International Review of Business Research Papers
Vol. 3 No. 3 August 2007 Pp.219-241

Performance Analysis and Staff Planning in a Telecommunications Contact Centre: A Queueing Theory Approach

Susila Munisamy* and Balambigai Balakrishnan**

A contact centre is a centralized office of a company that mainly handles incoming telephone calls from customers via telephone. The contact centre basically functions as a primary contact point between customers and their service providers. Contact centres are highly technology driven. However, surprisingly, most of the costs incurred in a contact centre are due to human resources. Customer service agents who handle the calls form most of the human resource component in a contact centre. An important goal of the contact centre is to provide a good level of customer service. A good customer service level will ensure customer satisfaction so that customers will return. The consequence of making customers wait too long may be lost profit from lost business opportunities. In this context, queuing models are important to determine the appropriate number of customer service agents that strike a balance between the two conflicting objectives of cost reduction and provision of good service. In this study, we build queuing models to evaluate the performance of a contact centre that belongs to one of Malaysia's leading telecommunication service provider and plan its staffing levels. The telephone calls coming in to the contact centre are treated as the customers in the queuing system, and the customer service agents are the servers. The telephone calls that arrived on a Monday during the peak period fitted the assumptions underlying the Erlang C or the M/M/s model in terms of arrival pattern and service time. The model was extended to M/M/s + M or the Erlang A model by including the patience variable. Both the models were then used to analyze and compare the contact centres' operating characteristics and to decide various numbers of staffs required to achieve different management objectives.

Field of Research: Management Science

1. Introduction

The contact centre (or, more generally, customer contact centre) is a centralized customer-service office that serves as foci of customer contact for business organizations[i]. Contact centres provide primarily tele-services, where they answer incoming calls from customers via telephone (Mandelbaum, Sakov and Zeltyn, 2001).

*Dr. Susila Munisamy, Department of Applied Statistics, University of Malaya , susila@um.edu.my
**Ms Balambigai Balakrishnan, Department of Applied Statistics, University of Malaya, balambigai@um.edu.my

Contact centres are established in a diverse range of businesses, in large, small and medium organizations. Mail order catalogue firms, utility companies, banks, departmental stores, insurance companies, airlines, emergency road service operators and many others get connected to their customers via the contact centre. Telecommunication service providers also uses the contact centre to answer customer enquiries regarding phone service billings, and the reporting of faulty services, ordering new features and benefits, changing customer profile (including address) and cancellation of services.

In a typical commercial contact centre, about 65 percent of costs are due to staffing, 25 percent of costs are for networking and communication, and the remaining 10 percent of costs are associated with maintenance and other overheads (Antipov and Meade, 2002). Thus, the major component of contact centre costs is staffing (Gans, Koole and Mandelbaum, 2003). In addition to having inaccurate number of customer service agents, the costs can also be incurred by the high contact centre agents' turnover rate. Poor contact centre management, increases the stress level and leads to drastic resignation of jobs among employees (Tuten and Neidermeyer, 2004). As a result, the contact centre faces increased expenses caused by ongoing training programmes to replace employees. The cost concern resulting from staffing levels can be tackled via the quantitative models which is generally analytical (Koole and Mandelbaum, 2002) and at times empirical (Whit, 2002). The stochastic model especially the queuing model has been the standard model used to investigate the quantitative aspect of the performance of a contact centre.

In this study, we build queuing models to evaluate the performance of a contact centre of a Malaysian telecommunication service provider and suggest appropriate staffing levels. In the past, staffing levels were based on a simple procedure, taking into account the previous week's call volume and pattern. This procedure may be inaccurate and therefore this study uses a scientific approach to analyze the performance of the call center and plan staffing levels. We investigate the performance in the context of the classical M/M/s model (also called the Erlang C) and the extended M/M/s + M model (also called the Erlang A) and compare the results. Another aspect of performance is service quality. In the contact centre industry, for example, a typical service quality goal is to ensure at least 80 percent of the customers do not wait more than 20 seconds in the telephone queue (Koole and Mandelbaum, 2002) i.e. known as the 80/20 rule. The analysis also looks at the performance of the contact centre and the staffing requirement when satisfying a combination of quality of service goals.

The layout of this paper is as follows. The next section provides information on the background of the contact centre of the telecommunication service provider in this study. This is followed by a literature review in Section 3 on the use of queuing theory in studies on contact centres. Section 4 explains the data collected and the methodology of the study, section 5 describes the two performance models used and their results, while Sections 6 compares the performance models. Section 7 concludes the paper and offers some recommendation.

2. The Contact Centre of the Telecommunication Service Provider

As mentioned afore, the contact centre in this study belongs to one of Malaysia's leading telecommunication service provider. The contact centre handles inbound and outbound calls, faxes, e-services and mail, and operates 7 days a week, 24 hours a day. The scope of the study is limited to only the incoming calls from the post paid customers via telephone and does not consider the role of the Interactive Voice Response (IVR). The incoming calls make up more than two–thirds of total calls handled at the contact centre. Figure 1 presents the flow of an incoming call through the system.

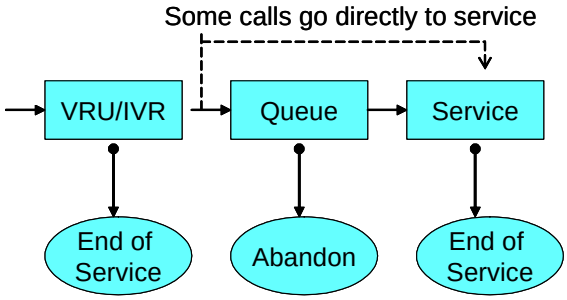


Fig. 1.Flow of an incoming call

An incoming call generally goes through three stages: Interactive Voice Response (IVR), queue and service, although some calls skip the queuing stage and go directly to the service stage. When a customer dials a given number they are instantly connected to an Interactive Voice Response (IVR) (also called Voice Response Unit or VRU). The IVR also enables customers to complete some self-service transactions (for example, for balance enquiries, customers may be told to "press five"). After completing the self-service transaction a customer may end the call at this point. If the customer/caller opts to speak to an agent or a Customer Service Representative (CSR), he presses the specified number on the telephone keypad and either gets connected immediately to an agent or joins the tele-queue waiting for an agent to become available. When a call exits from an IVR to join a queue it is recorded as calls arrived, entered or received. Impatient callers may not wait and leave the

queue abruptly. These calls are categorized as abandoned calls. The calls that reach the agent and receive the service are referred to as calls answered or served. Customers in a tele-queue are normally served on a first come first-served (FCFS) basis. However some callers classified as priority customers by the contact centre by-pass this system and get connected immediately to the next available agent. For example, a customer who spends more than RM 500 a month is classified as a priority customer at the contact centre under study.

