On multiple priority multi-server queues with impatience

Oualid Jouini, Alex Roubos

On Multiple-Priority Multi-Server Queues with Impatience

Oualid Jouini[1] & Alex Roubos[2]

[1]Ecole Centrale Paris, Laboratoire Génie Industriel,

Grande Voie des Vignes, 92290 Châtenay-Malabry, France

[2]VU University Amsterdam, Department of Mathematics,

De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

oualid.jouini@ecp.fr, a.roubos@vu.nl

May 8, 2014

***Journal of the Operational Research Society**, 65:616-632, 2014.*

Abstract

We consider Markovian multi-server queues with two types of impatient customers: high- and low-priority ones. The first type of customers has a non-preemptive 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 or she will abandon and be lost. We consider two cases where the discipline of service within each customer type is FCFS or LCFS. For each type of customers, we focus on various performance measures related to queueing delays: unconditional waiting times, and conditional waiting times given service and given abandonment. The analysis we develop holds also for a priority queue with mixed policies, i.e., FCFS for the first type and LCFS for the second one, and vice versa. We explicitly derive the Laplace-Stieltjes transforms of the defined random variables. In addition we show how to extend the analysis to more than two customer types. Finally we compare FCFS and LCFS and gain insights through numerical experiments.

Keywords: multi-server queues; queueing delays; abandonment; non-preemptive priority; FCFS; LCFS; Laplace-Stieltjes transforms.

1 Introduction

In this paper, we analyze queueing systems with multiple types of impatient customers. Customer abandonment (or also reneging) is an important feature in a wide variety of situations that may be encountered in telecommunication systems, manufacturing systems, and service systems such as call centers and health care systems. Theoretical models incorporating abandonment are therefore closer to reality and necessary to obtain more accurate analysis. Another important feature in practice is the differentiation in the service given to different customer types. A priority mechanism is 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 additionally known for their ease of implementation, explaining their prevalence in practice. Much of the queueing literature is devoted to analyzing priority queues. Most papers are restricted to two priority types. 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. On the other hand, a priority discipline is said to be non-preemptive if there is no interruption. A customer with higher priority just goes to the head of the queue and waits for his or her turn.

We consider a Markovian multi-server queueing system with two types of impatient customers: high- and low-priority ones. The high-priority type has non-preemptive priority over the other type. We assume common exponential distributions for service times as well as patience times for both customer types. We analyze two different systems by considering different disciplines of service within each queue. The discipline of service of a queue refers to the manner by which customers are selected for service when a queue has formed. The most common discipline that can be observed in everyday life is first-come first-served (FCFS). Some other in common usage are random order of service (ROS) and last-come first-served (LCFS), which is applicable to many inventory systems when it is easier to reach the nearest stored items which are the last in. In this paper, we consider FCFS and LCFS policies and derive various performance measures related to queueing delays. Our approach is based on the use of Laplace-Stieltjes transforms and on the characterization of the

virtual waiting time of a "virtual" infinitely patient customer. We also describe the procedure to extend the analysis to more than two customer types.

Our motivation for considering identical statistical behavior of customer types (service and patience times) relates to the type of models that motivate our analysis. We are considering firms where customers are segmented into different groups based on their value to the firm. This segmentation can be based on lifetime value or profitability. The company then provides different levels of service to these groups. This type of service-level differentiation is widely used in financial service, telecommunication call centers, and more. In the presence of this type of segmentation, the difference between customer types is not related to the statistical behavior of customers but to their importance for the company, which we capture through priorities. In concrete terms, we assume for our models that customer behavior and queries do not differ from one type to another. This is a reasonable assumption for such systems, see Zeltyn et al. (2009).

In what follows, we review some of the queueing literature related to this paper. We distinguish two streams of literature. The first deals with queueing models with impatient customers. The second focuses on priority queues. The literature on queueing models with abandonments focuses especially on performance evaluation. The importance of modeling abandonments in call centers is emphasized by Garnett et al. (2002), Gans et al. (2003), and Mandelbaum and Zeltyn (2009). Empirical evidence regarding abandonments in call centers can be found in Brown et al. (2005) and Feigin (2006). We refer the reader to Garnett et al. (2002), and references therein, for simple models assuming exponential patience. Garnett et al. (2002) suggest an asymptotic analysis of their Markovian abandonment model under the heavy-traffic regime. Their main result is to characterize the relation between the number of servers, the offered load, and system performance measures such as the probability of delay and the probability to abandon. This can be seen as an extension of the results of Halfin and Whitt (1981) by adding abandonments. A number of approximations for the probability to abandon are developed by Boxma and de Waal (1994). The authors have considered a multi-server queue with generally distributed service times and patience times. Brandt and Brandt (1999, 2002) consider a state-dependent Markovian multi-server queue with generally distributed patience times, in which the arrival rate depends on the number of customers in the system and

in which the service rate depends on the number of busy servers. They derive the steady-state distribution of the number of customers in the system and various waiting-time distributions. The impact of the patience distribution on the performance is studied by Mandelbaum and Zeltyn (2004). They observe an approximate linearity between the abandonment probability and the average waiting time. To analyze multi-server queues with generally distributed service times and patience times, Whitt (2005) develops an algorithm to compute approximations for standard steady-state performance measures. One of his conclusions is that the behavior of the patience distribution near the origin primarily affects the performance. Iravani and Balcıoğlu (2008a) propose two approximations that are based on scaling the single-server queue to obtain estimates for the waiting-time distributions. Other papers have treated the impatience phenomenon under various assumptions. Related studies include those by Baccelli and Hebuterne (1981), Altman and Borovkov (1997), Ward and Glynn (2003), and references therein.

Let us now briefly mention some of the literature dealing with priority queueing systems. We refer the reader to Davis (1966) and Kella and Yechiali (1985) for a simple Markovian non-preemptive queue where all customer types have the same service-time distribution. Wagner (1997) considers multi-server non-preemptive priority systems with a Markovian arrival process, service times having phase type distributions, and both cases of finite and infinite queueing spaces are considered. Other references considering more complicated models, but where abandonments are not allowed, include those by Kao and Wilson (1999), Takine (1999), and Sleptchenko (2003). As for preemption schemes, we refer the reader to Harchol-Balter et al. (2005), Sleptchenko and van der Heijden (2005), and references therein. Sleptchenko and van der Heijden (2005) derive 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 expected values of service times. Harchol-Balter et al. (2005) introduce a new technique to reduce the Markov chain dimensionality of an $M/PH/s$ model with an arbitrary number of preemptive-resume priority types. Some research on priority queues has been dedicated to systems with mixed priorities that combine the two disciplines (with and without preemption). Results for the single-server case can be found in Drekic and Stanford (2000), and for those in the multi-server case, we refer the reader to Zeltyn et al. (2009).

Although the two features of abandonment and priority have each received attention separately, there is limited literature that deals with both of them. We refer the reader to Choi et al. (2001), where the authors derive several performance measures for an $M/M/1$ queue with two types of impatient customers in which type 1 customers have impatience of constant duration, and type 2 customers have no impatience and low priority level. An extension of the latter model is addressed by Brandt and Brandt (2004) for general distributed patience times. For a healthcare application, Wang (2004) considers a single-server non-preemptive priority queue with two classes of impatient customers, exponential service times with identical rates, exponential patience times with possibly different rates, and FCFS policy for each customer type. He proposes an approximation for the probability to have an idle server, which allows to compute the expected values of the queue lengths and the unconditional waiting times. Rozenshmidt (2007) considers a similar model to ours (under FCFS) and derives expressions for the unconditional expected waiting times of all customer types. Here we extend that analysis by considering additional performance measures, by considering also LCFS, and by computing all moments of the random variables. We also refer the reader to an interesting paper by Iravani and Balcıoğlu (2008b), where the authors analyze different priority models: single-server models with general service times, and multi-server models with exponential service times and a call-back option. A more recent paper by Sarhangian and Balcıoğlu (2011) considers different priority models similar to Iravani and Balcıoğlu (2008b). One of their models is similar to the one analyzed in this paper. What is different is that they consider two customers types possibly with different abandonment rates, but only the FCFS policy for each customer type. They employ the level crossing technique to derive various performance measures as those analyzed in this paper. Our main contributions can be summarized as follows.

- We compute the Laplace-Stieltjes transforms of various random variables related to queueing delays: unconditional waiting times, and conditional waiting times given service and given abandonment. We do so for both high- and low-priority customers. Our approach is based on the computation of virtual waiting times. One can then easily numerically invert the Laplace-Stieltjes transforms in order to obtain the cumulative distribution functions of these random variables at any point of time, see Abate and Whitt (2006).

- The analysis is detailed for two different non-preemptive priority models. One where the discipline of service within each class is FCFS, and another one working under LCFS. Moreover, the analysis we develop holds for a priority queue with mixed policies, i.e., FCFS for the first type and LCFS for the second one, and vice versa. We also extend our approach to the case of more than two customer types.

- We numerically compare between the effects of the FCFS and LCFS policies on performance, and provide some insights to call center managers using such type of prioritization. A concrete motivation of this work, coming from the call center industry, is that of the Bouygues Telecom call center. Bouygues Telecom is a French mobile phone company where customers are grouped into different classes as a function of the type of their mobile phone contracts. In other words, customers are grouped into classes based on the monthly amount they are paying to Bouygues Telecom. All customers from all classes ask the same type of questions, and the segmentation is not related to their behavior. As mentioned above, it is then appropriate in such cases to assume models where customer classes are statistically identical in terms of service and abandonment times (Zeltyn et al. 2009).

The remainder of this paper is structured as follows. In Sections 2.1 and 2.2, we describe the basic two-class queueing models, and define the performance measures of interest, respectively. In Section 2.3, we then develop some preliminary results that would help us in the rest of the analysis. In Section 3.1, we provide the results of performance evaluation when high- and low-priority customers are served under the FCFS basis. Those when high- and low-priority customers are served under the LCFS basis are given in Section 3.2. In Section 3.3, we explain how the analysis can be extended to the case of more than two classes. In order to illustrate the results and compare between FCFS and LCFS, we give some numerical experiments in Section 3.4. Finally in Section 4, we provide some concluding remarks and directions for future research.

2 Preliminaries

We first describe the two basic multi-class queues (for FCFS and LCFS) that we will analyze in this paper. Then, we provide the definitions of the performance measures of interest. The performance measures are related to the queueing delays of customers. Finally, we present some preliminary derivations that we will need along the way.

