# Online scheduling policies for multiclass call centers with impatient customers

Oualid Jouini, Ger Koole, Yves Dallery

# Online Scheduling Policies for Multiclass Call Centers with Impatient Customers

Oualid Jouini[1] • Auke Pot[2] • Ger Koole[3] • Yves Dallery[1]

[1] *Laboratoire Génie Industriel, Ecole Centrale Paris, Grande Voie des Vignes, 92290 Châtenay-Malabry, France*

[2] *CCmath, Catharina van Clevelaan 2, 1181 BH Amstelveen, The Netherlands*

[3] *Department of Mathematics, VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands*

oualid.jouini@ecp.fr • pot@ccmath.com • koole@few.vu.nl • yves.dallery@ecp.fr

***European Journal of Operational Research**, 207:258-268, 2010*

**Abstract**

We consider a call center with two classes of impatient customers: premium and regular classes. Modeling our call center as a multiclass $GI/GI/s+M$ queue, we focus on developing scheduling policies that satisfy a target ratio constraint on the abandonment probabilities of premium customers to regular ones. The problem is inspired by a real call center application in which we want to reach some predefined preference between customer classes for any workload condition. The motivation for this constraint comes from the difficulty of predicting in a quite satisfying way the workload. In such a case, the traditional routing problem formulation with differentiated service levels for different customer classes would be useless. For this new problem formulation, we propose two families of online scheduling policies: queue joining and call selection policies. The principle of our policies is that we adjust their routing rules by dynamically changing their parameters. We then evaluate the performance of these policies through a numerical study. The policies are characterized by simplicity and ease of implementation.

**keywords**: call centers, queueing systems, abandonments, online scheduling policies

## 1 Introduction

A call center, or in general a contact center, is defined as a service system in which customer representatives (agents or servers) serve customers (callers), over telephone, fax, email, etc. Managing a call center is a diverse challenge due to many complex factors related to uncertain and time-varying demand for service. In this paper we focus on a real-time problem, namely customer routing and agent scheduling.

The routing problem is a control problem which has received a lot of attention as a call center application. We consider here a V-model according to the canonical designs presented in Garnett and Mandelbaum (2001). We assume that all agents are flexible enough to answer all requirements

of service. However, we divide customers into two different classes according to their importance, premium and regular customers. In addition, we assume that customers can abandon. It is natural in practice that a waiting customer is willing to wait for only a limited time, and will hang up within that time. Introducing abandonments in theoretical models is valuable, as ignoring abandonments leads to overstaffing and pessimistic estimations of queueing delays, see Duder and Rosenwein (2001). Garnett et al. (2002) show that models with and without abandonment tend to perform differently, even if the abandonment rate is low. Models including abandonments are therefore more realistic and will yield more accurate managerial insights.

Given a staffing level, our purpose is to develop control schemes for arrival calls and idle agents, subject to satisfying a constraint related to the probabilities of being lost (probability to abandon). We notice that the abandonment probability is one of the major indicators used in practice. Another widely used indicator is the 80/20 rule where we stipulate that at least 80% of the customers should wait less than 20 seconds. Our objective is to meet a target ratio constraint between the achieved abandonment probabilities of the two customer classes. This problem formulation was inspired by a real call center problem. The reason behind it is to translate a desired fairness between customer classes. In our call center case, the abandonment of any class of customers is equivalent to a loss of goodwill. Both classes are indeed valuable for the company with a particular preference to the premium class. Having nearly no abandonments of the premium class and a lot of abandonments of the regular one is not desirable. The call center would instead prefer to have more abandonments of premium calls and fewer abandonments of regular ones. This is captured through a ratio of the abandonment probabilities. The ratio would be typically between 0 and 1. A low value of this ratio would translate to a strict preference of the company to premium calls. A ratio close to 1 would however translate an equal preference between the two classes. A value in between would translate a certain degree of preference.

In practice, managers traditionally handle this problem by separately setting for each customer class a constraint on the probability to abandon. The drawback of such a formulation is that it is too much dependent on some predicted workload. Once the actual workload deviates from the predicted one, we are no longer able to meet the predefined performance constraints, and we could well reach an undesired preference between customer classes. Indeed we often use static strict priority rules, such as a static strict non-preemptive priority for one class over the other. So if the workload is underestimated, most of the capacity of the system will be dedicated to premium calls. We may then satisfy the performance constraint of premium calls, while having a heavily penalized one for regular calls. However if the workload is overestimated, the performance of premium calls will be very high and that of regular ones will not profit that much from the overcapacity.

Several studies, notably Jongbloed and Koole (2001) and Avramidis et al. (2004), have shown that the arrival process and the workload are hard to predict in call centers. A new formulation of the routing problem using a target ratio constraint between the service levels of the two classes would, as a consequence, be a better alternative. It allows to better control the different situations which may occur (under- or overestimation of the workload). Satisfying the constraint ratio enables to share as desired the capacity of the system between the two classes. In addition, this new formulation generalizes the traditional one where we have a target abandonment probability for each class.

We use online policies in order to reach our objective. An online policy is a discipline that is continuously updated as customers arrive and are processed. The main advantage of our control policies is that they require no information about the arrival processes in advance, i.e., the method is data-driven. If fluctuations in workload occur, the capacity allocated between the two classes is changed. The parameters of the policies are automatically adopted in such a way that the target on the service levels is met.

We focus on developing simple and useful online routing policies that are based on priority schemes. The provision of differentiated service levels often relies on the use of different priorities between customers. Schrage and Miller (1966) showed that scheduling policies similar to those in multiclass priority queues allow to achieve high performance measures, often nearly as good as those under optimal policies. We derive various schemes for online assignment of customers to queues, as well as for selection to service of waiting customers. The policies we propose are characterized to be workconserving (non-idling), which is natural for large service systems such as call centers. A policy is defined to be non-idling if there can be no idling servers when there are waiting customers. In our opinion, the restriction does not decrease the usefulness of the analysis because, in practice, Automatic Call Distributors (ACDs) do not often support idling policies, such as thresholds or reservations policies. Under a threshold policy, customers are selected for service by following some predefined rules that are based on the system state. Under a reservation policy, some servers are only allocated to a given type of customers. This kind of policies would force one or more servers to be idle while there are waiting customers.

This paper has three major contributions. The first contribution is the introduction of a new class of call routing policies which are robust to changes in the workload. The second is the analysis of the effect of scheduling policies on various performance measures which contributes to the literature on multiclass queueing systems. The third contribution is to propose simple and efficient scheduling policies that satisfy the target ratio constraint. This is especially important because it is known that in practice a non-intuitive solution with a complex structure would never

be used. The analysis yields quantitative insights, as well as useful principles and guidelines for the control problem.

The rest of the paper is organized as follows. In Section 2 we review the literature related to our work. In Section 3 we present the problem under consideration. Sections 3.1 and 3.2 are devoted to formulate the call center model and to define the notations, respectively. Section 3.3 concretely formulates the control problem. Section 3.4 gives some structural results that will help us later in the understanding of the behavior of the scheduling policies. In Section 4 we develop two families of online scheduling policies that allow to satisfy the target ratio constraint. The family of queue joining policies is addressed in Section 4.1, and the family of call selection policies is addressed in Section 4.2. In Section 5 we present and discuss simulation experiments of the proposed policies. The paper ends with some concluding remarks and highlight some directions for future research.

## 2 Literature Review

There is an extensive and growing literature on call centers, and in general on contact centers. We refer the reader to Akşin et al. (2007) for a survey of the recent literature on call center operations management. The literature related to this paper spans mainly two areas. The first deals with queueing systems with impatient customers. The second deals with the control of queueing systems in general, and the control of the V-model in particular.

In the following, we highlight some of the literature with regard to the first area. Queueing models incorporating impatient customers have received a lot of attention in the literature. To underline the importance of the abandonment modeling in the call center field, the authors in Mandelbaum and Zeltyn (2008) give some numerical examples that point out the effect of abandonment on performance measures. Brown et al. (2005) conduct an empirical study to characterize the distribution of abandonment times. The literature on queueing models with abandonments focuses especially on performance evaluation. We refer the reader to Ancker and Gafarian (1962), Garnett et al. (2002), and references therein for simple models assuming exponential abandonment times. Koole (2004) develops an algorithm for calculating tail probabilities of Cox distributions. This can be useful to compute many performance measures in a queueing system with abandonments. In Garnett et al. (2002), the authors study the subject of Markovian abandonments. They suggest an asymptotic analysis of their model under the heavy-traffic regime. Their main result is a relation between the number of agents, 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. Whitt (2004) establishes the Efficiency-Driven many-server heavy-traffic limits for a Markovian queue with abandonments and

limited waiting line. Other papers assumed abandonments to follow a general distribution. Whitt (2006) uses fluid models to approximate the steady state performance measures of queueing models with generally distributed interarrival, service and abandonment times. Other studies include those by Baccelli and Hebuterne (1981), Brandt and Brandt (2002), Ward and Glynn (2003), Whitt (2005), Avramidis et al. (2010), and references therein.

