

Queueing Models for Multiclass Call Centers with Real-Time Anticipated Delays

Oualid Jouini[†] Yves Dallery[†] Zeynep Aksin[‡]

[†] Ecole Centrale Paris
Laboratoire Génie Industriel
Grande Voie des Vignes
92295 Châtenay-Malabry Cedex, France
walid.jouini@ecp.fr, yves.dallery@ecp.fr

[‡] Koç University
College of Administrative Sciences and Economics
Rumeli Feneri Yolu
34450 Sariyer-Istanbul, Turkey
zaksin@ku.edu.tr

Corresponding author: Oualid Jouini (walid.jouini@ecp.fr, tel: +33141131166, fax: +33141131272.)

Submitted to *International Journal of Production Economics*, March 2007.

Abstract

In this paper, we consider two basic multiclass call center models, with and without reneging. We study the problem of announcing delays to customers upon their arrival. For the simplest model without reneging, we give a method to estimate virtual delays that will be used within the announcement step. For the second model, we first build the call center model incorporating reneging. The model takes into account the change in customer behavior that may occur when delay information is communicated to them. We develop a method based on Markov chains in order to estimate virtual delays of new arrivals for this model. Finally, some practical issues concerning delay announcement are discussed.

Keywords call centers, predicting delays, announcing delays, impatient customers, Markov chains, multiserver queues.

1 Introduction

Recent developments in technology and the business environment have dramatically increased the need for improved service systems. Well designed service systems allow to reduce costs and to promote user satisfaction. The service sector is gaining prominence both in terms of the number of researchers and practitioners working in it, and its contribution to the gross domestic product, see Artiba [6]. This paper deals with a well known service system, namely call centers. Call centers are

used to provide services in many areas and industries: emergency centers, information centers, help-desks, tele-marketing and more. A telephone service enables customers to obtain a fast response, with a minimal effort. Providing services via call centers, instead of a face-to-face service, usually translates into lower operational costs to the service provider. The call center industry has been steadily growing and it had been observed worldwide. Estimations indicate that around 70% of all customers transactions occur in call centers, see Nakibly [22]. Today, all Fortune 500 companies have at least one contact center. They employ an average of 4,500 agents across their sites. More than \$300 billion is spent annually on contact centers around the world, see McKinsey Quarterly [1].

The continued growth of call centers has brought with it a rich and interesting set of questions for both practitioners and academic researchers. This paper is motivated by a large French call center. As in many other organizations, the call center of our company constitutes the main point of contact with customers. Such centers have limited resources and face highly unpredictable demand that often result in long waits for customers. To improve customer satisfaction and alleviate congestion, call centers have recently started experimenting by informing arriving customers about anticipated delays, see Armony and Maglaras [4].

Information regarding anticipated delays is of a special importance in service systems with invisible queues, as in call centers. In such systems, the uncertainty involved in waiting is higher than that in visible queues, and it does not decrease over time. Customers have no means to estimate queue lengths or progress rate. So, feelings of frustration and anxiety increase during the waiting. We expect that delay information would avoid such situations, and make the waiting experience more acceptable. Zakay [31] stipulates that waiting information may distract customers' attention from the passage of time. Hence, they may perceive the length of the wait as short.

The goal of this paper is to study announcing delays under different call centers situations. We consider two different multiclass call centers models, with and without abandonments. Since we are dealing with stochastic systems, there is no possible way to predict exact waiting times. The best one can do is to estimate their distribution. The service provider should thereafter decide what is the exact information that will be provided to customers. For example, he may decide to provide the mean of the estimated waiting time or any other percentile of its distribution function. However, we should be careful: On the one hand, informing of a short waiting time, which is likely to underestimate the actual waiting, might reduce the reliability of the service provider in the eyes

of the customers. On the other hand, informing of a long waiting time, might result in longer perceived waiting times and in a decrease in satisfaction.

For each call center configuration, we develop a method to estimate virtual delays of new arrivals. Since the goal is to provide information which is relevant to a specific type of customer at a specific time, we focus on estimating the waiting time given the system state at the time of estimation. This is different from estimating the overall performance of the system, such as the average waiting time of all customers, which is usually done assuming a steady state. The models under consideration in this paper are characterized to be relevant in practice. The two major distinguishing features are priorities and the allowance for customers to be impatient. Priority mechanisms are a useful scheduling method that allows different customer types to receive differentiated performance levels. They are in addition known for their ease of implementation in practice. As for reneging, incorporating it in modeling is of great value in order to be as close as possible to reality. Waiting customers in call centers may naturally hang up once they feel that the time they spend in queue was too long, see Mandelbaum and Zeltyn [21].

The main contribution of this paper is the building of a multiclass call center model with impatient customers that incorporates delay information. We propose a model in which the original behavior of customers (reneging) is substituted by balking upon arrival. To fully characterize the new model, we compute for each type of customer the balking probabilities and derive closed form expressions for the moments of their distribution of virtual delays. We note that the analysis here may be viewed as an extension of the work of Whitt [27]. In the latter, the author addressed a similar problem for a call center model with a single class of impatient customers.

Here is how the rest of the paper is organized. In Section 2, we review some literature close to our work. In Section 3, we consider a multiclass call center model with infinitely patient customers. In Sections 3.1 and 3.2, we describe the model and develop a method for estimating virtual delays, respectively. This development would be used within the announcement step. We then move on and let customers renege while waiting in queue. In Section 4.1, we first describe the original model of the call center without delay information. In Section 4.2, we next focus on building the call center model assuming that it provides delay information to customers. In Section 4.3, we again develop a method to derive virtual delays for each type of customers. In Section 5, some practical issues are discussed. In Section 5.1, we question the need of announcing delays when queues are empty upon the arrival of a new customer. In Section 5.2, we present a useful approximation of virtual

delays in order to simplify the implementation of the announcement of delays. In Section 6, we present some concluding remarks and highlight some directions for future research.

2 Literature Review

The continued growth of both importance and complexity of modern call centers has led to an extensive and growing literature. Due to the uncertainty governing the call center environment (customers and agents behaviors), the literature has standardly addressed its issues using stochastic models, and in particular queueing models. Important related surveys are the paper of Koole and Mandelbaum [18] and its extended version Gans et al. [10] where the authors survey the literature dealing with the operations management of call centers. In addition, we recommend the overview of Whitt [29], and that of Mandelbaum [20] where the author provides a large number of research papers devoted to call centers issues.

The literature related to our work spans three main areas. The first is concerned with the analysis of multiserver systems motivated by call centers. The second deal with reneging phenomena. The third area is related to the problem of predicting and announcing delays information.