2.1 Modeling

Consider a queueing model with two types of customers: important customers denoted by type 1, and less important ones denoted by type 2. The model consists of two infinite-buffer queues for types 1 and 2, 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. Newly arriving customers of types 1 and 2 are assigned to queues 1 and 2, respectively. Customers of type 1 (waiting in queue 1) have priority over customers of type 2 (waiting in queue 2) in the sense that agents are providing assistance to type 1 customers first. The priority rule is non-preemptive, which simply means that a server currently serving a type 2 customer, while a new type 1 customer enters the system, will complete this service before turning to the queue 1 customer. Within each queue, we consider two cases for the discipline of service: FCFS and LCFS. Arrival processes of types 1 and 2 follow a Poisson process with rates λ_1 and λ_2, respectively. Let λ be the total arrival rate, $\lambda = \lambda_1 + \lambda_2$. Successive service times are assumed to be independent and identically distributed (i.i.d.), and follow a common exponential distribution with rate μ for both customer types.

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 the customer will abandon. Times before abandonment, for both customer types, are assumed to be i.i.d. and exponentially distributed with a common rate denoted by γ. We describe patience times by the random variable T. Finally, retrials are ignored, and abandonment is not allowed once a customer starts service. Following similar arguments, the behavior of the system can be viewed as a two-class

$M/M/s+M$ queueing system. The resulting model where the policy for each queue is FCFS (LCFS) is referred to as $\text{Model}_{\text{FCFS}}$ ($\text{Model}_{\text{LCFS}}$). Note that owing to abandonments, $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$ are unconditionally ergodic.

2.2 Notation

We denote by m the type of a customer, $m \in \{1, 2\}$. During the stationary regime, we define the following performance measures for $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$. To simplify the notations, we will not add indices to these quantities in order to refer to one of the models (we will add a clarification comment when it is necessary). In the remainder of this paper we refer to a customer as a she.

- W is the unconditional queueing delay of an arbitrary customer (regardless of her type).
- W_m is the unconditional queueing delay of a type m customer.
- $W_{m,s}$ is the conditional queueing delay of a type m customer, given that she will enter service.
- $P_{m,s}$ is the probability that a type m customer enters service.
- $W_{m,r}$ is the conditional queueing delay of a type m customer, given that she will abandon.
- $P_{m,r}$ is the probability that a type m customer abandons.
- $W_{m,d}$ is the conditional queueing delay of a type m customer, given that she has to wait.
- P_d is the probability of delay, i.e., the probability that a new arrival has to wait. Since $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$ are work conserving, P_d is independent of the customer type.
- $W_{m,d,s}$ is the conditional queueing delay of a type m customer, given that she was queued and that she will enter service. (We do not define a similar quantity for abandoned customers, since an abandoned customer is necessarily a delayed customer.)
- $P_{m,d,s}$ is the probability that a type m customer waiting in the queue will enter service.

To clarify the numerous definitions, we depicted in Figure 1 a schema of the performance measures of interest.

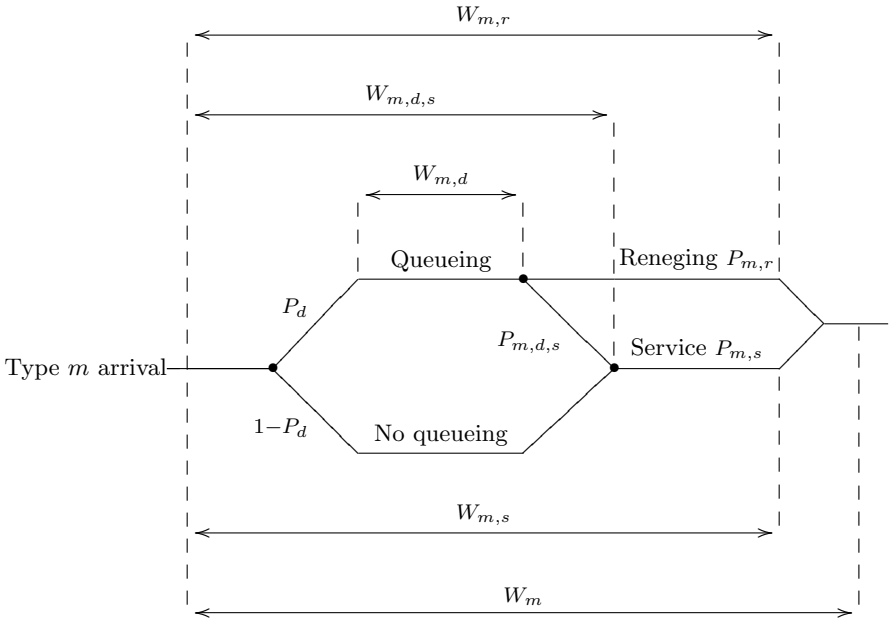

Figure 1. Performance measures for a type m customer.

In what follows, we provide some relations between the performance measures. For the remainder of the paper, we denote by $\mathbb{E}X^k$ the k-th order moment of a given random variable X, for $k \geq 1$. We also denote by $f_X(\cdot)$ and $F_X(\cdot)$ the probability density function (pdf) and the cumulative distribution function (cdf) of X. A customer who does not abandon will necessarily enter service, then $P_{m,s} + P_{m,r} = 1$. A customer who joins the queue has two possibilities: either she abandons, or she gets service, so $P_d = P_{m,r} + P_{m,d,s}$. Since the arrival processes are Poisson, the probability that a new arrival is of type m is λ_m/λ. Therefore,

$$\mathbb{E}W^k = \frac{\lambda_1}{\lambda}\mathbb{E}W_1^k + \frac{\lambda_2}{\lambda}\mathbb{E}W_2^k,$$

for $k \geq 1$. For type m customers, one may write

$$\mathbb{E}W_m^k = P_{m,s}\mathbb{E}W_{m,s}^k + P_{m,r}\mathbb{E}W_{m,r}^k, \tag{1}$$

for $k \geq 1$. Upon arrival, a customer is immediately addressed by one of the available servers, if any. If not, she has to wait and joins one of the queues (with probability P_d). Thus,

$$\mathbb{E}W_m^k = P_d\mathbb{E}W_{m,d}^k, \tag{2}$$

for $k \geq 1$. For customers that join the queue, we have

$$\mathbb{E}W^k_{m,d} = P_{m,d,s}\mathbb{E}W^k_{m,d,s} + P_{m,r}\mathbb{E}W^k_{m,r}, \tag{3}$$

which allows to determine $\mathbb{E}W^k_{m,d,s}$, for $k \geq 1$.

2.3 Preliminary Analysis

We start by computing the stationary probability distributions of the system states for $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$. At a given instant t, we denote by $n_1(t)$, $n_2(t)$, and $n(t) = n_1(t) + n_2(t)$ the number of type 1 customers in queue 1, that of type 2 in queue 2, and the total in both queues, respectively. Computing the stationary distribution of the process $\{n_2(t), t \geq 0\}$ or $\{(n_1(t), n_2(t)), t \geq 0\}$ is a complicated task. We only consider the processes $\{n_1(t), t \geq 0\}$ and $\{n(t), t \geq 0\}$ which are sufficient for the derivation of the performance measures. Recall that all stationary probabilities exist due to the ergodicity condition (which holds for any $\gamma > 0$).

Patience times are memoryless. Thus, as long as the scheduling policy within each queue is work conserving, the number of type 1 customers and type 2 customers in the system remain unchanged. Moreover, since patience as well as service times are identically distributed for both customer types, a work-conserving policy (priority between the queues or not) does not affect the total number of customers in the system. The following analysis holds for both $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$.

Let us consider the process $\{n(t), t \geq 0\}$. With regard to the total number of customers in the system ($\text{Model}_{\text{FCFS}}$ or $\text{Model}_{\text{LCFS}}$), our system is equivalent to a multi-server queue with a single type of customers. The arrival process is Poisson with intensity $\lambda = \lambda_1 + \lambda_2$. Hence, this system corresponds to the basic $M/M/s + M$ queueing system. The stationary probability distribution of i customers in the system, denoted by π_i for $i \geq 0$, is given by

$$\pi_i = \begin{cases} \dfrac{\lambda^i}{\mu^i i!}\pi_0, & 0 \leq i \leq s, \\ \dfrac{\lambda^i}{\mu^s s! \prod_{j=1}^{i-s}(s\mu + j\gamma)}\pi_0, & i > s, \end{cases}$$

with

$$\pi_0^{-1} = \sum_{i=0}^{s} \frac{\lambda^i}{\mu^i i!} + \sum_{i=s+1}^{\infty} \frac{\lambda^i}{\mu^s s! \prod_{j=1}^{i-s}(s\mu + j\gamma)}.$$

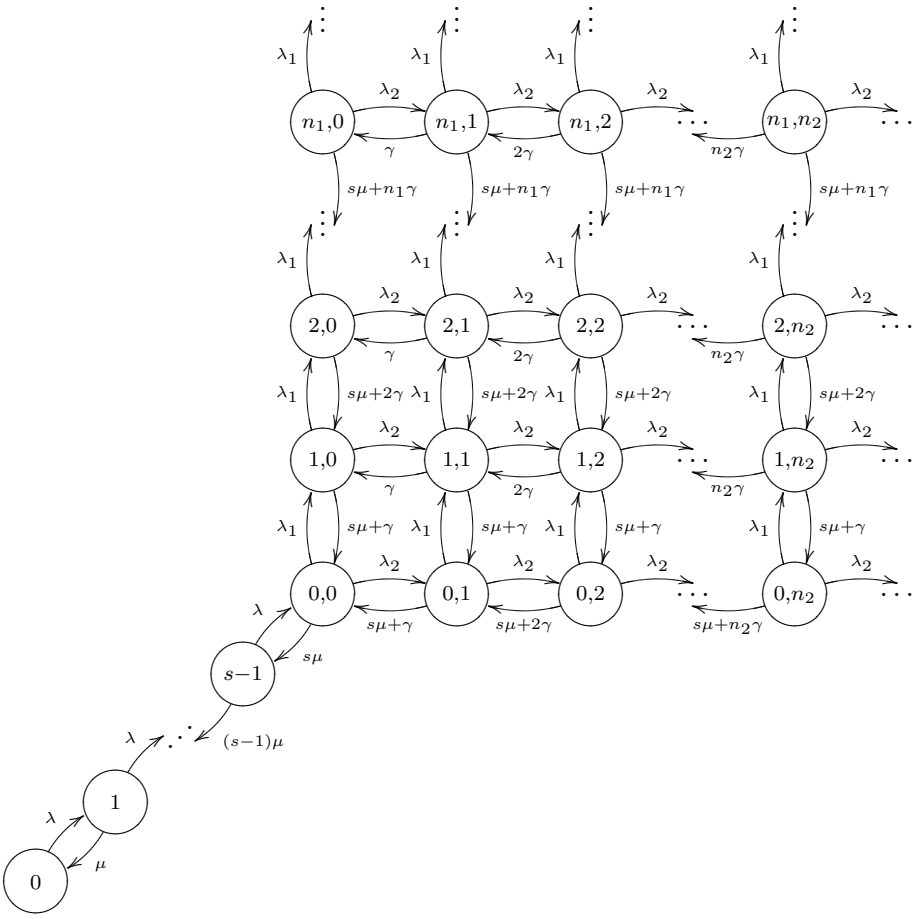

Figure 2. Markov chain for the number of customers in the queue.

