ESTIMATING AND ANNOUNCING WAITING TIMES IN MULTIPLE CUSTOMER CLASS CALL CENTERS

Oualid Jouini * **Yves Dallery** *

* *Laboratoire Génie Industriel, Ecole Centrale Paris, Grande Voie des Vignes, Châtenay-Malabry 92295 Cedex, France*

Abstract: Motivated by practices in call centers, we consider a multiclass, multiserver queueing system that offers information about anticipated delays to customers. First, We propose estimators of the state-dependent waiting time distribution of a new arrival customer. We next validate the estimators by empirical data. Second, we investigate the benefits of informing customers about anticipated delays, and we discuss different ways to communicate this information. In particular, we give attention to analyzing the announcement of the waiting time by steps.

Keywords: Stochastic Models, telephone call centers, Predicting Delays, Communicating Anticipated Delays

1. INTRODUCTION

A call center is a service system in which agents (servers) serve customers (callers), remotely over the phone. 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. Estimates (Nakibly, 2002) indicate that around 70% of all customers transactions occur in call centers.

This paper is motivated by *Bouygues Telecom* which is a major mobile telephony service provider based in France. The *Bouygues Telecom* call center employs about 2000 agents. It receives around 70000 calls per day, and constitutes the main point of contact with the customer.

As in *Bouygues Telecom*, many organizations use telephone call centers as an important channel of communication with their customers. Such centers have limited resources and face highly unpredictable demand that often result in long waits for their customers. To improve the customer service levels and alleviate congestion, call centers have recently started experimenting by informing arriving customers about anticipated delays (Armony and Maglaras, 2004).

Information regarding the anticipated waiting 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 in visible queues, and it does not decrease over time. Customers have no means to estimate queue lengths or progress rate, and the feelings of frustration and anxiety increase during the waiting. The goal of our work is to propose estimators of the waiting time in service systems, and in

call centers in particular. We use exact analytical methods and approximations in order to estimate the waiting time of our call center schema. Since the goal is to provide information which is relevant to a specific 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.

Since we are dealing with stochastic systems, there is no possible way to predict the exact waiting time. The best one can do is to estimate the waiting time distribution. The service manager should then 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 the distribution. On the one hand, informing on 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 on a long waiting time, might result in longer perceived waiting times and in decrease in satisfaction.

The remainder of this paper is structured as follows. We conclude this section with a short literature survey. In Section 2, we describe the model we consider in this paper: a non-preemptive priority multiserver queueing system with two classes of customers. In Section 3, we develop the state-dependent estimators of a new arrival call depending on its customer class. We distinguish two cases depending on whether the new arrival call finds an empty queues, or he finds customers waiting for service in the queues. We treat the first and the second case in Section 3.1 and in Section 3.2, respectively. In Section 4, we discuss what to announce to customers on the basis of the developed estimators. In Section 5, we present some concluding remarks and discuss extensions.

The literature related to our work spans two main areas. The first is concerned with the analysis of multiserver systems motivated by call center. Call centers can 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 (Gans *et al.*, 2003) is an international call center where incoming calls are in different languages. Related studies include those by (Garnett and Mandelbaum, 2001), (Aksin and Karaesmen, 2002), and references therein. Our concern in this paper is a full-flexible call center. We assume that all agents are flexible enough to answer all requirements of service. It is a plausible assumption for *Bouygues Telecom* as well as many other real cases where customer's questions do not require specific skills from the agents. *Bouygues Telecom* call center is unilingual, and the complete flexibility where all agents are able to process all tasks is not as difficult as in a multilingual call center where learning several languages for all agents is almost impossible. *Bouygues Telecom*'s customers request information from agents, mainly to ask for further information about an invoice, to pay or to claim something, to add or to remove options in their mobile phone service, etc. Full-flexible call centers were extensively studied. For example, see (Kolesar and Green, 1998), (Gans *et al.*, 2003), (Aguir, 2004), (Jouini *et al.*, 2004*b*) and references therein. In this paper, we are dealing in particular, with the transient analysis. Several papers have been proposed for the study of the transient behavior of queues, but in general, analytical solutions are extremely difficult to obtain. Since the transient solutions require solving sets of differential equations, numerical methods are often employed. For simple queueing systems, see (Kleinrock, 1975), and (Gross and Harris, 1998). For more complicated systems, interesting results are obtained in (Guillemin, 1999), and (Jouini and Dallery, 2006).

