A Modeling Tool to Minimize the Expected Waiting Time of Call Center's Customers with Optimized Utilization of Resources

Mohsin Iftikhar
Computer Science Department
College of Computer and Information Science
King Saud University, Riyadh, Saudi Arabia
miftikhar@ksu.edu.sa

Abstract — In order to deliver assured Quality of Service (QoS) to the customer in terms of minimizing the expected waiting time in the queue; the service providers are putting a lot of effort. The varying calling pattern of the customers during different times of the day poses many challenging problems in terms of allocating the number of employees in an efficient manner to handle the customer's calls. To be on the safe side, the service providers put more number of employees to handle customer calls, which also causes over provisioning of the resources and result high cost. On the other side, if service providers use less number of employees especially during the peak times, the customer has to wait for a long time in the queue before it gets service, which ultimately reduces the customer's satisfaction experience in terms of QoS. So there is a need to find out the best tradeoff between the number of employees allocated to serve the customers and the expected waiting time of the customer in the queue during different times of the day. To provide an efficient and assured QoS, here we present a novel and robust idea, which not only is targeting to reduce the expected waiting time of the customer in the queue but also provide an efficient mechanism to allocate varying number of employees during different time slots of the day according to the demand.

Keywords: QoS, Resource Utilization, Queuing and Scheduling

I. INTRODUCTION

Quality of Service is the ultimate aspiration of every customer at any communication platform that is often engaged with large array of requests in a queued fashion. Quick response with minimum or almost no waiting time and fair allocation of resources are the prominent factors to facilitate remarkable Quality of Service (QoS). To provide assured and guaranteed QoS in terms of expected waiting time for calling customers, the service providers over provision their resources to remain competitive in the market. This technique definitely incurs more cost. On the other side, if over provisioning is not done, the companies might start losing the customers, if they have to wait long time in the queues before getting service.

This study deals with two aspects: first to provide assured QoS to the customer in terms of minimizing the expected waiting time of a customer in a queue at any call center and second, utilizing the resources of a call center in an optimized fashion so that over provisioning should not be done. There are different types of call centers, namely, inbound call centers, outbound call centers, web enabled call centers, CRM call centers, telemarketing call centers and telephone call centers [1-5].

II. RELATED WORK

A queuing model to provide guaranteed QoS to the customer has been studied in [6]. The authors have modeled the time constrained incoming calls as type 1 customers. The type 2 calls represent the backlog of outgoing calls. The objective is to obtain a simple model that provides insight into the system behavior. Moreover, the aim is to derive simple scheduling policies that can be implemented in call center software. They proposed scheduling policies that keep part of the service capacity free for arriving time-constrained type 1 jobs. This policy combines the traffic from the two channels in such a way, which meets the expected waiting time constraint for type 1 customers while maximizing the throughput for type 2 jobs at the same time. In contrast to many other queuing models, they show that their policy is optimal for equal service time distributions. The experiments indicate that the optimal policy behaves nearly as a threshold policy. In [7] the authors study to systematically analyze the fit of the Erlang C model in realistic call center situations. They seek to understand the nature and magnitudes of the error associated with the model, and develop a better understanding of what factors influence prediction error. The fluid approximations for the mean number in the system have been derived for the two customers' classes using $M_t/M/n$ queuing model and preemptive priority resume logic [8]. Since the high priority customers can preempt the lower priority ones, hence, the low priority class customers will essentially receive service only as if no other type of customer (i.e. high priority) is present in the system. Thus, the high priority customer class results will be almost the same as the results for the single customer class.

The only difference is the dynamic priority process for the low priority customers, where these customers can abandon their queue and enter the high priority queue as a high priority customer. This process adds an extra term to the differential equations describing the process for the high priority customers. The authors suggest that for M_t/M_t/n, the number in system, or queue length, the process Q= {Q (t) I t> 0), as defined in Mandelbaum et al. (2000) for the single customer class case, must be defined for two customer classes [8]. The authors in [9] analyze different kinds of performance measures in call centers. The authors describe the model as two areas in the system i.e. one is waiting and service area and the other is retrial area. They propose a queuing model and extract the formulas for different kinds of performance measures such as mean number of customers in waiting and service area, mean number of busy servers, mean number of customers in retrial area, probability of blocking and loss probability. These indexes can be used for performance analysis of call centers.

In [10], the authors present two models. In each model, call center agents' assignments are different. In the first model, the agents are divided into two groups. One group of agents is assigned for handling inbound calls and another group of agent is assigned for taking care of outbound calls; whereas in the second model, agents handle both inbound and outbound calls simultaneously. The findings indicate that most of the time, the first model can gain more service effectiveness than second model. For example, in case of less expected waiting time (maximum one minute) and at less customer arrival rate, assigning inbound and outbound calls separately for separate group of agents gain more advantage in terms of service level.

Mandelbaum and Shimkin [11] developed a theoretical equilibrium-analysis of rational customers who compare their expected remaining waiting time with a subjective value they ascribe to service. This is equivalent to assuming a linear cost structure, and it implies that additional factors, such as the likelihood of "never" being served, are required to motivate a waiting customer to abandon the queue. Such a motivation is not needed as mentioned in [12] by Shimkin and Mandelbaum. They show that when waiting costs are nonlinear; an analogous equilibrium can be achieved. Zohar et al. [13] provide an empirical evidence for rational and, adaptive customers through presenting a simpler and analytically tractable form of the original, linear model. Armony and Maglaras [14, 15] and Whitt [16] both develop similar notions of abandonment and congestion as equilibrium phenomena.

The prior work have not considered the realistic arrival pattern of calls to a call center. Also, the queuing models the authors have considered to model the behavior of arriving pattern of customers are not realistic. Based on these factors, first of all, it is utmost important to consider the realistic arrival pattern of customer's calls to a call center. Also to model the behavior of the customer, we need to consider a traffic model that must be able to capture the bursty nature of the traffic because the traffic pattern of incoming calls of customers is not uniform. It varies during different times of the day and also over different days of the week. Hence, to cover this gap, this study provides guaranteed QoS by minimizing the waiting time of queued customer under optimal utilization of resources of any call center.

III. METHODOLOGY

The core idea of proposed model is to achieve best QoS that is having prime important for any service provider to remain competitive in the global market. The proposed research will provide a theoretical foundation for a variety of other research problems and have an impact on the communication network research community and telecommunication industry as a whole. We explain the methodology through a detailed example below, which we adopted to achieve the desired outcomes from this study. First of all, here we present some statistics that has been gathered for two consecutive months. This statistics present the arriving call pattern of the customers to a call center of a local telecommunication company in Australia (i.e. average number of calls coming from the customers during each half hour) over different time slots of the day and during different days of the week. Figures 1-7 present the calling pattern of the customers from Monday to Sunday respectively and Figures 8-9 show the calling pattern during weekdays and weekends respectively. The mean values in the graphs are presented with 95% confidence interval.