3. Review of Literature

Waiting in line is almost an everyday practice. Customers wait to receive services if it is worth. However, making customers, employees, jobs or even telephone calls to wait very long in a queue can have serious consequences. Thus, performance of waiting line or queue[ii] is an important concern to many organizations. Queuing theory, which was conceived by A.K. Erlang in 1917, is the theory that deals with performance of waiting line (Erlang, 1911, 1917). This theory uses queuing models which consists of mathematical formulas and relationship to represent the various types of queuing system. The queuing models are used to study the performance indicators of a waiting line. Performance of waiting line is normally measured in terms of average number of customers waiting in queue and the system (which includes customers in the queue and being served), the average time spent in the queue and the system, the percentage of time the servers are busy (utilization rate) and others (Anderson, Sweeney and Williams, 2003; Reid and Sanders, 2004; Hillier and Hillier, 2004). The main objective of a queuing model is to determine how much service capacity should be provided to a queue to avoid excessive waiting, which in turn will contribute to economic gain (Reid and Sanders, 2004; Hillier and Hillier, 2004). In another word, in designing queuing models organizations aim to achieve a balance between offering service quality to customers (short queues requiring many servers) and economic considerations (not too many servers).

Queuing is a common feature in a wide range of fields covering from many daily-life situations to more technical environments such as computers networks (Kleinrock, 1976) and telecommunications systems (Ross, 1995). Much of the early work was motivated by practical problems concerning telephone traffic. At later stages, queuing has been extensively applied to real problems arising in manufacturing (Buzacott and Shanthikumar, 1993) public transportation and service operations (Hall, 1991). The queuing literature has grown looking for theoretic and algorithmic tools, and mathematical models of queuing phenomena. As a result of this, despite of the potential applicability of queuing theory, the gap between theoretical developments and real applications also has grown.

Due to the explosive growth, during the recent decade, in the number of companies that have a contact centre to provide customer service, research on

call centers has gained tremendous popularity. A bibliography of research in call centers can be found in Mandelbaum (2004) which covers over 200 research papers. There is vast literature on statistical inference and forecasting, but relatively few studies have investigated stochastic processes and even less attention were devoted to queuing models in call centers.

A basic queuing system consists of the customer population (people or objects to be processed) and the process or service system (Reid and Sanders, 2005). Customers arrive individually or in bulk expecting to receive some service. If the customer is not served immediately, he/she then will join a queue and wait for service to begin. Those customers who are served immediately will never join a queue. Each customer will be served by either one or more servers and upon completion of the service the customer will leave the system (Hillier and Hillier, 2004).

Contact centre operations fit the description of a queuing system very well. In a queuing system of a contact centre, the 'calls' are the customer population or the incoming unit, that is, the unit that enters into a situation in which a queue could form. The queue, in which the calls form a waiting line with an expression or expectation of I want to be served next and the service rendered (includes number of agents and the method how the call was answered) that allows the callers to leave the system makes up the other component of the queuing system (Murdoch, 1978). To sum, the nature of the customers or the incoming unit, the queue and the service facility influences the type of queuing system formed (Hillier and Hillier, 2004).

Queuing models are developed based on various parameters or components that hold the queuing system (Koole and Mandelbaum, 2002). Murdoch (1978) listed input distribution (which looks at the call arrival patterns), service distribution, number of service channels and the number in the system (whether it is a closed, N limit, infinite system) as four important building blocks of a queuing model. Unlike Murdoch (1978) who used 4 building blocks, Kendall (1951) devised standard notation consisting of six characteristics (i.e. 1/2/3/4/5/6) to describe many queuing models. The first notation looks at the arrival distribution, the second specifies the nature of service time, the third discusses the number of parallel servers, fourth describes the queue discipline, fifth component illustrates the maximum allowable number of customers in the system and the final component gives the total size of the customer population (Winston, 1994). Kendall's notation is used hereafter in this paper.

According to Reid and Sanders (2005) the arrival process, i.e. the first component, can be defined in terms of patterns of arrival (whether it is controllable or uncontrollable), the size of the arrival (single, batch or constant), distribution of the arrival (Poisson, Erlang or Degenerate) and the degree of patience among the arrivals. The traditional queuing theory assumes the arrival times to follow what is called a Poisson process (Hillier and Hillier, 2004; Brown

and Zhao, 2002; Anderson et al., 2003; Brown et al., 2002; Mandelbaum et al., 2001; Koole and Mandelbaum, 2002).
Number of servers and service time distribution makes up the second and third component of the Kendall's notation. As far as the service-time distribution is concerned, the assumption of the traditional model is that the service time follows the exponential distribution (Green and Kolesar, 1989; Brown and Zhao, 2002; Anderson et al., 2003; Brown et al., 2002; Mandelbaum et al., 2001). In addition to the two characteristics mentioned, the service facility structure (single phase or multiple phase of service) and capacity of the server (single or batch serving) are important components of a queuing system (Reid and Sanders 2005).

The queue discipline, which is the fourth component of a Kendall's notation, describes the order in which customers are served. The most common queue discipline is the First Come First Serve (FCFS), in which the customers are served in the order of their arrival. LCFS (Last come first serve), SIRO (service in random order), reservations first, emergencies first are some of the many other queue disciplines available (Reid and Sanders, 2005). The maximum allowable number of customers in the system and the size of the population in which the customers are drawn complete the Kendall's notation (Winston, 1994).

In short, Kendall's notation concludes most queuing models available in the industry. Change in one characteristics of the notation result in a new theoretical framework and application opportunity. It is important to note that different queuing models can be derived as the parameters vary. Thus, when applied in the contact centre industry, queuing models are able cater to different types of customer behavior. In addition to the above, there are other additional features of a queuing system that can be included in building a queuing model. Balking (leaving the system immediately), abandoning or reneging (leaving the system after waiting for some time), and in both cases deciding to call back (retrial) in order to access service (Yang and Templeton, 1987; Falin, 1995), skill-based routing are some of additional parameters of a queue that can be modeled based on queuing theory.

Some of the theoretical researches were based on just one selected characteristics and some attempted to combine a few. Hoffman and Harris (1986) incorporated abandonment and retrial in a model which is also motivated by the problem of estimating real arrivals. Artalejo (1995) considers a multi-server system with balking and retrials. Customers that find all servers busy are assumed to balk or quit the system with a probability that depends on the number of customers waiting in the queue. This model does not consider abandonment behavior. On the other hand, the systems modeled by Whitt (1999) combine balking and abandonment behavior.

