CHAPTER ONE

INTRODUCTION

1.1 Background of the study

Vodafone call centre, also called Exceed Call Centre, was set up in 2005 for the erstwhile Ghana Telecommunications Company Limited (Vodafone Ghana is an operating entity of Vodafone, a part of the Global Vodafone Family) to serve as a means of communication with customers. The purpose was to take care of customer complaints, requests and also to serve as a medium for reaching out to the general public with product offerings. It is located at High Street, the heart of commercial activities in Accra. It has about 600 hundred customer service representatives (CSRs) who work in shifts. Our focus in this research would be the Vodafone Information Technology (IT) call centre called IT help desk. IThelpDesk is part of this group and are responsible for handling complaints of Vodafone internal IT customers. They are the first line of support for all IT related incidents. Their shift starts at 7am and closes at 7pm, Monday to Fridays and 8am to 4pm on Saturdays.

Most organizations with customer contact – private companies, as well as government and emergency services – have reengineered their infrastructure to include from one to many call centers, either internally managed or outsourced.

For many companies, such as airlines, hotels, retail banks, telecommunication, and credit card companies, call centers provide a primary link between customer and service provider.

At its core, a call center constitutes a set of resources – typically personnel, computers and telecommunication equipment – which enable the delivery of services via the telephone. The working environment of a large call center (Figure 1)

can be envisioned as an endless room, with numerous open-space cubicles, in which people with earphones sit in front of computer terminals, providing tele-services to phantom customers.

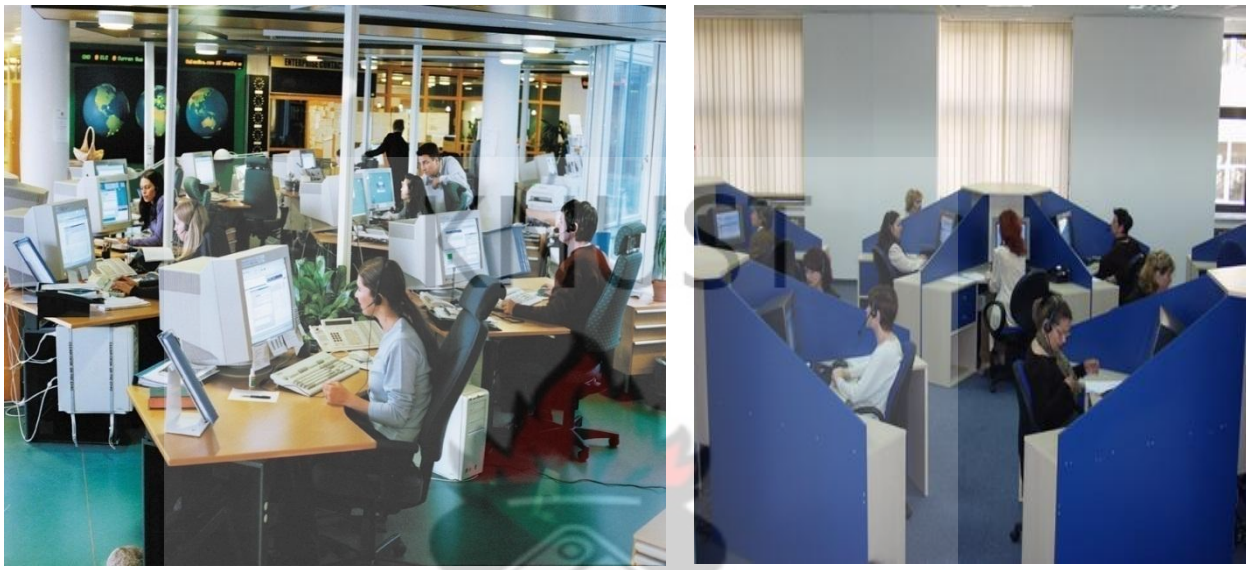

Figure 1: The Working Environment of a Call Center

The call center industry is vast and rapidly expanding, in terms of both workforce and economic scope. For example, a recent analyst's report, cited in Data monitor, and the U.S. Bureau of Labor Statistics, estimate the number of agents working in U.S. call centers to have been 1.55 million in 1999 - more than 1.4% of private-sector employment - and to be growing at a rate of more than 8% per year. In 1998, AT&T reported that on an average business day about 40% of the more than 260 million calls on its network were toll-free. One presumes that the great majority of these 104 million daily – "1-800" calls terminated at a telephone call center, as cited in Telephone Call Centers: Tutorial, Review, and Research Prospects (Gans, Koole, and Mandelbaum, 2003).

The quality and operational efficiency of these telephone services are also extraordinary. In a large, best-practice call center, many hundreds of agents can cater for many thousands of phone callers per hour; agent utilization levels can average between 90% to 95%; no customer encounters a busy signal and, in fact, about half of the customers are answered immediately; the waiting time of those delayed is measured in seconds, and the fraction that abandon while waiting varies from the negligible to a mere 1-2% (Gans, Koole, and Mandelbaum, 2003). Call centers, and service systems in general, give rise to many operational problems, one of which is the Staffing Problem: under an existing operational reality, finding the minimal-cost staffing levels that is required to meet some given Quality of Service (QoS) constraints. This problem has received a great deal of attention over the years, as it rightly deserves: staffing costs are estimated at about 70% of a call center's operational costs. Staffing "wisely" can thus result in substantial savings while achieving operational objectives (Gans, Koole, and Mandelbaum, 2003).

