Evaluating Erlang C and Erlang A Models for Staff Optimization: A Case Study in an Airline Call Center

K. Nag[1], M. Helal[2]
[1]Department of Mechanical Engineering, American University of the Middle East, Kuwait
[2]Department of Industrial Engineering, American University of the Middle East, Kuwait
(Kaushik.Nag@aum.edu.kw, Magdy. Helal@aum.edu.kw)

Abstract - **Call centers can easily be modeled as queuing systems where the calls arrive, wait in a virtual queue and are serviced by operators. The simplest representation of a call center is the M/M/N queue or the Erlang C model which in this study is tested on an airline call center. A further extension of the Erlang C model is the M/M/N+M queue or the Erlang A model which takes call abandonment into consideration, has also been evaluated and compared with the Erlang C and the simulation results. With the current system having a high number of abandoned calls, low number of agents and high utilization level, Erlang C estimates are markedly different to that of the real system and is pessimistically biased by a great extent. Erlang A predictions, on the other hand, although optimistically biased is a bit more accurate. The optimum staffing requirements for each model has been predicted.**

Keywords – **Erlang-C, Erlang-A, Call Center, Queueing Models, Staffing, Waiting Time**

-

I. INTRODUCTION

Call centers has become a primary mode of communication with customers for a majority of companies, primarily in service sectors. It has been estimated that around 3% of the US and UK workforce is involved with call centers and the industry is estimated to have an annual growth of 20% [1]. Among the service sectors in Kuwait, call centers in airline industry support a large number of customers on a daily basis primarily due to a huge proportion of expats living in the region, in addition to people often commuting to international destinations due to the small geographic area of Kuwait. As a result, the Global Communication Center (GCC) of a major airline has been chosen for the call center case study. The GCC serves as a single source globally for handling calls related to queries and telephonic booking for the Airways. It ensures that inbound calls or enquiries from airways customers, potential customers, and travel agencies regarding sales, services and products are responded to and attended promptly.

Although technology provides the major backbone of call centers, often excess of 70% of the operating costs are devoted to human resources costs [2]. Therefore staffing level optimization is a major area of focus for call center's performance and forms a basis of the current study. Call centers are often modeled as queuing systems, where arrived calls are put in a virtual queue and are then serviced by agents. One of the commonly used models in call centers is the M/M/N queuing system or the Erlang C model which is essentially a multiple server queue where the incoming calls can be distributed to any of the free agents. The Erlang C model assumes that customers arrive at a queuing system having n servers with an infinite capacity, i.e. every customer will be served at some point of time. With increasing traffic load, the system may become unstable as the number of customers who are waiting keeps increasing over time and then would follow FIFO (first in first out) service discipline. Erlang C model has been widely used to evaluate various system performances, primarily the proportion of callers that must queue before being served (ProbWait), Average speed to answer (ASA) and the Sevice Level, which is the fraction of calls serviced with a delay below a specified limit. One major limitation of the Erlang C model is the lack of monitoring the Abandonment Rate, which is the proportion of calls that abandon the queue before being served. Abandoned rates cannot be estimated directly from the model as it assumes no abandonment occurs. The Erlang C model also makes some other questionable assumptions which have proved to be problematic. The model assumes that the call arrival rate is known and follows a Poisson process. Some authors have questioned the arrival rate certainty suggesting that the arrival rate should be modeled as a stochastic process and further, a time-inhomogenous Poisson process fits better with the call arrival data [2]. Some studies have also addressed the staffing requirements when arrival rates are uncertain [3]. The model also assumes that service time is exponentially distributed which has found to be a poor fit and possibly a lognormal distribution would be a better fit [2,4].