Let us now focus on the second area of literature close to our work, i.e., the control of queueing systems. Scheduling policies have been studied in great depth within the context of queueing systems. A scheduling policy, or a discipline of service, prescribes the order in which customers are served. Randolph (1991) classifies scheduling policies into those using online schedule rules and those using static schedule rules. Each of the above classes of policies can be further classified into two major classes: agent scheduling and customer routing, see Garnett and Mandelbaum (2001) for more details. In the following, we present results about scheduling policies under the framework of V-models. Pekoz (2002) addresses the analysis of a multiserver non-preemptive priority queue with exponentially distributed interarrival and service times. She finds and evaluates the performance of an asymptotically optimal policy that minimizes the expected queueing delay for high priority customers. Guérin (1998) presents a model without waiting queues. The model contains a multiserver station, which receives low and high priority arrivals. He develops an admission policy for the low priority customers such that the fraction of blocked high priority customers is bounded and he analyzes the system under that policy. In the context of call centers, Gurvich et al. (2008) consider a large-scale system under the V-design and characterize asymptotically optimal scheduling and staffing schemes (as system load grows to infinity). The optimal scheduling and staffing schemes minimize the staffing costs subject to satisfying quality of service constraints for the different customer classes. Maglaras and Zeevi (2005) consider profit maximization for a loss system two-class V-model with pricing, sizing, and admission control. Milner and Olsen (2008) explore the role that service level constraints in outsourcing contracts play in settings where the contractor firm has both contractual and non-contractual customers. Another paper that consider a similar idea of contract and noncontract customers is Bhulai and Koole (2003). For a detailed survey of relevant papers considering the optimal control of the V-model, we refer the reader to Gurvich (2004). We finally refer the reader to Gurvich and Whitt (2008) and Dai and Tezcan (2008) for recent references on online policies developed within an economic context.

# 3 Framework

In this section, we first describe the basic model of our call center. Second, we define the various performance measures we are interested in. We then formulate the problem for which we want

to propose the scheduling policies. Finally, we develop some structural results about the relation between the performance measures under consideration and various scheduling policies.

## 3.1 Model Description

We model our call center as a queueing system with two customer classes: a premium customer class $A$, and a regular one $B$. The model consists of two infinite queues, say queues 1 and 2, and a set of $s$ parallel, identical servers representing the set of agents. All agents are able to answer all customer classes. The call center is operated in such a way that at any time, any customer can be addressed by any agent. Upon arrival, a customer is addressed by one of the available agents, if any. If not, the call joins one of the queues.

We consider two families of scheduling policies: queue joining and customer selection families. We will describe in details the proposed policies among the two families in Section 4. In what follows we describe their functioning in general. A policy belonging to the first family of queue joining policies determines the rule of assigning customers upon arrival to one of the queues. Upon arrival, a customer of any class can be sent to any queue. That is, each time an arrival $A$ or $B$ enters the system, an individual decision is made as a function of the system state: assign this new arrival to queue 1 or 2. In other words, we want to specify here that there is no an a priori fixed rule of assigning for example all customers $A$ to queue 1 and all customers $B$ to queue 2. So queue 1 or 2 may contain a mix of the two customer types. Finally note for this first family that customers waiting in queue 1 have a non-preemptive strict priority over those in queue 2.

A policy belonging to the second family of customer selection policies determines at each service completion which waiting customer class should start service. For this family, there is no fixed priority between the two queues. Upon arrival, all customers $A$ are sent to queue 1 and all customers $B$ are sent to queue 2. At each service completion, the server who has just become idle will choose to serve one of the queues, or equivalently one of the customer classes. This decision is not static and is dynamically made as a function of the system state. In other words, if it happens at some times that we would like to improve the performance of class $A$, so an idle server would tend to choose to serve customers waiting in queue 1 (customers $A$). The opposite is done if it happens during other times that we would like to improve the performance of class $B$.

For both families of policies, customers waiting in a given queue, are served in the order of their arrivals, i.e., under FCFS. Also, the priority rule between the queues is non-preemptive because it is not common in call centers to interrupt the service of a customer and serve another one with a higher priority.

Interarrival times and service times are assumed to be i.i.d. and follow a general distribution.

In certain cases, we shall consider the exponential distribution for successive service times. The mean service time rates of customer classes $A$ and $B$ are $\mu_A$ and $\mu_B$, respectively. In addition, we let the customers be impatient. After entering the queue, a customer waits a random length of time for service to begin. If service has not begun by this time she will abandon (leaves the queue). Patience times of classes $A$ and $B$ are assumed to be i.i.d. and exponentially distributed with rates $\gamma_A$ and $\gamma_B$, respectively. Assuming identical distribution of patience within each class, independently from their position in the queue, seems to be a plausible assumption for call centers, see Gans et al. (2003). Indeed, the tele-queueing experience in call centers is fundamentally different from that of a physical queue, in the sense that customers do not see others waiting and need not be aware of their "progress" (position in the queue) if the call center does not provide information about queueing delays. We impose the Markovian assumption on patience times in this paper. This is mainly to preserve tractability in our models. Though not all empirical studies from real call centers suggest that patience times are exponential (see for example Brown et al. (2005)), considering this assumption allows us at least to capture the uncertainty in patience times and helps us as a consequence to gain useful insights.

The system is workconserving, i.e., an agent is never forced to be idle while customers are waiting. Finally, retrials are ignored, and abandonments are not allowed once a customer starts her service. We also do not allow jockeying between separate queues. Following similar arguments, the behavior of this call center can be viewed as a modification of a $GI/GI/s+M$ queueing system. The symbol $M$ after the $+$ indicates the Markovian assumption for times before abandonments.

Note that owing to abandonments, the system is unconditionally stable. One proof which also holds for a more general model is as follows. Consider an arbitrary $GI/GI/s+GI$ queueing model (with general patience times) and let us prove that it is stable, i.e., the stochastic process representing the number of customers in system has a limiting distribution as time goes to infinity. The number of customers in this $GI/GI/s+GI$ model can be upper bounded by the number of busy servers in an associated $GI/GI/\infty$ infinite server model. The service times in the associated $GI/GI/\infty$ model have the same statistical distribution as that of the summation of service and patience times in the original $GI/GI/s+GI$ model. The idea of the upper bound comes from the fact that the waiting time of a queued customer in the $GI/GI/s+GI$ model is upper bounded by her patience threshold. So her sojourn time in system is upper bounded by her patience threshold plus her service time. The infinite-server model is known to be stable. We refer the reader to Whitt (1982) for a discussion and further references. This finishes the proof of the stability of the $GI/GI/s+GI$ model, and in particular our $GI/GI/s+M$ model.

## 3.2 Notations

We denote by $m$ the class of a customer, $m \in \{A,\ B\}$. We assume that at time zero the system starts empty. For a particular sample path, let $n^m(t)$ be the number of class $m$ arrivals during the time interval $[0,t]$, $t>0$. Consider now a scheduling policy $\pi$ and let $a_\pi^m(t)$ be the number of class $m$ customers that abandon the queue, and $b_\pi^m(t)$ the number of those that are receiving service and those that have been already served in $[0,t]$. We consider for each class $m$, the fraction of customers that abandon. We also consider this performance for all customer classes (an arbitrary customer). The fraction of class $m$ customers that abandon during $[0,t]$ is defined by $\mathbf{P}_\pi^m(t) = \frac{a_\pi^m(t)}{n^m(t)}$. The fraction of abandonments during $[0,t]$ for all classes is defined by $\mathbf{P}_\pi(t) = \frac{a_\pi^A(t)+a_\pi^B(t)}{n_\pi^A(t)+n_\pi^B(t)}$.

In the long run, the fraction of abandonments (probability to abandon) of class $m$ customers, say $\mathbf{P}_\pi^m$, and that of all classes, say $\mathbf{P}_\pi$, are given by $\mathbf{P}_\pi^m = \lim_{t\to\infty} \mathbf{P}_\pi^m(t)$ and $\mathbf{P}_\pi = \lim_{t\to\infty} \mathbf{P}_\pi(t)$. Recall that due to abandonments the system is stable so that the latter limits do exist (see the end of Section 3.1). We are now ready to define our main performance measure which we denote by $\mathbf{c}_\pi$. It is the ratio of the probability to abandon of class $A$ over that of class $B$, i.e.,

$$\mathbf{c}_\pi = \frac{\mathbf{P}_\pi^A}{\mathbf{P}_\pi^B}. \tag{1}$$

Similarly we define the expected waiting time in queue of class $m$ customers and that of all customer classes. Note that we only define these quantities for the customers who enter service. Under a given scheduling policy $\pi$, let $w_{q,\pi}^m(i,t)$ be the waiting time in queue of the $i^{th}$ class $m$ customer who enters service, $0 \le i \le b_\pi^m(t)$. As in the usual way, the expected waiting time in the queue, say $W_{q,\pi}^m(t)$, during $[0,t]$ of class $m$ customers is defined by $W_{q,\pi}^m(t) = \dfrac{\sum_{i=1}^{b_\pi^m(t)} w_{q,\pi}^m(i,t)}{b_\pi^m(t)}$. and that of all customer classes, say $W_{q,\pi}(t)$, by $W_{q,\pi}(t) = \dfrac{\sum_{i=1}^{b_\pi^A(t)} w_{q,\pi}^A(i,t) + \sum_{i=1}^{b_\pi^B(t)} w_{q,\pi}^B(i,t)}{b_\pi^A(t)+b_\pi^B(t)}$. In the long run, the expected waiting time in the queue of served class $m$ customers, say $W_{q,\pi}^m$, and the overall expected waiting time in queue of all served customers, say $W_{q,\pi}$, are given by $W_{q,\pi}^m = \lim_{t\to\infty} W_{q,\pi}^m(t)$, and $W_{q,\pi} = \lim_{t\to\infty} W_{q,\pi}(t)$.