Denote by $p(i)$ the stationary probability that all servers are busy and there are i customers in total in both queues, i.e., $p(i) = \pi_{s+i}$, for $i \geq 0$.

The probability of delay P_d is simply the probability that a new arrival finds all servers busy. It is then independent of the type of the new arrival. Moreover since the arrival process of a type m customer follows a Poison process, we use the PASTA property to state that the stationary probabilities seen by a new arrival coincide with those seen at an arbitrary instant. Thus, P_d is given by

$$P_d = 1 - \sum_{i=0}^{s-1} \pi_i.$$

Let us now characterize the stationary distribution of $\{n_1(t), t \geq 0\}$. To do so, we consider a two-dimensional Markov chain as shown in Figure 2. The state of the system is defined by the total

number of customers in the system (regardless of their type) if less than s customers are in the system (i.e., all customers are in service), and defined by the couple (n_1, n_2) denoting the number of queued customers of each type if s customers or more are in the system (i.e., all servers are busy). Let $p_1(i)$ denote the stationary probability that all servers are busy and i type 1 customers are in queue 1. By assembling all the states of each line in Figure 2, the balance equations lead to

$$p_1(i) = \frac{\lambda_1^i}{\prod_{j=1}^{i}(s\mu + j\gamma)} p_1(0), \tag{4}$$

for $i \geq 0$. To compute $p_1(0)$, we come back to the process $\{n(t), t \geq 0\}$. It is clear that the probability to be in state i, for $0 \leq i \leq s-1$, in the Markov chain of Figure 2 is equivalent to π_i. Next, the normalization condition gives

$$\sum_{i=0}^{s-1} \pi_i + \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} p_{1,2}(i,j) = 1, \tag{5}$$

where $p_{1,2}(i,j)$ is the stationary probability that all servers are busy, i type 1 customers are in queue 1, and j type 2 customers are in queue 2. Observe now that

$$p_1(i) = \sum_{j=0}^{\infty} p_{1,2}(i,j), \tag{6}$$

for $i \geq 0$. Combining thereafter Equations (4)–(6) leads to

$$p_1(0) = \left(1 - \sum_{i=0}^{s-1} \pi_i\right) \left(\sum_{i=0}^{\infty} \frac{\lambda_1^i}{\prod_{j=1}^{i}(s\mu + j\gamma)}\right)^{-1}.$$

Having in hand $p_1(i)$ and $p(i)$, for $i \geq 0$, the expected length of queue 1, say Q_1, and that of both customer types waiting in both queues, say Q, are therefore given by

$$Q_1 = \sum_{i=1}^{\infty} i p_1(i), \quad \text{and} \quad Q = \sum_{i=1}^{\infty} i p(i). \tag{7}$$

As a consequence, the stationary expected length of queue 2, say Q_2, is simply deduced by $Q_2 = Q - Q_1$.

We are now ready to compute the stationary probability to abandon and that to enter service for a new type m arrival. The probability $P_{m,r}$ can be viewed as the fraction of the stationary expected rate of type m abandoned customers over that of type m arrivals, seen at the epoch of a new type m arrival. Using PASTA and the memoryless property of patience times, we deduce that

the stationary expected rate of type m abandoned customers is γQ_m. So,

$$P_{m,r} = \frac{\gamma Q_m}{\lambda_m}.$$

The probability to enter service is only the complementary probability (no possible events of blocking or balking). Indeed, a customer who does not abandon will necessarily enter service,

$$P_{m,s} = 1 - P_{m,r}.$$

Finally, we also have

$$P_{m,d,s} = P_d - P_{m,r}.$$

3 Analysis of Queueing Delays

In this section, we characterize the distributions of the random variables W_m, $W_{m,d}$, $W_{m,s}$, $W_{m,r}$ and $W_{m,d,s}$. We do so by computing their k-th order moments, for $k \geq 1$. Although the stationary probabilities of the number of type m customers in the system, as well as P_d, $P_{m,s}$, $P_{m,r}$ and $P_{m,d,s}$ (computed in Section 2.3) are independent of the scheduling policy within each queue, the random variables of queueing delays do depend on the policy (FCFS or LCFS). We separately address the analyses for $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$ in Sections 3.1 and 3.2, respectively.

Our approach is based on the computation of first-passage times in various birth-death processes. As we will prove below, many of these random variables are equivalent to the length of an n-busy period in an FCFS $M/M/s+M$ queue, for $n \geq 0$. For $n \geq 1$, an n-busy period is defined as the elapsed time from the arrival of a customer to a busy $M/M/s+M$ system with $n-1$ waiting customers in the queue (n customers in the queue including the new arrival) until the epoch at which one server becomes idle. The 0-busy period reduces to the classical busy-period definition defined to begin with the arrival of a customer to a system with $s-1$ busy servers and to end when again one server becomes idle. We denote the length of an n-busy period by $BP_{n,\lambda}$, for $n \geq 0$. For an FCFS $M/M/1+M$ queue, one can obtain from Rao (1967) or Iravani and Balcıoğlu (2008b) the Laplace-Stieltjes transform of the pdf of $BP_{n,\lambda}$. Next, using Jouini (2012, Lemma 1) to state that the busy-period distribution is unchanged for all work-conserving policies, substituting the

expected service rate of a busy $M/M/1+M$ queue, μ, by that of an $M/M/s+M$ queue, $s\mu$, and denoting the Laplace-Stieltjes transform of the pdf of $BP_{n,\lambda}$ (for an $M/M/s+M$ queue with any work-conserving policy) by $\widetilde{F}_{BP_{n,\lambda}}(x)$, we obtain

$$\widetilde{F}_{BP_{n,\lambda}}(x) = \frac{\frac{s\mu}{x+s\mu} + \sum_{i=1}^{\infty}(-1)^i \left[\prod_{j=0}^{i-1}\left(1-\frac{s\mu}{x+s\mu+j\gamma}\right)\right]\frac{s\mu}{x+s\mu+i\gamma}\Theta(n,i)}{1+\sum_{i=1}^{\infty}\frac{\lambda^i}{i!\gamma^i}\left[\prod_{j=0}^{i-1}\left(1-\frac{s\mu}{x+s\mu+j\gamma}\right)\right]}, \tag{8}$$

with

$$\Theta(n,i) = \begin{cases} \displaystyle\sum_{j=0}^{i}\frac{(-1)^j\lambda^j}{j!\gamma^j}\binom{n}{i-j}, & 1 \le i \le n, \\ \displaystyle\sum_{j=i-n}^{i}\frac{(-1)^j\lambda^j}{j!\gamma^j}\binom{n}{i-j}, & i > n, \end{cases}$$

for $x \in \mathbb{R}^+$, and $n \geq 0$. We will later use Equation (8) to analyze queueing delays for low-priority customers in $\text{Model}_{\text{FCFS}}$, and both customer types in $\text{Model}_{\text{LCFS}}$. The analysis for high-priority customers in $\text{Model}_{\text{FCFS}}$ is in turn simpler by extending existing results in the literature.

3.1 Analysis of $\text{Model}_{\text{FCFS}}$

For high- and low-priority customers, we compute the k-th order moment of $W_{m,s}$ and $W_{m,r}$, which also allows to derive the k-th order moment of W_m, $W_{m,d}$ and $W_{m,d,s}$, for $k \geq 1$ and $m \in \{1,2\}$.

High-Priority Customers

Using an approach originally inspired by Whitt (1999a), Jouini et al. (2011a) derive all moments of $W_{1,s}$ and $W_{1,r}$ in the case of a finite multi-server queue with a single type of impatient customers. Here we further extend that approach for our priority queue. Consider a new type 1 arrival who finds all servers busy and n_1 waiting customers ahead of her in queue 1, $n_1 \geq 0$. It goes without saying that for the remaining cases (at least one server is idle), our customer will immediately enter service. Because of their lower priority, type 2 customers already waiting in queue 2, as well as those who will arrive later, will not affect the sojourn time in the queue of our new type 1 customer. Using Jouini et al. (2011a), we obtain

$$\mathbb{E}W_{1,s}^k = \frac{1}{P_{1,s}}\sum_{n_1=0}^{\infty} p_1(n_1)\Psi_{n_1+1}\mathbb{E}Y_{n_1+1}^k,$$

with

$$\Psi_{n_1} = \prod_{i=1}^{n_1} \left(1 - \frac{\gamma}{s\mu + i\gamma}\right) = \frac{s\mu}{s\mu + n_1\gamma},$$

for $n_1 \geq 1$, and Y_{n_1}, a random variable, is the summation of n_1 independent exponential distributions with parameters $s\mu + \gamma, s\mu + 2\gamma, \ldots, s\mu + n_1\gamma$. So, all moments of Y_{n_1} may be derived in a closed form. For example, its first two moments are

$$\mathbb{E}Y_{n_1} = \sum_{j=1}^{n_1} \frac{1}{s\mu + j\gamma}$$

and

$$\mathbb{E}Y_{n_1}^2 = \sum_{j=1}^{n_1} \frac{1}{(s\mu + j\gamma)^2} + \left(\sum_{j=1}^{n_1} \frac{1}{s\mu + j\gamma}\right)^2,$$

respectively, for $n_1 \geq 1$

Let us now focus on deriving $\mathbb{E}W_{1,r}^k$. For a new type 1 arrival who finds at least one idle server, $W_{1,r}$ is zero. Assume she is queued with n_1 waiting customers and that she will abandon while waiting in the queue. Let Z_{n_1+1} denote the random variable measuring her sojourn time in the queue before abandonment. Removing the condition on n_1, we obtain

$$\mathbb{E}W_{1,r}^k = \frac{1}{P_{1,r}} \sum_{n_1=0}^{\infty} p_1(n_1)\mathbb{E}Z_{n_1+1}^k.$$

Note that computing the moments of Z_{n_1}, for $n_1 \geq 1$, again involves summations of independent exponential random variables, and are easy to obtain. One may see that the probability to abandon at position j, for $1 \leq j \leq n_1$, is

$$\frac{\gamma}{s\mu + j\gamma} \prod_{l=j+1}^{n_1} \left(1 - \frac{\gamma}{s\mu + l\gamma}\right) = \frac{\gamma}{s\mu + n_1\gamma}.$$

Knowing that our customer will abandon at position j, the time to abandon, say $Z_{n_1}(j)$, is the sum of $n_1 - j + 1$ independent exponential random variables with parameters $s\mu + n_1\gamma, s\mu + (n_1 - 1)\gamma, \ldots, s\mu + j\gamma$. Averaging over all possibilities leads to

$$\mathbb{E}Z_{n_1}^k = \frac{\gamma}{s\mu + n_1\gamma} \sum_{j=1}^{n_1} \mathbb{E}Z_{n_1}^k(j).$$

For example, the expected value of Z_{n_1} may simply be written as

$$\mathbb{E}Z_{n_1} = \frac{1}{s\mu + n_1\gamma} \sum_{j=1}^{n_1} \frac{j\gamma}{s\mu + j\gamma}.$$

Using the results above combined with Equations (1)–(3), we obtain all moments of the random variables W_1, $W_{1,d}$ and $W_{1,d,s}$.