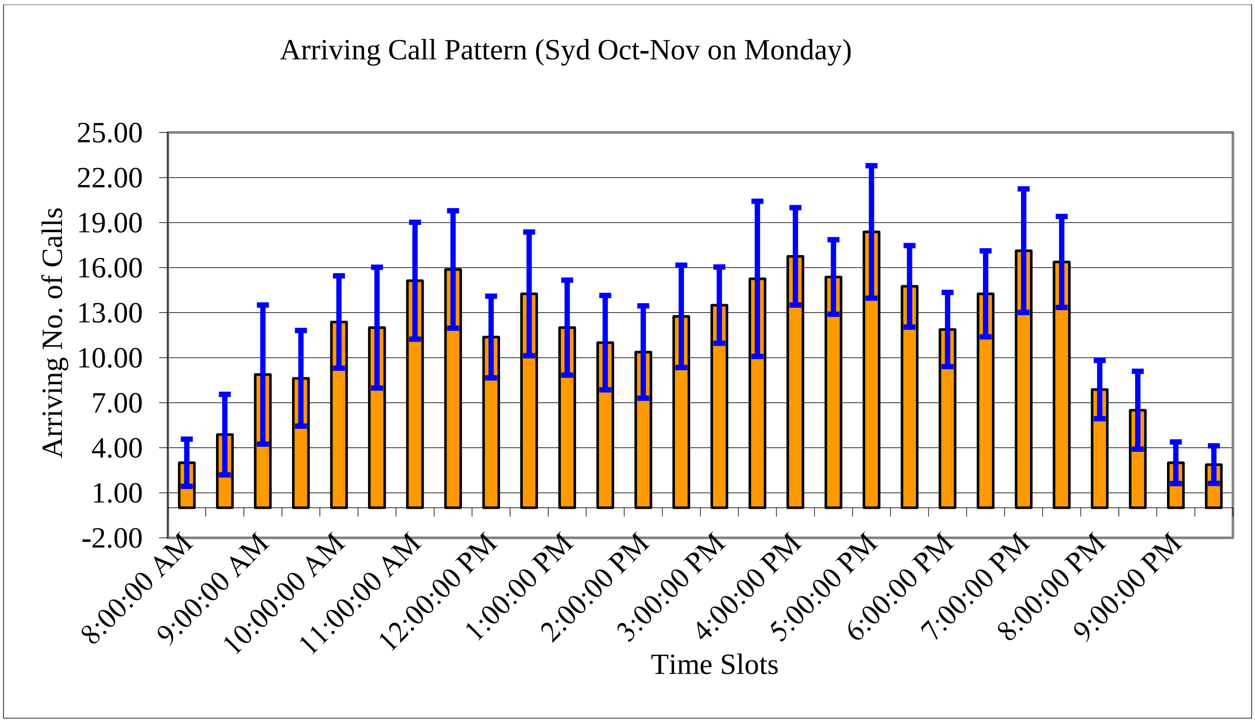

Figure 1. Arriving Call Pattern of the Customers on Mondays (During October-November)

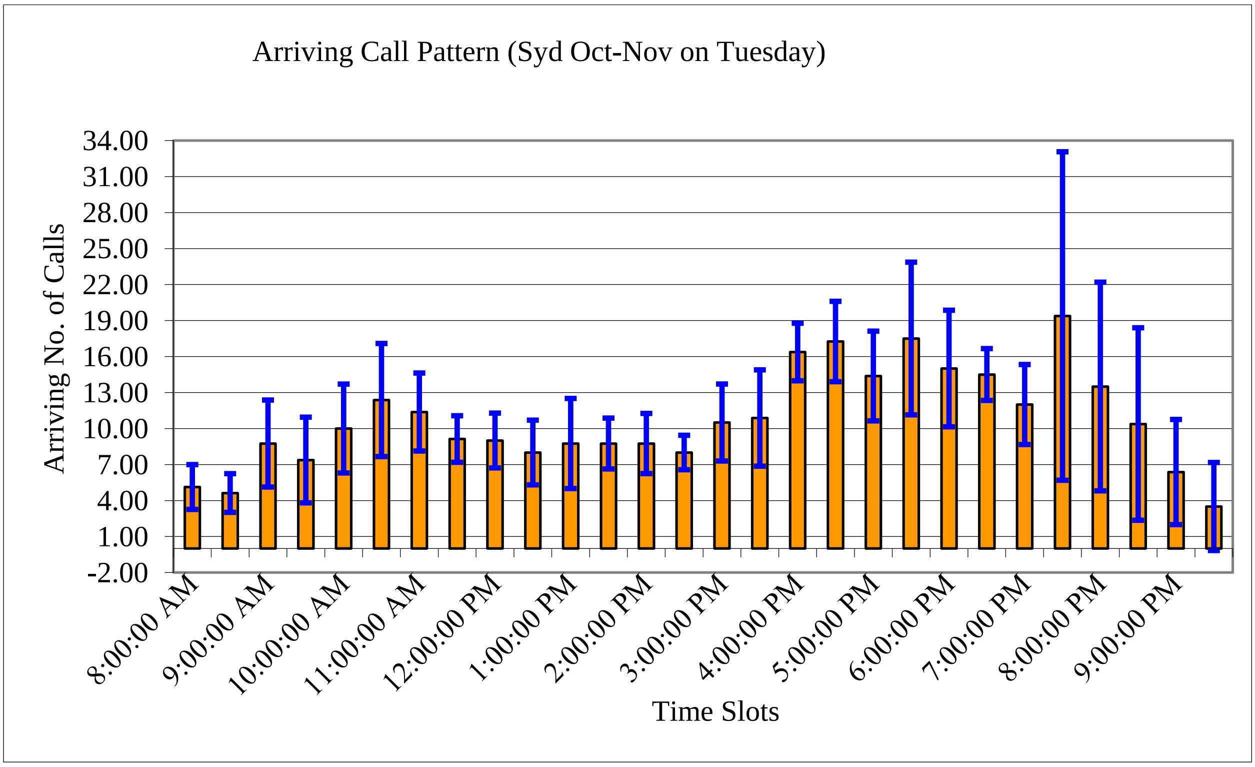

Figure 2. Arriving Call Pattern of the Customers on Tuesdays (During October-November)