The M/M/N+M model or the Palm's Erlang A model is found to be more appropriate in call centers since it can be applied to evaluate performance measures on the basis of delay probability, average waiting time, abandonment ratio, and service level [5]. The model assumes that each caller possess an exponentially distributed patience time (with mean θ^{-1}), beyond which the caller will abandon the call. When compared to Erlang C, Erlang A provides a queuing system which is much more stable since it takes into consideration a customer exiting or abandoning the queue when his/her waiting time has exceeded the patience time. Although complex in nature, the modeling and calculation for the Erlang A performance metrics has been put forward by some authors [5,6] along with approximations based on an asymptotic analysis of the queue. Due to the accuracy of the empirical fit of the Erlang A model, it is considered as a useful tool for the

978-1-5386-0948-4/17/$31.00 ©2017 IEEE

call center staff and management to use in the improvement of operational performance. Recently the basic queuing models are being researched in more details. Some authors have suggested that there is a need to retain the impatient customers, thereby proposing a Markovian queuing model with balking and reneging. The proposed model increases the probability of retaining impatient customers and the effect of keeping reneging customers on the overall profit has also been analyzed economically [7].

II. METHODOLOGY

A. System Description

The Airways' Global Communication Center consists of multiple cubical workstations with call operators working on ***three shifts***, each shift is of ***six hours*** duration. A single shift consists of ***5 operators***. The flow of calls is attended by an Interactive Voice Response Unit where the callers can select options and provide data input to the call center system where after each customer call, the system checks whether there is an agent available to take the call. If all the agents are busy, the customer would have to wait in the queue until an agent is ready to serve him. The agent use a database (AVIVA) where they input data related to the customer and receive required data related to booking and flight information for the customer as an output (fig.1). After the required service the agent ends the call to serve the next customer waiting in the queue.

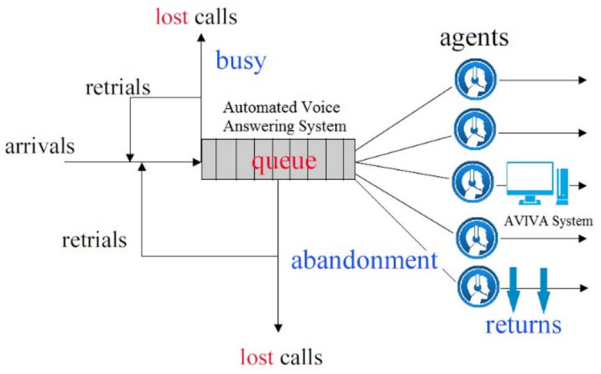

Fig.1. Schematic Representation of the Airline Call Center

Fig.2 shows the average number of calls for each month for the year 2016. The peak periods has been observed to be in the months of is January, June, July and December with an average of 33000 calls in each month. As a result, a particular shift of 6 hours for 1 day in the month of December 2016 was selected for the analysis.

The data was restricted to only ***one shift*** because of the unavailability of data in softcopies and compilation of data from hard copies usually takes a large amount of time. From the raw data of the entire shift (2-8 pm), a period of 30 minutes ***(1800 sec)*** interval was considered.

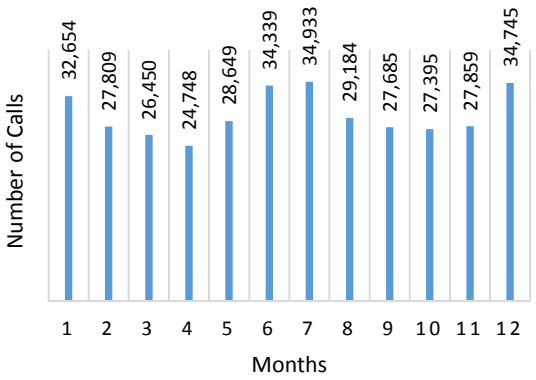

Fig. 2. Monthly Average Calls for the Call Center

The compiled data for the shift some of which have been utilized for our analysis are presented below:

Average Speed of Answer (in seconds): Average waiting time for the customer before being served by an agent. The average waiting time for an interval is calculated to be 112.16 sec.