The second area of literature that is close to our work is related to the problem of announcing waiting times. Close references are (Whitt, 1999*a*), (Whitt, 1999*b*), and (Nakibly, 2002). The literature on customers influenced by delay information begins with (Naor, 1969). An overview of customer psychology in waiting situations, including the impact of uncertainty, can be found in (Maister, 1984). (Taylor, 1994) showed that delays affect customers' service evaluations in an experiment involving airline flights. (Hui and Tse, 1996) conducted a survey on the relationship between information and customer satisfaction.

2. THE MODEL

We consider the queueing model of the call center. The company divides their customers into different classes according to their importance. There are tow classes type A and type B. In concrete terms, if the company own every month from one customer an amount of money crossing a given threshold, then that customer is of type A, else he is of type B. Extension to the general case of more than 2 classes of customers will be addressed later in this paper. Customers of type A have priority over customers of type B in the sense that agents are providing assistance to customers A first. The priority rule is non-preemptive, which simply means that an agent currently serving a B customer while an A customer joins the waiting queue will complete this service before turning to

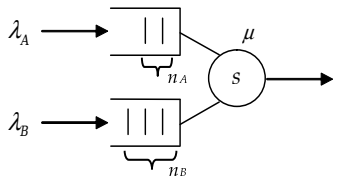

Fig. 1. The call center model

the A customer. The model consists of two infinite queues and a set of s identical servers representing the set of agents. All agents are able to answer all types of customers. Each type of customer has its own queue. Service times are assumed to be exponentially distributed and independent of each other with rate μ for both types of customers. The arrival process of customers is assumed to be Poisson. The arrival rate of customers A is λ_A, and that of customers B is λ_B. Following similar arguments, the behavior of this call center can be approximated by a simple non-preemptive priority $M/M/C$ queue with two classes of customers (A and B). Within each queue, the discipline is First Come, First Served (FCFS). Let the total arrival rate be λ_T, $\lambda_T = \lambda_A + \lambda_B$. Then, the server utilization ρ (proportion of time each server is busy) is $\rho = \lambda_T/s\mu$. Note that the condition for existence of a steady-state solution is $\rho < 1$; that is, the mean total arrival rate must be less than the mean maximal service rate of the system. The call center model is illustrated on Figure 1. In Section 3, we develop estimators for the state-dependent waiting time distribution of a new arrival call. The system state at a specific time is defined by the number of customers in the system. If the last is greater than the number of servers s, then the queues are not empty, and s customers are being served. Let n_A be the number of customers of type A in queue A, and n_B the one of customers B in queue B. Finally, let n_T be the total number of customers in the queues, $n_T = n_A + n_B$.

3. ESTIMATING THE STATE-DEPENDENT WAITING TIME DISTRIBUTION

In our study, we distinguish the case of empty queues, $n_T = 0$, from the case of non empty queues, $n_T > 0$. In the first case, a new arrival call faces two possibilities: The first possibility concerns the case when all the servers are busy, then the new arrival call have to wait that one of the s servers becomes idle in order to begin his service. The second possibility concerns the case when at least one of the servers is idle, then the call of interest begins his service immediately without having to wait. In the second case of non empty queues ($n_T > 0$), the call have to wait in all of the configurations.

3.1 Case of Empty Queues

Consider a new arrival call finding empty queues. Let p be the probability that the call of interest has to wait before beginning service. The probability p does not depend on the new arrival call type, and it represents the case that all the servers are busy. We are interested here on calculating p, in order to get some indications on its value in the working normal conditions of the call center. For instance, if the proportion p is too small, then most arriving calls begin service without waiting. So, there is little need for delay informations. However, if p is quite large, then a considerable proportion of new calls have to wait, and the prediction can be important.