Let us focus on the first area, i.e., call center modeling. Call centers may be broadly classified into two contexts: multi-skill call centers and full-flexible call centers. A multi-skill call center handles several types of calls, and agents have different skills. The typical example, see Gans et al. [10], is an international call center where incoming calls are in different languages. Related studies include those by Akşin and Karaesmen [2], Chevalier and Tabordon [9], and references therein. Our concern in this paper is full-flexible call centers. Therefore assistance to customers can be provided by any agent. This would be a plausible assumption for many real cases, especially for unilingual call centers where the complete flexibility is not as difficult as in multilingual call centers. Furthermore, we assume for the models under consideration that all agents are totally identical statistically. In other words, they can answer all questions coming from customers with the same efficiency, both quantitatively and qualitatively, even in case of different types of customers. Our motivation is related to the nature of the call center we are considering here, and which is the case for many other call centers applications. The difference between customer types is only qualitative, i.e., it is not related to the statistical behavior of customers but to their importance for the company. Full-flexible call centers were extensively studied in the literature. We refer the reader to Gans et al. [10], Jouini et al. [16], and references therein.

We now consider the second area of literature related to this paper, i.e., renegeing phenomena. Queueing models incorporating impatient customers have received a lot of attention in the literature. They are an important feature in a wide variety of situations that may be encountered in manufacturing systems of perishable goods, telecommunication systems, call centers, etc. To underline the importance of the abandonment modeling in the call center field, the authors in Gans et al. [10] and in Mandelbaum and Zeltyn [21] give some numerical examples that point out the effect of abandonments on performance. The literature on queueing models with renegeing focus especially on performance evaluation. We refer the reader to Ancker and Gafarian [3], Garnett et al. [11], and references therein for simple models assuming exponential renegeing times. In Garnett et al. [11], the authors study the subject of Markovian abandonments. They suggest an asymptotic analysis of their model under the heavy-traffic regime. Their main result is to characterize the relation between the number of agents, the offered load and system performances such as the probability of delay and the probability to abandon. Zohar et al. [32] investigate in their work the relation between customers renegeing and the experience of waiting in queue. Other papers have allowed renegeing to follow a general distribution. Related studies include those by Baccelli and Hebuterne [7], Brandt and Brandt [8], Ward and Glynn [26], and references therein.

In what follows, we mention some literature related to the third area, namely predicting and announcing delays to customers. Predicting virtual delays in Markovian models, as under consideration in this paper, deals in particular with the transient analysis of birth-death processes. Several papers have been proposed for the study of the transient behavior of queues using birth-death processes, but in general, analytical solutions are extremely difficult to obtain. A number of interesting results are derived in Whitt [28]. We also refer the reader to Jouini and Dallery [14], and Nakibly [22] for further results in various queueing situations. The next natural step afterwards is the announcement of predicted delays. We mention the relevant work of Whitt [27] where the author studies the effect of announcing delays on the performance of a single class call center. An extension of the latter work is addressed by Jouini et al. [15]. Other close references include those by Armony et al. [5], Guo and Zipkin [12], and references therein. In such studies, we should not ignore possible reactions by customers. Indeed, the announcement of delays would have a significant influence on customers. The literature on customers influenced by delay information begins with Naor [23]. An overview of customer psychology in waiting situations, including the impact of uncertainty, can be found in Maister [19]. Taylor [24] showed that delays affect customers' service

evaluations in an experiment involving airline flights. Hui and Tse [13] conducted a survey on the relationship between information and customer satisfaction.

3 Infinitely Patient Customers

In this section, we consider a call center model where customers are infinitely patient. The model is described in Section 3.1. Then in Section 3.2, we assume that the service provider gives delay information to customers upon arrival, and develop a method for estimating virtual delays. The latter would be used within the announcement step.

3.1 Model Description and Notations

Consider the queueing model of a call center with two classes of customers; valuable customers type A , and less valuable ones type B . The model consists of two infinite priority queues type A and B , and a set of s parallel, identical servers representing the set of agents. All agents are able to answer all types of customers. The call center is operated in such a way that at any time, any call can be addressed by any agent. So upon arrival, a call is addressed by one of the available agents, if any. If not, the call must join one of the queues. The scheduling policy of service assigns customers A (B) to queue A (B). Customers in queue A have priority over customers in queue B in the sense that agents are providing assistance to customers belonging to queue A first. The priority rule is non-preemptive, which simply means that an agent currently serving a customer pulled from queue B , while a new arrival joins queue A , will complete this service before turning to queue A customer. Within each queue, customers are served in FCFS manner.

Arrival processes of type A and B customers follow a Poisson process with rates λ_A and λ_B , respectively. Let λ_T be the total arrival rate, $\lambda_T = \lambda_A + \lambda_B$. Successive service times are assumed to be i.i.d., and follow a common exponential distribution with rate μ for both types of customers. Then, the server utilization ρ (proportion of time each server is busy) is $\rho = \lambda_T/s\mu$. The condition for stability is $\rho < 1$, that is to say that the mean total arrival rate must be less than the mean maximal service rate of the system. The resulting model, referred to as Model 1, is shown in Figure 1.

There are two reasons for considering common distributions for service times. The first one relates to the types of call centers that motivate our analysis. We are considering call centers where customers are segmented into different groups based on their value to the firm. This segmentation

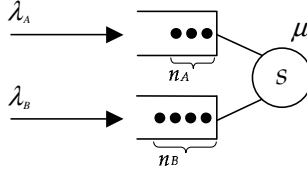


Figure 1: Call center model without reneging, Model 1

can be based on lifetime value or profitability. The call center then provides different levels of service to these groups. This type of service level differentiation is widely used in financial service and telecommunication call centers. 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 queries do not differ from one type of customer to another. The second reason is due to the complexity of the analysis when assuming different behaviors in the statistical sense. Our main purpose in this paper is to investigate simple but at the same time interesting models that allow us to better understand the system behavior and obtain practical guidelines.

In Section 3.2, we tackle the analysis of our call center by adding delay information. From the quantitative side, announcing delays would not affect our original model (Model 1), since customers are infinitely patient anyway. So, informing customers about their anticipated delays or not, leads to the same quantitative performance measures. However, announcing delays would ameliorate the waiting experience of customers by reducing uncertainty, and improve as a consequence their satisfaction.

3.2 Predicting and Announcing Virtual Delays

Consider a new arrival call. There are two possibilities: either at least one server is idle, or all servers are busy. In the former case, the customer enters service immediately without having to wait. So, the service provider does not announce any information to the customer. In the second case, he has to wait in queue for service to begin. In the following we give the distribution of the waiting time of a new arrival. This analysis will be used by the service provider afterwards, in order to inform customers about their delays.

As it is the case in most modern call centers, we assume that the technology allows us to know the system state at each new arrival epoch. The system state at each new arrival epoch is defined

by the number of customers in system. If the latter is larger than or equal to the number of servers s , then all servers are busy and the new arrival has to wait in queue. Let n_A be the number of type A customers in queue A seen by the new arrival, and n_B that of customers B in queue B , $n_A, n_B \geq 0$. Finally, let n_T be the total number of customers in queues seen by our new customer, $n_T = n_A + n_B$.

In what follows, we compute the mean and variance of the state-dependent waiting time distribution of each new customer type. At the epoch of each arrival, we assume that all servers are busy and $n_T = n_A + n_B \geq 0$ customers are waiting in queues. We separate the study depending on whether the call of interest is of type A or B . Type A customers observe a regular queue without priority, so estimation of their waiting time is easy to obtain. It is not the case for type B customers, because their waiting time is affected by future type A arrivals.