No. of ACD Calls: Total number of calls that the operators handled for the period. The average value for an interval is calculated to be ***25*** calls.

Average Call Duration or ACD Time (in seconds): The average service time per call. The average value for an interval is calculated to be ***181.833*** sec**.**

Aban Calls: Number of abandoned calls for the interval. The average value for an interval is calculated to be ***23*** calls.

Avg Aban time: the average time that the customers wait in the queue before abandoning the call. The average value for the interval is calculated to be 73.75 seconds.

Max Delay (in seconds): This represents the maximum time recorded for an abandoned call before the call was abandoned. The average value for an interval is calculated to be 334.33 seconds.

It is to be noted that the total number of calls for each interval can be calculated by adding the ACD calls and Aban calls for each interval. The average value for an interval is calculated to be ***48***.

B. Erlang C

Erlang C calculations are described below taking into consideration that call arrivals follow a Poisson process and are serviced with a fixed number of statistically identical agents (5 operators). It assumes that the caller will wait indefinitely until he gets served, and there is no abandonment of calls. Call shall enter an infinite length queue and are serviced on a First Come First Served (FCFS) basis. The target answer time for service level as suggested by the company is 30 seconds. Service time (T_s) follows an exponential distribution. All calculations have been performed in Excel and verified with an online available Erlang C calculator. It will allow us to calculate the minimum number of agents needed to satisfy a service level, based on the waiting times.

Call Arrival Rate: Arriving calls follows a Poisson distribution with a defined arrival rate.

$$\lambda = \frac{Number\ of\ calls}{The\ chosen\ period} = \frac{48}{1800} = .0267\ calls/second. \quad (1)$$

Traffic Intensity: Traffic intensity can be calculated as

$$\mu = \lambda * T_s = 0.0267 * 181.83 = 4.8488; \quad (2)$$

where λ is the arrival rate of calls per second, and T_s is the average call duration or the service time.

Average agent occupancy: The agent occupancy or utilization can be calculated as

$$p = \frac{\mu}{m} = \frac{4.8488}{5} = 0.9698\ or\ 96.98\ \%, \quad (3)$$

where μ is the traffic intensity and m is the no. of agents.

Probability of Waiting: The probability that the call is not answered immediately and has to wait.

$$\text{Prob (call has to wait)} = E_c(m, \mu) = \frac{\frac{\mu^m}{m!}}{\frac{\mu^m}{m!} + (1-p)\sum_{k=0}^{m-1}\frac{\mu^k}{k!}} \quad (4)$$

Average speed of answer (ASA): This is the estimated average waiting time of call and is calculated as

$$T_w = \frac{Ec(m,\mu) * Ts}{m*(1-p)} \quad (5)$$

Service level: The expected ratio of customers that waited less or equal to the acceptable waiting time (tt).

$$\text{Prob(waiting time} \le tt) = 1 - Ec(m, \mu) * e^{-(m,-\mu)*\frac{t}{Ts}} \quad (6)$$

C. Erlang A

The Erlang A queuing model assumes a poison arrival process, exponential service time and exponential patience time with a mean θ^{-1}. Estimating individual abandonment rate θ is not an easy task because of the fact that direct observations are censored, and the only thing that can be measured is the patience of the customers who abandoned the system before the service began. On the other hand, customers who are being served, their waiting time in the queue is considered as a lower level for their patience. According to Mandelbaum & Zeltyn, [6], many statistical tools for un-censoring data have been developed with the help of Survival Analysis. All these tools need call by call statistics which in our case is out of reach. However, exponentially distributed patience time allows us to use a very simple un–censoring method to calculate the average patience time, as below [2].

