Telephone Traffic Queues in a Customer Call Center

A Thesis

Presented for the

Master of Science Degree

The University of Tennessee at Chattanooga

Patrick Todd

April 2009

Abstract

The purpose of this paper is to evaluate the unemployment claims filing call center operated by the Tennessee Department of Labor. To do this we primarily utilize traditional Erlang models to analyze performance measure such as call blocking wait times and labor utilization. We find that some modifications to staffing levels at both down times and peak times would improve the aforementioned performance measures. Some limitations to this study are the limited availability of data, thus some assumptions were made. The data used is also form year 2008, so it is difficult to predict staffing levels necessary in the future, though methods are utilized to achieve this task. However, 2008 was year in which the call center experienced both consistently slow to moderate traffic loads in the first part of the year and very heavy traffic loads in the latter portion. As a result, 2008 is a good year in which to highlight the challenges managers face in adjusting capacity to meet swift fluctuations in demand.

Table of Contents

Introduction……………………………………………………..…………………1

Literature Review……………………………………………..………………...2

Case Study…………………………………………………...……………….5

Methodology and Analysis..…………………………………..…………..….6

Conclusion……………………………………………………………………18

References…………………………………………………………………….20

Tables and Figures

Table 1: Call Volume Based on Monthly Unemployment Rate…………..……...9

Figure 1: Hourly Call Volume………………………………………….....….10

Table 2: Optimized Capacity Jan-May/Jun-Dec…………..……..………......14

Table 3: 2009 Demand Forecast…………………..…………………………16

Introduction

Nearly everything waits. Whether it is people in a line waiting for service, or items at a production plant waiting for assembly, waiting is a fact of life. An accumulation of entities in waiting is termed a queue. Queues can be described in terms of their physical structure. How many different paths to service are there? How many services? How much service does a customer require before he or she determines the interaction complete? (Dickson et. al., 2005). In order to determine how much service should be available, queuing theory attempts to answer, through mathematical analysis, the questions: How long does a customer have to wait for service? How many people will form in the line (queue)?

With the advance of technology, service organization are increasingly finding more innovative ways to offer service more efficiently and more cost effectively. One of the fastest growing is the telephone call center. Call Centers have experienced consistent proliferation throughout the world (Dean, 2002). In the U.S., observers of the industry estimate that there are over 100,000 call centers with over 3 million customer service agents (Green et. al., 2003). Moreover, it is estimated that 3% of adult workers in the U.S. work in call centers (Jack et. al., 2006).

Literature Review

Jack et. al. (2006) categorize study of the service quality of call centers can be classified in 3 broad research areas:

1) The inputs (human and industrial psychology perspective)

2) The delivery process (operations management perspective)

3) The performance outcomes (service marketing perspective)

It is the second of these components on which this paper will focus. This area is concerned with the efficient use of labor, capacity, and delivery processes. Research in this area has yielded a number of models to reduce customer wait time and increase throughput and customer satisfaction. Furthermore, research has led to the development of scheduling techniques and optimization models that allow call centers to operate with greater efficiency (Jack et al, 2006).

Managing capacity at a call center can be a particularly difficult task. A call center's capacity is simply the total number of phone lines. Capacity decisions would be easy if the number of callers over a given period and service time of each caller remained constant, but this is never the case. Call centers often experience seasonal fluctuations in demand on a daily, monthly, or yearly basis and short term peaks in demand (Betts et. al., 2002). In order to achieve a high level of service, call available capacity must be able to efficiently meet fluctuating demand levels (Jack et. al., 2006). Betts et. al. (2002) identifies several strategies in the extant literature for medium term capacity management:

1) A level capacity approach that maintains a constant resource allocation.

2) A chase approach, which adjusts capacity to meet demand

3) Queuing systems can also be used to give an operation a chance to respond to demand.

Furthermore, Betts et. al. (2002) identify some limitations with these strategies. For instance, where a large difference between peak and steady demand levels exists, a level strategy may result in poor resource utilization or poor service levels. Queuing systems can also be limited in situations where demand is highly transitory.

In order to properly analyze any queuing system, six basic characteristics of queuing processes must be examined.