3.2.1 Virtual Delays for type A customers

Because of the strict priority, the waiting time of a new type A arrival does not depend on the number of type B customers waiting in queue. Also, it does not depend on future type A arrivals because queue A is working under the FCFS basis. Given n_A customers waiting in queue A , the call of interest has to wait until all n_A customers ahead of him enter service, plus the time it takes up to a service completion (of a customer in service). A service completion is exponentially distributed with mean $1/s\mu$ (s servers in parallel), and it is independent of the previous history because of the memoryless property. Hence, the waiting time of our customer is the sum of $n_A + 1$ i.i.d. exponential random variables each with mean $1/s\mu$, which has an Erlang distribution. Let us define $X_{n_A}^A$ as the conditional random variable measuring the waiting time of our customer, given the queue state n_A . The mean, $E(X_{n_A}^A)$, and variance, $Var(X_{n_A}^A)$, of $X_{n_A}^A$ are thereafter given by

$$E(X_{n_A}^A) = \frac{n_A + 1}{s\mu} \quad \text{and} \quad Var(X_{n_A}^A) = \frac{n_A + 1}{(s\mu)^2}. \quad (1)$$

We can also calculate the full probability distribution function (PDF) of $X_{n_A}^A$. The Erlang PDF is available in closed form, see for example Kleinrock [17]. We describe the ratio of the standard deviation, $\sigma(X_{n_A}^A) = \sqrt{Var(X_{n_A}^A)}$, by the mean, $\sigma(X_A)/E(X_A)$. This ratio has the remarkably simple form

$$\frac{\sigma(X_A)}{E(X_A)} = \frac{1}{\sqrt{n_A + 1}}, \quad (2)$$

independent of μ and s . We deduce that the waiting time distribution (conditional on all servers being busy) is highly concentrated about its mean, for large values of n_A . Note that the analysis above (for customers A) is still valid for the $GI/M/C$ queue. We only need to know the current state information.

3.2.2 Virtual Delays for type B customers

We focus on the waiting time distribution of a type B arrival who finds n_A type A customers and n_B type B customers waiting in queues. Before entering service, this customer has to wait for the queue to become empty of $n_T = n_A + n_B$ customers and of all future type A customers who arrive in between. Since all customers are statistically identical and the system is workconserving, we should observe that the waiting time of our customer does not depend on the order of service of the customers ahead of him. Hence, the duration of interest can be divided into two parts: The first is the busy period opened by the customer in service. The second part is the sum of $n_A + n_B$ busy periods, each one opened by one of the customer in queue. The busy period is the one of an $M/M/s$ queue with a Poisson arrival rate of λ_A . It is defined as the time from an arrival of a customer to the system with only one idle server until the first time one of the servers becomes idle.

Let us define $X_{n_T}^B$ as the conditional random variable measuring the waiting time of our customer, given queue state n_T . Knowing that the remaining service time of one customer is independent from the already finished work (exponential service times), then the distribution of the busy period opened by the customer in service is identically distributed as the busy period opened by one of the customers from the queue. Finally, the new type B customer has to wait for $n_A + n_B + 1$ i.i.d. busy periods of an $M/M/s$ queue with arrival rate λ_A . The probability density function (pdf) of the busy period of an $M/M/1$ queue can be found for example in Kleinrock [17]. Then, with a little thought it should be clear that the busy period pdf of interest here is obtained by only substituting the capacity of service μ (in the case of an $M/M/1$ queue) by $s\mu$ (in the case of an $M/M/s$ queue). To get a closed-form expression of the $X_{n_T}^B$ pdf, the mathematics becomes complicated ($n_T + 1$ -fold convolution of the busy period pdf). Fortunately, the mean of $X_{n_T}^B$ is fairly simple to obtain, by summing the busy periods means up to $n_T + 1$. The same approach is still valid for the variance computation of $X_{n_T}^B$ using in addition the independence between the random variables of the busy periods durations. To get the first two moments of the busy period,

we simply evaluate, respectively, the negative derivative and the positive second derivative at zero of its PDF Laplace transform in the time t . The mean, $E(X_{n_T}^B)$, and variance $\sigma(X_{n_T}^B)$, of $X_{n_T}^B$ are given by

$$E(X_{n_T}^B) = \frac{n_T + 1}{s\mu - \lambda_A} \quad \text{and} \quad \text{Var}(X_{n_T}^B) = (n_T + 1) \frac{s\mu + \lambda_A}{(s\mu - \lambda_A)^3}. \quad (3)$$

Equation (4) shows the ratio of the standard deviation, $\text{Var}(X_{n_T}^B) = \sqrt{\text{Var}(X_{n_T}^B)}$, by the mean,

$$\frac{\sigma(X_{n_T}^B)}{E(X_{n_T}^B)} = \sqrt{\frac{s\mu}{(n_T + 1)(s\mu - \lambda_A)}}. \quad (4)$$

Once again, we observe that the conditional waiting time distribution is highly concentrated about its mean, for large values of n_T .

It is not too difficult to extend the analysis of the conditional waiting time to the general case of an arbitrary number of customer classes. For example for the third priority class analysis, it is equivalent to aggregate the first two classes into one equivalent class. Thereafter, we use the same analysis as that conducted above for type B customers. This completes the study of announcing delays for call centers with infinitely patient customers.

4 Finitely Patient Customers

In this section, we consider a new major feature in our modeling. We let customers renege while waiting in queue. In Section 4.1, we first describe the original model of the call center without delay information. In Section 4.2, we next focus on building the call center model assuming that it provides delay information to customers. In Section 4.3, we finally derive for the latter model a number of performance measures related to queueing delays.

4.1 Model Description and Notation

We address the analysis of a call center with a single group of identical agents, serving two classes of impatient customers, high and low priority classes. The model is identical to that described in Section 3.1, however in addition we allow customers to be impatient. After entering the queue, a customer will wait a random length of time for service to begin. If service has not begun by this time he will renege and is considered to be lost. Times before renegeing for both types are assumed to be i.i.d., and exponentially distributed with a common rate γ for both customer types. We

motivate our consideration for a common distribution of impatience time as already mentioned in Section 3.1. We note that abandonments make our system unconditionally stable. The resulting model, referred to as Model 2, is shown in Figure 2.

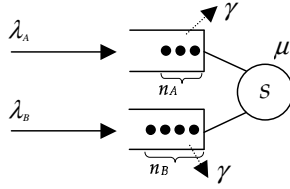


Figure 2: Call center model with reneging, Model 2

4.2 Call Center Modeling with Announcement

Assume moving from the call center described in Section 4.1 to a call center with delay announcement. On the contrary to a call center with infinitely patient customers, there is a modeling complexity when we provide delay information to customers, due to possible changes in their behavior. In this section, we investigate the impact of announcing delays on the customer abandonment experience. When we inform a customer about his anticipated delay, he will decide from the beginning, either to hang up immediately because he estimates that his delay is too long, or to start waiting in queue. In the latter case, there are two further possibilities. The first is that customers never abandon thereafter. The second possibility is that the customer patience will change, i.e., customers may abandon even if they had chosen to start waiting. It is easy to see that customers would have a patience behavior different from that in the original system (without announcement), depending on the information we provide to them. We refer the reader to Armony et al. [5] and Guo and Zipkin [12] for further details on the subject.