We also define the squared difference of waiting times of served class $m$ customers and that of all classes, in the transient regime as well as in the long run. The squared difference of the waiting time of served customers of class $m$, say $V_\pi^m(t)$, during $[0,t]$ is given by $V_\pi^m(t) = \dfrac{\sum_{i=1}^{b_\pi^m(t)} \left(w_{q,\pi}^m(i,t) - W_{q,\pi}^m(t)\right)^2}{b_\pi^m(t)}$ As for the squared difference of the waiting time in queue of all served customers within $[0,t]$, it

is defined by $V_\pi(t) = \dfrac{\sum_{m\in\{A,B\}} \sum_{i=1}^{b_\pi^m(t)} \left(w_{q,\pi}^m(i,t) - W_{q,\pi}(t)\right)^2}{\sum_{m\in\{A,B\}} b_\pi^m(t)}$ In the long run, the squared difference of class $m$ served customers, say $V_\pi^m$, and the overall squared difference of all served customers, say $V_\pi$, are given by $V_\pi^m = \lim_{t\to\infty} V_\pi^m(t)$, and $V_\pi = \lim_{t\to\infty} V_\pi(t)$. In the numerical experiments shown later, we will consider the standard deviation defined as the square root of $V_\pi^m$ and denoted by $\sigma_\pi^m$, and that of $V_\pi$ denoted by $\sigma_\pi$.

## 3.3 Problem Formulation

In this section we motivate and formulate our objective with regard to the scheduling policies we aim to develop. Due to the highly uncertain environment of call centers, it is usually hard to estimate the workload within a relative accuracy. We often end up in practice with either an underestimated, or an overestimated workload. Next, the common practice in call centers is to develop routing policies that aim to reach different service level constraints for different customer classes. In this paper we focus on service levels in terms of the abandonment probabilities. Having these statements in hand, one may obviously raise the following issue. In case of a fixed number of agents, the abandonment probability of each customer class will be affected by the forecasting error and will thereafter deviate from the predefined one. Furthermore, the actual service level deviations can be very different between the customer classes (for example when using strict priority scheduling policies). This behavior is undesirable for a call center manager because one would lose some given fairness between customer classes. For any work condition, a manager would like to reach a desired fairness between customer classes.

Based on this motivation, we formulate the following problem. We assume that scheduling has already taken place, such that the number of available agents is known. We aim to develop scheduling policies that satisfy a target ratio constraint, say $\mathbf{c}^*$, of the abandonment probabilities between of the two customer classes. In mathematical terms, we look, if at all possible, for $\pi \in \Pi$ subject to

$$\mathbf{c}_\pi = \mathbf{c}^*, \tag{2}$$

where $\Pi$ denotes the class of workconserving non-preemptive scheduling policies.

The target ratio $\mathbf{c}^*$ translates a desired preference between the two customer classes that we want to reach for any actual workload. As we will see later by means of Theorem 2, the $\mathbf{c}^*$ target formulation generalizes the traditional formulation where we have a service level constraint for each class. If the workload is quite correctly estimated, having a target ratio of abandonment

probabilities is equivalent to having a target abandonment probability for each class. In addition if the workload is incorrectly estimated (under- or overestimated), the capacity of the system is shared over customer classes in a way that allows to preserve (if possible) the predefined preference between customer classes. Specifically, we are interested in policies which are easy to implement in practice and which require as little information as possible about the arrival process and the capacity of service.

The expected waiting times of customer classes are not included in the problem formulation. The drawback of minimizing the waiting times of the served customers is that it could result in policies that are very similar to LCFS (per customer class), which is considered to be unfair. Further, we do not consider neither the abandonment probability, nor the expected waiting time in queue of an arbitrary customer. When considering workconserving policies, these quantities are unchanged for identical distributions of service times. They are furthermore not much sensitive when the distributions do differ. We will discuss these statements in the next sections in detail. Non-workconserving policies are ignored because they would make the routing problem even more complicated. Moreover, this kind of policies is not often seen in practice.

## 3.4 Structural Results

In this section we investigate the impact of scheduling policies on various performance measures. These results would be helpful in the understanding of the behavior of the policies we will develop and in assessing their efficiency.

Consider first queueing systems with infinitely patient customers, that is, a customer never leaves the queue. It is well known (see for example Gross and Harris (1998)) that the expected remaining workload in the queue is independent of the queueing discipline, assuming that the remaining total service or work required is order-of-service independent during any busy period. In other words, no service needs are created or destroyed within the system, implying: no abandonment in the middle of a service, no preemption when service times are not exponentially distributed, no forced idleness of servers, and so on. The proof can easily be done by comparing the diagrams of the cumulative work for two different queueing disciplines during the busy period, elsewhere both systems behave identically due to the workconserving property.

By means of Theorem 1, we motivate why workconserving policies are considered in this paper. Theorem 1 concerns a $GI/GI/s+M$ system, which has i.i.d. and generally distributed interarrival and service times, $s$ servers, and exponentially distributed patience times. The proofs of all results from this section are given in Section 1 of the supplementary material Jouini et al. (2010).

**Theorem 1** *Consider a non-preemptive $GI/GI/s+M$ queue and consider policies that do not*

*depend on service time realizations. Then there is a workconserving policy that minimizes the abandonment probability,* $\mathbf{P}$.

The memoryless condition of patience times in Theorem 1 is a necessary and sufficient condition when preemption is not allowed. Suppose that the abandonment occurs always during the beginning of the customer waiting, such under the decreasing failure rate distributions for example, then it can be optimal to wait for a new arrival and serve it before it abandons instead of serving a customer who has waited already for some time. In what follows, we derive some results related to workconserving policies. In Theorems 2 and 3, we investigate the conservation of the abandonment probability and the expected waiting time in the queue with respect to the scheduling policies, respectively. Some consequences are derived in Corollaries 1 and 2.

**Theorem 2** *Consider a* $GI/GI/s+M$ *queue. Patience times are assumed to be i.i.d. and exponentially distributed. Then the abandonment probability* $\mathbf{P}$ *is unchanged for all workconserving non-preemptive scheduling policies that are not allowed to depend on the service time realizations.*

Note that the result in Theorem 2 does not hold if service times depend on the order of service, or if we allow preemption when service times are not exponentially distributed, or if patience times are not identically and exponentially distributed. In the case of exponential service times, proving that $\mathbf{P}$ is unchanged for all workconserving policies with or without preemption can be obtained using Lemma 2 in Jouini and Dallery (2007). In the latter, the authors have proved this result for a $GI/M/s/K+M$ queue with a limited waiting space.

In Theorem 3, we focus on the waiting time in the queue with respect to workconserving non-preemptive scheduling policies. We again consider a $GI/GI/s+M$ queue, and focus on three different definitions of the expected waiting time. Let $W_q$ be the expected waiting time in queue of served customers. Let $W_q^{ab}$ be that of abandoning ones, i.e., the expected sojourn time in the queue before leaving the system without being served. Finally, we define $W_q^{tot}$ as the expected overall waiting time in the queue for all customers, i.e., served as well as abandoning customers. In Theorem 3, we prove an intuitive result for the conservation of $W_q^{tot}$. In addition, we give the scheduling policies that minimize and maximize $W_q$ and $W_q^{ab}$. Although the abandonment probability as shown in Theorem 2 does not vary for any workconserving non-preemptive scheduling policy, $W_q$ and $W_q^{ab}$ do vary.

**Theorem 3** *Consider a* $GI/GI/s+M$ *queue. Patience times are assumed to be i.i.d. and exponentially distributed. When considering the class of workconserving non-preemptive scheduling policies, the following holds*

1. *$W_q^{tot}$ does not depend on the scheduling policy.*

2. *$W_q$ and $W_q^{ab}$ depend on the scheduling policy.*

3. *The upper (lower) bound of $W_q$ is achieved under the FCFS (LCFS) discipline of service.*

4. *The upper (lower) bound of $W_q^{ab}$ is achieved under the LCFS (FCFS) discipline of service.*

Note that although the first moment $W_q^{tot}$ does not depend on the discipline of service, the second moment of the overall waiting time, and thus the full distribution, does depend on the discipline of service. As shown in Theorem 3, the maximum of the expected waiting time for served customers, $W_q$, is achieved under FCFS discipline. However we conjecture based on a well-known property from the literature (see for example Gross and Harris (1998)) that the minimum of its variance is also achieved under the FCFS policy. In practice, if the value of $W_q$ under the FCFS policy is almost equal to the value of other policies, a call center manager usually prefers the FCFS policy because of its fairness. We refer the reader to Avi-Itzhak and Levy (2004) for more details on the fairness property in queueing systems.

The result in Theorem 3 is still valid when considering also preemptive scheduling policies. However, service times have to be exponentially distributed. We finally comment that the second statement in Theorem 3 is still valid for generally distributed service times.

**Corollary 1** *Consider a $GI/GI/s+M$ queue with two customers classes $A$ and $B$. Service times as well as patience times are identically distributed for both customer classes. Then the overall abandonment probability $\mathbf{P}$ and the overall expected waiting time $W_q^{tot}$ are constant for any work-conserving non-preemptive scheduling policy.*

In the case of a two-class queue, we denote by $\pi_A$ and $\pi_B$ the policies that give strict non-preemptive priority to class $A$ and class $B$ customers, respectively.