1) Arrival pattern of customers

2) Service pattern of servers

3) Queue discipline

4) System capacity

5) Number of service channels

6) Number of service stages

The arrival pattern of customers is usually measured as the number of arrivals per a certain amount of time or the average time between successive arrivals. However, this is only completely valid for situations in which there is no uncertainty in the arrival pattern. If there is uncertainty, then the use of a probability distribution associated with this random process is necessary. Much the same is true of the service pattern of servers. However, unlike the arrival pattern, the service pattern assumes a customer is in the system to be serviced. If not, then the service facility is idle. An arrival pattern does not necessitate an occupied system. Queue discipline simply describes how customers are selected for service. In most cases a first in, first out (FIFO) approach is used. This is also the method used in the case study. Often a limited amount of waiting room is available in a queuing system. When this amount is reached, no more customers can enter the system. At this point, the system capacity has been reached. The number of service channels represents the total number of service station that can provide service simultaneously. The service stages describe the different places in a queuing system where service is received. It is often the case that only one stage of service exists such as customers waiting in a barber shop; however multiple stages can exist in scenarios such as a medical exam where patients may receive an x-ray at one stage and move to another stage to receive a blood test. Before any mathematical analysis is performed, it is important to describe the process intended for analysis. Knowledge of these six criteria is critical to that undertaking (Gross & Harris, 1974).

Case Study

In 2006, the Tennessee Department of Labor and Workforce Development (TDLWD) implemented a call center system for its unemployment claims filing service. Since its inception, the call center experienced problems stemming from a high volume of claimants calling in. At times, the system is bombarded with more callers than the system's capacity will allow. Furthermore, long waiting times are reported by many callers. The system also experiences cyclical patterns of increased volume with volume increasing on a monthly basis, particularly in January and July, when more claims are usually filed. Higher volume can also be expected at times when unemployment rises considerably. It is the purpose of this paper to analyze the current status of the claims center to determine where its major flaws are and to offer solutions to improve operating efficiency and service quality thereby increasing customer satisfaction. In particular the call center would likely be better served by implementing a more variable staffing strategy; that is, changing staffing levels throughout the day to meet varying levels of demand. However, this can be difficult to accomplish due to certain human factors such as the preferred time schedule of employees, providing employees periodic breaks, and scheduling off duty days and on duty days in a way most amenable to employees.

Methodology and Analysis

As noted, it is important to determine exactly what type of queuing system exists in order to determine the proper model to use to examine its efficiency. Numerous models exist for different types of systems. Queuing systems, for example, can be those wherein there are multiple stages of service or systems in which a line forms at each service channel among others. The system analyzed in this paper contains multiple service channels, but only one stage of service. The system serves a singular line in which customers are services on a First In First Out (FIFO) basis. That is, when a service channel becomes vacant (a customer leaves), then next customer in line proceeds to be serviced by that channel. The models to be used to analyze this system are Erlang models. These models will allow the user to see how this system is operating in terms of waiting times and blockages.

Before using any such models, the terms need to be defined which not already defined of the six outlined above. As noted above, the queue discipline is FIFO and only one stage of service. We will begin with a system capacity of 120 with 100 service channels. This simply means there are 100 actual human agents to serve customers with a total waiting capacity of 20. These are reasonable estimates of the actual figures in the current system. However, the accuracy of these numbers at this point is not of paramount importance. Through the analysis these numbers will be subject to manipulation to achieve more optimal values for performance metrics such as wait times and labor utilization rates. The arrival pattern of customers in the Erlang models we intend to use is assumed to be a Poisson process and the service time of customers to the system is assumed to be distributed exponentially. The arrival rate and service rate will be defined by the following variables:

Arrival rate = λ

Service rate = μ

Our system is thus best noted (Kendall notation) as:

M / M / k / k; where

Arrival process / service distribution / number of servers / waiting room

M – Markovian, the arrival process is Poisson and exponential is out service distribution.

The fist Erlang model to be used is the Erlang B model or Erlang loss function. This model is defined below:

$$E_k(A) = \frac{A^k/k!}{\sum_{n=0}^{k} A^n/n!}$$

This formula yields the probability that all positions in the system are occupied, all servers are busy, and all waiting positions are filled. A customer who arrives to this situation will not be allowed to enter the system. That customer is blocked, thus the Erlang B formula provides the blocking probability. The variable A in this formula is the Erlang unit and is evaluated thusly:

$A = 100(\lambda/\mu)$

In our case, μ will be constant at 200 callers/hour. That is, 30 minutes per call (2 per hour) for 100 servers. The arrival rate will vary based on time of day and month, as will be shown later.

It should be noted that the Erlang B formula assumes there is no waiting room in the system, which would seem to pose a problem in this situation with a waiting room of 20. In order to use this model we can treat the waiting room in our system as servers. In other words, there is a total system capacity of 120. If all 120 spaces are occupied, regardless of the nature of the space, then a caller will be blocked.