Low-Priority Customers

Our approach to derive the performance measures of type 2 customers is based on computing their virtual waiting time. Recall that the virtual waiting time is defined as the waiting time of an infinitely patient customer. For a new type 2 customer, we denote it by V_2. In what follows, we compute the k-th order moment of $W_{2,s}$ and $W_{2,r}$. Using the latter, all remaining performance measures are easily obtained thereafter.

Let us focus on the conditional waiting time of a type 2 customer given service, namely $W_{2,s}$. Recall that patience times are described by the random variable T. We have

$$F_{W_{2,s}}(t) = \frac{\mathbb{P}(V_2 < t, V_2 < T)}{\mathbb{P}(V_2 < T)},$$

for $t \geq 0$. First, observe that $\mathbb{P}(V_2 < T) = P_{2,s}$. Second, $\mathbb{P}(V_2 < t, V_2 < T) = \int_0^t e^{-\gamma x} f_{V_2}(x)\mathrm{d}x$. A new type 2 arrival who finds at least one idle server with probability $1 - P_d$, will immediately enter service. If not, assume that she finds $n = n_1 + n_2$ waiting customers. Thus, we may write for $t \geq 0$

$$\mathbb{P}(V_2 < t, V_2 < T) = (1 - P_d) \cdot 1 + \int_0^t e^{-\gamma x} \sum_{n=0}^{\infty} p(n) f_{V_{2,n}}(x)\mathrm{d}x, \tag{9}$$

where $V_{2,n}$ is the conditional virtual waiting time of a new type 2 customer, given that upon arrival she finds in total n waiting customers in both queues. Her virtual waiting time is not affected by all future type 2 arrivals because the discipline of service within queue 2 is FCFS. However, all future type 1 arrivals have to be considered because of their higher priority. Note also that this virtual waiting time does not depend on the couple (n_1, n_2) but on the total number of customers ahead of her $n = n_1 + n_2$ (common distribution of service and patience times for both customer types). As a consequence $V_{2,n}$ can be seen as the first-passage time at state -1 starting at state n in the birth-death process as shown in Figure 3. It is the time to empty the queue ahead of our customer and in addition one server becomes idle to handle her.

By considering a single-class $M/M/s + M$ queue (with mean arrival rate λ_1), one may see that $V_{2,n}$ is equivalent to the duration of an n-busy period, for $n \geq 0$. Let us denote the latter by BP_{n,λ_1},

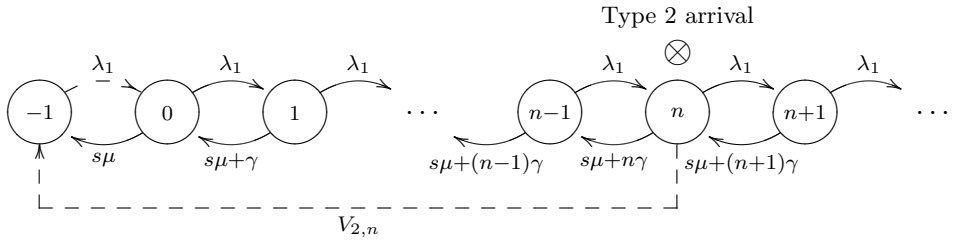

Figure 3. Virtual waiting time of a type 2 arrival finding n customers in queues 1 and 2, FCFS.

$V_{2,n} \equiv BP_{n,\lambda_1}$, for $n \geq 0$. Equation (9) then becomes

$$F_{W_{2,s}}(t) = \frac{1}{P_{2,s}} \left\{ 1 - P_d + \int_0^t e^{-\gamma x} \sum_{n=0}^{\infty} p(n) f_{BP_{n,\lambda_1}}(x) \mathrm{d}x \right\}, \tag{10}$$

for $t \geq 0$. Taking the derivative in t on both sides of Equation (10), we obtain

$$f_{W_{2,s}}(t) = \frac{1}{P_{2,s}} \sum_{n=0}^{\infty} p(n) e^{-\gamma t} f_{BP_{n,\lambda_1}}(t), \tag{11}$$

for $t \geq 0$. For the rest of the paper, we denote by $\widetilde{F}_X(x)$, for $x \in \mathbb{R}^+$, the Laplace-Stieltjes transform of the pdf $f_X(\cdot)$ of a random variable X. Note that the Laplace-Stieltjes transform of $e^{-\gamma t} f_{BP_{n,\lambda_1}}(t)$ is $\widetilde{F}_{BP_{n,\lambda_1}}(x+\gamma)$, for $x \in \mathbb{R}^+$. Applying next the Laplace-Stieltjes transform to Equation (11) implies

$$\widetilde{F}_{W_{2,s}}(x) = \frac{1}{P_{2,s}} \sum_{n=0}^{\infty} p(n) \widetilde{F}_{BP_{n,\lambda_1}}(x+\gamma), \tag{12}$$

for $x \in \mathbb{R}^+$. Using Equation (12), one can obtain any k-th order moment of $W_{2,s}$, for $k \geq 1$. It is given by

$$(-1)^k \frac{\mathrm{d}^k \widetilde{F}_{W_{2,s}}(x)}{\mathrm{d}x^k}\bigg|_{x=0},$$

for $k \geq 1$. Thus

$$\mathbb{E}W_{2,s}^k = \frac{(-1)^k}{P_{2,s}} \sum_{n=0}^{\infty} p(n) \widetilde{F}^{(k)}_{BP_{n,\lambda_1}}(\gamma),$$

where $h^{(k)}(\cdot)$ denotes the k-th derivative of a function $h(\cdot)$, for $k \geq 1$.

Let us now focus on the conditional waiting time of a type 2 customer given abandonment, $W_{2,r}$. We have

$$F_{W_{2,r}}(t) = \frac{\mathbb{P}(T < t, V_2 > T)}{\mathbb{P}(V_2 > T)},$$

for $t \geq 0$. First, observe that $\mathbb{P}(V_2 > T) = P_{2,r}$. Second, we may write

$$\mathbb{P}(T < t, V_2 > T) = \int_0^t \gamma e^{-\gamma x}(1 - F_{V_2}(x))\mathrm{d}x,$$

for $t \geq 0$. As a consequence, we obtain after some algebra

$$F_{W_{2,r}}(t) = \frac{1}{P_{2,r}} \left\{ 1 - e^{-\gamma t} - \int_0^t \gamma e^{-\gamma x} \left(1 - P_d + \sum_{n=0}^{\infty} p(n) F_{BP_{n,\lambda_1}}(x) \right) \mathrm{d}x \right\}, \tag{13}$$

for $t \geq 0$. Taking the derivative in t on both sides of Equation (13) leads to

$$f_{W_{2,r}}(t) = \frac{\gamma}{P_{2,r}} \left(P_d e^{-\gamma t} - e^{-\gamma t} \sum_{n=0}^{\infty} p(n) F_{BP_{n,\lambda_1}}(t) \right), \tag{14}$$

for $t \geq 0$. Using that the Laplace-Stieltjes transform of $F_{BP_{n,\lambda_1}}(t)$ is $\frac{1}{x}\widetilde{F}_{BP_{n,\lambda_1}}(x)$, for $x \in \mathbb{R}^+$, and applying the Laplace-Stieltjes transform to Equation (14) implies

$$\widetilde{F}_{W_{2,r}}(x) = \frac{\gamma}{P_{2,r}(x+\gamma)} \left(P_d - \sum_{n=0}^{\infty} p(n) \widetilde{F}_{BP_{n,\lambda_1}}(x+\gamma) \right),$$

for $x \in \mathbb{R}^+$. This finishes the characterization of $W_{2,s}$ and $W_{2,r}$. One can now use Equations (1)–(3) to obtain all moments of the remaining random variables W_2, $W_{2,d}$ and $W_{2,d,s}$.

To close the discussion, we note that one can obtain the expected queue lengths Q_1 and Q_2 (given by Equation (7)) by using the expressions for $\mathbb{E}W_1$ and $\mathbb{E}W_2$ derived in this section and by applying Little's law, $\lambda_m \mathbb{E}W_m = Q_m$, for $m \in \{1, 2\}$.

3.2 Analysis of Model$_{\mathbf{LCFS}}$

Similarly to the previous Section, we compute here for Model$_{\mathrm{LCFS}}$ the k-th order moment of $W_{m,s}$ and $W_{m,r}$, which also allows to derive the k-th order moment of W_m, $W_{m,d}$ and $W_{m,d,s}$, for $k \geq 1$ and $m \in \{1, 2\}$. We use the same approach based on the computation of the virtual waiting time of high- and low-priority customers.

High-Priority Customers

Let us consider a new "tagged" type 1 arrival and assume that she is infinitely patient. We denote her virtual waiting time by V_1. If she finds at least one idle server with probability $1 - P_d$, she immediately enters service. So, her virtual waiting time is zero. In the complementary case (all

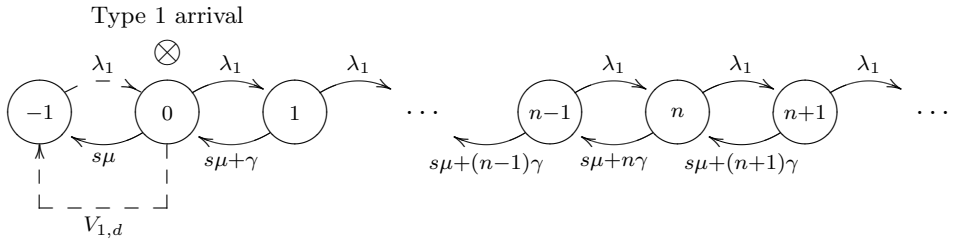

Figure 4. Virtual waiting time of a type 1 arrival, LCFS.

servers are busy), she is queued. Type 2 customers already waiting and those who arrive later are ignored because of their lower priority. Also, because the discipline of service within queue 1 is LCFS, type 1 customers already waiting in the queue are ignored. Thus, the conditional virtual waiting time, given delay for a new type 1 arrival is independent of the state of the two queues. Let us denote this conditional virtual waiting time by $V_{1,d}$, $V_1 = P_d V_{1,d}$. One can see that $V_{1,d}$ is the first-passage time at state -1 starting at state 0 in the birth-death process as shown in Figure 4.