In the modern call center scene, customers can be distinguished by the type of service they require. For instance, customers might be associated with different priority levels (VIP vs. Members) or different functional requirements (technical support vs. billing). Naturally, call centers address this situation by employing various types of servers, with varying sets of skills.

When the skill-level required to handle calls is low, a center may cross-train every employee to handle every type of call, and calls may be handled on first come-first-served (FCFS) or first in first out (FIFO) basis. In settings that require more highly skilled work, each agent may be trained to handle only a subset of the types of calls that the center serves, and “skills-based routing” may be used to route calls to appropriate agents.

The continued growth in both the economic importance and complexity of call centers has prompted increasingly deep investigation of their operations. This is manifested by a growing body of academic work devoted to call centers, research ranging in discipline from Mathematics and Statistics, through Operations Research, Industrial Engineering, Information Technology and Human Resource Management, all the way to Psychology and Sociology (Mandelbaum, 2002).

1.2 Statement of the Problem

In a typical call centre having too many agents leads to unnecessary costs, whilst too few leads to substandard service, and since typically 60-70% of all operating costs of a call centre are personnel costs (Koole, 2002), minimizing the number of personnel required to achieve the desired service level is one of the most important problems in call centres and is the problem that we will concentrate on in this study. In order to draw up an agent schedule, forecasts must first be developed for the rate at which calls arrive into the call centre and the average handling time. The average handling time or service time is not only the time that an agent spends on the telephone, but is the total time that the agent takes to deal with a call. Based on these forecasts, the number of agents required to achieve the service level can be determined.

Another important problem is how to define the grade of service that a call centre achieves. Criteria can be qualitative as well as quantitative. In this thesis, the focus is on quantitative service levels, since these are normally associated with the issue that we will be most interested in: how to calculate the minimum number of agents that achieve a desired service level. Quantitative service levels normally measure the accessibility of agents. The quantitative service level that is seen most often is related to waiting times and is stated as ensure that more than α% of callers

wait less than T seconds to be served, where typical values are α = 80 and T = 20 (Koole, 2002). In this case, we can define the achieved grade of service as the percentage of customers who actually wait less than T seconds. Another example of a quantitative service level is based on abandonments and stated as ensuring that less than B% of callers abandon their calls prior to them being answered, where typical values are B = 3% or B= 5% (Koole, 2002). Satisfying the service level based on abandonments is highly correlated to satisfying the service level based on waiting times: as waiting times increase, more customers will abandon.

1.3 Objective of the Study

The aim of this thesis is to

- Modell queueing in the Vodafone call center using queuing theories.
- Develop a forecast for the rate at which calls arrived at the call centre and the average handling time, from data collected.
- Develop a simulation method which would be used in determining the minimum number of agents required to operate at a minimum staff and waiting cost.

1.4 Methodology

The data used for analysis was from call records at the IThelpdesk section of the Vodafone call centre. They were obtained from performance reports, observation, and by interviewing agents. The data has been analyzed with arrival and service rates determined. This was fed into the simulation module. Excel functionalities like VLookup, isText, Match, Rand were used in developing the simulation module. The simulations were carried out varying the number of agents from two to five. The

output from hundred simulation runs for each number are further analyzed. Findings and recommendations are then presented.

1.5 Justification

Staffing adequately the Vodafone Call Centre to provide quality service at a minimum cost is a challenging problem for call centre managers to solve.

The growth of Call centers in both economic importance and complexity has prompted increasingly deep investigation of their operations.

The level of quality and operational efficiency required of these telephone services are extraordinary, requiring sound scientific principles thus necessitating more research.

Current analytical models have performed important roles in their management, but they leave much to be desired. More sophisticated approaches are needed to accurately describe the reality of call-center operations, and improve call-center performance significantly.

Design, management and optimization of the performance of a call centre is only possible as a result of system modeling and deep analysis of data supporting the model.

1.6 Limitations

Time and inadequate data availability have constrained the collection of data from many call units for analysis. The research would be limited to calls at one unit of the call centre, IThelpdesk. Our work would be limited to single skilled server type and single customer type. It would be limited to the quantitative analysis of the call operations and not the human behavior factors.

1.7 Outline of Thesis

This thesis consists of five chapters; Chapter One consists of the background to the study, statement of the problem, objective of the study, methodology, justification for and limitations to study. In Chapter Two, existing models for Call centers are reviewed. Chapter Three discusses the mathematical modelling of the Call center management, simulation and analysis of the data. Results from the analysis of the data are presented in Chapter Four. The conclusions and recommendations are finally discussed in Chapter Five.

CHAPTER TWO

Literature Review

Introduction

The problem of staffing the call centre at a minimum staffing and waiting cost is the problem we want to solve. In this chapter we reviewed various works done on call centers in relation to the technology and the staffing problem.