Several forms of delays information are possible. The best is that we give to a new customer his actual delay, which cannot be known in advance because it is random. The most natural in practice is that the service provider gives a certain percentile β of the virtual delay distribution to each new arrival. The virtual delay is the time it takes for a server to become free for the customer of interest. In other words, it is the time until all higher priority customers ahead of the arrival leave the queue plus the duration of a service completion. Whitt [27] has considered a similar problem for a single class call center. He proposed a model incorporating announcement by assuming that a new customer who finds all servers busy balks with a given probability. Once

a customer elects to wait in queue, he would never abandon thereafter. We assume that each new arrival comes with its own deadline of time patience, and paralleling to the model of Whitt [27], we stipulate that a new customer elects to join the queue with the probability that a server becomes free for him (his virtual waiting time) before he would renege. This is exact only if we assume that the customer acts as if the delay information was his actual delay, which is not the case. We do not let customers renege once they join the waiting line. This may be reasonable for high values of β , since the estimation of the anticipated delay should be fairly accurate in that case, so that ignoring renegeing would be valid.

Assume that a new arrival finds n_A waiting type A customers in queue A , and n_B waiting type B customers in queue B . Note that implicitly we are focusing on new arrivals finding all servers busy. If the number seen by an arrival is less than s , then the new arrival never balks and enters service immediately. Let us come back to a new arrival finding all servers busy. It should be clear that the probability of balking for a type A new arrival depends only on n_A (due to the priority rule), say $p_{bk}^A(n_A)$. However, the probability of balking for a new type B arrival depends on the couple (n_A, n_B) , say $p_{bk}^B(n_A, n_B)$. Furthermore, we should not fall in the confusion of only considering it as a function of $n_T = n_A + n_B$. Having different values of n_A and n_B , so that $n_T = n_A + n_B$ is held constant, would affect the virtual delay distribution of the customer of interest. The reason is that with delays information, the arrival rate of type A customers, seen by our new type B customer, is state of queue A dependent. As a consequence, not considering the couple (n_A, n_B) to compute the balking probability of that customer would lead to a wrong result.

Let $Y_{n_A}^A$ be the random variable measuring the state-dependent virtual delay for a new type A arrival finding n_A waiting customers ahead of him. Let $Y_{(n_A, n_B)}^B$ be the one for a new type B arrival finding n_A and n_B waiting customers ahead of him in queues A and B , respectively. Furthermore, let $G_{n_A}^A(t)$ and $G_{(n_A, n_B)}^B(t)$ for $t > 0$ be the PDF of $Y_{n_A}^A$ and $Y_{(n_A, n_B)}^B$, respectively. Then, the call center provides upon arrival the values

$$D_{n_A}^A = (G_{n_A}^A)^{-1}(\beta), \quad \text{and} \quad D_{(n_A, n_B)}^B = (G_{(n_A, n_B)}^B)^{-1}(\beta) \quad (5)$$

to type A and B customers, respectively. The balking probabilities are computed as follows. We denote by T the random threshold patience for both types (exponentially distributed with rate γ).

The probability for a new type A arrival to balk is thereafter

$$p_{bk}^A(n_A) = P(T < D_{n_A}^A). \quad (6)$$

The one for a new type B arrival is in turn given by

$$p_{bk}^B((n_A, n_B)) = P(T < D_{(n_A, n_B)}^B). \quad (7)$$

Having in hand the PDF of the exponential distribution, the following holds

$$p_{bk}^A(n_A) = 1 - e^{-\gamma \cdot D_{n_A}^A}, \quad \text{and} \quad p_{bk}^B((n_A, n_B)) = 1 - e^{-\gamma \cdot D_{(n_A, n_B)}^B}. \quad (8)$$

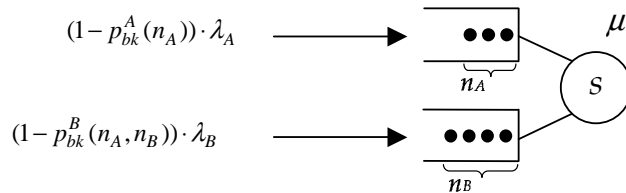


Figure 3: Call center model with delay information, Model 3

The resulting model of the call center, incorporating delay information, is shown in Figure 3, and is referred to as Model 3. Note that it is reasonable to assume that balking decisions are independent from one customer to another, so that arrivals still follow a Poisson process. What remains to be done in order to fully characterize Model 3 is to compute state-dependent arrival rates for each customer type, which in turn reduces to characterizing the distribution functions of $Y_{n_A}^A$ and $Y_{(n_A, n_B)}^B$. In the next section, we give closed-form expressions for their first two moments. Based on these results, we propose in Section 5.2, a helpful and practical approximation of their entire distributions.

4.3 Predicting and Announcing Virtual Delays

As in Section 3.2, we assume that the technology of our call center enables us to know when queues are empty, and whether there is an available agent for an upcoming customer, or not. If less than s customers are present in the system, the customer of interest gets service immediately.

If not, he has to wait in his corresponding queue for service to begin.

Knowing that all servers are busy, we focus on analyzing the conditional random variables $Y_{n_A}^A$ and $Y_{(n_A, n_B)}^B$, $n_A, n_B \geq 0$. We separate the analysis depending on whether the arrival call is of type A or B . Similar to Section 3.2, the priority scheme under consideration makes the analysis for type A customers less complicated than that for type B customers. The latter is indeed affected by future type A arrivals who have higher priority for service.

Let us recall that we are calculating virtual delays which will be used within a second step in order to compute balking probabilities. In other words, we are calculating the time it takes until a server becomes free for the customer of interest in case he elects to wait (does not balk). In what follows, we analyze these quantities for both customer types.

4.3.1 Virtual Delays for Type A Customers

Consider a new type A arrival who finds all servers busy, n_A waiting customers in queue A and n_B waiting customers in queue B . Owing to his higher priority, the virtual delay of a new type A arrival does not depend on the number of type B customers already present in system, see Figure 4. The customer has to wait until the n_A waiting customers leave the queue plus the time it takes for a service completion (when all servers are busy). By a customer who leaves the queue, we only mean a customer who enters service. In Model 3, there is no longer the possibility for customers to renege once they elect to join the waiting line.