From Figure 4, one can see that $V_{1,d}$ is equivalent to the duration of a 0-busy period in an $M/M/s+M$ queue with mean arrival rate λ_1, denoted by BP_{0,λ_1}. In a similar way as in Section 3.1, we characterize $W_{1,s}$ as follows. We have

$$F_{W_{1,s}}(t) = \frac{\mathbb{P}(V_1 < t, V_1 < T)}{\mathbb{P}(V_1 < T)},$$

for $t \geq 0$. We then obtain after some algebra

$$F_{W_{1,s}}(t) = \frac{1}{P_{1,s}} \left\{ 1 - P_d + P_d \int_0^t e^{-\gamma x} f_{V_{1,d}}(x) \mathrm{d}x \right\}, \tag{15}$$

for $t \geq 0$. Using $V_{1,d} \equiv BP_{0,\lambda_1}$ and taking the derivative in t on both sides of Equation (15) gives

$$f_{W_{1,s}}(t) = \frac{P_d}{P_{1,s}} e^{-\gamma t} f_{BP_{0,\lambda_1}}(t),$$

for $t \geq 0$, which by applying the Laplace-Stieltjes transform leads to

$$\widetilde{F}_{W_{1,s}}(x) = \frac{P_d}{P_{1,s}} \widetilde{F}_{BP_{0,\lambda_1}}(x+\gamma),$$

for $x \in \mathbb{R}^+$. Finally, we obtain

$$\mathbb{E}W_{1,s}^k = (-1)^k \frac{P_d}{P_{1,s}} \widetilde{F}^{(k)}_{BP_{0,\lambda_1}}(\gamma),$$

for $k \geq 1$. We now move to characterize $W_{1,r}$. We have

$$F_{W_{1,r}}(t) = \frac{\mathbb{P}(T < t, V_1 > T)}{\mathbb{P}(V_1 > T)},$$

for $t \geq 0$, which implies after some simplifications

$$F_{W_{1,r}}(t) = \frac{1}{P_{1,r}} \left\{ 1 - e^{-\gamma t} - \int_0^t \gamma e^{-\gamma x} \left(1 - P_d + P_d F_{BP_{0,\lambda_1}}(x) \right) \mathrm{d}x \right\}, \tag{16}$$

for $t \geq 0$. Taking the derivative in t on both sides of Equation (16) leads to

$$f_{W_{1,r}}(t) = \frac{P_d \gamma}{P_{1,r}} \left(e^{-\gamma t} - e^{-\gamma t} F_{BP_{0,\lambda_1}}(t) \right), \tag{17}$$

for $t \geq 0$. We now apply the Laplace-Stieltjes transform to Equation (17) and obtain

$$\widetilde{F}_{W_{1,r}}(x) = \frac{P_d \gamma}{P_{1,r}(x+\gamma)} \left(1 - \widetilde{F}_{BP_{0,\lambda_1}}(x+\gamma) \right),$$

for $x \in \mathbb{R}^+$. Finally, we close the discussion by mentioning that again one can use Equations (1)–(3) to obtain all moments of the remaining random variables W_1, $W_{1,d}$ and $W_{1,d,s}$.

Low-Priority Customers

Our approach again relies on determining the virtual waiting time of an infinitely patient type 2 customer. Consider such a customer. She will have a zero virtual waiting time with probability $1 - P_d$. With the complementary probability, she is queued. In the latter case, she will get priority over the type 2 customers already waiting. What matters for her virtual waiting time are the type 1 customers already waiting in queue 1 (denoted by n_1, $n_1 \geq 0$), as well as all future arrivals of types 1 and 2. Let us denote the conditional virtual waiting time of a type 2 customer, given a busy system and n_1 customers in queue 1, by V_{2,n_1}, $n_1 \geq 0$. One can see that V_{2,n_1} is the first-passage time at state -1 starting at state n_1 in the birth-death process as shown in Figure 5. It is then easy to see that V_{2,n_1} is equivalent to the duration of an n_1-busy period, say $BP_{n_1,\lambda}$, of an $M/M/s+M$ queue (with mean arrival rate $\lambda = \lambda_1 + \lambda_2$), $V_{2,n_1} \equiv BP_{n_1,\lambda}$.

Similarly to Equation (10), but by conditioning here on the state of queue 1, we obtain

$$F_{W_{2,s}}(t) = \frac{1}{P_{2,s}} \left\{ 1 - P_d + \int_0^t e^{-\gamma x} \sum_{n_1=0}^{\infty} p_1(n_1) f_{BP_{n_1,\lambda}}(x) \mathrm{d}x \right\}, \tag{18}$$

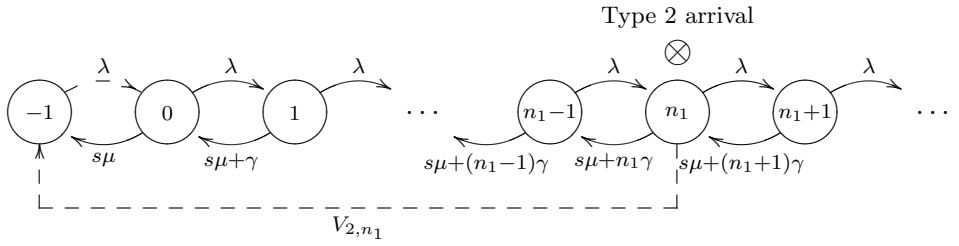

Figure 5. A new type 2 arrival arriving to an LCFS queue.

for $t \geq 0$. Taking the derivative in t on both sides of Equation (18) gives

$$f_{W_{2,s}}(t) = \frac{1}{P_{2,s}} \sum_{n_1=0}^{\infty} p_1(n_1) e^{-\gamma t} f_{BP_{n_1,\lambda}}(t),$$

for $t \geq 0$. Next, we may write

$$\widetilde{F}_{W_{2,s}}(x) = \frac{1}{P_{2,s}} \sum_{n_1=0}^{\infty} p_1(n_1) \widetilde{F}_{BP_{n_1,\lambda}}(x+\gamma), \tag{19}$$

for $x \in \mathbb{R}^+$. In a similar way as that for the FCFS case, but by using the random variable $BP_{n_1,\lambda}$ and averaging over all queue 1 states, we have

$$F_{W_{2,r}}(t) = \frac{1}{P_{2,r}} \left\{ 1 - e^{-\gamma t} - \int_0^t \gamma e^{-\gamma x} \left(1 - P_d + \sum_{n_1=0}^{\infty} p_1(n_1) F_{BP_{n_1,\lambda}}(x) \right) \mathrm{d}x \right\},$$

for $t \geq 0$, and

$$\widetilde{F}_{W_{2,r}}(x) = \frac{\gamma}{P_{2,r}(x+\gamma)} \left(P_d - \sum_{n_1=0}^{\infty} p_1(n_1) \widetilde{F}_{BP_{n_1,\lambda}}(x+\gamma) \right), \tag{20}$$

for $x \in \mathbb{R}^+$. Again, one can use Equations (1)–(3) to obtain all moments of the remaining random variables W_2, $W_{2,d}$ and $W_{2,d,s}$.

Note that for all cases analyzed above (any customer type and any discipline of service), one can check the relation $\mathbb{E}W_m^k = P_{m,s}\mathbb{E}W_{m,s}^k + P_{m,r}\mathbb{E}W_{m,r}^k$, for $k \geq 1$ and $m \in \{1,2\}$. In what follows, we do it for type 2 customers and $\text{Model}_{\text{LCFS}}$. It suffices to prove that $\widetilde{F}_{W_2}(x) = P_{2,s}\widetilde{F}_{W_{2,s}}(x) + P_{2,r}\widetilde{F}_{W_{2,r}}(x)$, for $x \in \mathbb{R}^+$. On the one hand, using Equations (19) and (20), we state that

$$P_{2,s}\widetilde{F}_{W_{2,s}}(x) + P_{2,r}\widetilde{F}_{W_{2,r}}(x) = \frac{\gamma P_d}{x+\gamma} + \frac{x}{x+\gamma} \sum_{n_1=0}^{\infty} p_1(n_1) \widetilde{F}_{BP_{n_1,\lambda}}(x+\gamma), \tag{21}$$

for $x \in \mathbb{R}^+$. On the other hand, we may write

$$F_{W_2}(t) = 1 - \mathbb{P}(\min\{V_2, T\} > t) = 1 - \mathbb{P}(V_2 > t)\mathbb{P}(T > t), \tag{22}$$

for $t \geq 0$. We also have

$$\begin{aligned} \mathbb{P}(V_2 > t) &= 1 - \left\{ (1 - P_d) \cdot 1 + \sum_{n_1=0}^{\infty} p_1(n_1)\mathbb{P}(V_{2,n_1} < t) \right\} \\ &= P_d - \sum_{n_1=0}^{\infty} p_1(n_1) F_{BP_{n_1,\lambda}}(t), \end{aligned} \tag{23}$$

for $t \geq 0$. Then, Equations (22) and (23) lead to

$$F_{W_2}(t) = 1 - P_d e^{-\gamma t} + e^{-\gamma t} \sum_{n_1=0}^{\infty} p_1(n_1) F_{BP_{n_1,\lambda}}(t),$$

for $t \geq 0$, which implies

$$f_{W_2}(t) = \gamma P_d e^{-\gamma t} + e^{-\gamma t} \sum_{n_1=0}^{\infty} p_1(n_1) f_{BP_{n_1,\lambda}}(t) - \gamma e^{-\gamma t} \sum_{n_1=0}^{\infty} p_1(n_1) F_{BP_{n_1,\lambda}}(t), \tag{24}$$

for $t \geq 0$. Finally, after some algebra, we deduce from Equation (24) that

$$\widetilde{F}_{W_2}(x) = \frac{\gamma P_d}{x + \gamma} + \frac{x}{x + \gamma} \sum_{n_1=0}^{\infty} p_1(n_1) \widetilde{F}_{BP_{n_1,\lambda}}(x + \gamma), \tag{25}$$

for $x \in \mathbb{R}^+$. By comparing Equations (21) and (25), we finish the proof.

3.3 More than Two Customer Types

The analysis in Sections 3.1 and 3.2 can be extended to a model with more than two customer types, for both FCFS and LCFS cases. In what follows, we provide indications about the approach to use. For FCFS or LCFS, consider the extended $M/M/s + M$ queueing model with k customer types, for $k > 2$. Type m has non-preemptive priority over type l, for $1 \leq m < l \leq k$. We assume that for all customer types, patience as well as service times are still statistically identical. Let us now focus on the performance measures of a type m customer with mean arrival rate λ_m, for $1 \leq m \leq k$.

