right)}{P_1(S)(c\mu - \lambda_1)^2}. \end{aligned}$$

Finally, once we have $E[W_i|S], E[W_i|A]$ in hand we can use Eq. (16) to find $E[W_i]$.

4 FCFS queues with impatient customers

In this section, we model first the single server, then the multi-server FCFS queueing systems with k classes of impatient customers. Class i customers arrive according to a Poisson process with rate λ_i, have exponential times-to-abandon with rates ω_i, $i = 1, \dots, k$.

We start with the single server case. As before, let $\overline{B}_i(x)$ be the complementary service time distributions for class i customers, and $\widetilde{\beta}_i(s)$ its LT. The following lemma provides the density function of the virtual waiting time and its LT when the reneging rate is the same for all classes.

Lemma 8. *In the multi-class FCFS $M/GI/1+M$ system, the density function of the virtual waiting time for all classes of customers is the same and satisfies the integral equation*

$$f(x) \quad = \quad P_0 \sum_{i=1}^{k} \lambda_i \overline{B}_i(x) + \sum_{i=1}^{k} \lambda_i \int_0^x \overline{B}_i(x-y) e^{-\omega_i y} f(y) dy. \tag{39}$$

If the reneging rate ω is the same for all classes, the LT of the virtual waiting time is given by

$$\widetilde{f}(s) = P_0 \sum_{j=0}^{\infty} \prod_{m=0}^{j} \left(\sum_{i=1}^{k} \lambda_i \widetilde{\beta}_i(s+m\omega) \right), \tag{40}$$

where

$$P_0 = \left(1 + \sum_{j=0}^{\infty} \prod_{m=0}^{j} \left(\sum_{i=1}^{k} \lambda_i \widetilde{\beta}_i(m\omega) \right) \right)^{-1}.$$

Proof. Note that Eq. (39) is very similar to Eq. (1), except that we have more than two classes and do not have different virtual waiting time density functions for each class. Rather, $f(x)$, is the density function of the virtual waiting time for all classes of customers.

When we take the LT of both sides of Eq. (39), we have

$$\widetilde{f}(s) = P_0 \sum_{i=1}^{k} \lambda_i \widetilde{\beta}_i(s) + \sum_{i=1}^{k} \lambda_i \widetilde{\beta}_i(s) \widetilde{f}(s+\omega_i). \tag{41}$$

We are unable to solve for $\widetilde{f}(s)$ unless we assume the same reneging rate for each class. By setting $\omega_1 = \ldots = \omega_k = \omega$, Eq. (41) becomes

$$\widetilde{f}(s+\omega) - (\sum_{i=1}^{k} \lambda_i \widetilde{\beta}_i(s))^{-1} \widetilde{f}(s) = -P_0,$$

and its solution (after employing Eq. (A.44)) is found as in Eq. (40), and P_0 is obtained using the normalizing equation $P_0 + \widetilde{f}(0) = 1$. □

In the FCFS multi-server case with c servers, as in Section 3, each server has a rate μ, and we assume that independent service times are exponentially distributed. Let $\lambda = \sum_{i=1}^{k} \lambda_i$, and as before, $\rho = \lambda/\mu$.

Proposition 4. *In the multi-class FCFS $M/M/c+M$ system, the density function of the virtual waiting time for all classes is given by*

$$f(x) = Be^{\{\sum_{i=1}^{k} \frac{\lambda_i}{\omega_i}(1-e^{-\omega_i x})-c\mu x\}}, \tag{42}$$

where

$$B = \lambda P_{c-1} = \frac{\lambda \rho^{c-1}}{(c-1)!} P_0,$$

and

$$P_0 = \left(\sum_{j=0}^{c-1} \frac{\rho^j}{j!} + \lambda \frac{\rho^{c-1}}{(c-1)!} \int_0^\infty e^{\{\sum_{i=1}^{k} \frac{\lambda_i}{\omega_i}(1-e^{-\omega_i x})-c\mu x\}} dx \right)^{-1}.$$

Proof. Eq. (39) can be re-written as

$$f(x) = \lambda P_{c-1} e^{-c\mu x} + \sum_{i=1}^{k} \lambda_i \int_0^x e^{-c\mu(x-y)} e^{-\omega_i y} f(y) dy. \tag{43}$$

Note that $f(x)$ satisfies Eq. (17) where we substitute $f(y)$ instead of $f_i(y)$.