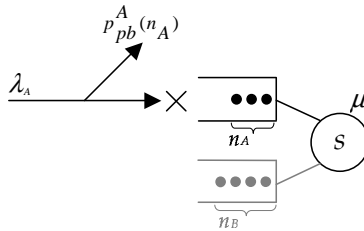


Figure 4: Virtual delay for a new type A arrival

Our customer A has to wait for the n_A customers ahead of him to enter service, and then he has to wait for a service completion. Overall, he has to wait for $n_A + 1$ service completions. Hence, the pdf of $Y_{n_A}^A$ is simply the convolution of the pdfs of $n_A + 1$ i.i.d. exponential random variables each with parameter $s\mu$. So, $Y_{n_A}^A$ has an $n_A + 1$ -Erlang distribution with parameter $s\mu$. The mean

and variance of $Y_{n_A}^A$ are, respectively, given by

$$E(Y_{n_A}^A) = \frac{n_A + 1}{s\mu}, \quad \text{and} \quad \text{Var}(Y_{n_A}^A) = \frac{n_A + 1}{s^2\mu^2}. \quad (9)$$

Having in hand the PDF of $Y_{n_A}^A$, it only remains to come back to Equations (5) and (6) in order to compute the balking probability $p_{bk}^A(n_A)$. Define now the standard deviation of $Y_{n_A}^A$ by $\sigma(Y_{n_A}^A) = \sqrt{\text{Var}(Y_{n_A}^A)}$, and the coefficient of variation by the ratio of the standard deviation over the mean, $cv(Y_{n_A}^A) = \sigma(Y_{n_A}^A)/E(Y_{n_A}^A)$. As shown in Equation (10), the ratio $cv(Y_{n_A}^A)$ is characterized to have a simple form independent of μ and s .

$$cv(Y_{n_A}^A) = \frac{1}{\sqrt{n_A + 1}}. \quad (10)$$

From Equation (10), we again note that for large values of n_A , the virtual delay of $Y_{n_A}^A$ is very concentrated about its mean. This implies that for large values of n_A , the mean value of $Y_{n_A}^A$ should provide a good approximation of the virtual delay.

4.3.2 Virtual Delays for Type B Customers

Knowing that all servers are busy, let n_A and n_B be the number of type A and B waiting customers seen by a new type B arrival, in queues A and B , respectively.

The random variable $Y_{(n_A, n_B)}^B$ is the time until the $n_T = n_A + n_B$ waiting customers start service, plus the time it takes for all future type A arrivals (during the wait of the customer of interest) to enter service, plus the duration for a service completion (when all servers are busy), see Figure 5.

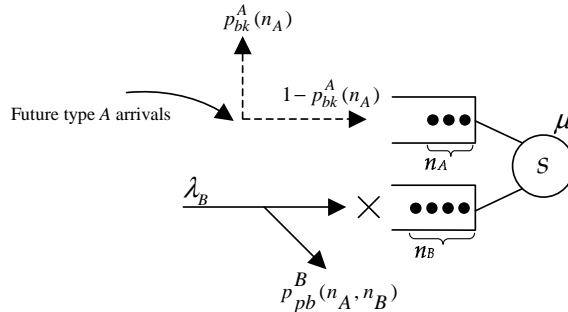


Figure 5: Virtual delay for a new type B arrival

To characterize $Y_{(n_A, n_B)}^B$, we ignore all future type B arrivals because the discipline of service

within queue B is FCFS. However, all future type A arrivals have to be considered because of their higher priority against the customer of interest. Recall that reneging is no longer possible. We only consider events of type A arrivals and service completions. Thereby, changes of queue states seen by our customer are as follows. As long as type A customers are waiting in queue, the number of type B waiting customers does not change, however that of type A customers increases by 1 further to a type A arrival or decreases by 1 further to a service completion. The number of type B waiting customers can not increase. It only decreases by 1 further to a service completion when no type A customers are waiting in queue. We should be careful not to forget that type A arrivals are state-dependent due to the balking decisions of customers upon arrival.

Based on the above explanation, we move on to employ the following two-dimensional Markov chain. Let the system state at a given random instant be (m_A, m_B) where m_A (m_B) is the number of type A (B) customers in queue A (B), $m_A, m_B \geq 0$. In addition, the Markov chain has an absorbing state denoted by (-1) . The system moves to (-1) subsequent to a service completion when both queues are empty. Reaching the latter state means that a server is available for the customer of interest. When m_A customers are waiting in queue A , we denote the state-dependent arrival rates of type A arrivals by

$$\lambda_A(m_A) = (1 - p_{bk}^A(m_A)) \times \lambda_A, \quad m_A \geq 0. \quad (11)$$

The non-zero transition rates are

$$\left\{ \begin{array}{l} q_{(m_A, m_B)(m_A+1, m_B)} = \lambda_A(m_A), \text{ for } m_A, m_B \geq 0, \\ q_{(m_A, m_B)(m_A-1, m_B)} = s\mu, \text{ for } m_A, m_B > 0, \\ q_{(0, m_B)(0, m_B-1)} = s\mu, \text{ for } m_B \geq 0, \\ q_{(0, 0)(-1)} = s\mu. \end{array} \right. \quad (12)$$

As shown in Figure 6, measuring $Y_{(n_A, n_B)}^B$ may be formulated as the downcrossing time until absorption in state (-1) , starting from state (n_A, n_B) .

The Markov chain we consider has a special structure allowing analytical solutions. From Figure 6, the random variable $Y_{(n_A, n_B)}^B$ may be rewritten as

$$Y_{(n_A, n_B)}^B = U(n_A) + V_{n_B-1} + \dots + V_0, \quad (13)$$

where $U(n_A)$ is the random variable measuring the downcrossing time until first passage at state

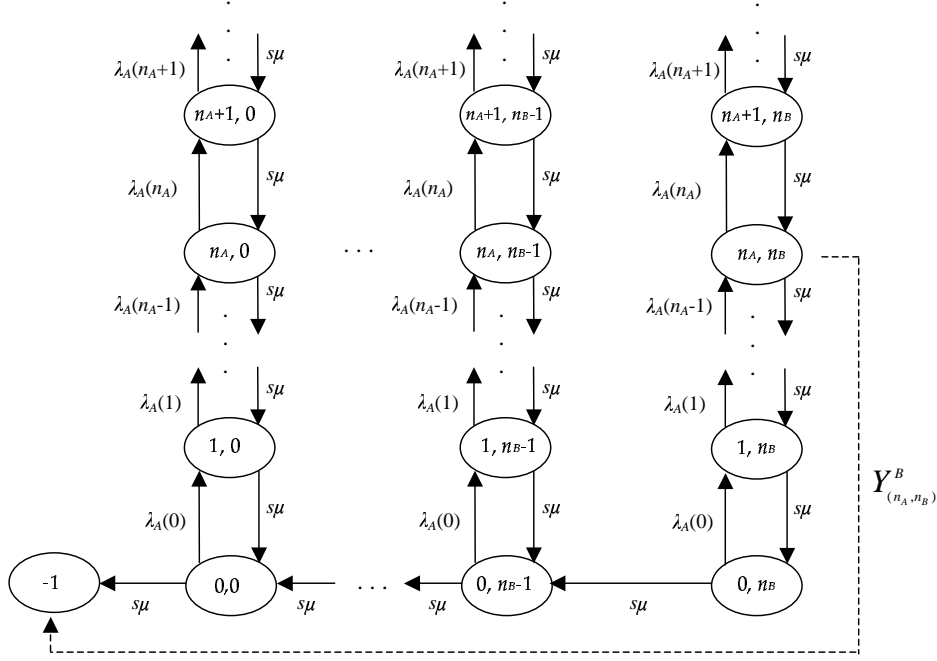


Figure 6: The random variable $Y_{(n_A, n_B)}^B$

$(0, n_B - 1)$ starting from state (n_A, n_B) , V_i is the random variable measuring the downcrossing time until first passage time at state $(0, i - 1)$ starting from state $(0, i)$ for $1 \leq i \leq n_B - 1$, and V_0 is the random variable measuring the downcrossing time until absorption in state (-1) starting from state $(0, 0)$.