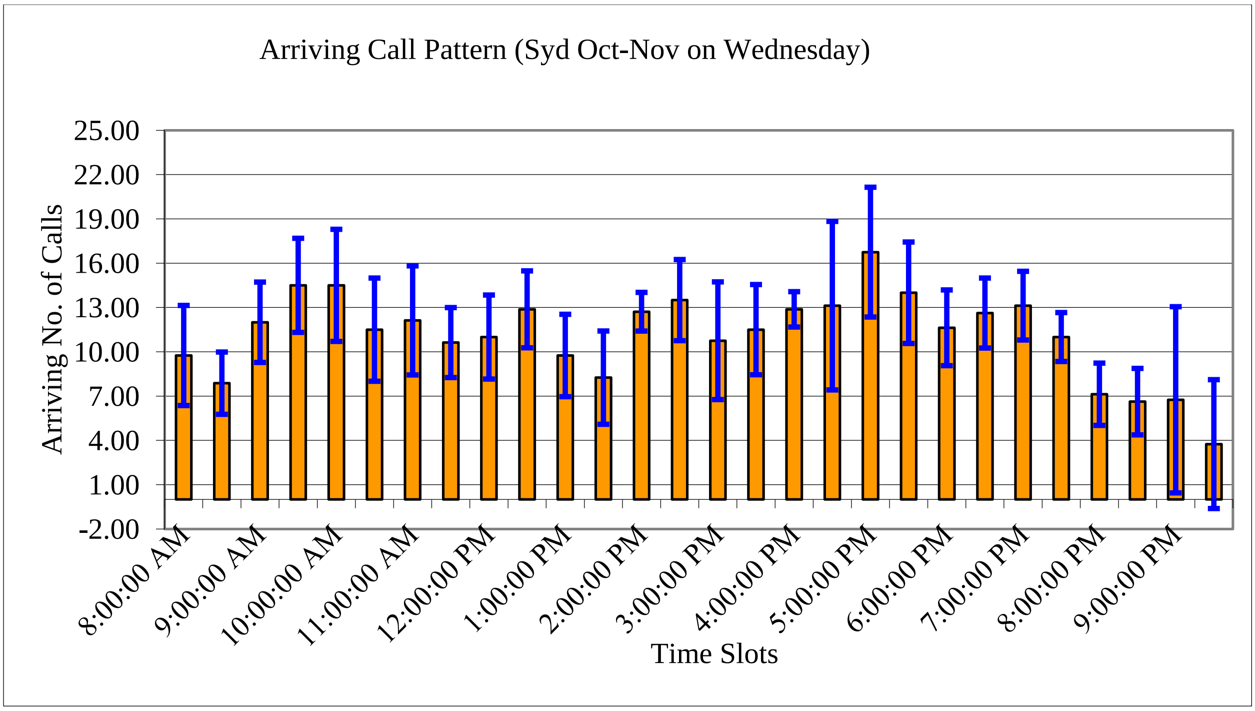

Figure 3. Arriving Call Pattern of the Customers on Wednesdays (During October-November)

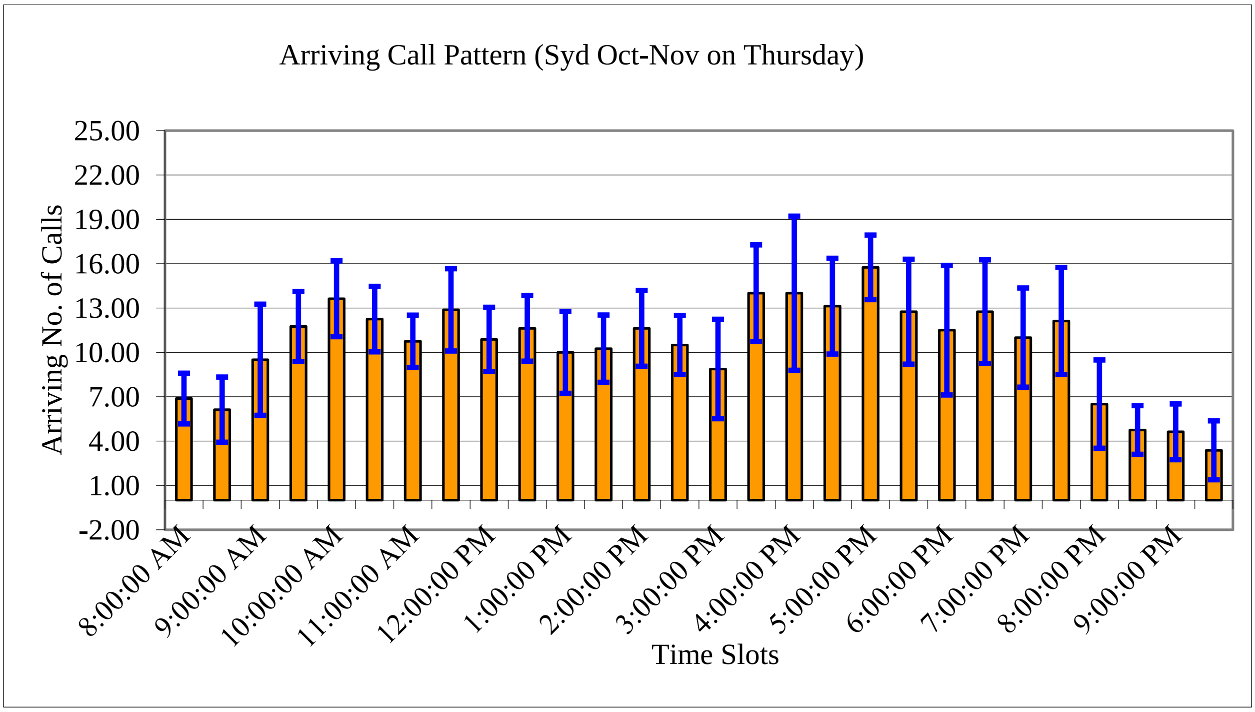

Figure 4. Arriving Call Pattern of the Customers on Thursdays (During October-November)