**Corollary 2** *Consider a $GI/GI/s+M$ queue with two of customers classes $A$ and $B$. Service and patience times are identically distributed for both customer classes. Then, for any workconserving non-preemptive policy $\pi$ the ratio of the abandonment probabilities $\mathbf{c}_\pi$ satisfies the following relation*

$$\mathbf{c}_{\pi_A} \leq \mathbf{c}_\pi \leq \mathbf{c}_{\pi_B}, \tag{3}$$

*where $\mathbf{c}_{\pi_A}$ and $\mathbf{c}_{\pi_B}$ are the ratios in the long run under $\pi_A$ and $\pi_B$, respectively.*

Note that values of $\mathbf{c}_\pi$ outside the interval $[\mathbf{c}_{\pi_A}, \mathbf{c}_{\pi_B}]$ may be achieved through non-workconserving policies, such as thresholds or reservations policies. These policies are indeed useful to discriminate

between customer classes. Under such policies, the lower bound for the abandonment probability $\mathbf{P}^A$ ($\mathbf{P}^B$) is equal to the value achieved under the policy that gives strict preemptive priority to class $A$ ($B$) customers. Obviously, the upper bound for $\mathbf{P}^A$ or $\mathbf{P}^B$ is 1. It is reached for a given class by simply refusing to serve customers of that class.

With the waiting time being defined as the waiting time of all customers (including both served and abandoned customers), we next give Theorem 4 which only holds for Markovian patience times. Note also that this theorem holds even for $GI/GI/s+M$ queues that are working under non-workconserving policies.

**Theorem 4** *Consider a $GI/GI/s+M$ queue, with two customer classes $A$ and $B$, and working under a given scheduling policy $\pi$. Service and patience times are identically distributed for both customer classes. Let $\mathbf{c}_\pi$ be the long run ratio of the abandonment probabilities. Then the long run ratio of the expected waiting times is also $\mathbf{c}_\pi$.*

# 4 Online Scheduling Policies

The call center we consider allows for flexible scheduling through dynamically sequencing the order of service, hereafter referred to as online scheduling. With the current technology, this is also doable for most call centers. However, an interesting and challenging problem is to design scheduling policies that are structurally simple and easy to implement.

In this section we develop online scheduling policies that allow, if at all possible, to the target ratio constraint $\mathbf{c}^*$. We consider techniques that do not anticipate the future and that are based on the history of the system, i.e., the achieved ratio $\mathbf{c}$ at the moment of the decision. These can be classified as online updating methods and online routing. Without loss of generality, we only consider objective ratios satisfying $\mathbf{c}^* < 1$. The case $\mathbf{c}^* = 1$ boils down to the FCFS discipline of service. The reason is that our model working under FCFS (within each class and for both classes) is simply equivalent to a single class queue working under FCFS. The case $\mathbf{c}^* > 1$ is not relevant for our analysis because class $A$ denotes the premium calls. Furthermore, even if we would like to investigate that case, it suffices to apply the analysis for the case $\mathbf{c}^* < 1$ by exchanging class $A$ by class $B$ and vice versa.

We propose online scheduling policies that belong to two different families. The first is the family of queue joining policies and is presented in Section 4.1. The second is the family of call selection policies and is presented in Section 4.2.

## 4.1 Queue Joining Policies

We propose three queue joining scheduling policies, denoted by $\pi_1$, $\pi_2$ and $\pi_3$. Upon a customer arrival, a policy determines a rule for the queue assignment. We assume that customers in queue 1 have a non-preemptive priority over customers in queue 2. Customers of a class can be sent to any queue. We recall that our policies do not anticipate on future events. They just react to the realization of the ratio that is determined by the history of the process. A comparison analysis of these policies is thereafter addressed in Section 5, using simulation experiments.

**Scheduling Policy $\pi_1$:** The scheduling policy $\pi_1$ starts identically to a strict priority policy that gives the higher priority to class $A$ customers. After the epoch at which the first class $B$ customer finishes her service, we apply the following assignment rule for any new arrival (denoted by the $k^{th}$ arrival). Let $d_k$ be the epoch of that arrival. Let $\mathbf{P}_k^A$ ($\mathbf{P}_k^B$) be the achieved service level until $d_k$ for class $A$ ($B$). Let $\mathbf{c}_k$ be the achieved ratio starting until $d_k$, $\mathbf{c}_k = \mathbf{P}_k^A/\mathbf{P}_k^B$. If $\mathbf{c}_k < \mathbf{c}^*$, then we give high priority to class $B$, i.e., if the new arrival is class $A$, it is routed to queue 2, otherwise, it is routed to queue 1. However if $\mathbf{c}_k \geq \mathbf{c}^*$, we give high priority to class $A$, i.e., if the new arrival is class $A$, it is routed to queue 1, and if it is class $B$, it is routed to queue 2. An illustration of $\pi_1$ is shown on Figure 1.

**Scheduling Policy $\pi_2$:** The scheduling policy $\pi_2$ starts identically to $\pi_1$ until the epoch at which the first customer $B$ finishes service. Following the same notations as in the last paragraph, let a new arrival enter the system. Under $\pi_2$, a customer $A$ is always routed to queue 1. However, the assignment rule of class $B$ customers is as follows. If $\mathbf{c}_k < \mathbf{c}^*$, a new class $B$ arrival is routed to queue 1, otherwise if $\mathbf{c}_k \geq \mathbf{c}^*$, it is routed to queue 2. An illustration of $\pi_2$ is shown on Figure 2.

**Scheduling Policy $\pi_3$:** The scheduling policy $\pi_3$ starts identically to policies $\pi_1$ and $\pi_2$ until the first customer $B$ finishes service. Following again the same notations as before, let a new arrival enter the system. Under $\pi_3$, a customer $B$ is always routed to queue 2. However, the assignment rule of customers of class $A$ is as follows. If $\mathbf{c}_k \geq \mathbf{c}^*$, then a new class $A$ arrival is routed to queue 1, otherwise if $\mathbf{c}_k < \mathbf{c}^*$, it is routed to queue 2. An illustration of $\pi_3$ is shown on Figure 3.

The scheduling policy $\pi_1$ can be immediately obtained intuitively. It allows the achieved ratio to be updated upon each arrival such that it converges in the long run to the objective. The idea behind $\pi_2$ is that we keep always customers of class $A$ in the high priority queue, however when it is necessary, we assign customers of class $B$ to this queue to improve their service level (which deteriorates the service level of customers of class $A$). Such a rule allows to increase the transient ratio and keep it close to the objective. As a consequence, the ratio converges in the long run to the desired value. Policy $\pi_3$ can be viewed as another variant. It sometimes allows to penalize class $A$ customers by assigning them to the low priority queue, which again allows to increase the

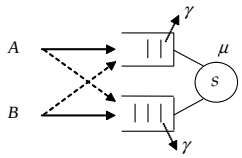


Figure 1: Scheduling policy $\pi_1$

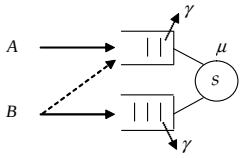


Figure 2: Scheduling policy $\pi_2$

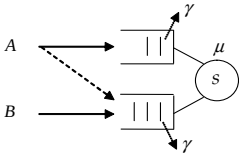


Figure 3: Scheduling policy $\pi_3$

transient ratio.

In Conjecture 1, we consider an $M/GI/s + M$ queue (Markovian interarrival times) and we conjecture some properties related to the achieved ratios.

**Conjecture 1** *Using the above notations, the following holds.*

1. $\pi_1$ *reaches* $\mathbf{c}^*$ *if and only if* $\mathbf{c}_{\pi_A} \leq \mathbf{c}^* \leq \mathbf{c}_{\pi_B}$.

2. $\pi_2$ *reaches* $\mathbf{c}^*$ *if and only if* $\mathbf{c}_{\pi_A} \leq \mathbf{c}^* \leq 1$.

3. $\pi_3$ *reaches* $\mathbf{c}^*$ *if and only if* $\mathbf{c}_{\pi_A} \leq \mathbf{c}^* \leq 1$.

In what follows, we give an intuitive explanation of why we expect this conjecture to hold. Let us consider the first statement of Conjecture 1 and let us take our basic model working under the scheduling policy $\pi_1$. Because we are considering workconserving non-preemptive policies, the lower and upper bounds of the achievable ratios are $\mathbf{c}_{\pi_A}$ and $\mathbf{c}_{\pi_B}$, respectively. It is clear that if we apply policy $\pi_A$ (strict priority to class $A$) starting at time zero, we will reach the ratio $\mathbf{c}_{\pi_A}$ in the long run. We will also reach $\mathbf{c}_{\pi_A}$ in the long run if we apply $\pi_A$ starting at any time $t > 0$. The reason is related to the fact that the number of customers in system, at the epochs of service completions, is a renewal process (Poisson process of arrivals and i.i.d. service times). The above remark is also valid for policy $\pi_B$ (strict priority to class $B$) and the upper bound $\mathbf{c}_{\pi_B}$. These are the two extremes. Consider now $\mathbf{c}^*$ ranging between $\mathbf{c}_{\pi_A}$ and $\mathbf{c}_{\pi_B}$. At time $t > 0$, if it happens that the achieved ratio is strictly lower than $\mathbf{c}^*$, then giving priority to customers of class $B$ (which is possible under $\pi_1$) allows necessarily to go beyond $\mathbf{c}^*$ after a given duration of time. Continuing in doing this, the ratio will converge to $\mathbf{c}_{\pi_B}$. In the other case, if it happens at time $t$ that the achieved ratio is strictly greater than $\mathbf{c}^*$, then giving priority to customers of class $A$ (which is possible under $\pi_1$) allows necessarily to go below $\mathbf{c}^*$ after a given duration of time. Continuing in

doing this, the ratio converges to $\mathbf{c}_{\pi_A}$. So, $\pi_1$ makes the achieved ratio fluctuating below and above $\mathbf{c}^*$. Continuing applying these manipulations, we conjecture that the ratio of the abandonment probabilities will converge to $\mathbf{c}^*$. This has been confirmed by simulation experiments.