The Markovian assumptions allow us to state that the random variables $U(n_A)$, V_0 , ..., and V_{n_B-1} are independent. From Figure 6, we see that V_0 , ..., and V_{n_B-1} are identically distributed. Let $E(Y_{(n_A, n_B)}^B)$ and $Var(Y_{(n_A, n_B)}^B)$ be the mean and variance of the random variable $Y_{(n_A, n_B)}^B$, respectively. Then, using the linearity property of expectations, we get

$$E(Y_{(n_A, n_B)}^B) = E(U(n_A)) + n_B \times E(V_0), \quad (14)$$

and from the independence between these random variables, the following holds

$$Var(Y_{(n_A, n_B)}^B) = Var(U(n_A)) + n_B \times Var(V_0). \quad (15)$$

Let us now focus on computing the means and variances of $U(n_A)$, V_0 , ..., and V_{n_B-1} . To do so, we define an intermediate birth-death process with discrete state space taking non-negative integer

values $\{0, 1, 2, 3, \dots\}$. The transition rates of the process are denoted by

$$\begin{cases} q_{0,1} = \lambda_A, \\ q_{m,m+1} = \lambda_A(m-1), \text{ for } m \geq 1, \\ q_{m,m-1} = s\mu, \text{ for } m \geq 1, \end{cases} \quad (16)$$

and $q_{m,n} = 0$ otherwise. The birth-death process is derived from the previous Markov chain and is shown in Figure 7.

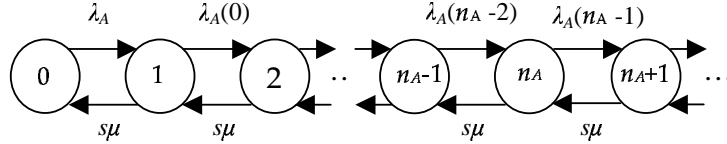


Figure 7: Intermediate birth-death process

One may intuitively see that the intermediate birth-death process allows us to compute the time it takes to empty the queue of a given number of waiting type A customers plus the time for a service completion, so that one of the waiting type B customer could enter service. The random variables $U(n_A)$, V_0 , ..., and V_{n_B-1} are defined on the intermediate birth-death process as follows. The random variable $U(n_A)$ is the downcrossing time until first passage at state 0, starting from state $n_A + 1$. As for the random variable V_i , $0 \leq i \leq n_B - 1$, it is only the first passage time at state 0, starting from state 1. Note that since we are calculating first passage times at state 0, the analysis is independent of the birth rate when the system is in that state, i.e., $q_{0,1} = \lambda_A$.

By considering a general birth-death process, the authors in Jouini and Dallery [14] give closed-form expressions for any moment of order $k \geq 1$ of several random variables related to first passage times. We use their results in our context here. To simplify the presentation, we introduce the quantities δ_m for $m \geq 0$. For $m = 0$ we let $\delta_0 = \lambda_A$, and for $m \geq 1$ we let $\delta_m = \lambda_A(m-1)$. Let us now define the potential coefficients of the intermediate birth-death process, say ϕ_m , as follows.

$$\phi_0 = 1, \text{ and } \phi_m = \frac{\prod_{j=0}^{m-1} \delta_j}{s^m \mu^m}, \text{ for } m \geq 1. \quad (17)$$

From Jouini and Dallery [14], the mean $E(U(n_A))$ and variance $Var(U(n_A))$ of $U(n_A)$ are thereafter given by

$$E(U(n_A)) = \sum_{m=1}^{n_A+1} \frac{1}{\delta_{m-1} \phi_{m-1}} \sum_{j=m}^{\infty} \phi_j, \quad (18)$$

$$\text{Var}(U(n_A)) = \sum_{m=1}^{n_A+1} \frac{2}{\delta_{m-1}\phi_{m-1}} \sum_{j=m+1}^{\infty} \frac{1}{\delta_{j-1}\phi_{j-1}} \left(\sum_{l=j}^{\infty} \phi_l \right)^2 + \sum_{m=1}^{n_A+1} \frac{1}{\delta_{m-1}^2\phi_{m-1}^2} \left(\sum_{j=m}^{\infty} \phi_j \right)^2. \quad (19)$$

It goes without saying that δ_0 (that is λ_A) could be eliminated from Equations (18) and (19), which agrees with our above claim. We only keep them here for presentation issues.

Concerning the mean $E(V_0)$ and variance $\text{Var}(V_0)$ of the random variable V_0 , they are given by

$$E(V_0) = \frac{1}{\delta_0} \sum_{m=1}^{\infty} \phi_m, \quad (20)$$

$$\text{Var}(V_0) = \frac{2}{\delta_0} \sum_{m=2}^{\infty} \frac{1}{\delta_{m-1}\phi_{m-1}} \left(\sum_{j=m}^{\infty} \phi_j \right)^2 + \frac{1}{\delta_0^2} \left(\sum_{m=1}^{\infty} \phi_m \right)^2. \quad (21)$$

Substituting Equations (18), (19), (20) and (21) back into Equations (14) and (15) leads to the expressions of the mean and variance of the random variable $Y_{(n_A, n_B)}^B$. Finally, using the results of Section 4.3.1 to compute the balking probabilities for type A arrivals, the mean and variance of $Y_{(n_A, n_B)}^B$ are thereafter fully characterized.

Note that one may derive all higher order moments of the virtual delay for both customer types, which allows us to derive their full distributions. However, the analysis would be cumbersome and numerically time consuming. We thereafter content ourself with only the first two moments, and propose a useful approximation of these distributions as we shall explain later in Section 5.2.

5 Some Practical Issues

In this section, we investigate some practical issues for an eventual implementation of delay information. We identify two points that may help practitioners. The first is discussed in Section 5.1, and evaluates the need of announcing delays when queues are empty upon the arrival of a new customer. The second is discussed in Section 5.2, and it deals with an approximation for computing the anticipated delay we communicate to each new arrival.

5.1 Empty Queues

In this section, we call into question the need for communicating delays to new customers who find the queues empty. We consider a new arrival call finding empty queues. We define q as the conditional probability that the call of interest has to wait before beginning service, given that he

finds the queues empty. The probability q does not depend on the type of the new arrival, and it concerns the case of having all servers busy knowing that the queues are empty. We are interested in calculating q , in order to get some indication on its value under normal working conditions of a call center. For instance, if the proportion q is too small, then most arriving calls begin service without waiting. So, there is little need for delay information. However, if q is quite large, then a considerable proportion of new calls have to wait, and the prediction would be important.

Another reason that prompts us to address this problem is related to the technology employed in some call centers. In most call centers we are able to know the state of the queues, i.e., the number of waiting customers, if any. If no customers are waiting in queue, the system may however be unable to know whether an agent is free for the new arrival or not. As a consequence, it would be interesting to study how the quantity q behaves.