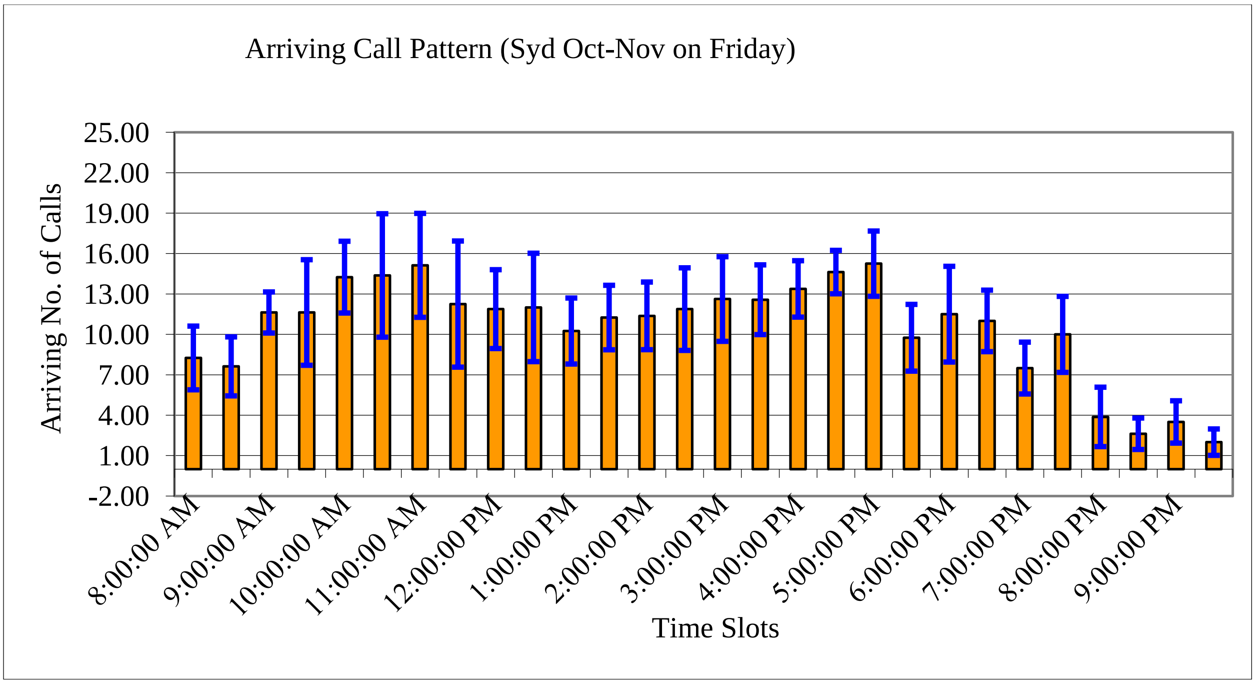

Figure 5. Arriving Call Pattern of the Customers on Fridays (During October-November)

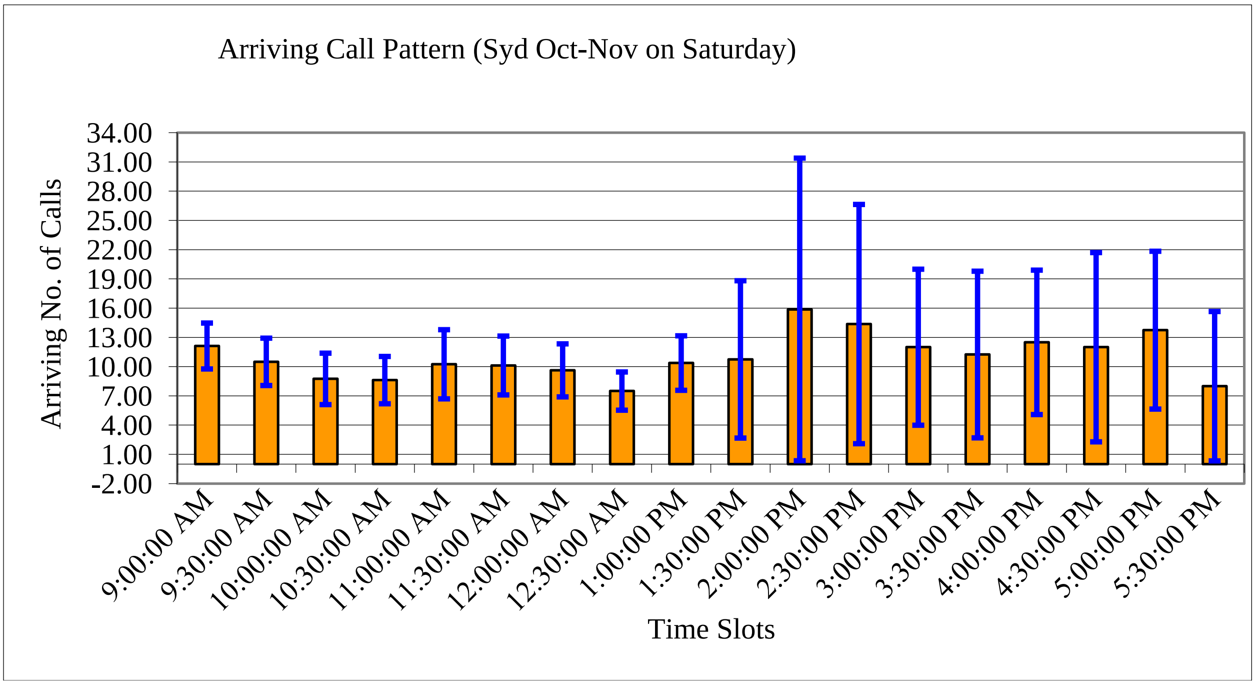

Figure 6. Arriving Call Pattern of the Customers on Saturdays (During October-November)