Consider now the second statement. Let $\mathbf{c}_{FCFS}$ be the achieved ratio under the FCFS policy. With the same explanation as above, we can see that any ratio ranging from $\mathbf{c}_{\pi_A}$ to $\mathbf{c}_{FCFS}$ can be reached by $\pi_2$. On the one hand, The lower bound for $\mathbf{P}^A$ is achieved when we give strict priority to class $A$. In addition at the same time, the upper bound for $\mathbf{P}^B$ is reached. Thus, the minimum possible achievable target ratio corresponds to policy $\mathbf{c}_{\pi_A}$. On the other hand, the lower bound for $\mathbf{P}^B$ is achieved by assigning all class $B$ arrivals to queue 1, which also allows to achieve the upper bound for $\mathbf{P}^A$. This corresponds to the FCFS policy for arrivals of all classes. Hence, the achieved ratio could not be worse than that under the FCFS policy, $\mathbf{c}_{FCFS} = 1$. Finally, we note that the explanation of the third statement is similar to that of the second one. This finishes the discussion about Conjecture 1.

One may construct several auxiliary policies similar to the ones above. For example, instead of changing the priority rule at each new arrival epoch, we only change it at the arrival epoch of the customer who finds all servers busy and both queues empty. Then, we continue that rule until the end of the current busy period. With regard to reaching the target ratio, the latter class of policies has the same properties as those of $\pi_1$, $\pi_2$ and $\pi_3$. One drawback could be that they are less reactive to correct the transient ratio. A further possibility is to construct similar policies by changing the priority rule cyclically; at a given arrival and based on the transient ratio, we decide for the priority rule and we apply it for a given fixed number of arrivals. Once the cycle finishes, we determine the priority rule at the epoch of the arrival that follows the cycle. Again, we keep that rule for the same given fixed number of new arrivals, and so on.

## 4.2 Call Selection Policies

In the following we focus on a dual family of policies, namely the family of call selection policies. We assign all class $A$ (class $B$) customers to queue 1 (queue 2). On the contrary to the previous class of policies, we do not implement a fixed priority discipline between the queues. Customers waiting in the queues are selected by using so-called waiting time factors, as described in Lu and Squillante (2004). These selections occur after each service completion of an agent, when both queues are non-empty. The choice about the next customer to serve depends on the waiting time factors, of which exactly one is associated with each queue. When selecting a customer for service, the first customer in each queue is considered. From these two customers, we serve the customer for which the product of its waiting time and the waiting time factor is the highest. The waiting

time factors are only queue-dependent.

This family of policies has appealing properties: it takes the actual waiting times of the customers into account, is fair with respect to the order of service within each class, and the reactivity of the policy is expected to be high. Note that setting both waiting time factors equal results in a policy that serves customers from the two classes within in a FCFS order, and by taking one factor equal to 0, one of the classes has full priority over the other, independent of the waiting times.

Let us denote the waiting time factors by $\alpha_A$ and $\alpha_B$ for queues 1 (waiting class $A$ customers) and 2 (waiting class $B$ customers), respectively. The waiting time factors are dynamically adjusted by a method having parameter $\beta$, with $\beta \in [0, 1]$. At each epoch of a service completion, say $t$, one of the waiting time factors is set to $\beta$, while the other one is set to 1. Thus, there are two possibilities at $t$, namely $(\alpha_A(t), \alpha_B(t)) \in \{(1, \beta), (\beta, 1)\}$. At time $t$, the abandonment probabilities of the two classes, $\mathbf{P}^A(t)$ and $\mathbf{P}^B(t)$, are computed by considering the historical events. Then, we can formulate the proposed policy as

$$(\alpha_A(t), \alpha_B(t)) \leftarrow \begin{cases} (1, \beta) \text{ if } \mathbf{P}^A(t) \geq \mathbf{c}^* \mathbf{P}^B(t) \\ (\beta, 1) \text{ if } \mathbf{P}^A(t) < \mathbf{c}^* \mathbf{P}^B(t) \end{cases} . \tag{4}$$

We do not claim that this is the best way to update the waiting time factors. This would require a more thorough analysis which out of the scope of this paper. An immediate advantage of the current method is a high response in fluctuating transient scenarios. More comments about this policy are given in Section 5. In the same way as that in the explanation of Conjecture 1, we conjecture here that waiting time factors as in Equation (4) reach any objective between the two limiting cases where one of the two classes has full priority, namely $\mathbf{c}_{\pi_A}$ and $\mathbf{c}_{\pi_B}$. In addition, we conjecture that for each feasible objective $\mathbf{c}^*$, there exists $\beta_{\mathbf{c}^*} \in [0, 1]$ so that choosing any value in $[0, \beta_{\mathbf{c}^*}]$ and assigning it to the parameter $\beta$ allows to reach $\mathbf{c}^*$. Any value beyond $\beta_{\mathbf{c}^*}$ could not allow to reach $\mathbf{c}^*$. Just think of what happens near the boundary $\beta = 0$. In this case, swapping the factors $\alpha_A(t)$ and $\alpha_B(t)$ will give each time the full priority to one of the classes such that the transient ratio fluctuates around $\mathbf{c}^*$. When $\beta$ increases, swapping the waiting time factors does no longer necessarily guarantee that the right class gets the priority so that the transient ratio moves in the right way. In the limit case when $\beta$ approaches 1, swapping will not allow to force the priority for one given class. So, the objective $\mathbf{c}^*$ can no longer be reached ($\mathbf{c}(t)$ will converge to 1 which corresponds to FCFS).

Alternatively, another method for updating the waiting time factors is to keep $\alpha_A$ fixed over time and to update $\alpha_B$ dynamically. Under the assumption that $\alpha_A$ is fixed, it is straightforward to decrease $\alpha_B$ when $\mathbf{P}^A$ is high, and increase $\alpha_B$ otherwise. A difficulty is to develop a method

that updates $\alpha_B$ in a simple but effective way, without many control parameters. We investigated several possibilities by means of numerical experiments. As a result, we decided to decrease the parameter $\alpha_B$ by multiplication by $\eta$ and increase it by division by $\eta$, with $0 < \eta \leq 1$. The factor $\alpha_B$ was updated after each event of service completion.

# 5 Simulation Experiments

The nature of the scheduling policies we are considering in this paper (online policies) makes a tractable analysis too complicated. Both analytical and numerical methods are hard to be derived. We thereafter resort to simulation experiments in order to prove the efficiency of the proposed policies and gain useful guidelines. We consider numerical examples of Markovian systems: exponential interarrival, service and patience times. Note that in practice, service times of premium calls are likely to be the largest because the call center would pay more attention to them by giving more personalized answers, $\mu_A < \mu_B$. For patience times, premium customers should be in most call center cases more patient than regular customers because the formers know that they will experience a good quality of answer after their waiting, $\gamma_A < \gamma_B$.

We consider 6 systems, denoted by System 1, ..., System 6. Systems parameters are chosen such that we get realistic scenarios. In all systems the number of servers is $s = 50$. The service rates are $\mu_A = 0.15$ and $\mu_B = 0.25$. The abandonment rates are $\gamma_A = 0.3$ and $\gamma_B = 0.35$. From one system to another, we vary the total arrival rate so as we get different "service utilizations", $\frac{\lambda_A+\lambda_B}{s\mu}$. We choose, $\lambda_A = \lambda_B = 5$, 6, 7, 9, 11, and 13, respectively. The "service utilization" is increasing starting from 100% in System 1 until 260% in System 6. On purpose we choose highly loaded systems in order to see the effect of abandonment. We consider balanced cases for the arrival processes in order to avoid including further aspects that may complicate the understanding of the simulation results. Recall that abandonments make our systems unconditionally stable. The simulations are done for the target ratios $\mathbf{c}^* = 0.5$, 0.7 and 0.9. We determined for each system the interval $[\mathbf{c}_{\pi_A}, \mathbf{c}_{\pi_B}]$, and we checked that the values $\mathbf{c}^* = 0.5$, 0.7 and 0.9 are ranging in all of these intervals. We notice for the balanced examples ($\lambda_A = \lambda_B$) considered here we have $\mathbf{c}_{\pi_A} = \frac{1}{\mathbf{c}_{\pi_B}}$.

For queue joining policies, we present the performance measures under policies $\pi_1$, $\pi_2$ and $\pi_3$, as well as those under policy $\pi_A$ (high priority for class $A$ customers). For call selection policies, we give the performance measures for $\beta$ equal to 0, 1/4, 1/2, 3/4, and 1.