The proportion p 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 the s servers busy knowing that the queues are empty.

$$\begin{aligned} p &= P\{all\ the\ servers\ are\ busy \mid n_T = 0\} \\ &= \frac{P\{all\ the\ servers\ are\ busy\ AND\ n_T = 0\}}{P\{n_T = 0\}}. \end{aligned}$$

Let k be the total number of customers in the system at the arrival moment of the call of interest. Then, we have

$$p = \frac{P\{k = s\}}{\sum_{i=0}^{s} P\{k = i\}}, \tag{1}$$

where $P\{k = i\}$ is the steady state probability that the number of customers in the system is i. We give the expression of p in Equation (2).

$$p = \frac{\frac{\lambda_T^s}{s!\mu^s}}{\sum_{i=0}^{s} \cdot \frac{\lambda_T^i}{i!\mu^i}} \tag{2}$$

Note that the above relation is equivalent to the one for an $M/M/C$ queue with the same parameters as here, with only one class of customers having an arrival rate of λ_T, and working under the FCFS discipline. Since service times for both customer types are identically distributed, then the latter system behaves identically as the one we consider in this paper, when considering the total number of customers in the system. This is due to the known work conserving property. In Figure 2, we plot the proportion p according to the server utilization ρ, for several call center's configurations. We see that p increases when ρ increases, that is, when the servers becomes more and more often busy. From Figure 2, we see also that for a fixed ρ, p is decreasing in the number of servers s. This is due to the pooling phenomenon, see (Jouini *et al.*, 2004*a*).

From the numerical examples reported above, we deduce some other qualitative results. We note that this numerical study concerns a range

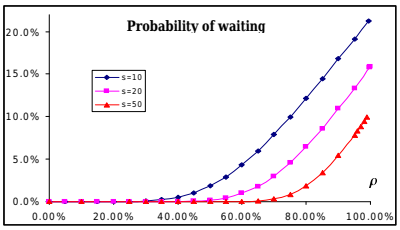

Fig. 2. Conditional probability of waiting, p

of typical parameters as those that would be encountered in real situations, that is, agents' team sizes ranging from 20 to 50, and server utilizations up to 98%. The insight here is that only a small proportion of new calls (who find no one in the queues) have to wait before beginning service. 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, with $s = 30$ and $\mu^{-1} = 5$ min, type A customers' expected waiting time is of 10 sec.

3.2 Case of Non Empty Queues

We consider now a new arrival customer who finds non empty queues. He finds $s + n_T$ customers in the system, $n_T = n_A + n_B$. 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 customers type B, because their waiting time is affected by future type A arrivals.

Waiting time of a type A customer

Because of the strict priority, the waiting time of a new arrival type A customer does not depend on the number of type B customers waiting in the system. Given n_A type A customers waiting in the queue, the call of interest has to wait first for the time until the first service completion (one service time of either type A or type B customer), and second, he has to wait for n_A independent service times of type A customers. The time until a service completion is exponentially distributed with mean $1/s\mu$, and it is independent of the previous history because of the memoryless property. Hence, the waiting time of the new call is the sum of $n_A + 1$ i.i.d. exponential random variables each with mean $1/s\mu$, which has an Erlang distribution. Conditioning on the state information, let X_A be the random variable describing the waiting time of interest. The mean and standard deviation of X_A are, respectively,

$$E(X_A) = \frac{n_A + 1}{s\mu} \quad \text{and} \quad \sigma(X_A) = \frac{\sqrt{n_A + 1}}{s\mu}. \quad (3)$$

We can also calculate the full cumulative distribution function (cdf) of X_A. The Erlang cdf is available in closed form, see (Gross and Harris, 1998). We describe the ratio of the standard deviation by the mean, $\sigma(X_A)/E(X_A)$. This ratio has the remarkably simple form

$$\frac{\sigma(X_A)}{E(X_A)} = \frac{1}{\sqrt{n_A + 1}}, \quad (4)$$

independent of μ and s. We deduce that the waiting time distribution (conditional on all servers being busy) is too much 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 need only to learn the current state information. The arrival process can be an arbitrary process, even non stationary.

Waiting time of a type B customer

We focus on the waiting time distribution of type B customer who finds n_A type A customers and n_B type B customers waiting in the queues. let X_B be the random variable describing this duration. We should observe that X_B does not depend on the order of service. The only requirement we have to respect is that the servers remain busy as long as customers are in the system. It is the well known "Work conserving" property. Hence, the duration of interest can be divided into 2 independent 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 the queue. The busy period is the one of an $M/M/C$ queue with a Poisson arrival rate of λ_A. It is defined as the time from an arrival of a customer to an empty system until the first time one of the servers becomes idle.