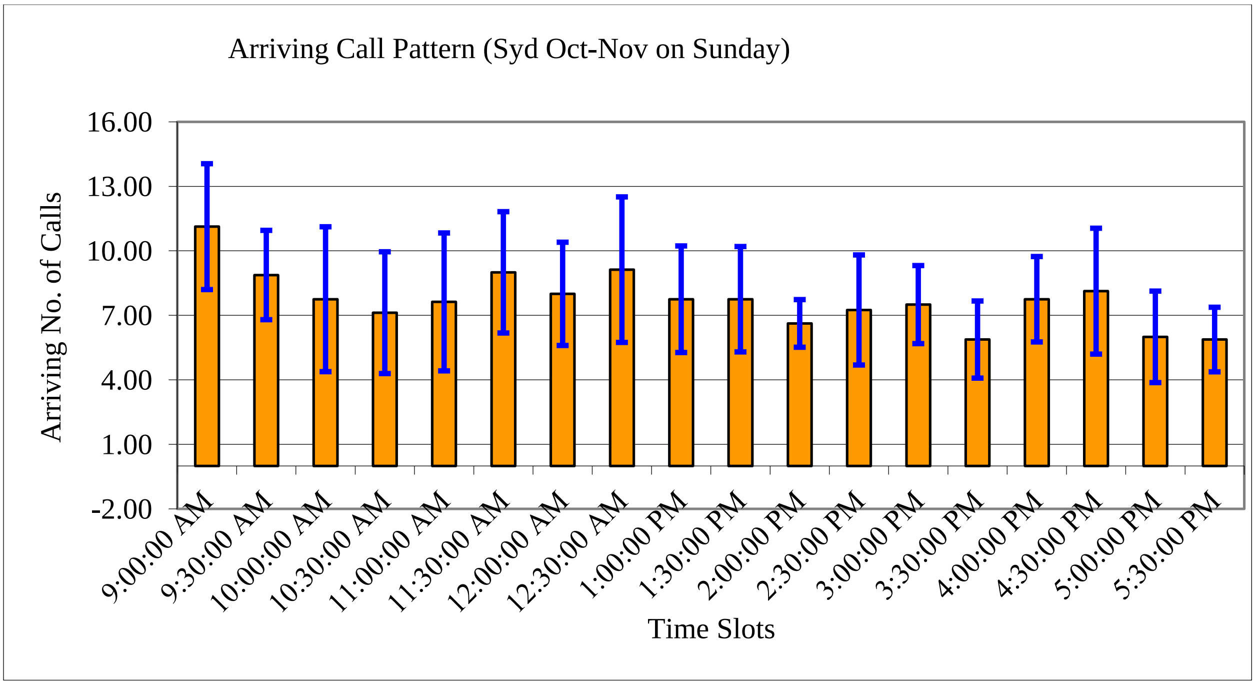

Figure 7. Arriving Call Pattern of the Customers on Sundays (During October-November)

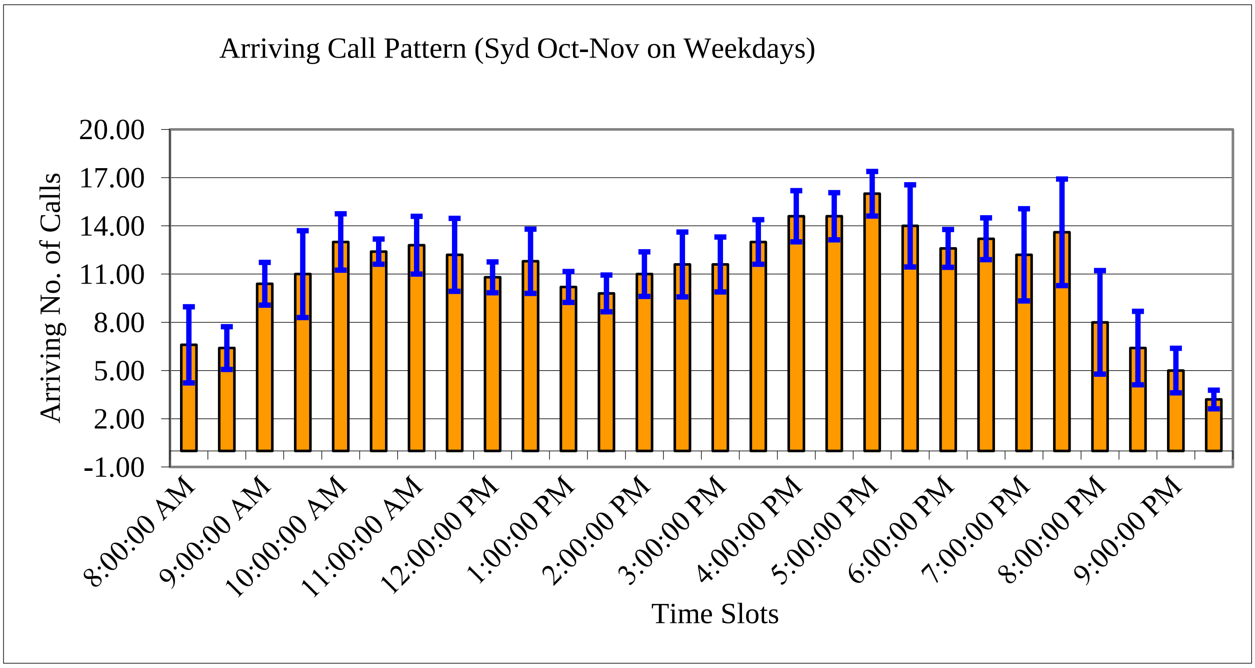

Figure 8. Arriving Call Pattern of the Customers on Weekdays (During October-November)

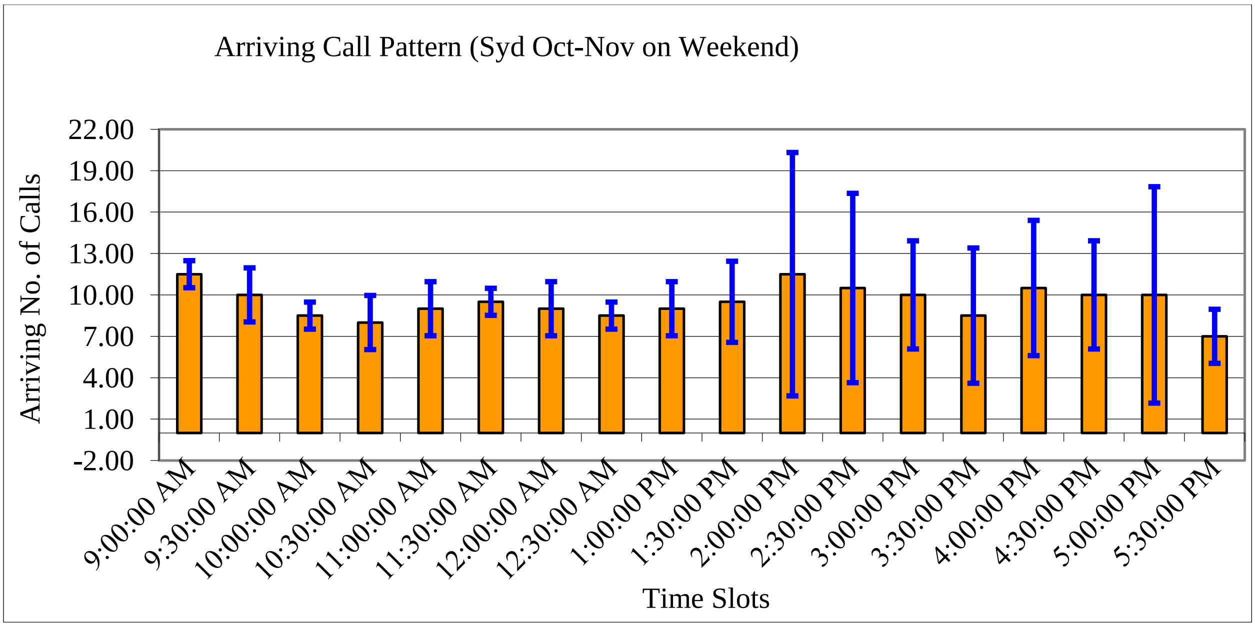

Figure 9. Arriving Call Pattern of the Customers on Weekends (During October-November)