## 5.1 Experiments for the Queue Joining Policies

The simulation results for $\mathbf{c}^* = 0.7$ are presented in Tables 1–6. In Tables 13–18 and 19–24 in Section 2 of the supplementary material Jouini et al. (2010), the results are presented for $\mathbf{c}^* = 0.5$ and 0.9, respectively. The rows, corresponding to the quantity $\mathbf{c}$, are to indicate the achieved long

run ratio under each scheduling policy.

| | $\pi_A$ | $\pi_1$ | $\pi_2$ | $\pi_3$ |
|---|---|---|---|---|
| $\mathbf{c}$ | 0.203 | 0.700 | 0.700 | 0.700 |
| $\mathbf{P}^A$ | 3.915% | 8.462% | 8.483% | 8.434% |
| $\mathbf{P}^B$ | 19.249% | 12.089% | 12.119% | 12.048% |
| $\mathbf{P}$ | 11.583% | 10.276% | 10.302% | 10.240% |
| $W^A$ | 0.128 | 0.254 | 0.273 | 0.270 |
| $W^B$ | 0.487 | 0.297 | 0.314 | 0.324 |
| $W$ | 0.292 | 0.275 | 0.293 | 0.296 |
| $\sigma^A$ | 0.187 | 0.484 | 0.373 | 0.383 |
| $\sigma^B$ | 0.754 | 0.567 | 0.483 | 0.467 |
| $\sigma$ | 0.557 | 0.527 | 0.431 | 0.427 |

Table 1: $\lambda_A = \lambda_B = 5$, $\mathbf{c}^* = 0.7$

| | $\pi_A$ | $\pi_1$ | $\pi_2$ | $\pi_3$ |
|---|---|---|---|---|
| $\mathbf{c}$ | 0.160 | 0.700 | 0.700 | 0.700 |
| $\mathbf{P}^A$ | 7.479% | 19.233% | 19.160% | 19.191% |
| $\mathbf{P}^B$ | 46.803% | 27.476% | 27.372% | 27.415% |
| $\mathbf{P}$ | 27.137% | 23.355% | 23.267% | 23.302% |
| $W^A$ | 0.249 | 0.585 | 0.667 | 0.658 |
| $W^B$ | 1.440 | 0.690 | 0.759 | 0.819 |
| $W$ | 0.684 | 0.635 | 0.711 | 0.734 |
| $\sigma^A$ | 0.253 | 0.827 | 0.522 | 0.565 |
| $\sigma^B$ | 1.286 | 0.980 | 0.705 | 0.681 |
| $\sigma$ | 0.986 | 0.904 | 0.617 | 0.628 |

Table 2: $\lambda_A = \lambda_B = 6$, $\mathbf{c}^* = 0.7$

| | $\pi_A$ | $\pi_1$ | $\pi_2$ | $\pi_3$ |
|---|---|---|---|---|
| $\mathbf{c}$ | 0.155 | 0.700 | 0.700 | 0.700 |
| $\mathbf{P}^A$ | 10.851% | 28.507% | 28.502% | 28.523% |
| $\mathbf{P}^B$ | 70.096% | 40.725% | 40.717% | 40.748% |
| $\mathbf{P}$ | 40.488% | 34.613% | 34.609% | 34.635% |
| $W^A$ | 0.365 | 0.864 | 1.062 | 1.040 |
| $W^B$ | 2.724 | 0.968 | 1.190 | 1.340 |
| $W$ | 0.958 | 0.911 | 1.120 | 1.176 |
| $\sigma^A$ | 0.327 | 1.142 | 0.602 | 0.696 |
| $\sigma^B$ | 1.750 | 1.396 | 0.799 | 0.815 |
| $\sigma$ | 1.377 | 1.264 | 0.701 | 0.767 |

Table 3: $\lambda_A = \lambda_B = 7$, $\mathbf{c}^* = 0.7$

| | $\pi_A$ | $\pi_1$ | $\pi_2$ | $\pi_3$ |
|---|---|---|---|---|
| $\mathbf{c}$ | 0.219 | 0.700 | 0.700 | 0.700 |
| $\mathbf{P}^A$ | 20.473% | 41.288% | 41.273% | 41.262% |
| $\mathbf{P}^B$ | 93.598% | 58.982% | 58.961% | 58.946% |
| $\mathbf{P}$ | 57.039% | 50.134% | 50.113% | 50.104% |
| $W^A$ | 0.723 | 1.227 | 1.691 | 1.633 |
| $W^B$ | 5.500 | 0.924 | 1.926 | 2.139 |
| $W$ | 1.079 | 1.102 | 1.788 | 1.841 |
| $\sigma^A$ | 0.509 | 1.536 | 0.740 | 0.889 |
| $\sigma^B$ | 2.628 | 1.803 | 0.840 | 0.957 |
| $\sigma$ | 1.526 | 1.658 | 0.791 | 0.951 |

Table 4: $\lambda_A = \lambda_B = 9$, $\mathbf{c}^* = 0.7$

We see from the experiments that the target ratio is always met by policies $\pi_1$, $\pi_2$ and $\pi_3$, which agrees with Conjecture 1. For each system, the value of the ratio under policy $\pi_A$ represents a lower bound for the achievable ratio under any workconserving non-preemptive scheduling policy. We can not do better when considering that class of policies.

| | $\pi_A$ | $\pi_1$ | $\pi_2$ | $\pi_3$ |
|---|---|---|---|---|
| $\mathbf{c}$ | 0.325 | 0.700 | 0.700 | 0.700 |
| $\mathbf{P}^A$ | 32.236% | 49.436% | 49.430% | 49.444% |
| $\mathbf{P}^B$ | 99.237% | 70.623% | 70.616% | 70.635% |
| $\mathbf{P}$ | 65.751% | 60.029% | 60.025% | 60.038% |
| $W^A$ | 1.237 | 1.581 | 2.161 | 2.119 |
| $W^B$ | 7.652 | 0.591 | 2.582 | 2.680 |
| $W$ | 1.309 | 1.217 | 2.316 | 2.325 |
| $\sigma^A$ | 0.624 | 1.543 | 0.864 | 0.933 |
| $\sigma^B$ | 3.343 | 1.274 | 0.749 | 0.816 |
| $\sigma$ | 0.981 | 1.526 | 0.849 | 0.932 |

Table 5: $\lambda_A = \lambda_B = 11$, $\mathbf{c}^* = 0.7$

| | $\pi_A$ | $\pi_1$ | $\pi_2$ | $\pi_3$ |
|---|---|---|---|---|
| $\mathbf{c}$ | 0.426 | 0.700 | 0.700 | 0.700 |
| $\mathbf{P}^A$ | 42.588% | 55.095% | 55.074% | 55.084% |
| $\mathbf{P}^B$ | 99.949% | 78.702% | 78.754% | 78.695% |
| $\mathbf{P}$ | 71.284% | 66.897% | 66.912% | 66.888% |
| $W^A$ | 1.784 | 1.974 | 2.542 | 2.534 |
| $W^B$ | 8.563 | 0.519 | 3.134 | 3.153 |
| $W$ | 1.791 | 1.506 | 2.732 | 2.733 |
| $\sigma^A$ | 0.659 | 1.365 | 0.918 | 0.932 |
| $\sigma^B$ | 4.263 | 0.608 | 0.623 | 0.633 |
| $\sigma$ | 0.701 | 1.358 | 0.880 | 0.895 |

Table 6: $\lambda_A = \lambda_B = 13$, $\mathbf{c}^* = 0.7$

We further analyze in this section the expected and the standard deviation of the waiting time in queue of each class of served customers. A rigorous proof of our claims is out of the scope of this paper. We only give some general ideas and intuitive explanations to support the claims we derive. We especially focus on the standard deviation of the waiting time. It has often been argued that a system with reasonable and predictable waiting times may be more desirable than a system with a lower expected waiting time but a higher variance, see Lu and Squillante (2004) for more details.

For class $A$ customers, starting from the lower value, most experiments show that the standard deviation values are ordered according to policies $\pi_A$, $\pi_2$, $\pi_3$ and $\pi_1$. The reason is basically related to the well-known property in queueing theory which claims that the FCFS discipline minimizes waiting time variance (time in queue and in system) when the queueing discipline is service time independent. We refer the reader to Randolph (1991) for a more extensive discussion. The best we can do for customers of class $A$ under a workconserving non-preemptive policy is not to give at any time the priority to customers of class $B$. Such a situation allows the realizations of class $A$ waiting times to be minimized. This is the case for policy $\pi_A$. Next, since the discipline of service within queue 1 is FCFS, $\pi_A$ should lead to the lower variance. With regard to the order of service of customers of class $A$, policy $\pi_1$ deviates more than $\pi_2$ and $\pi_3$ from the FCFS discipline. This tells us that $\pi_1$ has the highest variance. When comparing policies $\pi_2$ and $\pi_3$, one may see that on the contrary of policy $\pi_3$, policy $\pi_2$ respects the FCFS order for customers of class $A$, which indicates a lower variance than under policy $\pi_3$. The experiments for class $A$ show also that the expected waiting times in queue are ordered according to policies $\pi_A$, $\pi_1$, $\pi_3$ and $\pi_2$. The order $\pi_1$ then $\pi_3$ then $\pi_2$ is expected because of the general property that FCFS maximizes the expected waiting time of served customers. Policy $\pi_A$ is the best for the expected waiting time of served customers. An explanation would be related to the small values of waiting times achieved under that policy.

For class $B$ customers, starting from the lower value, we conclude from the majority of the experiments that the standard deviation values are structured for policies $\pi_3$, $\pi_2$, $\pi_1$ and $\pi_A$. When comparing policies $\pi_1$, $\pi_2$ and $\pi_3$, the explanation is identical to that conducted for the first comment. As for policy $\pi_A$, the only explanation we have is related to the waiting time values. The larger waiting times of class $B$ customers are achieved under policy $\pi_A$ because of their lower priority. This might explain why that policy has the highest variance for waiting times of class $B$ customers, because generally higher values have also higher variance. In addition to that, uncertainty in interarrival, abandonment and service times, allows to have a non-zero probability that some customers of class $B$ enter service without waiting or within only a short delay. This makes their waiting times variance to be the highest. From the experiments, we also see for class $B$ that the expected waiting times in queue of served customers are ordered according to policies $\pi_1$, $\pi_2$, $\pi_3$ and $\pi_A$. One may explain these results through the same arguments used above.