According to Gans, Koole, and Mandelbaum (2003), Call centers can be categorized along many dimensions for example by the functions that they provide, the size and geographic dispersion, the organization of work or even by the type of traffic, whether it handles only inbound or only outbound traffic or both. They state that the emergence of large-scale call centers has been enabled by technological advances in information and communications systems. According to them the technology works as follows; A public service telephone network (PSTN) provider uses the automatic number identification(ANI) and dialed number identification service (DNIS) to connect callers with a call center, which usually has its own private automatic branch exchange (PABX or PBX). Calls may be connected through the PABX to an interactive voice response (IVR) or voice response units (VRUs) that queries customers on their needs and resolves them or hands them over to an automatic call distributor (ACD) for routing to appropriate CSRs, based on a set criteria within the call center. Computer-telephone integration (CTI) “middleware” aids the routing of calls and also integrates a special information system, customer relationship

management (CRM) system, into the call center's operations. A CRM system tracks customer records and allows them to be used for decision making.

Gans et al, (2003) observed that call center goals are formulated as the provision of service at a given quality level, subject to a specified budget. The common practice is that upper management decides on the desired service level and then call center managers are called on to defend their budget. Similarly, costs can be associated with service levels (for example, toll-free services pay out-of-pocket for their customers' waiting), and the goal is to minimize total costs. It occurs, however, that profit can be linked directly to each individual call, for example in sales/mail-order companies, a direct trade-off can be made between service level and costs so as to maximize overall profit (Borst, Mandelbaum, and Reiman, 2000).

Operational service level is typically quantified in terms of some congestion or performance measures. Experience suggests a focus on abandonment, waiting and/or retrials, which underscores the natural fit between queueing models and call centers (Koole, 2002 & Feinberg et al., 2000).

Gans et al., (2003) argue that- call centers can be viewed, as queueing systems and the simplest and most used performance model is the stationary M/M/s queue, also known in call center circles as Erlang C. To them, it describes a single-type single-skill call center with s agents, operating over a short enough time-period so that calls arrive at a constant rate, yet randomly (Poisson); staffing level and service rates are also taken constant. They state that the assumed stationarity could be problematic if the system does not relax fast enough, for example due to events such as an advertisement campaign or a new-product release. They further state that the model assumes out busy signals, abandonment, retrials and time-varying conditions and that the reason for using the M/M/s queue is the fact that there exist closed form

expressions for most of its performance measures. M/M/s predictions could turn out highly inaccurate because reality often "violates" its underlying assumptions, and these violations are not straightforward to model Gans et al., (2003). For example, non-exponential service times leads one to the M/G/s queue which, in stark contrast to M/M/s, is analytically intractable Gans et al., (2003). One must then resort to approximations, out of which it turns out that service time affects performance through its coefficient-of variation $C = E/\sigma$. According to Koole, (2002) performance deteriorates/improves as stochastic variability in service times increases/decreases.

According to Koole, 2002 when modeling call centers, the useful approximations are typically those in heavy traffic, namely high agents' utilization levels at peak hours. He stated that in considering the M/G/s queue, for small to moderate number of agents's, Kingman's classical result asserts that waiting time is approximately exponential, with mean as given above. Large s, on the other hand, gives rise to a different asymptotic behavior. According to him, this was first discovered by Halfin and Whitt (1981) for the M/M/s queue, and recently extended to the multiclass GI/PH/N queue in the Halfin-Whitt regime by Puhalskii & Reiman, (2000). According to Koole (2000), the two key challenges for call center management are agent staffing and economies of scale.

The square-root safety-staffing principle, introduced formally in Dimensioning large call centers, by Borst, et al., (2000) but having existed long before, recommends a number of servers s given by

$$s = R + \Delta = R + \beta\sqrt{R}, \quad -\infty < \beta < \infty,$$

where $R = \lambda/\mu$ is the offered load (λ =arrival rate, μ =service rate) and β represents service grade. The actual value of β depends on the particular model and

performance criterion used, but the form of s is extremely robust and accurate. As an example, for the M/M/s queue analyzed in Dimensioning large call centers, β could be taking a positive function of the ratio between hourly staffing and delay costs. Δ is called the safety staffing. It is shown that the square root principle is essentially asymptotically optimal for large heavily-loaded call centers ($\lambda \uparrow \infty$, $s \uparrow \infty$), and it prescribes operation in the rationalized (Halfin-Whitt) regime.

The square-root principle is applicable beyond M/M/s (Erlang C). In the working paper, Designing a call center with impatient customers, , Mandelbaum, and Reiman (2001) verify that for the M/M/s model with abandonment, β can take also negative values, since abandonment guarantee stability at all staffing levels; for time-varying models, as in server staffing to meet time-varying demand. According to Jennings et al, (1996), β varies with time; and it used for skill-based routing (Borst and Seri, 2000).

Puhalskii and Reiman, support the principle for the M/G/s queue, given service times that are square integrable. (Extensions to heavy-tailed service times would plausibly give rise to safety staffing with power of R other than half.)

In all the extensions of Robust algorithms for sharing agents with multiple skills only the form $s = R + \beta\sqrt{R}$, R was verified, theoretically or experimentally, but the determination of the exact value of β, based on economic considerations is still an important open research problem (Borst and Seri, 2000).

On operational regimes and economies of scale, call centres are observed to have operated either quality-driven regimes or efficiency-driven regimes or a blend of the two. For instance, at the peak period of 10:00-11:00, of a large U.S. mail-catalogue retailer, a number of 765 customers called; service time was about 3.75 minutes on average with an after-call-work of 30 seconds and auxiliary work to the order of 5%

of the time; ASA was about 1 second and only 1 call abandoned. But there were about 95 agents handling calls, resulting in about 65% utilization - clearly a quality-driven operation. At the other extreme there are efficiency-driven call centers: with a similar offered work as above, ASA could reach many minutes and agents are utilized very close to 100% of their time (Koole, 2000).