$$\hat{\theta} = \frac{probability\ of\ abadonment}{average\ waiting\ time} = \frac{0.5}{112.166} = 4.4576 * 10^{-3} \quad (7)$$

$$Average\ patience\ time = \frac{1}{4.4576*10^{-3}} = 224.322\ sec \quad (8)$$

Erlang A can be used to determine the probability to wait and the average speed of answer. Firstly, calculating the performance metrics requires the evaluation of the incomplete gamma function

$$\gamma(x, y) \triangleq \int_0^y t^{x-1}e^{-t}dt, x > 0, y \ge 0 \quad (9)$$

The calculation for performance metrics for Erlang A model is provided by Mandelbaum & Zeltyn [6], the basic building blocks have been defined as

$$J = \frac{e^{\frac{\lambda}{\theta}}}{\theta}\left(\frac{\theta}{\lambda}\right)^{\frac{\mu\eta}{\theta}} * \gamma\left(\frac{\mu\eta}{\theta}, \frac{\lambda}{\theta}\right) \quad (10)$$

$$\text{and } \varepsilon = \frac{\sum_{j=0}^{n-1}\frac{1}{j!}\left(\frac{\lambda}{\mu}\right)^j}{\frac{1}{(n-1)!}\left(\frac{\lambda}{\mu}\right)^{n-j}} \quad (11)$$

The probability of waiting can be calculated as

$$P[wait > 0] = \frac{\lambda J}{\varepsilon + \lambda J}(1 - \theta) \quad (12)$$

The calculations for the above Erlang A model have been carried out with an online available calculator adopting a method outlined by Garnett, Mandelbaum and Reimann [5]. This queuing system is always stable because of customer abandonments. The performance measure calculated in the program are service level, delay probability, abandonment ratio and the average waiting time. The service level is the measure of the expected ratio of customers that waited less or equal to the acceptable waiting time. The authors suggested different ways to account the abandoned customers that waited the most time in the computation of the Service level (SL). Both exact (used for this case study) and approximate solution based on diffusion equations have been suggested. In the exact formulas, the waiting queue has limited capacity, whereas the capacity is unlimited in the approximate formula.

D. Simulation

A simple straightforward simulation model was created in Arena to compare the real system with the Erlang C and Erlang A estimates. When a call enters the trunk line, it will wait in the queue; if one of the trunk lines is free the call is served by the operator. In our system, 49% of the calls were abandoned calls. Since the call arrivals follow a Poisson process, interarrival times are independent and identically exponentially distributed. Number of replications was 4 and t-test was carried out in Minitab for validation.

III. RESULTS

A. Results of Erlang C

Table 1 presents the results of the Erlang C calculations as per the equations from 1 to 6. The inbound call activity can be modeled using very few variables like number of agents, number of calls, average call duration and the period length. The queue capacity was set to be an unlimited one. The results suggested that with an acceptable waiting time threshold of the service level of 30 seconds, the system will perform poorly with a service level of below 10%. The average speed of answer is very high with a delay probability of 92.75%. Table II and

Fig.3 suggests that increasing the number of operators from 5 to 7 will achieve the target of average waiting time of below 30 seconds with a service level of 80% and a delay probability of less than 30%. If the desired service level is beyond 90%, the number of operators should be increased to 8 which will also substantially reduce the average speed of answer and the probability of delay.

TABLE I
ERLANG C PERFORMANCE METRICS

Number of calls:	48
Period length (p): seconds	1800
Average call duration (t): seconds	181.8333
Number of agents (m):	5
Arrival rate (λ):	0.0267
Traffic intensity (μ):	4.8488
Agent occupancy:	96.98%
Probability to wait: Erlang C (m,μ)	92.75%
Avg speed of answer, sec: ASA (m,μ,t)	1162.69
Target answer time (tt): seconds	30
Service level: Erlang C (m,μ,t,tt)	9.80%

TABLE II
OPTIMUM SERVER REQUIREMENT BY ERLANG C

No. of Agents	Service Level (%)	Delay (%)	Avg speed of answer (secs)
4	0.00	100.00	infinity
5	9.80	92.75	1162.69
6	56.96	53.65	85.19
7	80.78	29.05	24.63
8	91.98	14.68	8.49
9	96.89	6.91	3.03
10	98.88	3.02	1.07