However, one point to note is that there is no single analytical model that accommodates all, or even most characteristics of modern contact centers (Koole and Mandelbaum, 2002). Most queuing models are developed based on steady state assumption in which the system operates in a stabilized phase. Performance indicators such as average waiting time are not dependent on the time under this assumption. The probability that the system is in certain state is completely independent of time when steady state (stationary) assumption is applied (Winston, 1994; Anderson et al., 2003). It is important to note the steady state assumption is achieved only when the service rate (μ) is greater than the arrival rate (λ). If the arrival rate is greater than the maximum service rate, the system never reaches steady state and the queue length is continually increasing. At this stage, the server utilization rate will be more than 1, which indicates the inability of the system to operate at the given situation. In short, different parameters of a queuing model and assumptions commonly used in building models have been seen based on queuing theory. In the next section, different types of queuing models available are explored.

The simplest used model in analyzing contact centre performance is the classical M/M/s also known in contact centre circles as Erlang C (see for e.g. Mandelbaum, 2000, and Mandelbaum and Zeltyn, 2005). The Erlang C is a basic model that does not account for busy signals, customer's impatience or serviced spanned over multiple visits. Several studies tried to capture these features in the queuing models. Whitt (1999) analyzed the performance of a call center using two M/M/s queuing models with balking and reneging. Garnett et al. (2002) attempted to capture customer impatience, which is a common phenomenon in contact centres, to develop the M/M/s + M model (also called Erlang A). This study provided the theoretical framework for the M/M/s + M model. The single study that contributed empirically and analytically in the area of contact centre was based on a unique full call-detail record of 450,000 telephone calls of a small call center of an Israel bank over a twelve month period (Brown et al., 2002). This analysis was guided by queuing theory and was divided into two sections. In the first section, several statistical techniques were developed for the analysis of the basic components of a queuing model. In the second section, Erlang A model was used to analyze the data which was drawn from three different period of time in a day.

But the modern contact centre is often a much more complicated queuing network. Srinivasan, Talim and Wang (2004) incorporated the role of the Interactive Voice Response Unit (IVR) prior to joining the tele-queue, Garnet and Mandelbaum (2002) accounted for multiple teams of specialized customer service agents, and Armony and Maglaras (2001) investigated models for calls from multi-type customers.

To sum up, although contact center offers ample of opportunities for research, it is surprising very little is available especially based on analytical model and validated by real data. Availability of data could have been one of the major

reasons for the lack of practical studies involving queuing theory generally and contact centre specifically.

4. Methodology

Secondary data was collected for a period of two years (2003 and 2004) and primary data detailing call by call history was collected for a period of one week (from 2nd February 2005 till 7th February 2005). This data was used to explore the underlying patterns of incoming calls to the contact centre to observe the hourly, daily, and weekly patterns. It was found that there are more calls on a weekday and weekend and during the office hours. It was also found the peak hour lied between 10 am to 12 pm on most days. For the performance analysis, the data for a period of one hour, between 11 am to 12 pm on a Monday was selected to estimate the parameters of the queuing models used in this study.

The most important parameters in a queuing model are the service pattern and the arrival pattern. The data on service time, call arrival time and call completion time were collected to calculate the mean service time and the mean inter-arrival time. We also tested the distribution of the arrivals and service times and found the arrivals to approximate the Poisson distribution and the service time to conform to an exponential distribution.

The M/M/s model is a multiple server model that assumes Poisson arrival rate and exponential service time. The M/M/s model ignores the abandonment factor in the contact centers. When, the callers loose their patience while waiting on the queue, they simply hang up or abandon the call. A contact centre that has substantial percentage of abandoned calls in its daily operations, should take this component into consideration in building a queuing model (Zeltyn and Mandelbaum, 2005). Ignoring abandonment of calls can cause either understaffing or overstaffing.

Palm (1957) was the first to formulate a model that takes abandonment factor into consideration. Since then, many other researchers have given it attention. Garnet et al. (2002) extended the M/M/s using the patience variable to derive the Erlang A model or the M/M/s + M model. The last symbol M, in M/M/s + M represents callers' patience which is exponentially distributed. It was assumed that each arriving calls to the contact centre is associated with an exponentially distributed random variable that quantifies the individual's patience. These patience variables are assumed to be independent and identically distributed with mean θ^{-1}, and they are independent of all the other parameters of the model as well. The positive θ is identified as the abandonment rate (Garnett et al., 2002).

We use the above two queuing model to analyze the data drawn during a peak hour on a Monday and evaluate the call centers performance. The formulas for performance indicators for Erlang C and Erlang A models (Mandelbaum and

Zeltyn, 2003; Mandelbaum and Zeltyn, 2005; Mandelbaum and Zeltyn, 2005(b); Garnett et al., 2002(b)) are presented presented in Appendix 1. The performance models were implemented using Queuing ToolPak 4.0 (http://www.bus.ualberta.ca/aingolfsson/qtp) which is a Microsoft Excel add-in tool and the 4CallCentersv2.23 software (http://iew3.technion.ac.il/serveng/4CallCenters) developed by Ofer Garnett from Technion, Israel Institute of Technology.

5. Results of the Performance Models

a) M/M/s or Erlang C

First, we use the M/M/s or Erlang C queuing model to analyze the performance of the contact centre in this study. The data fitted the assumption of Poisson arrivals and exponentially distributed service time required by this model. The contact centre received a total of 954 calls in an hour which is at a rate of 15.9 calls per minute, had a service rate of 16.2 calls per hour (for every single server) and a total of 60 customer service agents called servers hereinafter. The queue capacity and the population were found to be infinite and the callers were treated on a first come first serve (FCFS) basis. The utilization factor, ρ, at the contact centre with 60 servers are used is 0.981. The utilization factor calculates the fraction of time each server is busy in a queuing system and must be below 1 to ensure the system is stable i.e. to prevent the queue from growing indefinitely. Queuing Toolpak 4.0 was used to measure the performance indicators of the contact centre based on the Erlang C model.

Since, an objective of this study is to equip the contact centre under study with right number of agents or servers; first, we attempt to look at the effect of the different number of servers on the performance indicators. The other performance indicators for a range of number of servers beginning from 59 to 85 are presented in Table 1. When the contact centre uses 59 servers, the utilization factor is found to be 1 which is an indication that the contact centre will be operating at full capacity. However, when the total servers used are below 59, many of the performance indicators were not available due to the fact the system becomes unstable and the queue grows indefinitely.

Firstly, we explore the relationship between number of servers and the utilization factor. A negative linear relationship is found between the utilization factor and the number of servers (as indicated by the formulae for utilization factor). As the number of servers increase, the utilization factor decreases steadily. This indicates holding the arrival rate and service rate constant, higher number of servers reduces the workload and relaxes the contact centre environment.