The proportion q represents the conditional probability that a new call has to wait knowing that the queues are empty. Or equivalently, the probability that a new call finds s servers busy knowing that the queues are empty. For the call center incorporating announcement of delays of Section 3 (Model 1, without reneging) and that of Section 4 (Model 3, with reneging), we have

$$\begin{aligned} q &= Pr\{all\ servers\ busy \mid n_T = 0\} \\ &= \frac{Pr\{all\ servers\ busy\ AND\ n_T = 0\}}{Pr\{n_T = 0\}}. \end{aligned} \tag{22}$$

This implies

$$q = \frac{Pr\{k = s\}}{\sum_{i=0}^s Pr\{k = i\}}, \tag{23}$$

where $Pr\{k = i\}$ is the steady state probability that a new arrival finds i customers (regardless of their type) are in system (Model 1 or Model 3). To compute q , our approach is based on system state probabilities seen by a randomly chosen new arrival. From the PASTA property (Poisson Arrivals See Time Averages), these probabilities coincide with those seen by an outside random observer, i.e., simply the probabilities that the system is in a given state at a random instant. The PASTA property is based on the memoryless property of the Poisson process, which allows to generate a sequence of arrivals that take a random look at the system. We refer the reader to Kleinrock [17] for further explanation, and Wolff [30] for a rigorous proof.

It suffices therefore to derive the steady state probabilities of Models 1 and 3 at a random instant. To do so, we employ their associated Markov chains. For each model, we denote the

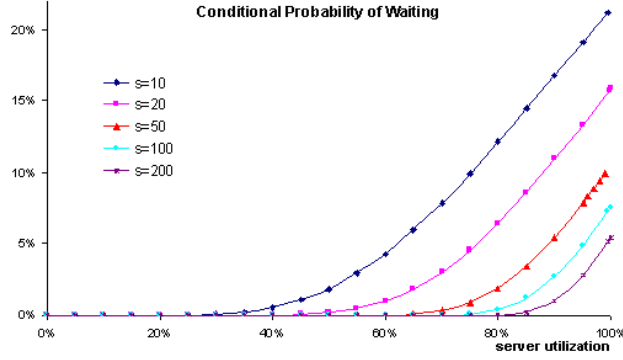


Figure 8: Conditional probability of waiting, given empty queues

system state by the total number i of customers present in system, $i \geq 0$. We denote the steady state probability of having i customers in Model 1 (Model 3) by $p_1(i)$ ($p_2(i)$). As we shall explain, we don't need to fully characterize the Markov chain of our models. For both models, the transition rate to be in state i , starting in state $i - 1$, is λ_T , $1 \leq i \leq s$. The transition rate to be in state $i - 1$, starting in state i , is $i\mu$, $1 \leq i \leq s$. Hence

$$p_j(i) = \frac{\lambda_T}{i\mu} p_j(i - 1), \text{ for } j \in \{1, 2\} \text{ and } 1 \leq i \leq s, \quad (24)$$

which allows us to deduce that $p_j(i) = \frac{(\lambda_T)^i}{i!\mu^i} p_j(0)$, and gives the following simple and common expression of q for both models

$$q = \frac{\lambda_T^s}{s!\mu^s} \times \left(\sum_{i=0}^s \frac{\lambda_T^i}{i!\mu^i} \right)^{-1}. \quad (25)$$

In Figure 8, we plot the proportion q according to the server utilization $\rho = \frac{\lambda_T}{s\mu}$. We choose a large range of parameters typical of those that would be encountered in real situations. Agent team sizes are ranging from 10 to 200, and server utilizations up to 99.95%. As one would expect, we see that q increases when ρ increases (servers becomes more and more often busy). From Figure 8, we also see that for a fixed ρ , q is decreasing in the number of servers s . This is due to the well known pooling phenomenon. Furthermore, this pooling effect is clearly present when moving between call centers with low number of servers ($s = 10$ to $s = 20$), but is less and less noticeable when moving between large call centers ($s = 100$ to $s = 200$). The reason is that beyond a certain limit, a large system has a sufficient level of pooling, so that moving to a larger system will not have that high improvement in performances.

The central insight is that only a small proportion of new calls (who find empty queues) have to

wait before beginning service. Under the normal conditions of work, the proportion is of the order of 10%. As a consequence, there is not a real need to inform customers about their anticipated delays. In addition, even if one customer has to wait, his waiting time is sufficiently short. For example, consider Model 1 with $s = 50$, $\lambda_A = 4.9 \text{ min}^{-1}$, $\lambda_B = 4.9 \text{ min}^{-1}$, and $\mu = 0.2 \text{ min}^{-1}$ (so $\rho = 98\%$), the conditional expected waiting time of a new type A customer, given empty queues is $\frac{1}{s\mu} = 6 \text{ sec}$. The one for a new type B customer is the mean of the busy period of an $M/M/s$ queue ($s = 50$ servers, mean service rate $\mu = 0.2 \text{ min}^{-1}$, and mean arrival rate $\lambda_A = 4.9 \text{ min}^{-1}$), which is given by $\frac{1}{s\mu - \lambda_A} = 11.8 \text{ sec}$. Taking now Model 3 with the same parameters and adding abandonments (with a certain rate $\gamma > 0$) implies the same value for a new arrival A (in case he elects to wait), only 6 sec. As for a new type B customer (in case he elects to wait), the quantity of interest is lower than that of Model 1. The explanation is as follows. In Model 3, fewer customers are accepted due to balking decisions of arrivals. This makes the busy period duration of Model 3 shorter than that of Model 1.

5.2 Normal Approximation of Virtual Delays

Given the system state upon each arrival and given a coverage probability β , the service provider has to compute the value of the anticipated delay. This numerical computation operation is characterized to be too heavy. In fact for the low priority customers, exact moment expressions of the virtual delays involve infinite summations. This would imply some numerical difficulties especially since we need to conduct such real-time operations for each arrival!

From a practical point of view, a normal distribution provides a satisfactory approximation of virtual delays. Since the random variables of virtual delays we consider here deal with summations of independent random variables, the Normal approximation should work well, see Whitt [28] and Ward and Whitt [25]. This claim is supported by theoretical results based on the Law of Large Numbers and the Central Limit Theorem. The Normal approximation should especially work well for new arrivals who find a large number of waiting customers in queue.

We only need the mean and standard deviation of the state-dependent virtual delay in order to get its full distribution (approximately). Thus, we propose to use the Normal distribution by only picking up the means and variances we derived in Sections 3 and 4. We should however point out that for small values of β , such distributions may lead to negative values of anticipated delays. To be judicious, we may adapt a given normal distribution by truncating it. For instance, let $h(t)$

and $H(t)$, $-\infty < t < +\infty$, be the pdf and PDF of the original Normal distribution, respectively. Also, let $h_{tr}(t)$ and $H_{tr}(t)$, $0 < t < +\infty$, be those for the truncated Normal distribution. The pdf of $h_{tr}(t)$ is calculated as $h_{tr}(t) = \frac{h(t)}{1-H(0)}$. By doing so, we even out the area of the negative region ($t < 0$) over that of the positive region, so that we build an appropriate distribution. Note that this transformation should not really affect the original Normal distribution. The reason is that the quantity $1 - H(0)$ is low for very small numbers of waiting customers in queue, and may be reasonably neglected otherwise.