Now to allocate the number of employees in an efficient manner, first of all, we need to find out the expected waiting time that a new arriving customer can experience during each time slot (half hour). To proceed with this, we make the following assumptions. We denote the arrival rate (i.e., average number of calls coming from the customers during each half hour) by λ . When a new customer arrives and if he finds some customers in the queue, we denote the expected number of customers already waiting in the queue by $E[N]$. The average service time required to handle each customer is denoted by $E[S]$. Based on the gathered statistics of these two months, we have found that an employee of internet help desk takes approximately 12 minutes to handle each call on average. The expected waiting time, a customer can experience before it starts getting service is denoted by $E[W]$. We denote the number of employees providing service to the customers during each particular time slot by A. When a new customer arrives and finds some customers already waiting in the queue then this new arriving customer can only start getting service until all the customers waiting in front of it leave the waiting queue and join the service. The expected waiting time of this new arriving customer can be found from:

$$E[W] = E[S] + \frac{E[N]E[S]}{A} \qquad \textbf{(1)}$$

Figure 10 shows the expected waiting time of a new arriving customer in a block diagram.

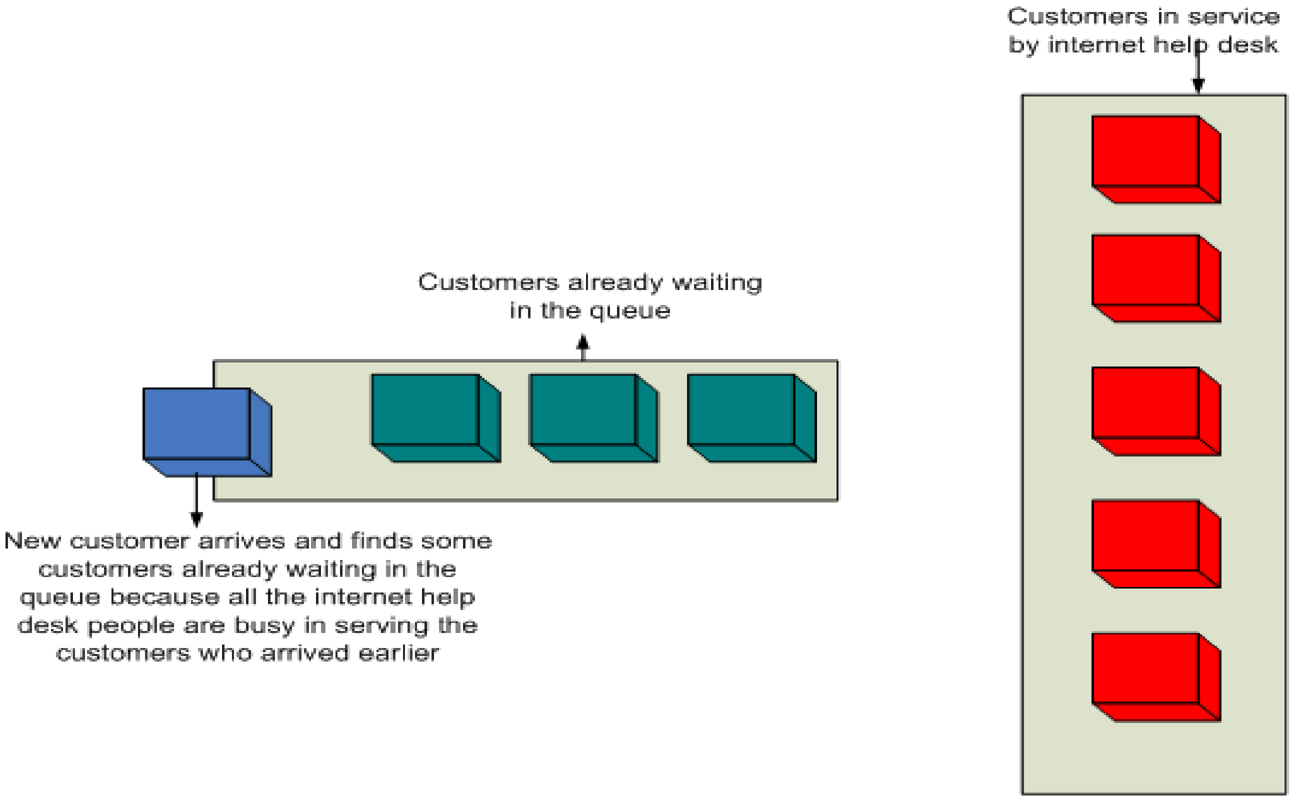
Customers in service
by internet help desk
Customers already waiting
in the queue
New customer arrives and finds some
customers already waiting in the
queue because all the internet help
desk people are busy in serving the
customers who arrived earlier

Figure 10. A new customer arrives and finds some customers already waiting in the queue because all of the internet help desk people are busy in providing the service to the customers who arrived earlier