Table 1 :Performance indicators for ATSP with λ = 954 calls an hour and μ = 16.2 calls per hour and varying number of servers

Servers	**59**	**60**	**61**	**62**	**63**	**64**	**65**	**66**	**67**
Utilization factor (❒	1.00	0.98	0.97	0.95	0.93	0.92	0.91	0.89	0.88
Probability of no calls in the system (P_0)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average number of calls in the queue (L_q)	520.76	44.29	19.69	11.21	7.06	4.70	3.22	2.25	1.59
Average number of calls in the system (L)	579.65	103.18	78.58	70.10	65.95	63.58	62.11	61.14	60.48
Average time in the queue (W_q) (seconds)	1965.14	167.14	74.32	42.30	26.66	17.72	12.15	8.49	6.00
Average time in the system (W) (seconds)	2187.36	389.36	296.54	264.53	248.88	239.94	234.37	230.71	228.23
Probability that an arriving call has to wait to be served (P_w)	0.98	0.84	0.71	0.59	0.49	0.41	0.33	0.27	0.22
Percentage of calls answered within 20 seconds	3	24	42	55	66	74	81	86	89
Servers	**68**	**69**	**70**	**71**	**72**	**73**	**74**	**75**	**76**
Utilization factor (❒	0.87	0.85	0.84	0.83	0.82	0.81	0.80	0.79	0.77
Probability of no calls in the system (P_0)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average number of calls in the queue (L_q)	1.13	0.81	0.58	0.41	0.29	0.21	0.15	0.10	0.07
Average number of calls in the system (L)	60.02	59.70	59.47	59.30	59.18	59.10	59.04	58.99	58.96
Average time in the queue (W_q) (seconds)	4.27	3.05	2.18	1.55	1.11	0.78	0.55	0.39	0.27
Average time in the system (W) (seconds)	226.49	225.27	224.40	223.78	223.33	223.01	222.78	222.61	222.49
Probability that an arriving call has to wait to be served (P_w)	0.18	0.14	0.11	0.08	0.07	0.05	0.04	0.03	0.02
Percentage of calls answered within 20 seconds	92	94	96	97	98	99	99	99	100
Servers	**77**	**78**	**79**	**80**	**81**	**82**	**83**	**84**	**85**
Utilization factor (❒	0.76	0.75	0.75	0.74	0.73	0.72	0.71	0.70	0.69
Probability of no calls in the system (P_0)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Average number of calls in the queue (L_q)	0.05	0.03	0.02	0.02	0.01	0.01	0.00	0.00	0.00
Average number of calls in the system (L)	58.94	58.92	58.91	58.90	58.90	58.90	58.89	58.89	58.89
Average time in the queue (W_q) (seconds)	0.19	0.13	0.09	0.06	0.04	0.03	0.02	0.01	0.01
Average time in the system (W) (seconds)	222.41	222.35	222.31	222.28	222.26	222.25	222.24	222.23	222.23
Probability that an arriving call has to wait to be served (P_w)	0.02	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00
Percentage of calls answered within 20 seconds	100	100	100	100	100	100	100	100	100

Next, the relationship between the utilization factor and average time in queue or queue in seconds is investigated. Up until about 80 percent of utilization of the capacity, the waiting time for a call remains close to 0 seconds. As soon as it goes beyond 80 percent the amount of time each call spends in the queue grows rather rapidly. This can be tackled by increasing the number of servers or reducing the time spent on each call or by controlling the number of calls arriving at the contact centre.

The relationship between the number of servers and the probability of an arriving call has to wait to be served is found to be negative exponential in nature. When there are 80 servers, each call arriving at the contact centre has 0 probability of having to wait to be served, in other words, absolutely no calls has to wait when there are 81 servers waiting to serve the incoming calls. In a nutshell, the M/M/s model indicates, holding the arrival rate and service rate constant, different number of servers can leave different impact on the service provided. At the beginning, as the number of servers increase the performance indicators improved tremendously. However, after reaching the optimal point (in this case approximately at 65 servers), the impact of the additional servers reduces tremendously.

Service rate is influenced by the service time, which is the amount of time spent on each call and the server utilization rate. Service time tend to vary depending on the complexity of the call received. Thus, it is important to look at number of servers required for various service times. Using Erlang C or the M/M/s model, the number of servers required was calculated for various service time ranging from 120 seconds to 360 seconds. It was found, the relationship between the number of servers required and average service time is linear for all the three scenarios. However, there are more servers required for the same amount of service time when the agent utilization rate is lower. For example a call that takes 6 minutes to answer requires 136 servers for 70 percent agent utilization, 119 for 80 percent agent utilization and 106 servers for 90 percent agent utilization. The difference between 70 percent and 90 percent agent utilization is 30 servers per hour. This indicates the vulnerability of the contact centre and the tremendous effect of different number of servers on the smooth running of contact centre on an hourly basis.

As we have seen, the number of servers, service rate (and time) and arrival rate are very important parameters in the M/M/s model. With the current 60 servers, the contact centre is operating at its full capacity leaving a lot of room for improvement in the level of service provided to customers.

b) M/M/s + M or Erlang A model

Here, the performance of the contact centre is evaluated using the M/M/s + M or Erlang A model. This model includes the abandonment factor. The average caller's patience (ACP) is derived from the division of average wait in queue in seconds by the percentage of abandoned calls (Mandelbaum and Zeltyn, 2005). θ^{-1} is the symbol used to represent the ACP. The average wait in queue at the contact centre under study is 32.4 seconds and 21 percent of total calls received are abandoned during 11 am to 12 pm on Monday, 7th February 2005. Based on this it was found the ACP at the contact centre under study is 154.28 seconds or 2 minutes and 34 seconds and is assumed to be exponentially distributed.

The M/M/s + M model assumes exponentially distributed patience time in addition to Poisson arrival and exponential service time. The same values are used for parameters, for example, the arrival rate is 954 calls per hour, the average service time equals 3 minutes and 42 seconds, and the number of servers is 60. In addition, the average caller's patience is 2 minutes and 34 seconds. Since the Queuing Toolpak 4.0 did not have the abandonment feature, the 4callcenters version 2.23 (http://iew3.technion.ac.il/serveng/4CallCenters) was used to develop the performance indicators of the contact centre under study in this section.

What-if analysis was carried out to see the impact of changes to these input parameters on the performance indicators using the Erlang A model. Table 2 exhibits the performance indicators for servers ranging from 12 to 79. It was found that the relationship between average time in the queue in seconds for calls arriving at the contact centre and the percentage of abandoned calls is linear in nature. An increase in average waiting time increases the percentage of abandoned calls. Almost 80 percent of the calls will be abandoned when the server equals to 12 and the average time in the queue is 123 seconds. When the servers were reduced below 12, 4callcentre reported an extremely overloaded contact centre and the performance indicators were not made available. It is important to note that this could be due to violation of the assumption that the utilization rate is below 1.