Within the quality-driven regime, almost all customers are served immediately upon calling. At the efficiency-driven regime, on the other hand, essentially all customers are delayed in queue(Koole, 2000). According to Borst and Seri, (2000), as explained in, Robust algorithms for sharing agents with multiple skills, and elaborated on momentarily, well-managed large call centers operate within a rationalized regime, where quality and efficiency are balanced in the face of scale economies. The rationalized regime was first identified in practice by Sze (1984). According to Koole, (2000), each caller within a call center occupies a trunk-line. When all the lines are occupied, a calling customer gets a busy signal. According to him, a manager could eliminate all delays by dimensioning the number of lines to be equal to the number of agents, in which case M/M/s/s, or Erlang-B ("B" for Blocking) becomes the "right" model. But then there would typically be ample busy-signals. Moreover, prevailing practice goes in fact the other way: it is to dimension ample lines so that a busy signal becomes a rare event. But then customers are forced into long delays. This is costly for the call center (thinking about the 1-800 costs) and possibly also for the customers – who might well prefer a busy-signal over an information-less delay, and hence they abandon the tele-queue before being served (Koole, 2000).

According to Koole (2000), the busy-signal vs. delay vs. abandonment tradeoff has not yet been formally and fully analyzed.

A simulation study of M/M/s/B presented in Performance characteristics of automated call distribution systems (Feinberg, 1990) where B stands for the overall number of lines ($B \geq s$); it is argued that only 10% lines in excess of agents provides good performance: more lines would give rise to too much waiting and fewer to too many busy signals. A more appropriate framework would be the M/M/s/B+G queue, where +G indicates arbitrarily distributed patience (following the notation and results of, "On queues with impatient customers", by Baccelli & Hebuterne, 1981).

According to Koole, 2000, an analytically tractable model is the M/M/s/B+M, in which patience is assumed exponential. Procedures for estimating the mean patience, as an input parameter to performance analysis, are cited in, "Designing a call center with impatient customers", by Garnett, Mandelbaum, & Reiman (2002), and Empirical analysis of a call center by Mandelbaum et al., (2000). According to them, mean patience could alternatively be used as a tuning parameter, where its value is determined to establish a fit between practice and theory.

Gans, et al. (2003), comment that in cases of abrupt changes in the arrival rate, or when the system is overloaded during one or more time intervals, the system can be far from stationary and this non-stationarity must be accounted for. Whilst uncertainty in the inter-arrival times is modeled explicitly by assuming a Poisson process, the arrival *rate* is assumed known. This is far from true and arrival rates which exceed forecasts by 10% are not unheard of (Koole, 2002). He notes that the inclusion of abandonments is particularly valuable in call centres where the load is high compared to the number of agents. He is also argued that if abandonments are assumed to be exponentially distributed with a constant average patience, γ then "estimating this parameter is a non-trivial statistical problem", but that the *Kaplan-Meier estimator* can be used.

Gans et al., (2000) recommend using a combination of analytical models and simulation; analytical models for “insight and calibration" and simulation for “fine tuning".

According to Jongbloed & Koole, (2000) the Erlang C method, assuming a homogeneous Poisson arrival process and exponential service times, the call volume data studied suggests that the number of arrivals in a time period is often overdispersed, meaning that the variance is larger than the mean (if the assumption of a homogeneous Poisson arrival process were true, then the variance would be equal to the mean). They respond to this by suggesting that the arrival rate can be modeled by a random variable, so that the arrival process is *doubly Poisson*. According them it makes sense; since there will always be variability in the actual arrival rate, no matter how accurate a forecast is and this extra variability needs to be taken into account. Then every data point can be viewed as being generated in two steps. Firstly, the arrival rate, λ, is drawn from the *mixing distribution* (the distribution of the random variable used to model the arrival rate). The arrival count then follows a Poisson process with that rate. If the mixing distribution is known, then a confidence interval for the arrival rate can be found, using this mixing distribution. The number of agents predicted by the Erlang C formula is increasing in the arrival rate; hence the upper and lower bounds from the confidence interval can be inserted into the Erlang C formula to give a confidence interval for the number of agents required.

They also tackle the subject of estimating the mixing distribution and both parametric and nonparametric methods are discussed, with the parametric method employing a Gamma distribution to model the arrival rate.

Two different methods of staffing are considered by Jongbloed and Koole, (2001). The first assumes that the number of agents cannot be altered during a time interval. The upper bound on the number of agents required then becomes important since a worst case scenario must be assumed in order to satisfy the service level as often as possible. However, the lower bound can be used to show how far of the upper bound can be, which is important as overstaffing translates into unnecessary costs. The second method is preferable; here it is assumed that the number of agents is flexible. The number of fixed agents can then be assigned according to the estimate given by the lower bound and a number of flexible agents according to the difference between the two, so that the number of agents answering telephones can be varied between the upper and lower bounds according to real time operating conditions. In a contact centre, flexible agents may be assigned tasks such as answering emails, which are not as urgent, so those agents can be used to answer telephone calls as necessary.

A simulation model for inbound call centres is developed by Robbins et al., (2006), which includes time varying and uncertain arrival rates as well as varying staffing levels. The authors show that different call centre staffing models are highly sensitive to uncertainties in arrival rates and that performance levels can differ significantly from the target levels, when the arrival rate varies from the forecast. The authors argue that, even though it is a common practice, models which assume a known arrival rate are suspect and far from robust.