To solve Eq. (43), we take the derivative with respect to x, which gives us the following first order differential equation:

$$f^{'}(x) \equiv \frac{df(x)}{dx} = \sum_{i=1}^{k} \lambda_i e^{-\omega_i x} f(x) - c\mu f(x),$$

the solution of which has the form given in Eq. (42). Note that the constant B is found by setting $x = 0$ in Eqs. (43) and (42). Finally, P_0 can be computed from Eq. (17). □

With $\widetilde{f}(s)$ in Eq. (40) for the single server case or $f(x)$ in Eq. (42), we can use Eqs. (10)-(16) to compute the waiting time distributions, the fraction of customers served and the mean waiting times for each class (and by replacing P_0 with $P(W=0)$ in these equations for the multi-server case). Note that in the single-server case with identical reneging rates, all these measures (e.g., $E[W_i|S]$) will be the same for each class. The difference will be in the mean system time of served customers due to different mean service times added to $E[W_i|S]$.

An alternative approach to obtain $f(x)$ for the multi-server case is to use the single class $M/M/c+H_k$ model (see Baccelli and Hebuterne, [3]) with Poisson arrival rate λ and service rate μ where H_k is a k-stage Hyperexponential r.v., which is an exponential r.v. with rate ω_i with probability λ_i/λ.

5 Conclusion

In this paper, we model delay systems that are inspired from health-care systems, customer contact centers, and communication networks. In all systems, multiple classes of impatient customers are served. The first system involves a single server attending to two classes of impatient customers in accordance with the non-preemptive priority policy. Our contribution is incorporating general and class-specific service time distributions in this setting. The second system has identical parallel servers receiving classes of impatient customers that require the same type of service. Although we assume the same exponential service time distribution for all customers, we permit different classes to exhibit different times-to-abandon characteristics by assuming class-specific deterministic patience limits or class-specific reneging rates when patience times follow exponential distributions. The third system is operating under the FCFS rule. Although, we introduce general and class-specific service time distributions for multiple classes of customers, we can design a tractable solution only when all customers have the same exponentially distributed patience limits. In all models, we employ the level-crossing technique to express the virtual waiting time density functions and obtain their LT's. We relate these transforms to the classical system performance measures.

Acknowledgements

This work was supported in part by Natural Sciences and Engineering Research Council (NSERC) of Canada. The authors thank the anonymous referee who provided invaluable suggestions to improve the manuscript.

References

[1] Abate, J., & Whitt, W. (1995). Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal of Computing* 7: 36–43.

[2] Ahghari, M., & Balcıoğlu, B. (2009). Benefits of cross-training in a skill-based routing contact center with priority queues and impatient customers. *IIE Transaction* 41: 524–536.

[3] Baccelli, F., & Hebuterne, G. (1981). On queues with impatient customers. In F.J. Kylstra (ed.), *Performance '81.* North Holland, Amsterdam 1981), pp. 159–179.

[4] Baccelli, F., Boyer, P., & Hebuterne, G. (1984). Single-server queues with impatient customers. *Advanve Applied Probability* 16: 887–905.

[5] Boots, N. K., & Tijms, H. (1999). A multiserver queueing system with impatient customers. *Management Science* 45: 444–448.

[6] Boxma, O., Perry, D., Stadje, W., & Zacks, S. (2010). The busy period of an $M/G/1$ queue with customer impatience. *Journal of Applied Probability* 47: 130–145.

[7] Brandt, A., & Brandt, M. (1999). On a two-queue priority system with impatience and its applications to a call center. *Methodology and Computing in Applied Probability* 1: 191–210.