Table 2:

Performance indicators for ATSP with λ = 954 calls an hour, μ = 16.2 calls per hour, θ^{-1} is 2.34 minutes for varying number of servers

Number of Agents	12	13	14	15	15	16	17	18	19	20	21	22	23	24
Utilization rate (□)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average number of calls in the queue (L_q)	32.49	31.79	31.10	30.40	30.40	29.71	29.02	28.32	27.63	26.94	26.24	25.55	24.86	24.16
Average time in the queue (W_q) (seconds)	122.59	119.97	117.35	114.74	114.74	112.12	109.50	106.88	104.26	101.65	99.03	96.41	93.79	91.17
Percentage Abandoned	0.80	0.78	0.76	0.75	0.75	0.73	0.71	0.69	0.68	0.66	0.64	0.63	0.61	0.59
% of calls answered within 20 seconds	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Number of Agents	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**
Utilization rate (□)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Average number of calls in the queue (L_q)	24.16	23.47	22.77	22.08	21.39	20.69	20.00	19.31	18.61	17.92	17.23	16.53	15.84	15.15
Average time in the queue (W_q) (seconds)	91.17	88.56	85.94	83.32	80.70	78.09	75.47	72.85	70.24	67.62	65.01	62.39	59.78	57.17
Percentage Abandoned	0.59	0.58	0.56	0.54	0.52	0.51	0.49	0.47	0.46	0.44	0.42	0.41	0.39	0.37
% of calls answered within 20 seconds	0.00	0.00	0.00	0.1	0.1	0.2	0.2	0.4	0.5	0.8	1.1	1.6	2.2	3
Number of Agents	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**	**46**	**47**	**48**	**49**	**50**	**51**
Utilization rate (□)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99	0.99	0.98
Average number of calls in the queue (L_q)	14.46	13.77	13.09	12.40	11.72	11.05	10.38	9.71	9.06	8.42	7.78	7.17	6.57	5.98
Average time in the queue (W_q) (seconds)	54.57	51.97	49.38	46.80	44.24	41.69	39.16	36.66	34.19	31.76	29.37	27.04	24.78	22.58
Percentage Abandoned	0.35	0.34	0.32	0.30	0.29	0.27	0.25	0.24	0.22	0.21	0.19	0.18	0.16	0.15
% of calls answered within 20 seconds	4	5.2	6.7	8.5	10.6	13.1	15.9	19	22.5	26.3	30.4	34.7	39.3	43.9
Number of Agents	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**	**61**	**62**	**63**	**64**	**65**
Utilization rate (□)	0.98	0.98	0.97	0.97	0.96	0.96	0.95	0.94	0.93	0.93	0.92	0.91	0.90	0.89
Average number of calls in the queue (L_q)	5.42	4.88	4.37	3.89	3.43	3.01	2.61	2.25	1.93	1.63	1.37	1.14	0.94	0.77
Average time in the queue (W_q) (seconds)	20.46	18.43	16.49	14.66	12.95	11.34	9.86	8.50	7.27	6.16	5.17	4.31	3.55	2.89
Percentage Abandoned	0.13	0.12	0.11	0.10	0.08	0.07	0.06	0.06	0.05	0.04	0.03	0.03	0.02	0.02
% of calls answered within 20 seconds	48.7	53.4	58.1	62.6	67	71.2	75.1	78.6	81.9	84.8	87.4	89.7	91.7	93.3
Number of Agents	**66**	**67**	**68**	**69**	**70**	**71**	**72**	**73**	**74**	**75**	**76**	**77**	**78**	**79**
Utilization rate (□)	0.88	0.87	0.86	0.85	0.84	0.82	0.81	0.80	0.79	0.78	0.77	0.76	0.75	0.74
Average number of calls in the queue (L_q)	0.62	0.50	0.39	0.31	0.24	0.18	0.14	0.10	0.08	0.06	0.04	0.03	0.02	0.01
Average time in the queue (W_q) (seconds)	2.34	1.87	1.48	1.16	0.90	0.68	0.52	0.39	0.29	0.21	0.15	0.11	0.08	0.05
Percentage Abandoned	0.02	0.01	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
% of calls answered within 20 seconds	94.7	95.9	96.8	97.6	98.2	98.6	99.0	99.3	99.5	99.6	99.7	99.8	99.9	99.9

The relationship between number of servers and number of abandoned calls are negative exponential in nature. The relationship between percentage of abandoned calls, percentage of answered calls, percentage of utilization and varying number of servers are also tested. As the number of servers increases, the percentage of abandoned calls decreases until it reaches 0 and it remains at 0 after 71 servers. The percentage of calls answered increases steadily and responds positively to the increase in the number of servers. However, at 71 servers and onwards the percentage of calls answered remains at 100 percent. Apparently the contact centre will be able to answer all calls the moment it starts using 71 servers.

6. Comparisons between M/M/s (Erlang C) and M/M/s + M model (Erlang A)

The M/M/s and M/M/s + M models are compared in Table 3. Table 3 provides a summary of the performance indicators of both the Erlang A and Erlang C. It is evident, taking the abandonment factor into consideration, the Erlang A,

produces shorter waiting time and shorter queue given the same parameters. The agent's also enjoy a lower utilization rate.

Table 3
Comparing performance indicators of Erlang C and Erlang A for 60 servers
λ = 15.9 calls per minute, μ = 16.2 calls per hour and θ^{-1} is 2.34 minutes

Performance indicators	Erlang C (without abandonment)	Erlang A (with abandonment)
Fraction abandoning	-	4.7%
Average time in the queue (seconds)	167.14	7.27
Average number of calls in queue	44.29	1.93
Agent's utilization	98 %	93%

Now, we turn to the performance of the contact centre and the staffing requirement when satisfying a combination of quality of service goals. As mentioned afore, one of the performance objectives of the contact centre under study is to answer 80 percent of the incoming calls within 20 seconds. Using Erlang C, it was found 64 servers were required to answer arriving calls within 20 seconds on average and 65 servers were required to answer at least 80 percent of the calls within 20 seconds. On the other hand, from Table 4, Erlang A only requires 53 servers to reduce the average waiting time to 20 seconds and below. However, the 80 percent target to answer calls within 20 seconds is only achieved when 60 servers are placed on duty. To achieve both the objective of 80/20 and 80 percent agent utilization rate at least 73 servers are required when Erlang A model is used compared to 74 in the Erlang C model.