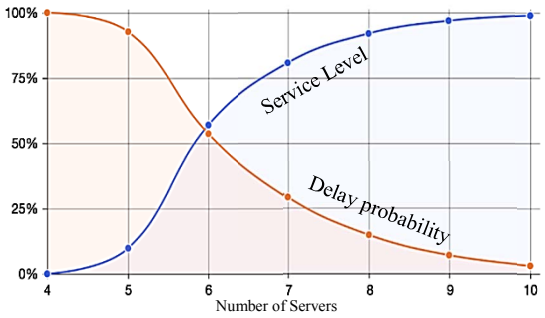

Fig.3. Relationship between no. of agents, delay probability and service level in Erlang C model

Fig.4. Relationship between no. of agents and average waiting time in Erlang C model

B. Results of Erlang A

Table III presents the results of the Erlang A calculations taking into account of call abandonment. The results suggested that with an acceptable waiting time threshold of the service level of 30 seconds, the system will perform far better when compared to that of Erlang C results with a service level of around 63%. The average speed of answer is around 35 sec which is close to the target value of 30 seconds, with a delay probability of 55%.

TABLE III
ERLANG A PERFORMANCE METRICS

Number of calls:	48
Period length (p), seconds:	1800
Average call duration (t), seconds:	181.8333
Number of agents (m):	5
Arrival rate (λ):	0.0267
Trunk line	150
Average patience time, sec, $\frac{1}{\theta}$	224.322
Probability to wait:	55.49%
Average speed of answer, seconds:	34.89
Target answer time (tt), seconds:	30

TABLE IV
OPTIMUM SERVER REQUIREMENT BY ERLANG A

No. of Agents	Service Level (%)	Delay (%)	Abandon-ment (%)	ASA (secs)
4	42.48	74.04	26.53	59.50
5	63.25	55.49	15.55	34.89
6	79.31	37.20	8.37	18.78
7	89.60	22.39	4.15	9.31
8	95.28	12.18	1.90	4.26
9	98.05	6.03	0.81	1.81
10	99.26	2.73	0.32	0.72

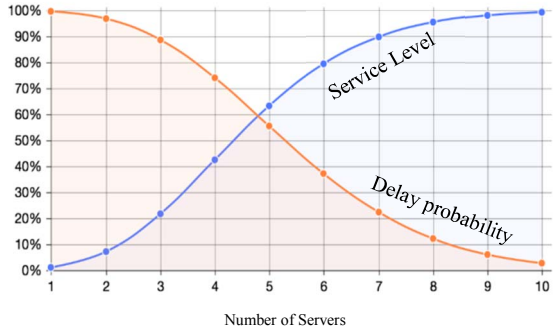

Fig.5. Relationship between no. of agents, delay probability and service level in Erlang A model

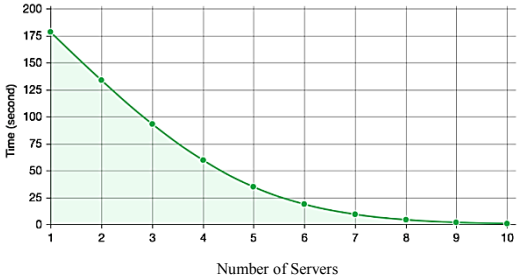

Fig.6. Relationship between no. of agents and average waiting time in Erlang A model

`Fig.7. Relationship between no. of agents & abandonment ratio

Table IV and Fig. 5 suggests that increasing the number of operators from 5 to 6 will reduce the average speed of answer to only 18.5 sec with a service level of around 80% and a delay probability of less than 37%. The call abandonment has found to be reduced from 15% to 8%.