[8] Brandt, A., & Brandt, M. (2004). On the two-class $M/M/1$ system under preemptive resume and impatience of the prioritized customers. *Queueing Systems* 47: 147–168.

[9] Brill, P. H. (1975). *System Point Theory in Exponential Queues*, Ph.D. Thesis, Department of Industrial Engineering, University of Toronto.

[10] Brill, P. H., & Posner, M. J. M. (1977). Level crossings in point processes applied to queues: Single-server case. *Operations Research* 25: 662–674.

[11] Brill, P. H. (1979). An embedded level crossing technique for dams and queues. *Journal of Applied Probability* 16: 174–186.

[12] Brill, P. H. (2008). *Level Crossing Methods in Stochastic Models.* Springer.

[13] Choi, B. D., Kim, B., & Chung, J. (2001). $M/M/1$ queue with impatient customers of higher priority. *Queueing Systems* 38: 49–66.

[14] Cohen, J. W. (1977). On up- and downcrossings. *Journal of Applied Probability* 4: 405–410.

[15] Cohen, J. W., & Rubinovitch, M. (1977). On level crossings and cycles in dam processes. *Mathematics of Operations Research* 2: 297–310.

[16] Daley, D. J. (1965). General customer impatience in the queue $GI/G/1$. *Journal of Applied Probability* 2: 186–205.

[17] Daley, D. J., & Servi, L. D. (1998). Idle and busy periods in stable $M/M/k$ queues. *Journal of Applied Probability* 35: 950–962.

[18] Ho, T., & Zheng, Y. S. (2004). Setting customer expectations in service delivery: An integrated marketing-operations perspective. *Management Science* 50: 479–488.

[19] Iravani, F., & Balcıoğlu, B. (2008a). Approximations for the $M/GI/N+GI$ type call center. *Queueing Systems* 58: 137–153.

[20] Iravani, F., & Balcıoğlu, B. (2008b). On priority queues with impatient customers. *Queueing Systems* 58: 239–260.

[21] Jagerman, D. L. (1982). An inversion technique for the Laplace transform. *Bell System Technical Journal.* 61: 1995–2002.

[22] Jagerman, D. L. (2000). *Difference Equations with Applications to Queues*, Marcel Dekker, Inc., New York.

[23] Jouini, O., & Dallery, Y. (2007). Stationary delays for a two-class priority queue with impatient customers. *Proceedings of the 2nd International Conference on Performance Evaluation Methodologies and Tools*, Nantes, France.

[24] Kleinrock, L. (1975). *Queueing Systems Volume I: Theory*, John Wiley & Sons, New York.

[25] Liu, L., Kulkarni, V. G. (2008a). Busy period analysis for $M/PH/1$ queues with workload dependent balking. *Queuing Systems*, 59: 37–51.

[26] Liu, L., & Kulkarni, V. G. (2008). Balking and reneging in $M/G/s$ system exact analysis and approximations. *Probability in the Engineering and Informational Sciences* 22: 355–371.

[27] Milner, J. M., & Olsen, T. L. (2008). Service level agreements in call centers: perils and prescriptions. *Management Science* 54: 238–252.

[28] Palm, C. (1953). Methods of judging the annoyance caused by congestion. *Tele* 2: 1–20.

[29] Perry, D., & Asmussen, S. (1995). Rejection rules in the $M/G/1$ queue. *Queueing Systems* 19: 105–130.

[30] Rao, S. S. (1967). Queueing with balking and reneging in $M/G/1$ systems. *Metrika* 12: 173–188.

[31] Rozenshmidt, L. (2008). *On priority queues with impatient customers: stationary and time-varying analysis.* Master's Thesis, Technion - Israel Institute of Technology, Haifa, Israel.

[32] Shantikumar, J. G. (1981). On level crossing analysis of queues. *The Australian Journal of Statistics* 23: 337–342.

[33] Saltzman, S. M., & Mehrotra, V. (2001). A call center uses simulation to drive strategic change. *Interfaces* 31: 87–101.