According to Whitt, (2005), "the queueing model *M/GI/s/k+GI* has long been regarded as appropriate for call centres". He then shows that the *M/M/s/k+M*(*n*) model often provides an excellent approximation to the *M/GI/s/k+GI* queueing model. This, he says, is extremely helpful; whilst the latter is relatively intractable, the former is not. *M*(*n*) refers to the fact that abandonments are exponentially

distributed and the abandonment rate is *state dependent*, so the rate at which customers abandon is allowed to depend on their position in the queue. This behavior can certainly be imagined in call centres that provide customers with information about where they are in the queue, or give customers an expected waiting time. A customer who is told that she has to wait longer, or is further down the queue, is probably more likely to abandon the queue than if she is nearer to the front.

Complete call-by-call data over the duration of a year is examined by Brown et al, (2005). The data, obtained from a call centre belonging to a bank, included all calls from customers who wished to speak to an agent - about 450 000 in total. The analysis supports the assertion that the arrival process is an inhomogeneous Poisson process with additional randomness in its arrival rate, as suggested by Jongbloed & Koole, (2001). However, they find that, rather than being exponentially distributed service times tend to be lognormally distributed.

Time to abandonment was curiously found to have two peaks. The first occurred after a few seconds, whilst the second occurred after around 60 seconds. This corresponded to the customer being played a message informing them that they were in a queue, causing many to give up and hang up. Another interesting result was that the Erlang A model was found to describe the performance of this call centre well and predictions made using it "proved surprisingly robust". This is echoed by Whitt, (2006) who stated that the Erlang A model is certainly superior to Erlang C".

In heavy traffic, even a small fraction of busy-signals or abandonment could have a dramatic effect on performance, and hence must be accounted for. This demonstrated via the M/M/s+M model cited in, "Methods of judging the annoyance caused by congestion" by Palm, (1953), "On queues with impatient customers", by

Baccelli & Hebuterne, (1981) and "Designing a call center with impatient customers", by Garnett, Mandelbaum, & Reiman(2001), which add an abandonment feature to M/M/s (Erlang C): specifically, one models customers' patience as exponentially distributed, independently of everything else; customers abandon if their patience expires before they reach an agent. The M/M/s +M queue is referred to as Erlang A, "A" for Abandonment, and for the fact that this model interpolates between Erlang B and Erlang C.

A model for a call center with busy-signals should be M/M/s/B +M, to account for the existence of B lines. Performance analysis of the M/M/s/B +M queue has been implemented at www.4callcenters.com. In this example, there were sufficiently many lines so that the busy signal phenomenon was negligible, thus the use of Erlang A is recommended.

In their book, "Modelling Daily Arrivals to a Telephone Call Center"; Avramidis, Deslauriers, & L'Ecuyer, (2004) develop stochastic models of time-dependent arrivals with application to call centres specifically in mind. The focus is on reproducing the behavior that has been observed in recent empirical studies of call centre arrival data. Firstly, the total daily demand is over dispersed compared to the Poisson distribution as observed by Jongbloed & Koole, (2001). Secondly, there are large changes in the arrival rate as the time of day varies as shown by Tanir & Booth, (1999). Thirdly, arrival counts in different time periods are correlated, and finally, arrival counts on successive days are also correlated as shown in "Statistical Analysis of a Telephone Call Center": A Queueing-Science Perspective by Brown et al., (2005).

The authors develop three models of a time-dependent arrival process, two of which are similar to the doubly stochastic Poisson process suggested by Jongbloed &

Koole, (2001). They examine data obtained from a Bell Canada call centre and find that it exhibits every type of behaviour suggested above. One of the concerns about the doubly stochastic Poisson models is that, although they capture a time-varying arrival rate, they do not support correlation between arrival counts in different time periods.

Fluid approximations to queueing processes have been considered by Gans et al., (2006).

The authors allow time varying parameters and not only abandonments, but also retrials. Numerical results show that the simple fluid approximation suggested is fairly accurate.

On performance over multiple intervals and overload, according to Green & Kolesar, (1991), to make the translation to intra-day performance, and thus to inhomogeneous Poisson arrivals, (weighted) sums of interval performances are taken, where for each interval another call arrival rate is taken. They call this the pointwise stationary approximation and that, an alternative idea would be to take the average arrival rate, and use this as input for a performance model. This can give extremely bad results, even if the occupancy is constant (Green & Kolesar, 1989&1991)

According to Koole, (2000) standard modeling applications for call centers use stationary performance measures for each interval, say of 30 minutes duration and this works in general pretty well. But exceptions arise with abrupt significant changes in arrival rate, particularly when overload occurs during one or more intervals. Then a backlog is built up, and nonstationarity has to be accounted for. Such a behavior could arise from an external event, such as advertising a telephone number on TV, or when the call center opens in the middle of the day. Such abrupt

overloads can be modeled with the help of fluid models by Mandelbaum et a., (1999). These results are extended by Mandelbaum et a., (2000), however, these fluid approximations work less well in underload situations, as has been argued about by Altman, Jiménez, and Koole, (2001).

According to Bouzada (2009) a few call center characteristics make it difficult to apply analytical formulas from the Queue Theory for its modeling, including: generic distribution for the handling time, time-varying arrival rates, temporary overflows and abandonment.