Specific to this scenario, due to the unavailability of data, some assumptions had to be made in order to do an analysis of the current state of operation in the call center. It is known through the news media that management reports a 23% increase in average monthly call volume from 19,729 calls in 2007 to 24,238 calls in 2008. From the 2008 the total annual volume will be 290,856 (24,238 x 12). From this the state's monthly unemployment rates, which are widely available in the public domain, can be used to extrapolate an expected level of volume for each month. A high correlation between unemployment rate and call volume is expected. The figures in the Table 1 below are more than consistent with that expectation yielding a perfect correlation ($R^2 = 1$) between the unemployment rate and call volume.

From here it is necessary to reduce the data from a monthly scale to a daily scale. The number of workdays each month can be known simply from excluding weekends and holidays from the total days in each month. Then the total call volume each month can be divided by the number of workdays. It is also known that the call center operates on a 7.5 hours workday (8 A.M. to 4:30 P.M). It can also be reasonably assumed that there is some variation in call volume from hour to hour. While data is not directly available from management, we can logically assume a volume pattern where volume is least during early hours, increasing to peak at some point during the middle before decreasing in the later hours. This data, along with the unemployment and call data, is displayed in the Table 1.

Table 1: Call Volume Based on Monthly Unemployment Rate

Month	Unemployment Rate	% of yearly claims	Calls/month	Workdays	Calls/Day
Jan	5.4	7.1%	20531	21	978
Feb	5.8	7.6%	22052	20	1103
March	5.8	7.6%	22052	20	1103
April	5.1	6.7%	19390	22	882
May	5.9	7.7%	22432	21	1069
June	6.8	8.9%	25854	21	1232
July	7	9.2%	26614	22	1210
Aug	6.6	8.6%	25093	20	1255
Sep	6.9	9.0%	26234	21	1250
Oct	6.7	8.8%	25474	23	1108
Nov	6.9	9.0%	26234	17	1544
Dec	7.6	9.9%	28895	20	1445
	76.5		290856	20	1534

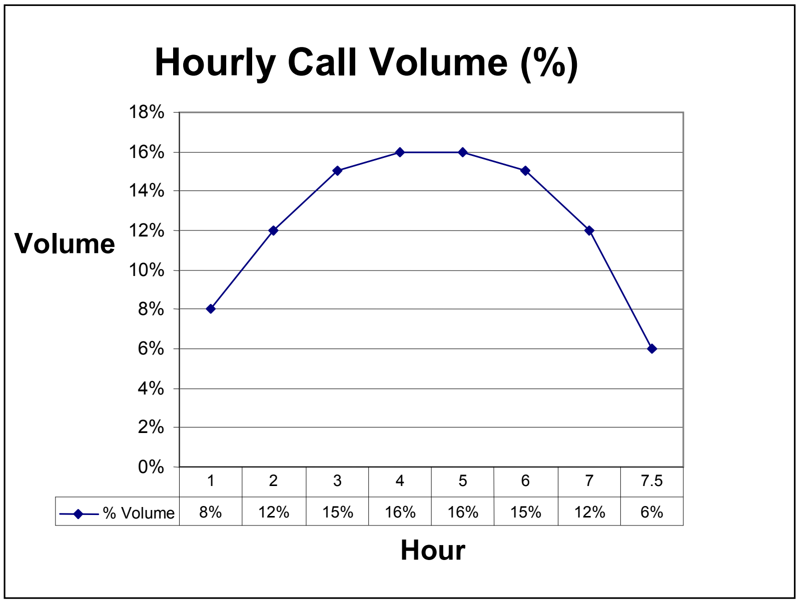

Figure 1: Hourly Call Volume

Based on the data in Table 1, it is possible to use the Erlang B model to determine in general the blocking probabilities for each hour of every month in 2008. For example, look at the results for the busiest hours (4 and 5) for November.

$$A = 100(247/200) = 124$$

$$E_{120}(124) = \frac{124^{120}/120!}{\sum_{n=0}^{120} 124^{n}/n!} = .089$$

For the busy hours in November, the blocking probability is basically 9%. That almost 1 in 10 callers attempting to enter the system is blocked, is a major problem. This was the most severe blocking rate experienced the entire year. Other times that experienced significant blockages (>1%) are as follows:

Hours 3 and 6	Hours 4 and 5
November = 5%	November = 8.9%
December = 2.2%	December = 5%

The most obvious solution to this problem is to hire more agents to take calls during these busy times. However, it’s better to have a discussion of possible solutions after other operating metrics for this scenario are analyzed.