[34] Stanford, R. E. (1979). Reneging phenomenon in single channel queues. *Mathematics of Operations Research* 4: 162–178.

[35] Stolletz, R., & Helber, S. (2004). Performance analysis of an inbound call center with skills-based routing. *OR Spectrum* 26: 331–352.

[36] Takács, L. (1974). A single server queue with limited virtual waiting time. *Journal of Applied Probability.* 11: 612–617.

[37] Talreja, R., & Whitt, W. (2008). Fluid models for overloaded multiclass many-server queueing systems with first-come, first-served routing. *Management Science* 54: 1513–1527.

[38] Wang. Q. (2004). Modeling and analysis of high risk patient queues. *European Journal of Operational Research* 155: 502–515.

[39] Xiong, W., Jagerman, D. L., & Altiok, T. (2008). $M/G/1$ queue with deterministic reneging times. *Performance Evaluation* 65: 308–316.

[40] Zeltyn, S., & Mandelbaum, A. (2005). Call centers with impatient customers: many-server asymptotics of the $M/M/n+G$ queue. *Queueing Systems* 51: 361–402.

Appendix A The solution form and the convergence of the series for Eqs. (3) & (4)

In this Appendix, we will show that the solution for the difference equations given in Eqs. (3) and (4) is of the form

$$u(s) = \sum_{j=0}^{\infty} \frac{-b(s+j\omega)}{a(s)a(s+\omega)...a(s+j\omega)}, \tag{A.44}$$

and that the series is uniformly convergent in s.

We start by showing the uniform convergence of the series in the solution of Eq. (4). Assuming the existence of the LT's $\widetilde{g}_0(s)$ and $\widetilde{\kappa}(s)$, we note that $\widetilde{g}_0(s) = (1-\tilde{l}_0(s))/s$, and $\widetilde{\kappa}(s) = (1-\tilde{h}(s))/s$, where s has a positive real part, i.e., $\mathcal{R}e(s) > 0$. Identifying $a(s) = 1/(\lambda_2\widetilde{\kappa}(s))$ and $b(s) = -P_0(1+\lambda_1\widetilde{g}_0(s))/(\lambda_2\widetilde{\kappa}(s))$ from Eq. (4), the jth term of the series in Eq. (5), which we denote by $U_j(s)$, is

$$\begin{aligned}
U_j(s) &= \frac{P_0 + \lambda_1\widetilde{g}_0(s+j\omega)(\lambda_2\widetilde{\kappa}(s+j\omega))^{-1}P_0}{[\lambda_2\widetilde{\kappa}(s)\lambda_2\widetilde{\kappa}(s+\omega)...\lambda_2\widetilde{\kappa}(s+j\omega)]^{-1}} \\
&= P_0\lambda_2^{(j+1)}\prod_{m=0}^{j}\widetilde{\kappa}(s+m\omega) + P_0\left[\lambda_1\widetilde{g}_0(s+j\omega)\lambda_2^j\prod_{m=0}^{j-1}\widetilde{\kappa}(s+m\omega)\right].
\end{aligned}$$

Now observe that for s with $\mathcal{R}e(s) > 0$, and all $m \geq 0$

$$\widetilde{\kappa}(s+m\omega) = \int_0^\infty e^{-(s+m\omega)x}\overline{H}(x)dx \leq \int_0^\infty e^{-(m\omega)x}\overline{H}(x)dx = \widetilde{\kappa}(m\omega),$$

and similarly $\widetilde{g}_0(s+m\omega) \leq \widetilde{g}_0(m\omega)$. Hence, letting

$$M_j = P_0\lambda_2^{(j+1)}\prod_{m=0}^{j}\widetilde{\kappa}(m\omega) + P_0\left[\lambda_1\widetilde{g}_0(j\omega)\lambda_2^j\prod_{m=0}^{j-1}\widetilde{\kappa}(m\omega)\right],$$

we have $U_j(s) \leq M_j$ for all s. We now claim that $\sum_{j=0}^{\infty} M_j$ is absolutely convergent.