We notice that we only gave some directions to compare the policies with regard to the expected value and standard deviation of waiting times. For instance, the comparison should lie in the values of $\lambda_A$ and $\lambda_B$, also in the target ratio constraint $\mathbf{c}^*$. A target ratio close to 1 would make our policies work in a similar manner than that of the FCFS discipline, whereas an objective far from 1 (being

under or beyond) would make the policies similar to the strict priority policy. Based on the analysis here, we can not distinguish a best policy. However, one may recommend policy $\pi_2$. First, it reaches the objective ratio. Second, it gives (in most cases) the lowest variance of waiting times of customers $A$. Third, it allows to have a "good" variance for customers $B$, as well as for all customers.

## 5.2 Experiments for the Call Selection Policies

The simulation results are presented in Table 7–12 below for $\mathbf{c}^* = 0.7$. In Tables 25–30 and 31–36 of the supplementary material Jouini et al. (2010), the results are presented for $\mathbf{c}^* = 0.5$ and 0.9, respectively. The rows (corresponding to the quantity $\mathbf{c}$) display the achieved long run ratio under different values of $\beta$. The rest of this section is devoted to discuss the simulation results.

| $\beta$ | 0 | 0.25 | 0.5 | 0.75 | 1.0 |
|---|---|---|---|---|---|
| $\mathbf{c}$ | 0.699 | 0.700 | 0.700 | 0.720 | 0.875 |
| $\mathbf{P}^A$ | 0.084 | 0.085 | 0.084 | 0.086 | 0.093 |
| $\mathbf{P}^B$ | 0.121 | 0.121 | 0.120 | 0.119 | 0.107 |
| $\mathbf{P}$ | 0.102 | 0.103 | 0.102 | 0.102 | 0.100 |
| $W^A$ | 0.258 | 0.270 | 0.272 | 0.278 | 0.301 |
| $W^B$ | 0.303 | 0.322 | 0.325 | 0.325 | 0.294 |
| $W$ | 0.280 | 0.295 | 0.298 | 0.301 | 0.298 |
| $\sigma^A$ | 0.463 | 0.398 | 0.366 | 0.358 | 0.398 |
| $\sigma^B$ | 0.547 | 0.489 | 0.458 | 0.440 | 0.392 |
| $\sigma$ | 0.506 | 0.446 | 0.415 | 0.401 | 0.395 |

Table 7: $\lambda_A = \lambda_B = 5$, $\mathbf{c}^* = 0.7$

| $\beta$ | 0 | 0.25 | 0.5 | 0.75 | 1.0 |
|---|---|---|---|---|---|
| $\mathbf{c}$ | 0.700 | 0.700 | 0.700 | 0.720 | 0.882 |
| $\mathbf{P}^A$ | 0.192 | 0.192 | 0.192 | 0.194 | 0.213 |
| $\mathbf{P}^B$ | 0.274 | 0.274 | 0.274 | 0.269 | 0.241 |
| $\mathbf{P}$ | 0.233 | 0.233 | 0.233 | 0.232 | 0.227 |
| $W^A$ | 0.606 | 0.652 | 0.667 | 0.683 | 0.751 |
| $W^B$ | 0.719 | 0.804 | 0.833 | 0.829 | 0.736 |
| $W$ | 0.660 | 0.724 | 0.746 | 0.752 | 0.744 |
| $\sigma^A$ | 0.763 | 0.595 | 0.508 | 0.479 | 0.537 |
| $\sigma^B$ | 0.921 | 0.744 | 0.660 | 0.611 | 0.533 |
| $\sigma$ | 0.843 | 0.674 | 0.591 | 0.551 | 0.535 |

Table 8: $\lambda_A = \lambda_B = 6$, $\mathbf{c}^* = 0.7$

| $\beta$ | 0 | 0.25 | 0.5 | 0.75 | 1.0 |
|---|---|---|---|---|---|
| $\mathbf{c}$ | 0.700 | 0.700 | 0.700 | 0.727 | 0.888 |
| $\mathbf{P}^A$ | 0.285 | 0.285 | 0.285 | 0.290 | 0.316 |
| $\mathbf{P}^B$ | 0.407 | 0.408 | 0.407 | 0.399 | 0.356 |
| $\mathbf{P}$ | 0.346 | 0.346 | 0.346 | 0.345 | 0.336 |
| $W^A$ | 0.921 | 1.028 | 1.065 | 1.103 | 1.214 |
| $W^B$ | 1.057 | 1.313 | 1.385 | 1.377 | 1.197 |
| $W$ | 0.983 | 1.157 | 1.210 | 1.228 | 1.206 |
| $\sigma^A$ | 1.036 | 0.738 | 0.574 | 0.512 | 0.574 |
| $\sigma^B$ | 1.292 | 0.958 | 0.765 | 0.665 | 0.572 |
| $\sigma$ | 1.161 | 0.857 | 0.686 | 0.603 | 0.573 |

Table 9: $\lambda_A = \lambda_B = 7$, $\mathbf{c}^* = 0.7$

| $\beta$ | 0 | 0.25 | 0.5 | 0.75 | 1.0 |
|---|---|---|---|---|---|
| $\mathbf{c}$ | 0.700 | 0.700 | 0.700 | 0.750 | 0.900 |
| $\mathbf{P}^A$ | 0.412 | 0.413 | 0.413 | 0.426 | 0.460 |
| $\mathbf{P}^B$ | 0.589 | 0.590 | 0.590 | 0.568 | 0.511 |
| $\mathbf{P}$ | 0.501 | 0.501 | 0.501 | 0.497 | 0.486 |
| $W^A$ | 1.351 | 1.614 | 1.709 | 1.809 | 2.003 |
| $W^B$ | 1.126 | 2.135 | 2.382 | 2.313 | 1.986 |
| $W$ | 1.258 | 1.828 | 1.986 | 2.025 | 1.995 |
| $\sigma^A$ | 1.420 | 0.957 | 0.643 | 0.526 | 0.582 |
| $\sigma^B$ | 1.718 | 1.417 | 0.946 | 0.686 | 0.581 |
| $\sigma$ | 1.553 | 1.196 | 0.849 | 0.650 | 0.582 |

Table 10: $\lambda_A = \lambda_B = 9$, $\mathbf{c}^* = 0.7$

| $\beta$ | 0 | 0.25 | 0.5 | 0.75 | 1.0 |
|---|---|---|---|---|---|
| $\mathbf{c}$ | 0.700 | 0.700 | 0.700 | 0.772 | 0.910 |
| $\mathbf{P}^A$ | 0.494 | 0.495 | 0.494 | 0.517 | 0.553 |
| $\mathbf{P}^B$ | 0.706 | 0.707 | 0.706 | 0.669 | 0.608 |
| $\mathbf{P}$ | 0.600 | 0.601 | 0.600 | 0.593 | 0.581 |
| $W^A$ | 1.692 | 2.062 | 2.201 | 2.381 | 2.635 |
| $W^B$ | 0.769 | 2.733 | 3.275 | 3.075 | 2.617 |
| $W$ | 1.353 | 2.309 | 2.596 | 2.663 | 2.627 |
| $\sigma^A$ | 1.501 | 1.084 | 0.668 | 0.531 | 0.583 |
| $\sigma^B$ | 1.336 | 1.889 | 1.126 | 0.694 | 0.582 |
| $\sigma$ | 1.510 | 1.469 | 1.008 | 0.692 | 0.582 |

Table 11: $\lambda_A = \lambda_B = 11$, $\mathbf{c}^* = 0.7$

| $\beta$ | 0 | 0.25 | 0.5 | 0.75 | 1.0 |
|---|---|---|---|---|---|
| $\mathbf{c}$ | 0.700 | 0.700 | 0.700 | 0.791 | 0.918 |
| $\mathbf{P}^A$ | 0.550 | 0.551 | 0.551 | 0.582 | 0.618 |
| $\mathbf{P}^B$ | 0.786 | 0.787 | 0.787 | 0.736 | 0.674 |
| $\mathbf{P}$ | 0.668 | 0.669 | 0.669 | 0.659 | 0.646 |
| $W^A$ | 2.007 | 2.423 | 2.601 | 2.863 | 3.157 |
| $W^B$ | 0.571 | 3.193 | 4.149 | 3.713 | 3.141 |
| $W$ | 1.545 | 2.671 | 3.099 | 3.192 | 3.150 |
| $\sigma^A$ | 1.361 | 1.125 | 0.651 | 0.539 | 0.583 |
| $\sigma^B$ | 0.703 | 2.324 | 1.267 | 0.703 | 0.581 |
| $\sigma$ | 1.36611 | 1.651 | 1.152 | 0.735 | 0.582 |

Table 12: $\lambda_A = \lambda_B = 13$, $\mathbf{c}^* = 0.7$

From the experiments, we obviously see that all quantities increase when the workload increases. If $\beta$ is high, for example 3/4 or 1, then the ratio $\mathbf{c}^*$ is not reached. The reason is that the weights of the waiting times are almost equal, such that the discipline of service of the customers waiting

in the two queues is close to a FCFS policy. Such high values of $\beta$ do not allow to sufficiently discriminate one class of customers to the detriment of the other. Both classes have almost equal priority of service. In addition, the parameter $\beta_{\mathbf{c}^*}$ (introduced in Section 4.2) is increasing in $\mathbf{c}^*$. The explanation is as follows. As $\mathbf{c}^*$ increases, giving the priority to one class over the other one is less and less needed. This allows $\beta_{\mathbf{c}^*}$ to be higher. In the limit case (for $\mathbf{c}^* = 1$), $\beta_{\mathbf{c}^*}$ reaches its upper bound ($\beta_{\mathbf{c}^*} = 1$).