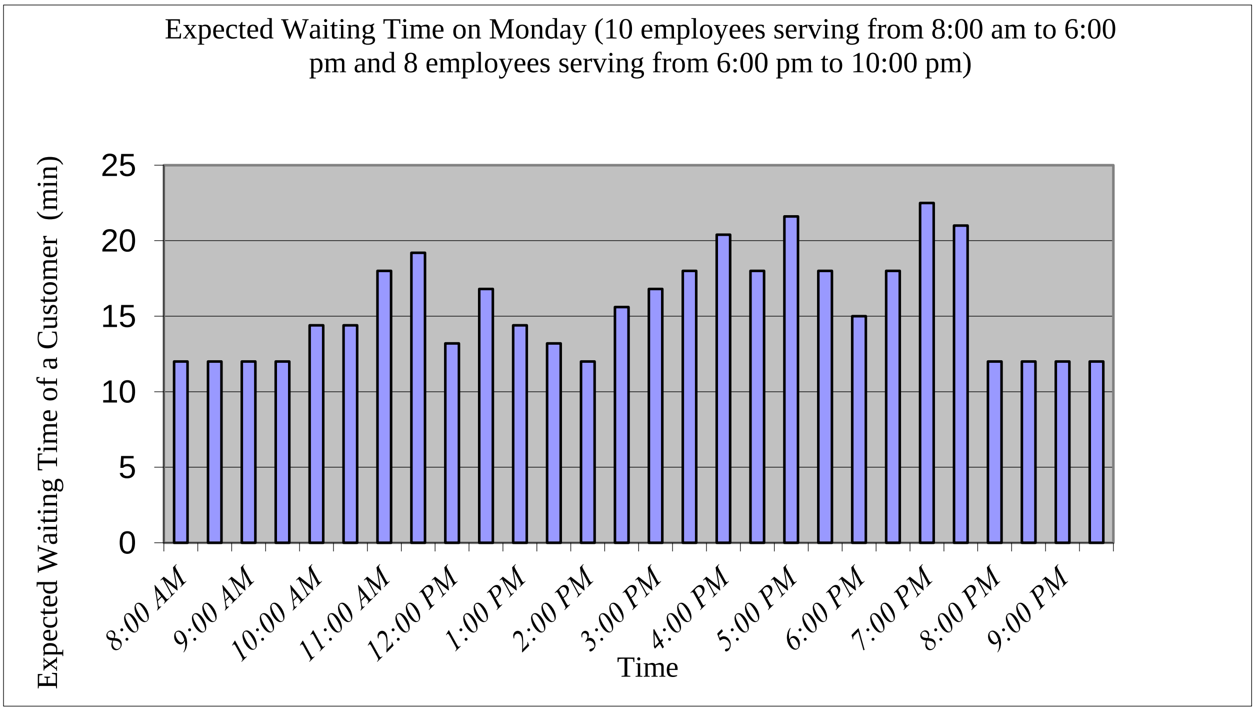

Figure 11. Expected Waiting Time of the Customer based on the gathered statistics on Mondays during October-November

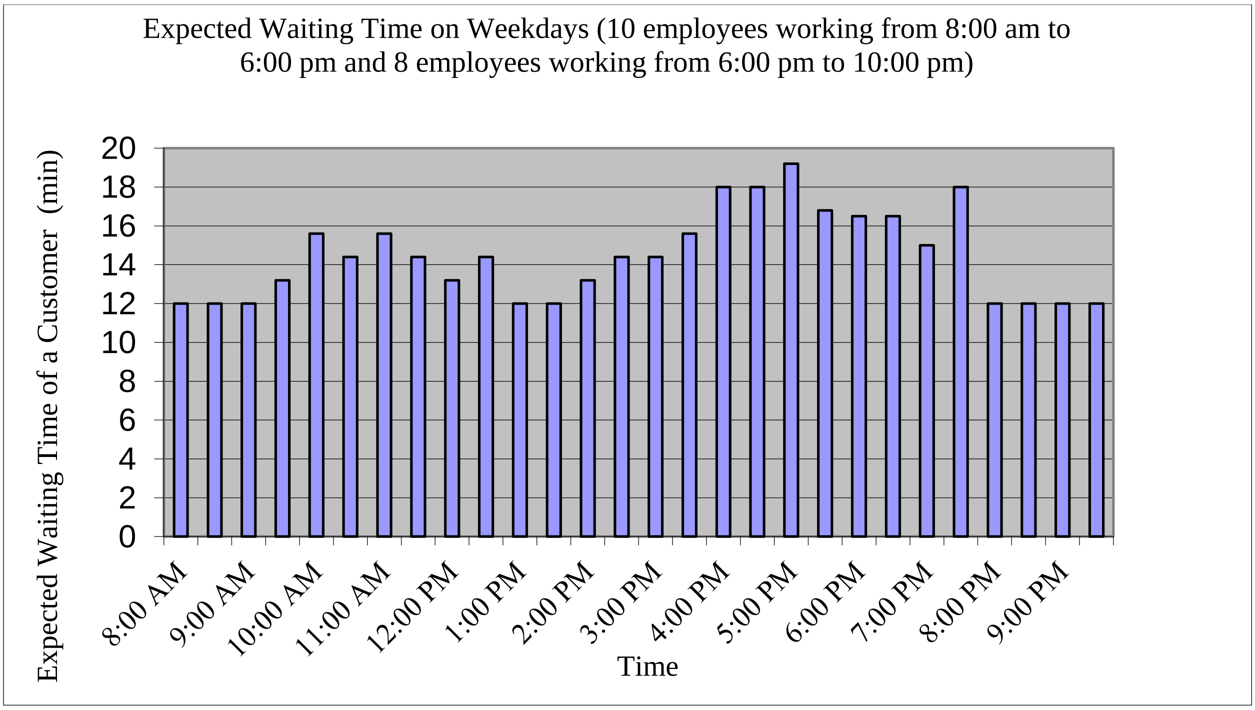

Figure 12. Expected Waiting Time of the Customer based on the gathered statistics on Weekdays during October-November

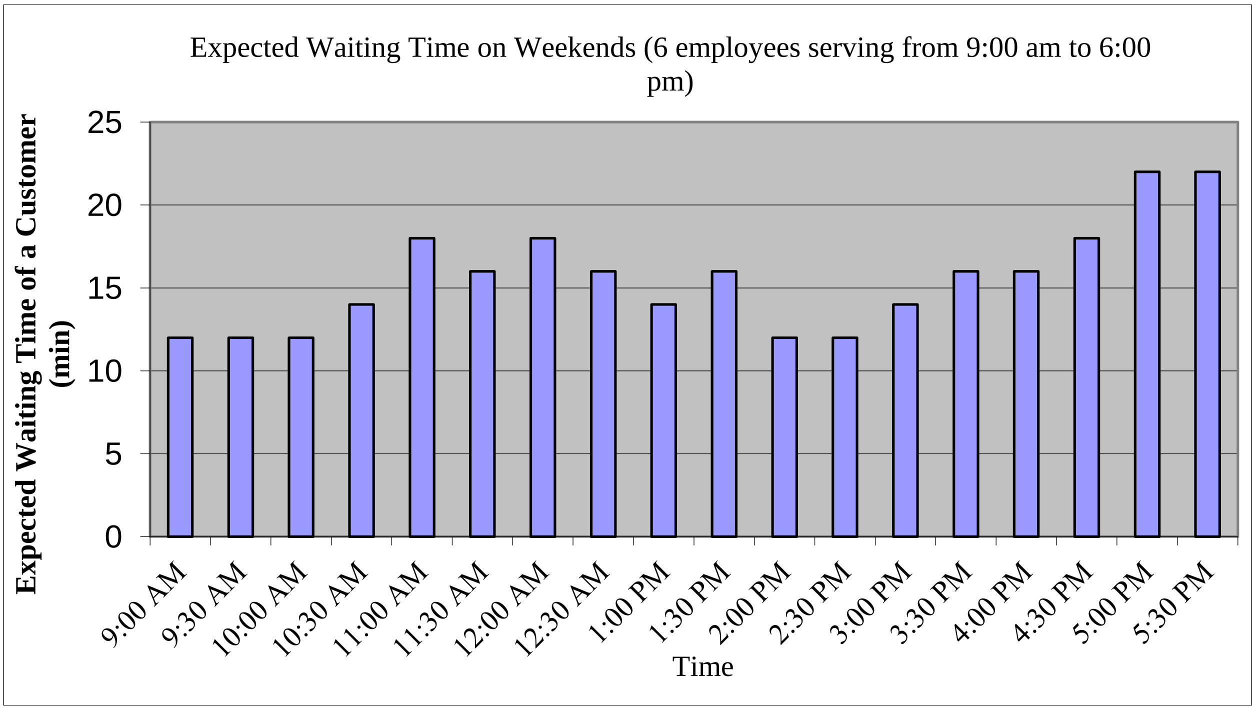

Figure 13. Expected Waiting Time of the Customer based on the gathered statistics on Weekends during October-November