Knowing that the remaining service time of one customer is independent from the 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 queues. Finally, the new type B customer has to wait for $n_A + n_B + 1$ i.i.d. busy periods. The probability density function (pdf) of the busy period of an $M/M/1$ queue is found, for example, in (Kleinrock, 1975). Then, with a little thought it should be clear that the busy period pdf of interest here is obtained by just replacing the capacity of service μ (in the case of an $M/M/1$ queue) by $s\mu$ (in the case of an $M/M/C$ queue). To get the analytic expression of the X_B pdf, the mathematics becomes extremely complicated. Even the numerical calculus takes too long time, being a $n_A + n_B$-fold convolution of the busy period pdf. Fortunately, the mean of X_B is fairly simple to obtain, by summing the busy periods' means up to $n_A + n_B + 1$. The same approach is still valid for the variance computation of X_B using, in

addition, the independence between the random variables of the busy periods duration. 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 cdf Laplace transform in the time t. Equations (5) and (6) show, respectively, the mean and the standard deviation of X_B,

$$E(X_B) = \frac{n_A + n_B + 1}{s\mu - \lambda_A}, \quad (5)$$

$$\sigma(X_B) = \sqrt{(n_A + n_B + 1)\frac{s\mu + \lambda_A}{(s\mu - \lambda_A)^3}}. \quad (6)$$

Equation (7) shows the ratio of the standard deviation by the mean,

$$\frac{\sigma(X_B)}{E(X_B)} = \sqrt{\frac{s\mu}{(n_A + n_B + 1)(s\mu - \lambda_A)}}. \quad (7)$$

Once again, we deduce that the conditional waiting time distribution is too much concentrated about its mean, for large values of $n_A + n_B$.

It is not too difficult to extend the conditional waiting time estimation to the general case with k classes of customers, $k > 2$. For example, from the third class side, it is equivalent to aggregate the first two classes into one equivalent class by summing the corresponding arriving flows. Hence, in an analogue fashion as above, we can get easily the quantities of interest.

Normal Approximation

From a practical point of view, a normal distribution provide a satisfactory approximation. As discussed in (Whitt, 1999*b*), given that we exploit system state information, the normal approximation should works well. We only need the mean and the standard deviation of the conditional waiting time to get all of its moments, and equivalently its cumulative distribution function (approximately). For the rest of the paper, we use this approximation for both type A and type B customers, by using the exact means and standard deviations obtained above.

To validate the approximations, we did a comparaison with real waiting times from the call center database. For each customer type, and for many real configurations of the system state, we plotted the real waiting time distributions (we used the database history to get the real waiting times data). Then, we compared these real distributions with the approximated ones. We did the comparaison for several configurations of the call canter. We do not present the statistical comparaison here, but we give, in the following, some indications on the approximations' errors. It is natural that we will not announce the expected waiting time, for a new arrival customer. The call center manager will fix a given percentile depending on the customers' requirements. Then, on the basis of the required cdf, we will announce the anticipated waiting time to the customer. For percentiles ranging from 90% to 95%, we find that the deviations are ranging from 5% to 20%. The results appear to be attractive in the sense that the announcement is often done with a high accuracy (a percentile of 95% for *Bouygues Telecom*). We should note that beyond the normal approximation, some assumptions take part in the errors' values, such as the exponential approximation of the service time, ignoring the abandonments, etc.

4. ANNOUNCING ABOUT ANTICIPATED DELAYS BY STEPS

In this section, we focus on what we announce to customers. Given the waiting time estimators, the next question is how we should profit from these predictions? There is not a single best method for all circumstances. In the following, we discuss some elements addressing this issue. We do not think that it is interesting for the customer to get a high accuracy of his delay. There is no a remarkable difference between being informed that he will wait for example 1.64 min or 1.79 min. Also, there is no need to tell the customer a large delay. If the anticipated delay is greater than 5 min, we should just inform the customer that he will wait more than 5 min. A delay beyond 5 min seems to be excessive in our call center case, and any value in this range is perceived with the same fashion. This value may not be reasonable for an emergency call center case.

The idea we propose here is to inform about the delays by steps. For example, we define the following steps: < 30 sec (the delay will be less than 30 sec), < 1 min, < 2 min, < 3 min, < 4 min, < 5 min, and a final step > 5 min. Now, for a given percentile, we evaluate the waiting time, then we pick among the predefined intervals the one the evaluated time is belonging, and we announce this interval to the customer. One interesting advantage of this method should be noted. In the cases when the estimated delay and the real one are in the same interval of time, the error do not exist!. For example, if the first is 1.21 min and the second is 1.85 min, then in both of the cases we will announce that the delay is less than 2 min. So the approximated delay and the real one coincide. We did many experiments, we found that there is no errors between approximation and reality in 71% of the cases. In the rest of the cases, the error is of 1 min (we announce the step below the real step or the one above). Another advantage is that informing by steps improves the accuracy, because the step we announce is always an upper bound. For example, if with a chance of 90%, the waiting time is less than 55 sec, then we will announce 1

min. Hence, the accuracy increases from 90% to 94%.