Table 4
Number of servers required to achieve various objectives by Erlang C and Erlang A for 60 servers
λ = 15.9 calls per minute, μ = 16.2 calls per hour and θ^{-1} is 2.34 minutes

Objective	Erlang C	Erlang A	Difference
Time in queue less than 20 seconds	64	53	11
80/20 – at least 80 percent of the calls are answered within 20 seconds	65	60	5
80 percent utilization rate	73	74	1
80/20 rule and 80 percent utilization rate	74	73	1

In short, with the first two objectives (to answer incoming calls within 20 seconds and the 80/20 rule) it appears that Erlang C requires more servers than Erlang A. However, the difference in the number of servers required reduced tremendously for the last objective, which is to achieve the 80/20 rule with only 80 percent agent utilization. According to Garnett et al. (2002), the model that ignores the abandonment factor tends to overstaff and the model that includes the abandonment factor tend to under staff. Table 4 confirms his claim, whereby the Erlang C requires more servers than Erlang A for at least three objectives listed.

The 80 percent utilization only objective produces unique results. More servers are required to achieve this objective using Erlang A than Erlang C model. This is an interesting issue that should be focused on in the future research. The final objective, however, provides a very small difference in number of servers required. An important factor to note is, the agent utilization plays a major role in influencing the number of server required despite including the abandonment factor into consideration.

The percentage abandoned is only unique to the Erlang A as the Erlang C does not take the abandonment rate into consideration in performance modeling. Unlike the Erlang C (which becomes unstable when the utilization factor hits 100 percent and more), the Erlang A seems to be stable at all times (except when below 12 servers are used). In fact the system continues to work despite the 100 percent agent occupancy in Erlang A due the fact the abandonment factor is taken into consideration.

The relationship between the average wait in queue, average queue length and agent utilization for both the Erlang A and Erlang C model are compared for different number of servers (refer to Figure 1 to 3). For all three figures, note that Erlang C and Erlang A does not produce any result when below 60 servers and 12 servers are used respectively. This is because, at this stage, the arrival rate is greater than the service rate, thus the system continues to grow and never clears out.

The average time in queue and the length of the queue for Erlang C and Erlang A for various levels of servers (in Figure 1 and 3) exhibits a distance between the graphs. Calls in Erlang C model wait longer and join a longer queue compared to when the Erlang A model is used. However, the performance of both the models equalizes when approximately 70 servers are used. This is because there are enough servers under both the models to take all incoming calls with no delay, thus there is no abandonment and the Erlang A behaves like the Erlang C.

Figure 1
Comparison between Erlang A and Erlang C – average time in queue

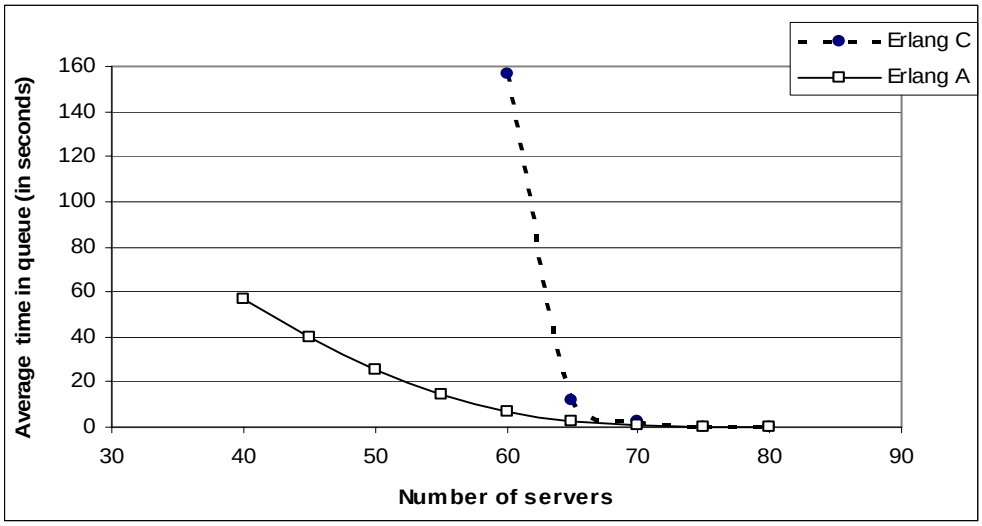
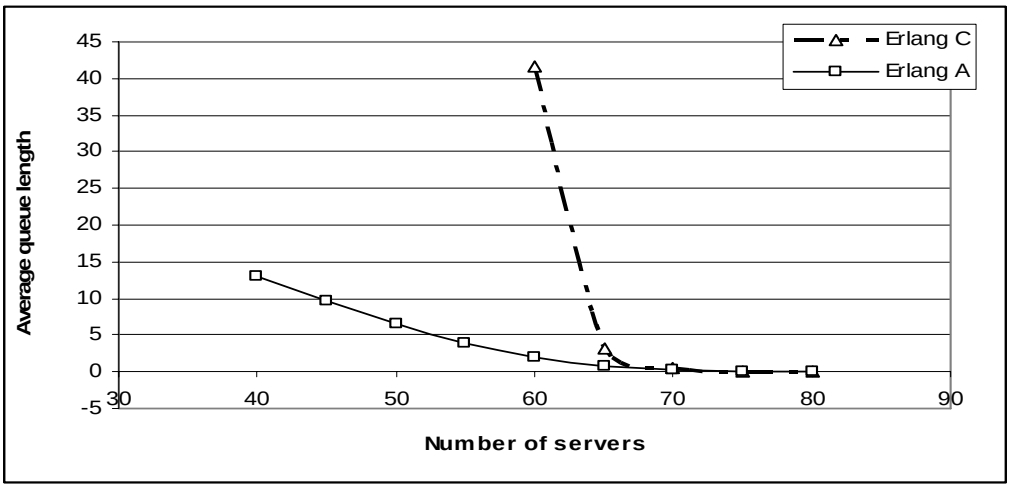
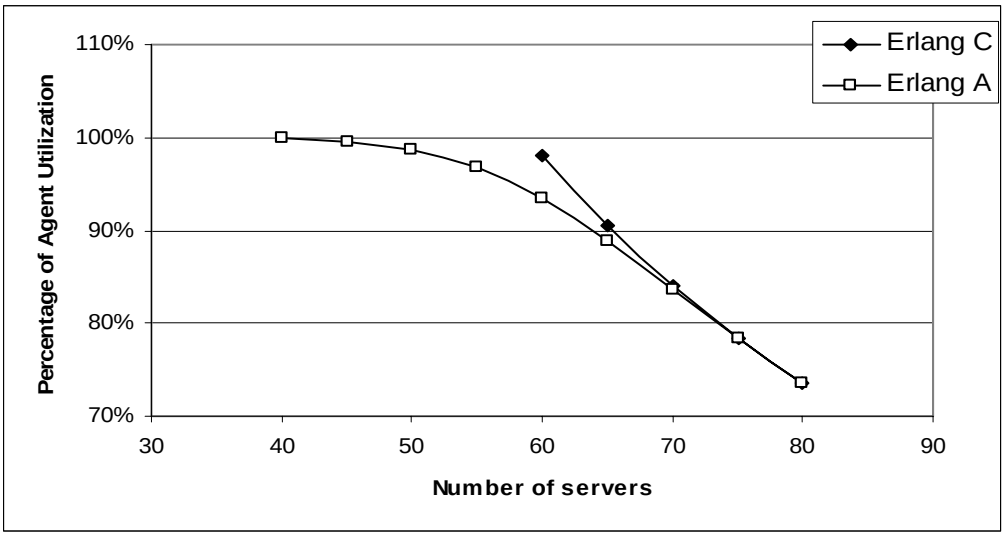