6 Conclusions and Further Extensions

We focused on a fundamental problem in the operations management of call centers, namely the issue of informing customers about their queueing delays. Predicting delays is especially important when customers do not have direct access to information about the state of the system. In such a case, it has indeed been recognized that customers become dissatisfied with the service provider when they are forced to wait for an unknown delay. Announcing delays would reduce the undesirable phenomenon of customers abandonments, which allows to decrease the variability of waiting times. As a consequence, we improve the quality of delay information we give to customers, which in turn reduces customer renegeing even more, and so on.

In the first part of the paper, we considered a multiclass priority call center where customers are assumed to be infinitely patient. We thereafter developed a method for estimating virtual delays for each type of customers. These virtual delays are to be used within the implementation step of delay announcement. Note that assuming no renegeing, announcing delays would not affect customers behavior from the queueing model perspective. In the second part of the paper, we again considered a setting with customer types having different priorities, and we let in addition customers be impatient. The analysis extended that of Whitt [27] where a single class model is considered. This constitutes a relevant case in practice, representing call centers that aim to provide differentiated service levels to customers with different value for the firm. Our study is the first to model customer reactions to delay announcements in such a setting with priorities, and to provide an analysis for this case. Finally, we discussed some helpful issues for practitioners such as the need of announcing delays when a new customer finds the queues empty.

Worrying about accurate and practical results, one may continue our work by empirically describing customer reactions in response to delay announcements, in order to validate our modeling

of that reaction herein. An ambitious extension is to consider non-stationary arrivals which would be an important issue in practice. A further topic for future research is to analyze more complex systems: more than two customer classes and general distributions for service times and times before renegeing. It should be also of value to extend our work to the case of different statistical behaviors of different customer types. In other words, service times as well as times before renegeing are not identically distributed for different customer types.

References

1. McKinsey Quarterly: The Online Journal of McKinsey & Co. Available at www.mckinseyquarterly.com, 2006.
2. O. Z. Akşin and F. Karaesmen. Designing Flexibility: Characterizing the value of Cross-Training Practices. 2002. Under revision.
3. C. J. Ancker and A. Gafarian. Queueing with Impatient Customers Who Leave at Random. *Journal of Industrial Engineering*, 13:84–90, 1962.
4. M. Armony and C. Maglaras. Contact Centers with a Call-Back Option and Real-Time Delay Information. *Operations Research*, 52:527–545, 2004.
5. M. Armony, N. Shimkin, and W. Whitt. The Impact of Delay Announcements in Many-Server Queues with Abandonment. 2005. Working paper, New York University.
6. A. Artiba. Performance Evaluation, Planning and Control, of Productive Systems. *International Journal of Production Economics*, 85:1–2, 2003.
7. F. Baccelli and G. Hebuterne. On Queues With Impatient Customers. *Performance'81 North-Holland Publishing Company*, pages 159–179, 1981.
8. A. Brandt and M. Brandt. Asymptotic Results and a Markovian Approximation for the $M(n)/M(n)/C + GI$ System. *Queueing Systems: Theory and Applications (QUESTA)*, 41:73–94, 2002.
9. P. Chevalier and N. Tabordon. Overflow Analysis and Cross-Trained Servers. *International Journal of Production Economics*, 85:47–60, 2003.

10. N. Gans, G. Koole, and A. Mandelbaum. Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management*, 5:73–141, 2003.
11. O. Garnett, A. Mandelbaum, and M. Reiman. Designing a Call Center with Impatient Customers. *Manufacturing & Service Operations Management*, 4:208–227, 2002.
12. P. Guo and P. Zipkin. Analysis and Comparaison of Queues with Different Levels of Delay Information. 2004. Working paper, Duke University.
13. M. Hui and D. Tse. What to Tell Customer in Waits of Different Lengths: an Integrative Model of Service Evaluation. *Journal of Marketing*, 60:81–90, 1996.
14. O. Jouini and Y. Dallery. Moments of First Passage Times in General Birth-Death Processes. 2006. Submitted for publication.
15. O. Jouini, Y. Dallery, and O. Z. Akşin. Modeling Call Centers with Delay Information. 2006. Working paper, Ecole Centrale Paris and Koç University.
16. O. Jouini, Y. Dallery, and R. Nait-Abdallah. Analysis of the Impact of Team-Based Organizations in Call Centers Management. *Management Science*, 2007. To appear.
17. L. Kleinrock. *Queueing Systems, Theory*, volume I. A Wiley-Interscience Publication, 1975.
18. G. Koole and A. Mandelbaum. Queueing Models of Call Centers An Introduction. *Annals of Operations Research*, 113:41–59, 2002. abridged version.
19. D. Maister. Psychology of Waiting Lines. *Harvard Business School Cases*, pages 71–78, 1984.
20. A. Mandelbaum. Call Centers (Centres): Research Bibliography with Abstracts. 2002. Version 3, 137 pages. Downloadable from ie.technion.ac.il/serveng/References/ccbib.pdf.
21. A. Mandelbaum and S. Zeltyn. Staffing Many-Server Queues with Impatient Customers: Constraint Satisfaction in Call Centers. 2006. Working paper, Thechnion, Haifa, Israel.
22. E. Nakibly. *Predicting Waiting Times in Telephone Service Systems*. 2002. Ph.D. Thesis, The Senate of the Technion.
23. P. Naor. The Regulation of Queue Size by Levying Tolls. *Econometrica*, 37:15–24, 1969.

24. S. Taylor. Waiting for Service: The Relationship Between Delays and Evaluations of Service. *Journal of Marketing*, 58:56–69, 1994.
25. A. R. Ward and W. Whitt. Predicting Response Times in Processor-Sharing Queues. *Proceedings of the Fields Institute Conference on Communication Networks*, 2000.
26. A.R. Ward and P.W. Glynn. A Diffusion Approximation for a Markovian Queue with Reneging. *Queueing Systems: Theory and Applications (QUESTA)*, 43:103–128, 2003.
27. W. Whitt. Improving Service by Informing Customers about Anticipated Delays. *Management Science*, 45:192–207, 1999.
28. W. Whitt. Predicting Queueing Delays. *Management Science*, 45:870–888, 1999.
29. W. Whitt. Stochastic Models for the Design and Management of Customer Contact Centers: Some Research Directions. 2002. Working paper, Columbia University.
30. R.W. Wolff. Poisson Arrivals See Time Averages. *Operations Research*, 30:223–231, 1982.
31. D. Zakay. An Integrated Model of Time Estimation. *Times and Human Cognition: A Life Span Perspective*, 1989. Iris Levin and Dan Zakay, eds, Amsterdam: North Holland.
32. E. Zohar, A. Mandelbaum, and Shimkin N. Adaptive Behavior of Impatient Customers in Tele-Queues: Theory and Empirical Support. *Management Science*, 48:566–583, 2002.