We see also from the experiments that the overall expected waiting time of the served customers increases as $\beta$ increases. The reason pertains to the fact that as $\beta$ increases, the scheduling policy approaches the FCFS policy. As shown in Theorem 3, the FCFS policy maximizes the expected waiting time in queue of served customers. Meanwhile, the expected waiting time of the abandoned customers decreases. Finally, the variance of the waiting times (for all classes) is decreasing in $\beta$. One intuitive explanation is due to the known property of the FCFS policy, which minimizes the waiting times variance. As $\beta$ increases, the behavior of our policy approaches that of the FCFS policy, and as a consequence, the overall variance decreases. In the limit case ($\beta = 0$), we have an alternation for service selection which is the most distant from the FCFS policy. Not surprisingly, its corresponding variances are the highest.

It is also interesting to observe that $\beta_{\mathbf{c}^*}$ is independent of the workload. One intuitive explanation would be related to the balanced cases we consider here, $\lambda_A = \lambda_B = \lambda$. In such cases, increasing $\lambda$ will increase the number of arrivals of both classes by the same factor. For a given objective ratio $\mathbf{c}^*$, choosing one $\beta$ in the same interval $[0, \beta_{\mathbf{c}^*}]$ allows thereafter to increase the number of abandonments of both classes by the same factor, which allows to reach $\mathbf{c}^*$. The reason is that we keep unchanged the way we are giving priorities to customer classes. If we choose $\beta$ beyond $\beta_{\mathbf{c}^*}$, we have no longer the sufficient flexibility to discriminate customer classes so as we increase the abandonments of both classes in the same way as the arrivals.

## 5.3 Comparison

In what follows we address the comparison of the two families of policies: queue joining and call selection policies. We use the simulations results to draw the basic conclusions.

Not surprisingly the expected waiting times of the served customers are shorter for the queue joining policies. This conclusion is valid for each class of customers and for all classes. The reason has to do with the order in which customers are scheduled for service. Under the call selection policies, customers of each class are served in the order of their arrival. This makes the waiting times larger under that type of policies (see Theorem 3). As a consequence, the expected waiting time of all served customers is also the highest under the call selection policies. The second conclusion

is that the variances (equivalently the standard deviations) of the waiting times are lower for the call selection policies. The explanation is identical to the previous one and is again pertaining to the order of service of customers.

One may also compare both types of policies from the reactivity perspective. We mean by reactivity, the speed of a given policy to make the transient ratio $\mathbf{c}(t)$ close to the objective. Although both types of policies allow to satisfy the target ratio constraint in the long run, we comment that they have different reactivities. The call selection policies are more reactive. One intuitive explanation is as follows. When for example the transient ratio is lower than the objective, both policies give high priority to class $B$ in order to increase this transient ratio. A queue joining policy as $\pi_2$ will assign a new customer $B$ to the queue with higher priority, however it may happen that there are type $A$ waiting customers in that queue ahead of the customer $B$ of interest. In such a case, the transient ratio would thereafter still decrease. It would increase once we finish to serve the type $A$ customers waiting in the highest priority queue. This case does not occur under the call selection policies. Depending on the circumstances, the latter policies immediately allow to increase or decrease the transient ratio, by scheduling one given class for service to the detriment of the other class.

The reactivity of the call selection policy strongly depends on the value of $\beta$. It is maximized for $\beta = 0$ such that customers of the class that allows to push the transient ratio to $\mathbf{c}^*$ is served first. For higher values of $\beta$ the waiting times of the customers are more taken into account such that customers that have waited longer than others would have a higher priority for service. This can yield actions that are sometimes not optimal with respect to the ratio constraint. By means of the waiting time factors, a trade-off is made between serving the customer with the longest waiting time and serving the customer that improves the transient ratio of abandonment probabilities. Sometimes the optimal actions of both objectives will be the same but not in all cases.

# 6 Conclusions

We focused on a fundamental real-time problem of call centers management. We considered a two-class call center and developed online scheduling policies subject to satisfying a target ratio constraint of the abandonment probabilities of the two customer classes. This new formulation of the control problem is robust with respect to the system workload. Furthermore, it generalizes the traditional formulation where we have a target abandonment probability for each class. The policies are argued to be relevant in practice; they are efficient, easy to understand for managers, predictable and easy to implement.

First, we gave some structural results in order to better understand the impact of different

scheduling policies on the performance measures of interest. Second, we proposed several online scheduling policies allowing to meet the considered constraint. Third, we conducted a simulation study in order to compare the proposed policies with regard to several performance measures such as the expected and the variance of the waiting time in queue of served customers.

In this paper, we focused the analysis on a given period of the day. In future research we would like to focus on intervals of a day. Then it would be interesting to find a method that translates the whole day objective into a set of objectives per period of the day. It would be also interesting to extend the analysis to more than two customer classes and agents with different skill sets.

# 7 Acknowledgements

The authors would like to express their gratitude to Rabie Nait-Abdallah from Bouygues Telecom for several useful discussions which were at the origin of the problem analyzed in this paper.

# References

Akşin, O., Armony, M., and Mehrotra, V. (2007). The Modern Call-Center: A Multi-Disciplinary Perspective on Operations Management Research. *Production and Operations Management*, 16:665–688.

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

Avi-Itzhak, B. and Levy, H. (2004). On Measuring Fairness in Queues. *Advances In Applied Probability*, 36:919–936.

Avramidis, A., Chan, W., Gendreau, M., l'Ecuyer, P. and Pisacane, O. (2010). Optimizing Daily Agent Scheduling in a Multiskill Call Center. *European Journal of Operational Research*, 200:822–832.

Avramidis, A., Deslauriers, A., and l'Ecuyer, P. (2004). Modeling Daily Arrivals to a Telephone Call Center. *Management Science*, 50:896–908.

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

Bhulai, S. and Koole, G. (2003). A Queueing Model for Call Blending in Call Centers. *IEEE Transactions on Automatic Control*, 48:1434–1438.

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

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2005). Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. *Journal of the American Statistical Association*, 100:36–50.

Dai, J. and Tezcan, T. (2008). Optimal Control of Parallel Server Systems with Many Servers in Heavy Traffic. *Queueing Systems*, 59:95–134.

Duder, J.C. and Rosenwein, M.B. (2001). Towards "Zero Abandonments" in Call Center Performance. *European Journal of Operational Research*, 135:50–56.

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

Garnett, O. and Mandelbaum, A. (2001). An Introduction to Skills-Based Routing and its Operational Complexities. Teaching notes, Technion.

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

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

Guérin, R. (1998). Queueing-Blocking System with two Arrival Streams and Guard Channels. *IEEE Transactions on Communications*, 36:153–163.

Gurvich, I. (2004). Design and Control of the M/M/N Queue with Multi-Class Customers and Many Servers. Masters thesis, Technion.

Gurvich, I., Armony, M., and Mandelbaum, A. (2008). Service-Level Differentiation in Call Centers with Fully Flexible Servers. *Management Science*, 54:279–294.

Gurvich, I. and Whitt, W. (2008). Service-Level Differentiation in Many-Server Service System via Queue-Ratio Routing. *Operations Research*. To appear.

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

Jongbloed, G. and Koole, G. (2001). Managing Uncertainty in Call Centers using Poisson Mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318.

Jouini, O. and Dallery, Y. (2007). Monotonicity Properties for Multiserver Queues with Reneging and Finite Waiting Lines. *Probability in the Engineering and Informational Sciences*, 21:335–360.

Jouini, O., Pot, A., Koole, G. and Dallery, Y. (2009). Supplementary Material to "Online Scheduling Policies for Multiclass Call Centers with Impatient Customers". *European Journal of Operational Research* To appear.

Koole, G. (2004). A Formula for Tail Probabilities of Cox Distributions. *Journal of Applied Probability*, 41:935–938.

Lu, Y. and Squillante, M. (2004). Scheduling to Minimize General Functions of the Mean and Variance of Sojourn Times in Queueing Systems. Working Paper, IBM Research Division.

Maglaras, C. and Zeevi, A. (2005). Pricing and Design of Differentiated Services: Approximate Analysis and Structural Insights. *Operations Research*, 53:242–262.

Mandelbaum, A. and Zeltyn, S. (2009). Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers. *Operations Research*, 57:1189–1205.

Milner, J. and Olsen, T. (2008). Service-Level Agreements in Call Centers: Perils and Prescriptions. *Management Science*, 54:238–252.

Pekoz, E. (2002). Optimal Policies for Multi-Server Non-Preemptive Priority queues. *Queueing Systems*, 42:91–101.

Randolph, W. H. (1991). *Queueing Methods for Services and Manufacturing*. Prentice Hall.

Schrage, L. and Miller, L. (1966). The Queue M/G/1 with the Shortest Remaining Processing Time Discipline. *Operations Research*, 14:670–684.

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

Whitt, W. (1982). On the Heavy-Traffic Limit Theorem for GI/G/infinity Queues. *Advances in Applied Probability*, 14:171–190.

Whitt, W. (2004). Efficiency-Driven Heavy-Traffic Approximations for Many-Server Queues with Abandonments. *Management Science*, 50:1449–1461.

Whitt, W. (2005). Engineering Solution of a Basic Call-Center Model. *Management Science*, 51:221–235.

Whitt, W. (2006). Fluid Models for Multiserver Queues with Abandonments. *Operations Research*, 54:37–54.