Figure 2
Comparison between Erlang A and Erlang C – average queue length



The percentage of agent utilization for Erlang C and Erlang A are compared below. As the number of servers increases the agent's occupancy rate decreases for both the models. However, the number of servers required to achieve certain level of agent occupancy differ among the two models. For example, to achieve 90 percent agent occupancy Erlang A requires approximately 65 servers and Erlang C 67 servers. In this example, adding 2 or 3 (from 65 to say 68) servers to M/M/S or Erlang C would result in M/M/S+A or Erlang A. Nonetheless, since personnel costs are the major operational costs in operating contact centers, even 1 to 2 percent changes in the costs leaves economically significant impact.

Figure 3
Comparison between Erlang A and Erlang C – Percentage of agent utilization



7. Conclusion and Recommendation

In this research we have attempted to build queuing models to evaluate the performance of the contact center of one of Malaysia's leading telecommunication service provider and help plan its staffing levels to improve service quality while keeping costs down.

The empirical data gathered from this contact centre between 11 am to 12 pm on Monday, 7th February 2005 was used for the purpose of building a performance models in the context of two models i.e. the Erlang C and the Erlang A. Applying the Erlang C model, the results showed that with the current number of 60 servers, the contact centre is not performing up to the standard required in the industry i.e. 80 percent of the calls to be answered in 20 seconds. With 60 servers, only 24 percent of the calls were answered within 20 seconds and the average number of calls in the queue was 44 and each call on average spent 2 minutes and 44 seconds waiting in the queue. What-if-

analysis was conducted with a range of servers and it was found that 65 servers were needed to improve the service to the required standard of 80/20.

It is important to note that the Erlang C model does not take the abandonment factor into consideration. This in turn, leads the organization to either under staff or over staff resulting in economical consequences. Thus, a second more advanced model which includes the abandonment factor, The M/M/s + M or the Erlang A was used to evaluate the performance of the contact centre. The additional parameter, namely the average caller's patience (θ^{-1}) was as an input to Erlang A model. With 60 servers, the contact centre performs better in Erlang A, where, the queue is shorter (1.9 calls), and the average time in queue (7.27) is also quicker and the agent utilization rate is lower compared to Erlang C. 60 servers are enough for the contact centre to provide a good level of service under the Erlang A model. The inter-relationship between the arrival rate, service time, and number of servers were explored extensively for both the models.

This study also compares the performance of the contact centre using Erlang C and Erlang A models. The results of the Erlang A are more accurate as it represents the contact centre more realistically. The performance indicators from Erlang A shows good performance as 60 servers were required to achieve the 80/20 service level. Erlang C indicates overstaffing of 5 servers to achieve the 80/20 industry target. As soon as the number of servers reached 71, there were no abandonment and the Erlang A behaved like Erlang C. It is important to note, that this results obtained using the peak hour data between 11 am and 12 pm on a Monday, thus conclusion based on this study are limited to peak hour on a Monday only.

In addition, the contact centre under study can attempt to increase the mean service rate which is currently at 16.2 customers per hour (for every single server). This can be done by making a creative design change with the application of technology. More questions can be included in the IVR to enable the servers and the callers to have a clear idea of the purpose of the call. This enables the servers to focus on the problem solving rather than spending time on understanding the problem. The contact centre also can attempt to reduce the number of arrival or more importantly distribute the call arrivals fairly throughout the day. Off-peak hour promotions can be carried out to encourage the callers to call after office hours. The important point to note, the cost of promotional efforts must be less than employing new customer service representatives and should be able to improve customer satisfaction. More customers can be encouraged to use the IVR, thus reducing the burden on the customer service representatives fully. Last but not least, currently, there are ample data available in the contact centre. These data needs to be managed to

realize its full benefits. Thus, it is important to employ a team of analyst, with contact centre and management science knowledge to reap the benefit from this data.

References

Anderson, D., Sweeney, D and Williams, T. 2003, An Introduction to Management Science Quantitative Approaches to Decision Making. Tenth Edition. Thomson, Australia.

Antipov, A and Meade, N. 2002. "Forecasting call frequency at a financial services call center". *The Journal of Operational Research Society,* Vol. 53 No. 9, pp. 953-960.

Armony, M and Maglarac, C. 2001. Customer Contact Centers with Multiple Service Channels, Working Paper.

Artalejo, J,R. 1995. "A queuing system with returning customers and waiting line". *Operations Research Letters 17: 191–19*

Brown, L. D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. 2002. Statistical Analysis of a Telephone Call Center: A Queuing Science Perspective. Technical report, University of Pennsylvania. Retrieved 9th December 2004. (Available online). *http://iew3.technion.ac.il/serveng/References/references.html*.

Brown, L and Zhao, L. 2002. "A New Test for the Poisson Distribution". *Sankhya, 64, 611–625.*

Buzacott, J.A. and Shanthikumar, J.G. 1993. Stochastic Models of Manufacturing Systems, Prentice Hall.

Erlang, A.K(1917. "Solutions of Some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges". Electroteknikeren, 13, 5–13 [in Danish]. *Nyt Tidsskrift Mat. B*, 20, 33–39.

Erlang, A.K. (1911). The Theory of Probability and Telephone Conversations.

Falin,G. (1995). Estimation of retrial rate in a retrial queue. *Queuing Systems 19:231–246*

Gans, N. Koole, G and Mandelbaum, A. 2003. "Telephone call centers: a tutorial and literature review". Invited review paper. *Manufacturing and Service Operations Management*. Vol. 5 No. 2, pp. 79 – 141.

Garnett, O and Mandelbaum, A. 2002. "An introduction to skills-based routing and its operational complexities". Technion, Israel. Retrieved January 9, 2005, (Available: Online) http://iew3.technion.ac.il/serveng.

Garnett, O., Mandelbaum, A and Reiman, M. 2002b, "Designing a Call-Center With Impatient Customers". *Manufacturing and Service Operations, Management*, Vol. 4, pp. 208–227.