First, we need to compute the stationary probabilities to have all servers busy and i waiting customers in queues $1, 2, \ldots, m$, denoted by $p_{1\to m}(i)$, and those to have i waiting customers in all queues, denoted by $p(i)$, for $i \geq 0$ and $1 \leq m \leq k - 1$. To compute these probabilities, it suffices to use the two-class analysis of Section 2.3 by transforming the k-class $M/M/s + M$ queue into

a two-class one. We do so by aggregating the first m types into a one type with mean arrival rate $\sum_{j=1}^{m} \lambda_j$, and the rest of types into a second one with mean arrival rate $\sum_{j=m+1}^{k} \lambda_j$, for $1 \leq m \leq k-1$. This allows to compute P_d and also the expected number of customers in queues $1, 2, \ldots, m$, denoted by $Q_{1\to m}$, for $1 \leq m \leq k-1$, and that in all queues, denoted by $Q_{1\to k} = Q$. Thus, the expected length of queue m is $Q_m = Q_{1\to m} - Q_{1\to m-1}$, for $1 \leq m \leq k$. We then obtain $P_{m,r} = \frac{\gamma Q_m}{\lambda_m}$, and $P_{m,s} = 1 - P_{m,r}$, for $1 \leq m \leq k$. In what follows, we focus on characterizing the random variables $W_{m,s}$ and $W_{m,r}$, which allows also to characterize the remaining random variables W_m, $W_{m,d}$ and $W_{m,d,s}$, for $1 \leq m \leq k$. We use a similar approach as in the previous Sections, with some changes that we mention next. Each time, the approach consists of finding an equivalent two-class queue.

Consider the k-class model working under FCFS. For $m = 1$, we aggregate types $2, \ldots, k$ into one type. We then apply the same analysis as for high-priority customers in Section 3.1. For $2 \leq m \leq k$, we aggregate types $1, \ldots, m$ into one high-priority type, and types $m+1, \ldots, k$ into one low-priority type. We thereafter use the stationary probabilities $p_{1\to m}(i)$, and duration of the i-busy period of a single-class $M/M/s+M$ queue with mean arrival rate $\sum_{j=1}^{m-1} \lambda_j$, for $i \geq 0$.

Consider now the k-class model working under LCFS. For $m = 1$, what we need is P_d and the duration of the 0-busy period in a single-class $M/M/s+M$ queue with mean arrival rate λ_1. For $2 \leq m \leq k$, we in turn aggregate types $1, \ldots, m-1$ into one high-priority type, and types $m, \ldots, k$ into one low-priority type. We thereafter use the stationary probabilities $p_{1\to m-1}(i)$, and the duration of the i-busy period of a single-class $M/M/s+M$ queue with mean arrival rate $\sum_{j=1}^{m} \lambda_j$ (since for a new type m arrival, future type m arrivals have priority over her), for $i \geq 0$. This closes the discussion about the extension to a model with more than two customer types.

Remark 1. In what follows, we discuss the extension of the analysis to a mixed model similar to the basic one described in Section 2.1. The difference is that we allow the discipline of service in one of the two queues to be different from the one in the other queue. For example, type 1 customers are served under FCFS, while type 2 customers are served under LCFS (or the opposite case). The extension is very easy to do. All the expressions for the stationary probabilities in Section 2 hold for the mixed model. Consider a given type. If it is served under FCFS (LCFS), then it suffices

to apply the same analysis as shown for that type in Section 3.1 (Section 3.2). This finishes the characterization of the mixed model.

3.4 Numerical Illustration

In this section we present numerical results to illustrate the usefulness of the performance evaluation models described in the previous sections. In particular, we compare between the performance measures of $\text{Model}_{\text{FCFS}}$ and $\text{Model}_{\text{LCFS}}$ and examine the effect of pooling on staffing levels. Furthermore, we discuss how our results are valid for more general models with different rates of service and general patience times.

Comparison between FCFS and LCFS

In Figure 6, we plot the conditional expected waiting times given service and given abandonment of each customer type, as a function of the staffing level s. We observe that the conditional expected waiting time given service is better under LCFS that the one under FCFS. The opposite is however true for the conditional expected waiting time given abandonment. For a single-class $M/M/s+M$ queue, Jouini (2012) proved that FCFS (LCFS) maximizes (minimizes) the conditional expected waiting time given service, and minimizes (maximizes) the one given abandonment. Note that it is easy to extend these results to each customer type in our multiple-type models here.

On the one hand, this observation concretely implies that a call center manager in practice would prefer to use LCFS in order to improve the waiting time before service of a given customer type. This is unfair from a customer perspective. On the other hand, Figure 6 also reveals that in contrary to $\mathbb{E}W_{m,r}$, $\mathbb{E}W_{m,s}$ is not highly impacted by the policy in queue m, for $m \in \{1,2\}$. Therefore, an appropriate decision for a manager would be to use FCFS for each customer type. First, it allows to preserve fairness between customers with the same level of priority. Second, it allows to achieve a good $\mathbb{E}W_{m,s}$ not far from the optimal one. Third, it is optimal in order to minimize the conditional waiting times given abandonment.

Let us focus on the behavior of $\mathbb{E}W_{m,s}$ and $\mathbb{E}W_{m,r}$, for $m \in \{1,2\}$. As mentioned above, the experiments show that, in the contrary to $\mathbb{E}W_{m,r}$, $\mathbb{E}W_{m,s}$ is quite insensitive to the scheduling

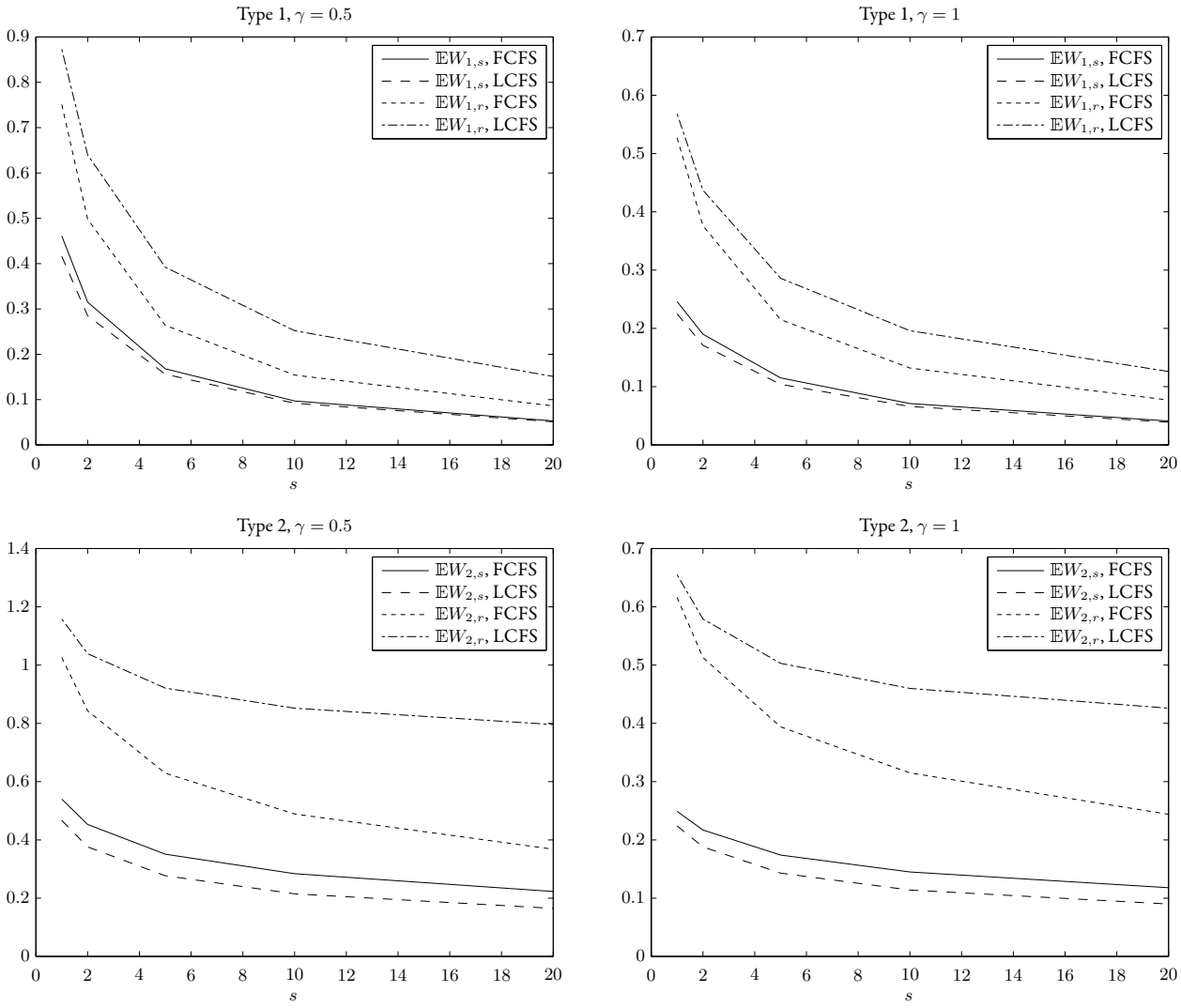

Figure 6. Conditional expected waiting times given service and given abandonment ($\mu = 1$, $s = 2\lambda_1 = 2\lambda_2$).

policy (LCFS and FCFS give the lower and upper bounds, respectively). This is however true for a range of system parameters for which the performance is not too deteriorated, namely with a probability of service higher 70%. A possible explanation is as follows. Let us denote by $W_{m,s,\mathrm{FCFS}}$ ($W_{m,r,\mathrm{FCFS}}$) and $W_{m,s,\mathrm{LCFS}}$ ($W_{m,r,\mathrm{LCFS}}$) the conditional waiting times given service (abandonment) under FCFS and LCFS for customers type $m \in \{1, 2\}$, respectively. Consider for example type 1 customers. Since $\mathbb{E}W_1$, $P_{1,s}$ and $P_{1,r}$ are unchanged for all work-conserving policies, Equation (1) implies

$$P_{1,s}\mathbb{E}W_{1,s,\mathrm{FCFS}} + P_{1,r}\mathbb{E}W_{1,r,\mathrm{FCFS}} = P_{1,s}\mathbb{E}W_{1,s,\mathrm{LCFS}} + P_{1,r}\mathbb{E}W_{1,r,\mathrm{LCFS}},$$

or equivalently

$$\mathbb{E}W_{1,s,\mathrm{FCFS}} - \mathbb{E}W_{1,s,\mathrm{LCFS}} = \frac{P_{1,r}}{P_{1,s}} \left(\mathbb{E}W_{1,r,\mathrm{LCFS}} - \mathbb{E}W_{1,r,\mathrm{FCFS}}\right). \tag{26}$$

To understand the intuition behind the observation, consider now the two extreme cases as follows. The first case is one where $P_{1,s}$ is very high (close to 1). This means that almost all customers enter service. Our system is then very similar to the one with no abandonment. In the latter, a sojourn in queue automatically ends with a service, and $\mathbb{E}W_1$ is unchanged in the scheduling policy. Then, in our case with abandonments, $\mathbb{E}W_1$ is very close to the expected conditional waiting time given service. In other words, $\mathbb{E}W_{1,s,\mathrm{FCFS}} - \mathbb{E}W_{1,s,\mathrm{LCFS}}$ is very small. The observation that $\mathbb{E}W_{1,r,\mathrm{LCFS}} - \mathbb{E}W_{1,r,\mathrm{FCFS}}$ is higher than $\mathbb{E}W_{1,s,\mathrm{FCFS}} - \mathbb{E}W_{1,s,\mathrm{LCFS}}$ can be now explained using Equation (26), since $P_{1,r}/P_{1,s} < 1$.