According to Bapat and Pruitte Jr. (1998), the premises adopted by the studies based on Queue Theory analytical models are extremely limited when based on call centers current context because: (i) the incoming calls are all of the same kind; (ii) from the moment a call enters a queue, it never leaves it, and this usually overestimates the labor needed, increasing the personnel costs for the company; (iii) the attendants handle the calls following the FIFO ("first in, first out") discipline; and (iv) each operator handles all calls the same way. These premises rarely work at the environment in which call centers are inserted, since, according to the mentioned authors −depending on the individual tolerance for waiting his turn to be handled – a client may disconnect the call, if queued. Furthermore, the operators normally differ in relation to their own skills and to the handling time. Additionally, the clients' needs are very different and, sometimes, a priorization that can offer a better service might be necessary.

Simulation, according to Mehrotra (1997), explicitly shapes the interaction between calls, routes and agents, as well as the random individual incoming calls and the also random duration of the handling service. Through the use of simulation, managers and analysts translate the call centers gross data (call forecast, distribution of the

handling times, schedule hours and the agents abilities, call route vectors, etc.), in handling information on the service levels, clients abandonment, use of agents, costs and other important performance measures of a call center.

According to Chokshi (1999) and Klungle and Maluchnik (1997), the use of simulation to help management decisions in a call center allows the following benefits: (i) to visualize future processes and be used as a communication tool; (ii) to validate the processes premises before its implementation; (iii) to analyze the impact of the changes (scenario studies) in detail; (iv) to foresee the aggregated needs of resources and to schedule the working team; (v) to measure the performance indicators; and (vi) to estimate impacts on costs and economies.

One of the usages of the simulation in a call center, as said by Hall and Anton (1998), is the evaluation when one may verify "where the call center is". The key-question is "how efficient is the operation nowadays?" The goal of this evaluation is to establish a point of departure (and reference) for the change.

In accordance to Mehrotra, Profozich and Bapat (1997), Yonamine (2006), Gulati and Malcolm (2001), Bapat and Pruitte Jr. (1998) and Paragon (2005), a simulation model can be used (and has been used more frequently than ever) – besides normally allowing graphics and animations – to contemplate a few other critical aspects of the modern receptive centers of all sizes and types, such as: (i) a specific service level; (ii) flexibility on the distribution of time between incoming calls and of handling time; (iii) consolidation of the central offices; (iv) skill-based routing; (v) multiple types of calls; (vi) simultaneous lines; (vii) call disconnect patterns; (viii) call returns; (ix) overflow and filling of capacity; (x) waiting lines prioritization; (xi) call

transference and teleconferences; (xii) operators preferences, proficiency, time learning and schedule. The outputs model can emerge in shape of waiting time, call disconnecting average amount, (both with the possibility of differentiation on the call types) and level of the operators utilization (with possibility of the operator types differentiation). And, due to the applicability of this approach to the real and complex characteristics of call centers, the Simulation can make its dimensioning and management more reliable.

In accordance to Mehrotra, Profozich and Bapat (1997), Steckley, Henderson and Mehrotra (2005), Paragon (2005), Mehrotra (1997), Klungle and Maluchnik (1997), Pidd (1998) and Tanir and Booth (1999), the traditional methods most often used to manage and size a call center (intuitive estimatives, unprepared computations, worksheets and Erlang queue theoretical models) are becoming significantly limited due to the variability of the incoming calls, routes and handling time, to the operators skills and priorities, to the call heterogeneity and the interaction among them and the line trunks, to the dynamic of the call disconnections, to the recent tendencies (such as the skill-based routing, electronic channels and interactive calls handling) and, in general, to the sophistication and complexity more and more evidently noticed in the call centers systems. For example: analytical models usually assume that the clients arrival follows a Poisson process when, as a matter of fact, the call centers' data constantly reject this premise. In addition, worksheets and Erlang models overestimate the number of agents, besides having not much precision for call centers with different handling for each kind of client.

The simulation enlarges the capacity of the analytical tools and consists of a superior approach when there is no workable theoretical model capable to provide a reasonable system representation and when the means are not sufficient, the

accuracy is important, the operation is detailed, the demand varies too much, bottlenecks and processes design changing needs must be identified, or else an animation is necessary to improve the communication of a change to the company's board. The industry recent tendencies demand more sophisticated approaches and the simulation provides the necessary techniques to acquire the insights about these new tendencies and helps to shape its present and future designs, consisting in the only analysis method able of modeling a call center efficiently and accurately, throughout an approach much more practical,

flexible in terms of inputs and outputs, and capable of allowing the inclusion of important details, of representing much better the reality (without great needs of simplifications as theoretical models do), of enabling a better and deeper understanding concerning the call center processes and of generating much more robust results regarding the call center performance, allowing its optimization in a more reliable way (Paragon, 2005; Riley, 2005; Mehrotra, 1997; Klungle; Maluchnik, 1997; Tanir; Booth, 1999; Saliby, 1989; Hillier; Lieberman, 1995; Hertz, 1980; Mehrotra; Profozich; Bapat, 1997; Bapat; Pruitte Jr., 1998; Chokshi, 1999; Klungle, 1999; Worthington; Wall, 1999; Ragsdale, 2001; Mehrotra; Fama, 2003).