An other idea is to announce a lower bound and an upper bound of the estimated delay, instead of only an upper bound. The idea is of value when the cdf of the conditional waiting time is too close to zero for small times values. This case occurs for a new arrival call who finds a large number of customers in the queues.

5. CONCLUSION AND EXTENSIONS

We focused on a fundamental problem in the management of call centers, that is, customer-delay information. Predicting delays is especially important when customers do not have direct access to system state information. In the first part of the study, we developed estimators of the conditional waiting time distribution in a call center working under the strict priority policy. We found the exact distribution functions, but we used the normal approximation because the first ones are too complicate to implant in practice. The normal distribution is supported by theoretical results based on the central limit theorem. Next, we showed that estimators can be easily extended in the case of k classes of customers, $k > 2$. We validated our estimators by real waiting times data. In the second part of the study, we focused on an important question dealing with the way of the delay announcement. We proposed to announce the anticipated delays by steps, and we showed that this method has many advantages to reduce the errors' approximations.

In a future study, we aim to extend the estimators by considering abandonments and limited waiting lines, and more general service-time distributions to get more accurate analysis. By considering the abandonments, we may reduce the delays' approximations, especially when a new customer finds a large number of calls waiting in the queue. In such cases, many customers may abandon from the queues, which reduce the waiting time of the customer of interest. Up to now, we do not take into account this phenomenon. It would be also interesting to investigate the impact of the announcement on the customer abandonment experience. In fact, when we inform one customer about his delay, he will decide from the beginning, either to wait until starting his service without abandoning, or to hang up immediately.

REFERENCES

Aguir, M. S. (2004). *Modeles stochastiques pour laide a la decision dans les centres dappels.* Ph.D. Thesis, Ecole Centrale Paris.

Aksin, O. Z. and F. Karaesmen (2002). Designing flexibility: Characterizing the value of cross-training practices. Working paper, INSEAD.

Armony, M. and C. Maglaras (2004). Contact centers with a call-back option and real-time delay information. *Operations Research* **52**, 527–545.

Gans, N., G. Koole and A. Mandelbaum (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* **5**, 73–141.

Garnett, O. and A. Mandelbaum (2001). An introduction to skills-based routing and its operational complexities. Teaching note, Technion.

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

Guillemin, F. (1999). Excursions of birth and death processes, orthogonal polynomials, and continued fractions. *Journal of Applied Probability* **36**, 752–770.

Hui, M. and D. Tse (1996). What to tell customer in waits of different lengths: an integrative model of service evaluation. *Journal of Marketing* **60**, 81–90.

Jouini, O. and Y. Dallery (2006). Moments of first passage times in general birth-death processes. Submitted to Stochastic Models.

Jouini, O., Y. Dallery and R. Nait-Abdallah (2004*a*). Analysis of the impact of team-based organizations in call centers management. Submitted to Management Science.

Jouini, O., Y. Dallery and R. Nait-Abdallah (2004*b*). Stochastic models of customer portfolio management in call centers. *Proceedings of the German Operations Research Society.*

Kleinrock, L. (1975). *Queueing Systems, Theory.* Vol. I. A Wiley-Interscience Publication.

Kolesar, P. J. and L. V. Green (1998). Insights on service system design from a normal approximation to erlang's delay formula. *Production Operation Management* **7**, 282–293.

Maister, D. (1984). Psychology of waiting lines. *Harvard Business School Cases* pp. 71–78.

Nakibly, E. (2002). *Predicting Waiting Times in Telephone Service Systems.* Ph.D. Thesis, The Senate of the Technion, Haifa, Israel.

Naor, P. (1969). The regulation of queue size by levying tolls. *Econometrica* **37**, 15–24.

Taylor, S. (1994). Waiting for service: The relationship between delays and evaluations of service. *Journal of Marketing* **58**, 56–69.

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

Whitt, W. (1999*b*). Predicting queueing delays. *Management Science* **45**, 870–888.