From the gathered statistics, we can easily find out the number of waiting customers in the queue during each time slot. By putting the values in equation (1), we can easily find out the expected waiting time of a new arriving customer during each time slot. Here to conduct the worst case analysis, we assume that all of the customers during each time slot (half hour) arrive in first two-three minutes. Then on the basis of equation (1), we can find out the expected waiting time of a customer. Figure 10 is showing the average expected waiting time of a customer on Mondays during different times of the day. Similarly, Figures 11 and 12 present the average expected waiting time of a customer during weekdays and weekends respectively during different times of the day.

IV. PROPOSED MODEL

As shown through the above simple example, we can easily observe that by gathering the real data (i.e. the number of calls coming to a call center during 24 hours a day and over the entire week), it becomes easy to model it. Also there is another important point to note that customer never call on a uniform pattern but instead, there is a varying pattern of incoming calls throughout the day and most of the time, the incoming calls reflect a bursty nature especially during the peak hours. Hence it is utmost important to consider the following points in call center modeling.

(1) We need to rely on realistic data of incoming calls of customer to call center
(2) To model the behavior of incoming calls, we cannot use traditional Poisson model, instead, we need to consider a traffic model that must be able to capture the bursty nature of incoming calls of customers. The most common models that can capture the bursty nature of traffic are On/Off, Fractional Brownian Motion (FBM) and Levy motion.
(3) Then we need to combine this traffic model with a queuing system. And we need to extract the closed form expressions for different QoS parameters such as queuing delay, queue length, call drop rate in case of full queue etc.
(4) Simulate the analytical framework
(5) Develop a tool for optimizing the utilization of resources of a call center while at the same time minimizing the customer's expected waiting time in a queue.

The selection of the right traffic model is the most important step in call center modeling because, it must be able to capture the exact nature of incoming call patterns of the customers. Further, the traffic model must also exhibit the following characteristics. It must be analytical and solvable (when we feed it into some queuing system, we must be able to extract the QoS parameters from it). It must be less time consuming for simulation studies. It must exhibit accuracy, which is the most important characteristic particularly from business point of view.

On the behalf of Figures 1-9, we can forecast the arriving pattern of the customer's calls in a very good manner. Also on the basis of above analysis, we can provide the best service to the customer in terms of minimizing their waiting time in the queue by allocating the exact required number of employees during each time slot of the day. Overall, it will improve the system efficiency by allocating the exact number of resources (i.e. the required number of internet help desk people) during each time slot. Based on the proposed model, not only we can minimize the expected waiting time of the customers in the queue but we can also improve the system utilization by efficient allocation of resources along with avoiding the over provisioning thus saving the cost as well.

V. CONCLUSION

Customer satisfaction and new customer attraction are greatly reliant on QoS for any platform. The expected waiting time of a customer mainly depends upon the average number of customers already waiting in the queue plus the number of internet help desk people busy in providing the service along with their average call handling time. Furthermore, QoS management relies on boundary conditions of parameters such as average and maximum delays in queue of customers as well as the dynamic and optimized allocation customer care representative particularly during peak hours. The proposed model is enriched to distinguish various kinds of call patterns in different day of week and is fully able to allocate resources (customer care representatives and hardware resources) according to the situation dynamically. Our philosophy negates the static allocation of customers because this is not an optimal approach. The dynamic differentiation and allocation of resources can minimize waiting time with fair allocation of resources in order to get an optimal Quality of Service at call center.

REFERENCES

[1] Mandelbaum and N. Shimkin. A model for rational abandonments from invisible queues. Queueing Systems, 36:141–173, 2000. 6.3.4, 7.3

[2] N. Shimkin and A. Madelbaum. Rational abandonment from tele-queues: nonlinear waiting costs with heterogeneous preferences. Working Paper, Technion, 2002.

[3] M. Armony and C. Maglaras. On customer contact centers with a call-back option: customer decisions, routing rules, and system design. Working Paper, Columbia University, 2001

[4] M. Armony and C. Maglaras. Contact centers with a call-back option and real-time delay Information. Working Paper, Columbia University, 2002.

[5] W. Whitt. How multi-server queues scale with growing congestion-dependent demand. Working Paper, AT&T Research, 2001

[6] Sandjai Bhulai and Ger Koole, " A Queueing Model for Call Blending in Call Centers ", 1434 IEEE transactions on automatic control, vol. 48, no. 8, august 2003.

[7] Thomas R. Robbins, D. J. Medeiros "DOES THE ERLANG C MODEL FIT IN REAL CALL CENTERS?" Simulation Conference (WSC), Proceedings of the 2010, IEEE

[8] Ahmad D. Ridley, Michael C. Fu "FLUID APPROXIMATIONS FOR A PRIORITY CALL CENTER WITH TIME VARYING ARRIVALS" Proceedings of the 2003 Winter Simulation Conference, IEEE

[9] Renxiang Zhu, Yijun Zhu " Performance Analysis of Call Centers Based on M/M/s/k+G Queue with Retrial, Feedback and Impatience", IEEE International Conference on Grey Systems and Intelligent Services, November 10-12, 2009, Nanjing, China

[10] Panichapat Chuchual, Narissa Chongpravatisakul, Teerasarn Kusolmanomai, Somrote Komolavanij "Inbound and Outbound Calls Assignment for an Efficient Call Center", 978-1-4244-6487-6/10/$26.00 ©2010 IEEE

[11] A. Mandelbaum and N. Shimkin. A model for rational abandonments from invisible queues. Queueing Systems, 36:141–173, 2000.

[12] N. Shimkin and A. Madelbaum. Rational abandonment from tele-queues: nonlinear waiting costs with heterogeneous preferences. Working Paper, Technion, 2002.

[13] E. Zohar, A. Mandelbaum, and N. Shimkin. Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. Management Science, 48:566-583, 2002.

[14] M. Armony and C. Maglaras. On customer contact centers with a call-back option: customer decisions, routing rules, and system design. Working Paper, Columbia University, 2001

[15] M. Armony and C. Maglaras. Contact centers with a call-back option and real-time delay

[16] W. Whitt. How multi-server queues scale with growing congestion-dependent demand. Working Paper, AT&T Research, 2001.