The opposite is true in another extreme case with a very high $P_{1,r}$. Most of the customers abandon. Then any scheduling policy would lead to a conditional waiting time given abandonment very similar to the unconditional one. Since the latter is unchanged in the scheduling policy, $\mathbb{E}W_{1,r,\mathrm{LCFS}} - \mathbb{E}W_{1,r,\mathrm{FCFS}}$ is very small. Using again Equation (26) and the fact that in this case $P_{1,r}/P_{1,s} > 1$, we state that $\mathbb{E}W_{1,r,\mathrm{LCFS}} - \mathbb{E}W_{1,r,\mathrm{FCFS}}$ is lower than $\mathbb{E}W_{1,s,\mathrm{FCFS}} - \mathbb{E}W_{1,s,\mathrm{LCFS}}$. More generally, $\mathbb{E}W_{1,s,\mathrm{FCFS}} - \mathbb{E}W_{1,s,\mathrm{LCFS}}$ is smaller than $\mathbb{E}W_{1,r,\mathrm{LCFS}} - \mathbb{E}W_{1,r,\mathrm{FCFS}}$ if $P_{1,r} < P_{1,s}$ and vice versa. For the case $P_{1,r} = P_{1,s}$, the two quantities coincide. Figure 7 illustrates how the difference $\mathbb{E}W_{1,s,\mathrm{FCFS}} - \mathbb{E}W_{1,s,\mathrm{LCFS}}$ may considerably vary as a function of the system performance. We consider in Figure 7 three different systems: the first with $s = 2\lambda_1 = 2\lambda_2$, the second with $s = \lambda_1 = \lambda_2$, and the third with $s = \lambda_1/2 = \lambda_2/2$. Then the first system ($P_{1,s}$ around 0.9) performs better than the second ($P_{1,s}$ around 0.7), which in turn performs better than the third ($P_{1,s}$ around 0.4). We observe that $\mathbb{E}W_{1,s,\mathrm{FCFS}} - \mathbb{E}W_{1,s,\mathrm{LCFS}}$ is the lowest for the first system, then for the second one, then for the third one, which agrees with the explanations above.

The same reasoning holds for customers of type 2. Moreover, since $P_{1,s} > P_{2,s}$, $\mathbb{E}W_{1,s,\mathrm{FCFS}} - \mathbb{E}W_{1,s,\mathrm{LCFS}}$ is smaller than $\mathbb{E}W_{2,s,\mathrm{FCFS}} - \mathbb{E}W_{2,s,\mathrm{LCFS}}$ as we observe from Figure 6. In summary, for real-life parameters for which the probability of service of a given customer type is likely to be higher than 70%, it is appropriate to use FCFS as a scheduling policy. The results of the standard deviations of queueing delays give another support to this idea. As one can see, Table 1 gives further

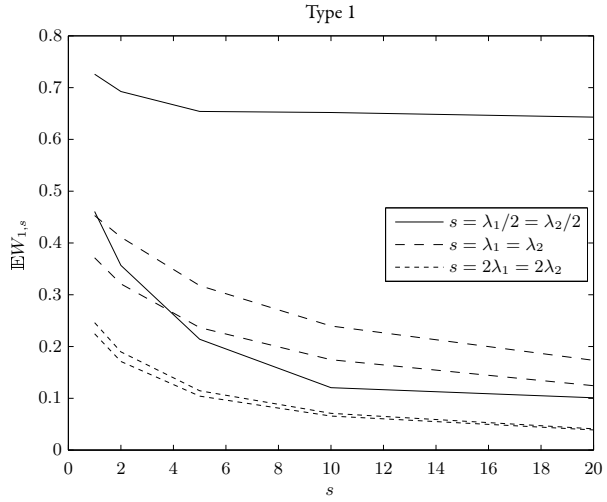

Figure 7. Behavior of $\mathbb{E}W_{1,s}$ for FCFS (upper lines) and LCFS (lower lines) ($\mu = 1$, $\gamma = 1$).

arguments in favor of FCFS. Values of standard deviations are indeed lower for FCFS than those for LCFS, except for the single-server case for type 2.

Staffing Levels

As expected we see from Figure 6 that performance improves in the system size, due to pooling effects. The same observation holds for the extreme cases of the heavily loaded systems in Figure 7. Pooling has however a diminishing return. The benefits in terms of the reduction of expected conditional waiting times, given service or abandonment, are more apparent for very small systems than for bigger ones.

For fixed customer arrival rates, we observe from Figure 8 as expected that waiting times decrease in the number of servers. However, we again see a diminishing return. Waiting times very quickly decrease when adding a server to a very small staffing level, but they are not reduced much from a higher staffing level around $s = 4$ for instance. The system manager can then choose a close-to-optimal staffing level while having an appropriate service level.

Also, we see that performance in terms of queueing delays improves in the abandonment rate γ. Therefore, staffing levels decrease in the abandonment rate. As γ increases, patience times decrease, so fewer customers are present in the system, and as a consequence virtual delays improve. For each type, although the expected conditional waiting times given service and given abandonment (see Figure 6) do vary with the scheduling policy (FCFS, LCFS, etc.), the unconditional expected

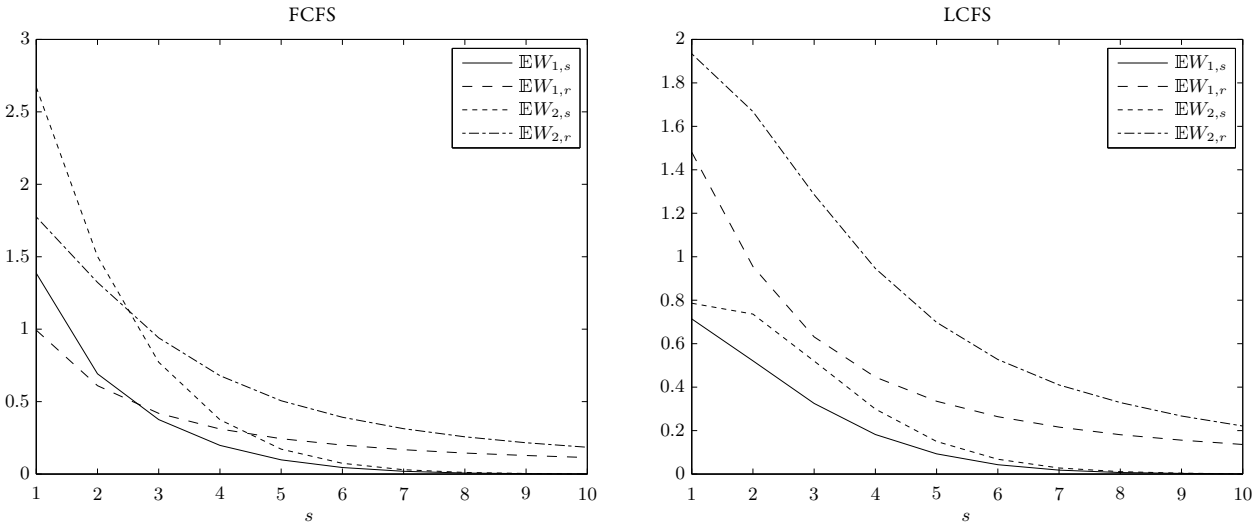

Figure 8. Effect of the staffing level on the expected waiting times ($\mu = 1$, $\gamma = 0.5$, $\lambda_1 = \lambda_2 = 2$).

	Type 1, FCFS				Type 2, FCFS			
s	$\mathbb{E}W_1$	$\sigma(W_1)$	$\sigma(W_{1,s})$	$\sigma(W_{1,r})$	$\mathbb{E}W_2$	$\sigma(W_2)$	$\sigma(W_{2,s})$	$\sigma(W_{2,r})$
1	0.539	0.720	0.702	0.728	0.713	0.977	0.910	1.017
2	0.347	0.474	0.468	0.477	0.563	0.795	0.752	0.831
5	0.177	0.249	0.247	0.253	0.408	0.589	0.570	0.611
10	0.100	0.144	0.143	0.148	0.316	0.457	0.448	0.466
20	0.054	0.080	0.079	0.083	0.241	0.346	0.342	0.343
	Type 1, LCFS				Type 2, LCFS			
s	$\mathbb{E}W_1$	$\sigma(W_1)$	$\sigma(W_{1,s})$	$\sigma(W_{1,r})$	$\mathbb{E}W_2$	$\sigma(W_2)$	$\sigma(W_{2,s})$	$\sigma(W_{2,r})$
1	0.539	0.807	0.719	0.927	0.713	1.069	0.887	1.216
2	0.347	0.569	0.513	0.711	0.563	0.923	0.755	1.121
5	0.177	0.327	0.303	0.467	0.408	0.765	0.614	1.033
10	0.100	0.201	0.189	0.315	0.316	0.662	0.524	0.985
20	0.054	0.116	0.111	0.197	0.241	0.570	0.446	0.948

Table 1. Comparison between standard deviations of queueing delays ($\mu = 1$, $s = 2\lambda_1 = 2\lambda_2$, $\gamma = 0.5$).

waiting times are as expected unchanged (see Table 1).