Hall, R.W. (1991). Queuing Methods: For Services and Manufacturing, Prentice Hall.

Hillier, F and Hillier, M. (2004). Introduction to Management Science: A Modeling and Case Study Approach with spreadsheet. 2nd edition. McGraw-Hill.

Hoffman, K,L , Harris, C,M. 1986. "Estimation of a caller retrial rate for a telephone
information system". *European Journal of Operational Research 27: 207–214*

Queuing Toolpak, http://www.bus.ualberta.ca/aingolfsson/qtp, Retrieved October 9, 2005. (Available online).

Kendall, D,G. 1951. Some problems in the theory of queues. J. Royal. *Statist Society. Ser B, Vol 13, pp 151.*

Kleinrock, L.(1976). Queuing Systems: Computer Applications. Vol. 2, JohnWiley & Sons, New York.

Koole, G and Mandelbaum, A. 2002. "Queuing models of call centers, an introduction". *Annals of Operations Research,* Vol. 112, pp. 41–59.

Mandelbaum, A and Zeltyn, S. 2005. "Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers". Draft, March 2005. Retrieved September 27, 2005. (Available online). http://iew3.technion.ac.il/serveng/References/references.html.

Mandelbaum, A. Sakov, A and Zeltyn, S. 2001. "Empirical analysis of call center". Technical report. Retrieved July 9, 2005. (Available: Online). http://www.ie.technion.ac.il/serveng/course/096324

Mandelbaum, A. and Zeltyn, S. 2003, "The Impact of Customers Patience on Delay and Abandonment: Some Empirically-Driven Experiments With the M/M/N+G Queue," submitted to *OR Spectrum*, Special Issue on Call Centers. Retrieved: July 9, 2005. (Available online). http://iew3.technion.ac.il/serveng/ References/references.html.

Mandelbaum, A. 2004, "Call Centers (Centres) research Bibliography with Abstract". (Available: Online). http://iew3.technion.ac.il/serveng/References/ccbib.pdf

Mandelbaum, A and Zeltyn, S. 2005. "Service Engineering in Action: The Palm/Erlang-A Queue, with Applications to Call Centers". Draft, March 2005. Retrieved September 27, 2005. (Available online). http://iew3.technion.ac.il/serveng/References/references.html.

Mandelbaum, A and Zeltyn, S. 2005(b). "The M/M/n+G Queue: Summary of Performance Measures". Retrieved November 27, 2005. (Available online). http://iew3.technion.ac.il/serveng/References/references.html.

Murdoch, J. 1978. Queuing theory, worked examples and problems. Macmillan Press Limited. Wing King Tong. Hong Kong.

Palm, C. 1957. "Research on Telephone Traffic Carried by Full Availability Groups". *Tele, vol.1, 107 pp*. (English translation)

Ragsdale, C. 2001. Spreadsheet Modeling and Decision Analysis: A Practical Introduction to Management Science. 3rd Edition. South-Western College Publishing.

Reid, D and Sanders, R. 2004. Operations Management: An Integrated Approach. John Wiley and Sons, New York.

Ross, K. W. 1995. Multiservice Loss Models for Broadband Telecommunication Networks. Springer-Verlag, London.

Srinivasan, R. Talim, J and Wang, J. 2004. "Performance Analysis of a call center with interactive voice response units". *Top*. Vol. 12, No 1, pp. 91 – 110.

Tuten, L. and Neidermeyer, E. 2004. "Performance, satisfaction and turnover in call centers: The effects of stress and optimism". *Journal of Business Research.* Vol. 57 No. 1, pp. 26-34.

Winston, L,W. 1994. Operations Research: Application and Algorithms. 3rd Edition. Duxbury Press. California.

Yang, T and Templeton, J,G,C. 1987. A survey on retrial queues. *Queuing Systems 2: 201–233.*

Whitt, W. 2002. Stochastic-Processes Limits. Springer. New York.

Whitt, W. 1999. "Improving Service by Informing Customers Anticipated Delays". *Management Science.* Vol. 45 No. 2, pp. 192-207.

4callcenterv2.23. (2002). Retrieved 18 April 2005. (Available online) http://iew3.technion.ac.il/serveng/4callcenters/Downloads.html.

Appendix 1

Notations:

λ = the mean number of arrivals per time period (the mean arrival rate)
μ = the mean number of services per time period (the mean service rate)
s = number of servers
ρ = Utilization factor
P_0 = The probability that no units are in the system
L_q = The average number of units in the waiting line
L = The average number of units in the system
W_q = The average time a unit spends in the waiting line
W = The average time a unit spends in the system
P_w = The probability that an arriving unit has to wait for service
P_n = The probability of n units in the system
$P\left[Ab \middle| W > 0\right]$ = Probability to abandon of delayed customers

$E\left[W \middle| W > 0\right]$ = Average waiting time of delayed customers

M/M/S Model or Erlang C: Performance Indicators

$$\rho = \frac{\lambda}{s\mu}$$

$$P_0 = \frac{1}{\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\left(\frac{s\mu}{s\mu-\lambda}\right)}$$

$$L_q = \frac{(\lambda/\mu)^s \lambda\mu}{(s-1)!(s\mu-\lambda)^2} P_0$$

$$L = L_q + \frac{\lambda}{\mu}$$

$$W_q = \frac{L_q}{\lambda}$$

$P_n = \frac{(\lambda/\mu)^n}{n!} P_0$ for n ≤ s

$P_n = \frac{(\lambda/\mu)^n}{s!\, s^{(n-k)}} P_0$ for n > s

M/M/S + M Model or Erlang A: Performance Indicators

$$y = H(x) = \frac{1}{\theta}(1 - e^{-\theta x})$$

$$A(x, y) = \frac{ye^{xy}}{(xy)^{y}}\gamma(y, xy)$$

$$\gamma(x,y) \approx \int_0^y t^{x-1}e^{-t}dt \text{ , } x > 0, y > 0$$

$$\rho \approx \frac{\lambda}{s\mu} \qquad P\left[Ab|W > 0\right] = \frac{1}{\rho A(\frac{s\mu}{\theta}, \frac{\lambda}{\theta})} + 1 - \frac{1}{\rho}$$

$$E\left[W|W > 0\right] = \frac{1}{\theta}[\frac{1}{\rho A(\frac{s\mu}{\theta}, \frac{\lambda}{\theta})} + 1 - \frac{1}{p}]$$

[i] Initially set up to answer traditional phone calls, call centres have grown to handle other media such as faxes, email, e-services, and instant messaging, and are more generally known as contact centres. In this study both the terms are used interchangeably.

[ii] The management scientist refers to a waiting line as a queue (Hillier and Hillier, 2004). In this study we use both the term interchangeably.