Now turn to another Erlang model, Erlang C. This model, M / M / k , will determine the probability that a customer entering the system will have to wait for service.

$$C_k(A) = \frac{A^k/ k! \ (k/(k-A))}{\sum_{n=0}^{k-1} (A^n/n!) + (A^k/k!(k/k-A))}$$

In this model, as the carried traffic in Erlangs (A) approaches the number of servers, k, the probability of waiting approaches one 1, or 100%. In other words when the traffic is equal to the number of servers, waiting is certain for all new arrivals. However, in cases where A ≥ k, this formula fails. This is illustrated when we compute the model for hours 4 and 5 in November.

$$C_k(124) = \frac{124^{100}/ 100! \ (100/(100-124))}{\sum_{n=0}^{k-1} (124^n/n!) + (124^{100}/100!(100/100-124))} = -10.48$$

In this case, the formula yields a value that cannot be interpreted in terms of probability. In such a case the system is considered unstable; the arrivals rates consistently exceed service rates. If the formula is evaluated for a situation in which $A < k$, such as hours 4 and 5 in June ($A = 98.5$), then a wait probability of about .82 is produced. Though the system is unstable in situations like that of hours 4 and 5 in November where $A \geq k$, it is clear from the computation of hours 4 and 5 in June that the waiting on service in November is a virtual guarantee. In fact, if the formula is computed for a situation in which $A = 99.9$, it generates a wait probability of almost 99%. Furthermore, in this situation it is not possible for the waiting line to extend indefinitely as the Erlang C formula assumes, since there is a limited waiting capacity. Calls that arrive after all waiting room is occupied will be blocked from entering the system.

However, the Erlang C calculations do allow the computation of waiting times using the following formula:

$W = C_k(A)(\mu) / (1-U)k$; where U is the utilization rate defined by A/k.

This formula also yields untenable results in cases where $A \geq k$, but again, situations where A is only slightly than k can provide information about cases where $A \geq k$. For instance in June hours 4 and 5, the formula yields a wait time, $W = 1042.8$ seconds, or over 17 minutes, certainly unacceptable. It is reasonable to assume that cases where $A \geq k$, will only be worse.

It is quite clear that this system does not function very efficiently with a total capacity of 120. During busy periods callers experience long waits, if they enter the system at all. Furthermore, during slower periods utilization rates are low indicating significant levels of idleness. It should be noted that the utilization rates used are computed differently from those in the waiting time formula above. Those computations pertain to the Erlang C model. A different utilization computation is used for an Erlang B. The rates do not vary significantly in this system, however since the system is more closely resembles that of an Erlang B model (M / M / k / k), this computation is preferred. The formula is as follows:

$$U = (1-E_k(A))A / k$$

The year can be divided up in two sections based on the significant increase in call volume occurring in June, January – May, and June – December. The utilization rates in January – May are often below 50% and only during the busy hour of February, March, and May does the utilization rate exceed 70%. Clearly, the capacity is much larger than necessary for these months. In order to find a more optimal capacity, the following constraints must be set:

$E_k(A) < 1\%$

$W < 3$ minutes

Under present condition blocking probabilities and wait times are virtually nonexistent such that there is plenty of room for those to increase without becoming problematic for overall system efficiency. Based on the above constraints, the following results were attained.

Table 2: Optimized Capacity Jan-May/Jun-Dec

Jan-May			Jun-Dec		
Hour	Lines	Workers	Hour	Lines	Workers
1	59	49	1	77	67
2	82	71	2	110	98
3	100	88	3	134	122
4	106	94	4	142	129
5	106	94	5	142	129
6	100	88	6	134	122
7	82	71	7	110	98
7.5	83	72	7.5	110	99

These results indicate significant differences in needed capacity on an hour to hour basis as well as a difference between the two segments of the year.

In the 2008 data definite trends were observed, particularly the steady rise in call volume later in the year. From this data it is important to begin planning capacity needs for future months in 2009. Due to its relatively easy use and the limited amount of data needed, exponential smoothing is a frequently used method of demand forecasting. There are some variations of this method applicable to different situations. For the purposes of this paper the trend-adjusted exponential smoothing method is most reasonable given the trendy nature of the data. This method is expressed by the following formulae:

$A_t = \acute{\alpha}D_t + (1-\acute{\alpha})(A_{t-1} + T_{t-1})$ => exponentially smoothed average of series in period t

$T_t = \beta(A_t - A_{t-1}) + (1-\beta)T_{t-1}$ => exponentially smoothed average of the trend in period t

$F_{t+1} = A_t + T_t$ => forecast for period t+1

$\acute{\alpha}$ = smoothing parameter for the average ($0<\acute{\alpha}<1$)

β = smoothing parameter for the trend ($0<\beta <1$)

D_t = demand for period t

To forecast January 2009, take the demand (call volume) from the final 3 months of 2008. Using $\acute{\alpha} = .8$ and $\beta = .5$ the data yields the following results:

$A_o = 26{,}868$ => simple 3-month average

$T_o = 1{,}711$ = 3-month trend

$A_t = 28{,}832$

$T_t = 1{,}838$

$F_t = 30{,}670$

Month	F_t
Jan	30670
Feb	32858
Mar	34853
Apr	36938
May	38982
Jun	41044
Jul	43098
Aug	45156
Sep	47212
Oct	49269
Nov	51325
Dec	53382

Table 3: 2009 Demand Forecast

A forecast for January of 30,670 constitutes an increase in demand of 1,775. This is a very reasonable forecast given the present trend. In fact, it is known that the unemployment rate increased to 8% in January, thus an increase in demand as forecasted is precisely what should be observed. It is possible to produce a forecast for subsequent months by assuming the forecast for January is the actual demand. Using the exact same parameters we compute a demand forecast of 32,858 for February. It is important to note that producing long term forecasts based on forecasted data (i.e. forecasting February by using the January forecast) is not recommended. Such forecasts will tend to always assume a consistent upward or downward trend depending on the trend present for the initial month's forecast. The demand forecast for all of 2009 can be computed to illustrate this. Clearly the December forecast is far greater than the January forecast. In this scenario this would only happen if unemployment continued to increase throughout the year. Though possible, it is a highly unlikely scenario.

Since demand in this scenario is directly related to the rate of unemployment, a more likely scenario would be short term increases in unemployment followed by stabilization and perhaps decline. However, to know with greater certainty one should conduct or refer to other economic forecasts. Such forecasts are beyond the scope of this study.

The best method, rather than using one month's forecast to produce another, is to take the actual demand figure for January, once it is known, and produce the February forecast. This allows one to track actual demand with forecast demand thus allowing managers to refine forecasting parameters to reflect better the current trends.

Conclusion

Many things could be further analyzed which are beyond the intended scope of this paper, such as how best to schedule employees so as to provide optimal capacity or whether it would be more feasible to constantly employ the minimum capacity necessary to handle busy hour traffic only changing on a monthly rather than hourly basis. Might it be necessary to implement a virtual queuing strategy as described by Dickson et al (2005) in which callers are given an opportunity to call back at a certain time and provided with a code that insures the call is handled promptly? Given the fluctuations in demand on this system are directly related to fluctuation in the economy (unemployment). It is very difficult, even for the best economists, to predict future economic whims and thus exceedingly difficult to plan an efficient process which depends upon those whims.

What is clear from the analysis conducted throughout the course of this paper is that this system as presently operating is in need of some changes that can improve operating efficiency. Certainly, as unemployment continues to increase, current times are much more reminiscent of those in June-December, particularly November and December, if not worse. Our forecast suggests an increase in demand at least in the short term, if not longer, thus the capacity used for 2008 will be inadequate for 2009.

While the precise data in some cases is not known, hence some assumptions were made, it is simply a matter of one using the methods of this paper with more precise data before we can know truly how the system is functioning.

References

Betts, A., Meadows, M. and Walley P. (2000), "Call Center Capacity Management", *International Journal of Service Industry Management* Vol. 11 No. 2, 2000, pp.185-196

Dickson, D., Ford, R.C. and Laval B. (2005), "Managing Real and Virtual Waits in Hospitality and Service organizations", *Cornell Hotel and restaurant Administration Quarterly* Vol. 46 No. 1, 2005, pp. 52-68

Dean, A.M. (2002), "Service Quality in Call Centers: Implications for Customer Loyalty", *Managing Service Quality* Vol. 12 No. 6, 2002, pp. 414-423

Green, L.V., Kolesar, P.J. and Soares, J. (2003) "An Improved Heuristic for Staffing Telephone Call Centers with Limited Operating Hours", *Production and Operations Management* Vol. 12 No. 1 2003, pp. 46-61

Gross, D. and Harris, C.M. (1974), *Fundamentals of Queuing Theory,* John Wiley and Sons, New York, NY

Jack, E.P., Bedics, T.A. and McCary, C.E. (2006) "Operational Challenges in the Call Center Industry: a Case Study and Resource-based Framework" *Managing Service Quality* Vol. 16 No. 5 2006, pp. 477-500

Krajewski, L.J, Malhotra, M.K. and Ritzman, L.P. (2007) *Operations Management Processes and Value Chains,* Prentice-Hall, Upper Saddle River, NJ.

Zukerman, M. (2008) *Introduction to Queuing Theory and Stochastic Teletraffic Models*, Available at: http://www.ee.unimelb.edu.au/staff/mzu/classnotes.pdf