C. Results of Simulation

In the simulation model, the total number of calls that entered the system was 574 for the entire shift; 308 of those calls were served by the 5 operators and the remaining 266 of the calls were abandoned. The average waiting time in queue was found to be 1.76 minutes (105.6 seconds) with a maximum waiting time of 10.4 minutes. The utilization level of the agents was found to be 83% with an average service time of 2.7 min (162 sec).

IV. DISCUSSION

Both Erlang C and Erlang A results showed marked dissimilarity with the simulation results. The Erlang C model is subjected to a reasonably large error over the range of parameter values. The Erlang C has found to be highly pessimistically biased as the real system has performed much better than the predicted value. This high level of measurement error between Erlang C and the simulated value can be attributed to the high level of abandonment in the real system. Further, with low number of agents and high utilization level, the measurement errors can be even greater. This is in line with some previous studies where it was observed that the Erlang C model is most accurate when the number of agents is large and agent utilization is low.

Erlang A measurement errors compared with the real system has found to be relatively small. The predicted error in the average speed of answer and probability to wait is relatively small when compared with the real system. However, the model is still optimistically biased which means that the predicted measures are better than the measures in the real system. For example the average speed of answer with 5 agents was found to be around 35 sec compared to the simulation result of 105 sec. The utilization level was also lower with 63% compared to the simulation value of 83%. Increasing the number of agents from 5 to 6; both the simulated model and the Erlang A model shows a reduction of the average speed of answer below the acceptable limit for the company (18.78 sec for Erlang A, 31 sec for simulation). Abandoned rate was also found to be below 10%. It is to be noted that *abandonment rate* θ using a single value for all intervals, as is the case in this study, would ignore possible variations in customers' patience time over the time of day [2]. The optimum staffing requirement of 6 agents for the call center as per the more accurate Erlang A model has been identified which may influence several factors like waiting time, agent utilization level and others.

V. CONCLUSION

The paper presented a comparative analysis of Erlang C and Erlang A models in a simple inbound airlines call center set up. Erlang A model was found to be more appropriate than the Erlang C model in this situation considering that the abandoned call rate is relatively high in the current scenario. Erlang C, although a relatively simple model, with low number of agents and high utilization level has shown a very high level of measurement error when compared to that of the simulation results

ACKNOWLEDGMENT

The authors would like to acknowledge the contribution of S. AlOtaibi, A. Ahmad, A. AlRamadhan & E. AlKhabbaz; graduating students of the Industrial Engineering department at AUM, Kuwait.

REFERENCES

[1] G. Koole & A. Mandelbaum, "Queueing Models of Call Centers: An Introduction", *Annals of Operations Research*, vol. 113, no. 1, pp 41-59, 2002.

[2] L. Brown, N. Gans, A Mandelbaum & A. Sakov et al., "Statistical Analysis of a Telephone Call Center: A Queuing- Science Perspective", *Journal of the American Statistical Association*, vol 100, no 469, pp 36-50, 2005.

[3] A. Bassamboo, J. M. Harrison & A. Zeevi, "Design and Control of a Large Call Center: Asymptotic Analysis of an LP-based method", *Operations Research*, vol 54, no 3, pp 419-435, 2006.

[4] N. Gans, G. Koole & A. Mandelbaum, "Telephone Call Centers: Tutorial, Review and Research Prospects", *Manufacturing and Service Operations Management*, vol. 5, no. 2, pp 79-141, 2003.

[5] O. Garnett, A. Mandelbaum & M. Reiman, "Designing a Call Center with Impatient Customers", *Manufacturing and Service Operations Management*, vol. 4, no. 3, pp 208-227, 2002.

[6] M. Avishai & S. Zeltyn, "Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers", *Advances in Services Innovations*, Part 1, pp 17-45, Springer Berlin Heidelberg, ISBN 978-3-540-29858-8, 2007.

[7] R. Kumar & S. K . Sharma, "An M/M/c/N queuing system with reneging and retention of reneged customers", *International Journal of Operational Research*, vol 17, no 3, pp 333-344, January 2013.