Different Service Rates

Our analysis relies on the assumption that for all customer types, the service rates are identical. This assumption is not unrealistic when customers are segmented into different groups based only on their value, and not on type-dependent statistical behavior. However, we present here numerical

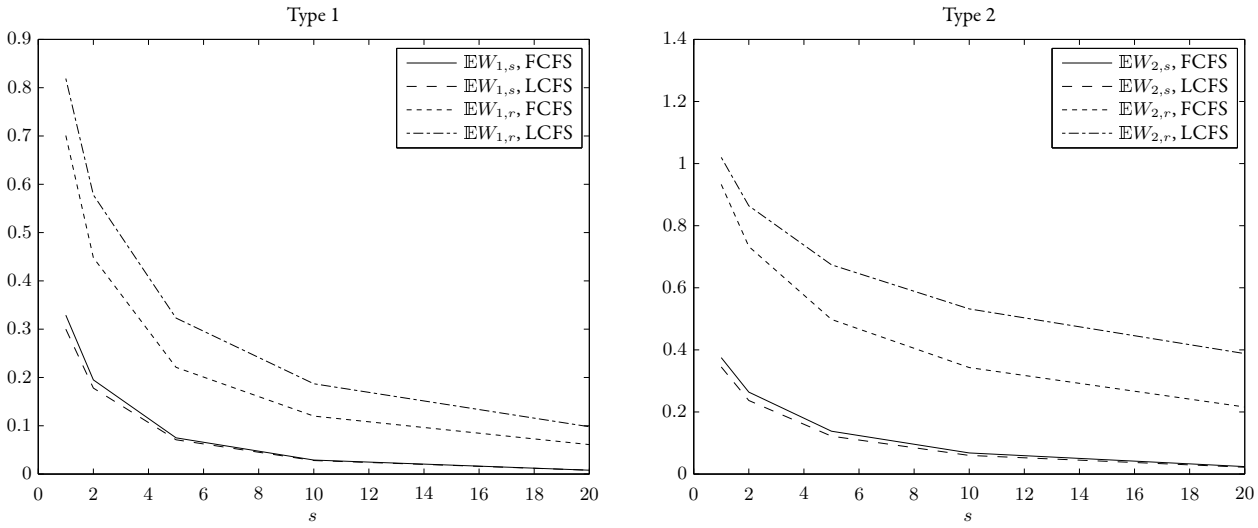

Figure 9. Conditional expected waiting times given service and given abandonment ($\mu_1 = 1$, $\mu_2 = 2$, $\gamma = 0.5$, $s = 2\lambda_1 = 2\lambda_2$).

results where the service rates do differ between customer classes. These results are therefore necessarily based on simulations.

First of all, we may easily extend the results of Jouini (2012, Theorem 1), who proved that FCFS (LCFS) maximizes (minimizes) the conditional expected waiting time given service, and minimizes (maximizes) the conditional expected waiting time given abandonment. These results also hold for any multi-class $G/G/s + M$ queue where interarrival times or service times differ from class to class. The only requirement is that patience times are exponential.

Let us denote by μ_1 the service rate of type 1 customers, and by μ_2 the service rate of type 2 customers. We depict in Figure 9 the conditional expected waiting times for each customer type with $\mu_1 = 1$ and $\mu_2 = 2$. This figure confirms the previous findings. Compared with Figure 6, the waiting times of both types of customers are lower due to the increased service rate of type 2 customers. Because the probability of service has increased, the difference between FCFS and LCFS has also decreased.

General Patience Times

We mentioned earlier that the multiple-priority queues under consideration in this paper are motivated by applications for the operations management of call centers. In our models, customer

patience times are restricted to be exponentially distributed. However, empirical evidences reveal that customer patience times are not exponentially distributed. We refer the reader for example to Brown et al. (2005), Feigin (2006), and references therein. In what follows, we discuss how our analysis is still of value for call centers in practice. All the arguments below concerns single-class call centers, but they still hold for multiple-class models.

To analyze multi-server queues with generally distributed service and patience times, Whitt (2005) develops an algorithm to compute approximations for standard steady-state performance measures. One of his main conclusions is that the behavior of the patience distribution near the origin primarily affects the performance. This is coherent with work by Mandelbaum showing that the Erlang A is a robust model even though patience times are not necessarily exponential in practice. Under the Quality-Efficiency-Driven regime where delays are short, Zeltyn and Mandelbaum (2005) point out that the patience distribution near the origin determines the behavior of the system.

More recently, Dai and He (2011) and Ward (2012) confirm that conclusion in the context of an $M/M/s+G$ queue. Under the heavy-traffic regime, they propose to approximate the performance of an $M/M/s+G$ queue by that of an $M/M/s+M$ queue with patience parameter equal to the probability density function of the general distribution evaluated at $t=0$. We refer also the reader to Jouini et al. (2011a) and Jouini et al. (2011b) where the authors consider call centers relatively smaller than those considered in the papers cited above. In the context of a call center with delay announcement, for which the patience distribution is far from being exponential, the authors again confirm the conclusion that what really matters is the patience distribution near the origin rather than in the tail.

Finally, consider call centers with customer balking, which is the case in many call centers in practice. In such cases, a newly arriving customer that finds all agents occupied may balk, i.e., immediately leave the system without service. An appropriate modeling of this behavior is to define a balking probability that is independent of any other event, see for example Whitt (1999b). Under this balking modeling, one may use the analysis in this paper, by simply multiplying the arrival rates in some appropriate places by this balking probability.

In summary, the exponential approximation for the distribution of times before abandonment

seems quite good. The analysis developed in this paper can then be applied for real-life call centers where patience times are not often exponentially distributed.

4 Conclusion

We considered multi-server non-preemptive priority queueing systems in which customers wait for service for a limited time only and leave the 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 considered two models: one where the discipline of service within each class of customers is FCFS, and another one where it is LCFS. For each customer type, we explicitly derived the Laplace-Stieltjes transforms of the unconditional waiting times, the conditional waiting times given service, and the conditional waiting times given abandonment. Numerical inversion methods for Laplace-Stieltjes transforms can be then used in order to obtain the cdf values of these random variables at any point of time. Moreover, we described the approach to extend the analysis to more than two customer types. The analysis in this paper holds also for a priority queue with mixed policies, i.e., FCFS for the first type and LCFS for the second one, and vice versa. Finally, we provided some numerical experiments in which we showed how FCFS would be preferred by a manager in practice.

There are various ways for future research. A challenging and interesting step is to extend our approach to the case of many customer types with different mean service and patience times. It is also interesting to consider general service-time distributions. Another useful extension would be to consider protocols with mixed priorities, i.e., both preemptive and non-preemptive priorities.

References

J. Abate and W. Whitt. A unified framework for numerically inverting Laplace transforms. *INFORMS Journal on Computing*, 18(4):408–421, 2006.

E. Altman and A.A. Borovkov. On the stability of retrial queues. *Queueing Systems*, 26(3/4): 343–363, 1997.

F. Baccelli and G. Hebuterne. On queues with impatient customers. In *Performance '81*, pages 159–179. North-Holland, 1981.

O.J. Boxma and P.R. de Waal. Multiserver queues with impatient customers. In J. Labetoulle and J.W. Roberts, editors, *Proceedings of the 14th International Teletraffic Congress*, pages 743–756, 1994.

A. Brandt and M. Brandt. On the $M(n)/M(m)/s$ queue with impatient calls. *Performance Evaluation*, 35(1):1–18, 1999.

A. Brandt and M. Brandt. Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s + GI$ system. *Queueing Systems*, 41(1/2):73–94, 2002.

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(1/2):147–168, 2004.

L.D. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50, 2005.

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

J. G. Dai and S. He. Queues in Service Systems: Customer Abandonment and Diffusion Approximations. *Tutorials in Operations Research*, pages 36–59, 2011.

R.H. Davis. Waiting-time distribution of a multi-server, priority queuing system. *Operations Research*, 14(1):133–136, 1966.

S. Drekic and D.A. Stanford. Threshold-based interventions to optimize performance in preemptive priority queues. *Queueing Systems*, 35(1/4):289–315, 2000.

P.D. Feigin. Analysis of customer patience in a bank call center. Working paper, 2006.

N. Gans, G.M. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141, 2003.

O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.

S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981.

M. Harchol-Balter, T. Osogami, A. Scheller-Wolf, and A. Wierman. Multi-server queueing systems with multiple priority classes. *Queueing Systems*, 51(3/4):331–360, 2005.

F. Iravani and B. Balcıoğlu. Approximations for the $M/GI/N + GI$ type call center. *Queueing Systems*, 58(2):137–153, 2008a.

F. Iravani and B. Balcıoğlu. On priority queues with impatient customers. *Queueing Systems*, 58 (4):239–260, 2008b.

O. Jouini. Analysis of a last come first served queueing system with customer abandonment. *Computers & Operations Research*, 39:3040–3045, 2012.

O. Jouini, O.Z. Akşin, and Y. Dallery. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management*, 13(4):534–548, 2011a.

O. Jouini, Y. Dallery, and O.Z. Akşin. Suppelemntary material to call Centers with Delay Information: Models and Insights. *Manufacturing & Service Operations Management*, 2011b. Available online, http://msom.pubs.informs.org/ecompanion.html.

E.P.C. Kao and S.D. Wilson. Analysis of nonpreemptive priority queues with multiple servers and two priority classes. *European Journal of Operational Research*, 118(1):181–193, 1999.

O. Kella and U. Yechiali. Waiting times in the non-preemptive priority $M/M/c$ queue. *Communications in Statistics. Stochastic Models*, 1(2):257–262, 1985.

A. Mandelbaum and S. Zeltyn. The impact of customers' patience on delay and abandonment: some empirically-driven experiments with the M/M/n+G queue. *OR Spectrum*, 26(3):377–411, 2004.

A. Mandelbaum and S. Zeltyn. Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205, 2009.

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

L. Rozenshmidt. On priority queues with impatient customers: Stationary and time-varying analysis. Master's thesis, Technion, Israel Institute of Technology, 2007.

V. Sarhangian and B. Balcıoğlu. Waiting Time Analysis of Multi-Class Queues with Impatient Customers. 2011. Working paper. University of Toronto.

A. Sleptchenko. Multi-class, multi-server queues with non-preemptive priorities. Technical report, EURANDOM, Eindhoven University of Technology, 2003.

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(1):81–107, 2005.

T. Takine. The nonpreemptive priority $MAP/G/1$ queue. *Operations Research*, 47(6):917–927, 1999.

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(1):67–84, 1997.

Q. Wang. Modeling and Analysis of High Risk Patient Queues. *European Journal of Operational Research*, 155:502 – 515, 2004.

A.R. Ward. Asymptotic Analysis of Queueing Systems with Reneging: A survey of Results for FIFO, Single Class Models. *Surveys in Operations Research and Management Science*, 17:1–14, 2012.

A.R. Ward and P.W. Glynn. A diffusion approximation for a Markovian queue with reneging. *Queueing Systems*, 43(1/2):103–128, 2003.

W. Whitt. Improving service by informing customers about anticipated delays. *Management Science*, 45(2):192–207, 1999a.

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

W. Whitt. Engineering solution of a basic call-center model. *Management Science*, 51(2):221–235, 2005.

S. Zeltyn and A. Mandelbaum. Call Centers with Impatient Customers: Many-Servers Asymptotics of the M/M/n+G Queue. *Queueing Systems*, 51:361–402, 2005.

S. Zeltyn, Z. Feldman, and S. Wasserkrug. Waiting and sojourn times in a multi-server queue with mixed priorities. *Queueing Systems*, 61(4):305–328, 2009.