Bouzada (2006) explains that this happens because, amongst other reasons, the handling capacity dimensioning consists of a critical activity in the reaching of the efficiency and effectiveness of the operation. And the simulation tool usually fits better to the dimensioning of more complex operations (as that related to modern call centers, for instance), since it can model very well the real world, presenting more accurate and relatively precise results. It is true that these results are not as precise as the theoretical ones obtained by analytical methods, but they are usually

pretty close to them. The precept of the Simulation says that "it is better to have a rough solution for a very realistic model than an exact solution for a model with several approximations".

As studied by Mehrotra and Fama (2003), and Hall and Anton (1998), the call centers are interesting objects for the simulation studies, because: (i) they cope with more than one type of call, where each type represents a line; (ii) the calls received in each line arrive by chance − as time goes by; (iii) in a few cases, agents make calls proactively (especially in telemarketing or charging calls), or as a return for a call received; (iv) the duration of each call is random, as well as the work that the agent executes after the call (collecting of data, documentation, research...); (v) the progress on the systems which route the calls for the agents, groups or locals, make the logics behind the call center even more sophisticated; (vi) agents can be trained to answer only one type of call, several types of calls or all types of calls with different priorities and preferences specified for the routing logics; and (vii) the great amount of money invested in call centers, on both forms, capital and work, is capable to justify the use of this so powerful tool.

Hall and Anton (1998) said that call centers may use the Simulation tool to test (and eventually justify its implementation) whether some changes can prove or not to be able to improve the system before its implementation. The best call centers use this tool effectively to design the system, manage the operation and plan ahead, in the face of potential scenarios.

According to Mandelbaum and Zeltyn (2005) the classical M/M/n queueing model (Erlang-C), is the most frequently used in workforce management of call centers. Erlang-C assumes Poisson arrivals at a constant rate λ, exponentially distributed

service times with a rate μ, and n independent statistically-identical agents, however Erlang-C does not allow abandonment. This they say is significant deficiency because customer abandonment is not a minor, let alone a negligible, aspect of call center operations.

CHAPTER THREE

METHODOLOGY

3.0 Introduction

In this chapter we model three kinds of queues that occur in the operation of a call centre. We derive expressions for determining the minimum number of agents required to serve calls at the minimum staff cost and other relevant performance formulas for each of the three queues. We also developed a simulation approach for deriving the minimum number of agents required to serve calls at our case study, IThelpdesk.

3.1 Background to Call Centers as Queueing Systems

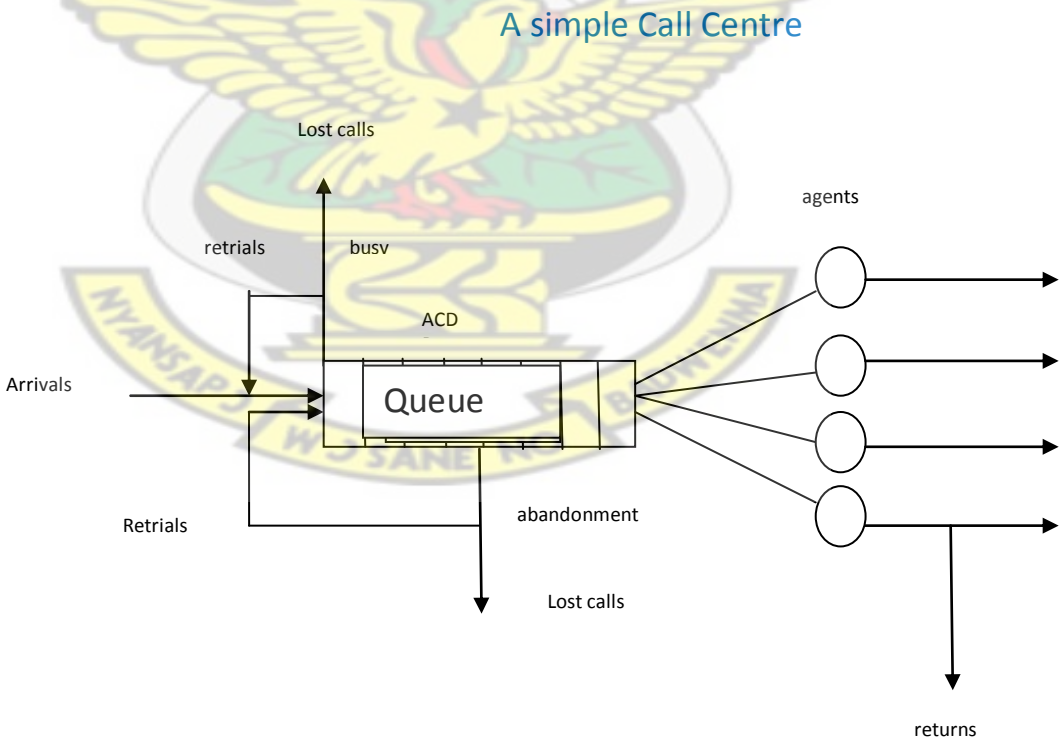

Figure 2 is an operational scheme of a simple call center showing the relationship between call centers and queueing systems.