Using the ratio test and noting that $\widetilde{h}(s) \to 0$ and $\widetilde{l}_0(s) \to 0$ as $\mathcal{R}e(s) \to \infty$, we have

$$\begin{aligned}
\limsup_{j\to\infty}\left|\frac{M_{j+1}}{M_j}\right| &= \limsup_{j\to\infty} \frac{P_0\lambda_2^{(j+2)}\prod_{m=0}^{j+1}\widetilde{\kappa}(m\omega) + P_0\left[\lambda_1\widetilde{g}_0((j+1)\omega)\lambda_2^{(j+1)}\prod_{m=0}^{j}\widetilde{\kappa}(m\omega)\right]}{P_0\lambda_2^{(j+1)}\prod_{m=0}^{j}\widetilde{\kappa}(m\omega) + P_0\left[\lambda_1\widetilde{g}_0(j\omega)\lambda_2^{j}\prod_{m=0}^{j-1}\widetilde{\kappa}(m\omega)\right]}\\
&= \limsup_{j\to\infty} \frac{\lambda_2^2\widetilde{\kappa}(j\omega)\widetilde{\kappa}((j+1)\omega) + \lambda_1\widetilde{g}_0((j+1)\omega)\lambda_2\widetilde{\kappa}(j\omega)}{\lambda_2\widetilde{\kappa}(j\omega) + \lambda_1\widetilde{g}_0(j\omega)}\\
&= \limsup_{j\to\infty} \frac{\lambda_2^2\frac{1-\widetilde{h}(j\omega)}{j\omega}\frac{1-\widetilde{h}((j+1)\omega)}{(j+1)\omega} + \lambda_1\lambda_2\frac{1-\widetilde{l}_0((j+1)\omega)}{(j+1)\omega}\frac{1-\widetilde{h}(j\omega)}{j\omega}}{\lambda_2\frac{1-\widetilde{h}(j\omega)}{j\omega} + \lambda_1\frac{1-\widetilde{l}_0(j\omega)}{j\omega}}\\
&= \limsup_{j\to\infty} \frac{\lambda_2^2(1-\widetilde{h}(j\omega))\frac{1-\widetilde{h}((j+1)\omega)}{(j+1)\omega} + \lambda_1\lambda_2\frac{1-\widetilde{l}_0((j+1)\omega)}{(j+1)\omega}(1-\widetilde{h}(j\omega))}{\lambda_2(1-\widetilde{h}(j\omega)) + \lambda_1(1-\widetilde{l}_0(j\omega))}\\
&= 0 < 1,
\end{aligned}$$

and hence the series is indeed absolutely convergent. Therefore, since $U_j(s) \leq M_j$ for all s, by Weierstrass M-Test, the series $\sum_{j=0}^{\infty} U_j(s)$ is uniformly convergent in s. Similarly, by choosing $a(s) = 1/(\lambda_1\widetilde{\beta}_1(s))$ and $b(s) = -(P_0 + \lambda_2\widetilde{\beta}_2(s)(P_0 + \widetilde{f}_2(\omega_2)))/(\lambda_1\widetilde{\beta}_1(s))$ from Eq. (3), the jth term of the series in Eq. (5), which we denote by $Q_j(s)$, is obtained. Then, one can show that $\sum_{j=0}^{\infty} Q_j(s)$ is uniformly convergent in s as well. Finally, we observe that for any $j \geq 0$, $\lim_{s\to\infty} U_j(s) = \lim_{s\to\infty} Q_j(s) = 0$, and since both series are uniformly convergent in s, we also have $\lim_{s\to\infty}\sum_{j=0}^{\infty} U_j(s) = \lim_{s\to\infty}\sum_{j=0}^{\infty} Q_j(s) = 0$. Therefore, the boundary conditions $\tilde{f}_1(\infty) = \tilde{f}_2(\infty) = 0$, are satisfied for $c = 0$, and hence by uniqueness of the density functions $f_1(x), f_2(x)$ (and their LT's) we can conclude that the solution of Eqs. (3) and (4) has the form given in Eq. (A.44).