Our case study call centre can be depicted by the figure 2 and has the following setup. A set of **k** trunk lines connects calls to the center. There are **w** ≤ k work **(**w**)** stations, often referred to as seats, at which a group of N ≤ w agents (N) serve incoming calls. An arriving call that finds all k trunk lines occupied receives a busy signal and is blocked (Case *1)* from entering the system. Otherwise it is connected to the call center and occupies one of the free lines. If fewer than N agents are busy, the call is put immediately into service. If it finds more than N but fewer than k calls in the system, the arriving call waits in queue (Case *2)* for an agent to become available. Customers who become impatient hang up, or abandon, before being served (Case *3)*. For the callers that wait and are ultimately helped by a customer service representative (CSR), the service discipline is first-come, first-served.

Once a call exits the system it releases the resources it used – trunk line, work station, agent – and these resources again become available to arriving calls. The remaining blocked and abandoned calls are lost. Thus, the number of trunk lines k acts as an upper bound on the number of calls that can be in the system, either waiting or being served, at one time. Similarly, the number of CSRs taking calls, N ≤ w, provides an upper bound on the number of calls that can be in service simultaneously.

For any fixed N, one can construct an associated queueing model in which callers are customers, the N CSRs are servers, and the queue consists of callers that await service by CSRs. When N changes, (k − N), the number of spaces in queue, changes as well.

3.2 Notation

The notations we would be using are standard notations in queueing theory.

The number of customers in the system refers to both the customers currently in the queue and those currently being served.

λ – arrival rate: the rate parameter if a homogeneous Poisson arrival process is assumed;

μ - service rate: the rate parameter if service times are assumed to be exponentially distributed ($1/\mu$ is the average duration of service);

$\beta = \mu^{-1}$ - the average service time if service times are assumed to be exponentially distributed;

γ - individual abandonment rate($1/\gamma$ is the average patience of a caller if abandonment times are assumed to be exponential);

s - number of servers/agents;

$a = \lambda/\mu$ - the offered load;

$\rho = \lambda/s\mu$ - the load to the system or load per agent;

π - the stationary distribution of the number of customers in the system (if it exists);

W_Q - the time that an arbitrary customer spends waiting in the queue, if the system is in a stationary situation;

L_Q(or j)- the random number of customers in the queue, if the system is in a stationary situation.

Note that a and ρ are dimensionless, but are measured in Erlang, which is a measure of telecommunications traffic.

t - patience time is the time that the customer is willing to wait for service.

3.3 General assumptions

We assume that we have a multi-agent single-skill inbound call centre and model the queueing process by case *1*, an *M/M/s (or* case *2*, *M/M/s/s or* case *3*,

M/M/s+M as the case may require) queue, which means that arrivals follow a Poisson process with constant arrival rate λ and service times are exponentially distributed with constant rate 1/μ. The assumption that the arrival rate follows a Poisson process means that the *inter-arrival times* are independent and identically distributed as exponential random variables with rate $^1/_\lambda$. There are also *s* agents serving the calls. If the assumptions above hold, then several useful formulae exist. The *offered load is given by,* $a = \frac{\lambda}{\mu}$ and that of the *load per agent is given by,* $\rho = \frac{\lambda}{s\mu}$

Calls are independent of each other, the service discipline is first come first served and the system is assumed to be in a stationary state.

Suppose also that $\rho < 1$ and let *i* denote the number of customers in the *system* (those currently being served as well as those in the queue if any).

We would now examine the various scenarios (cases) one at a time.

3.4 Case 1: Queue with call blocking(Erlang B)

If in addition to the general assumptions the number of trunks k (in the call centre) equals the number of agents serving the calls s, i.e. k=s, then,

i. The system assumes an *M/M/s/s* queueing system which means that a queue is not formed

ii. if a customer calls and there is not a free agent, the call is blocked from entering the system. The customer only hears a busy signal and that call is lost.

The fraction of arriving customers who find all the servers busy (the probability of blocking, or loss probability) can be determined by the Erlang loss or Erlang B, (Cooper, 2000), formula,

$$B(s,a) = \frac{\frac{a^s}{s!}}{\sum_{i=0}^{s-1} \frac{a^i}{i!}} \qquad (3.1)$$

Formula (3.1) is hard to calculate directly from its right-hand side when *s* and *a* are large, but is easy to calculate numerically using the following iterative scheme:

$$B(n,a)=\frac{aB(n-1,a)}{n+a(aB(n-1,a))} \quad \text{(given that n=1,2,3…s; } B(0,a)=1\text{, (Cooper, 2000)} \quad (3.2)$$

In this case, the service level normally defines an acceptable probability that a call is blocked. The minimum number of agents required to achieve this can then be calculated using equation (3.2).

3.5 Case 2: Queue with no abandonment (Erlang C)

If, in addition to the general assumptions, k > s, there is an unlimited number of available places in the queue, which means that calls are never blocked (i.e. a caller always has a place in the queue), and in addition callers do not abandon their calls whilst they are in the queue; (i.e. they will always wait to be served).

Then according to Clark, (2007) the stationary distribution for the number of customers in the system is given by

$$\pi(i)=\begin{cases}\frac{a^i}{i!}\pi(0) & \text{if } i<s \text{ and} \\ \frac{a^i}{s!s^{i-s}}\pi(0) & \text{otherwise)}\end{cases} \quad (3.4)$$

where

$$\pi(0)^{-1}=\sum_{i=0}^{s-1}\frac{a^i}{i!}+\frac{a^s}{(s-1)!(s-a)} \quad (3.5)$$

Also from Appendix A (A. 5i) the probability that a caller has to wait at all in the queue before being served is given by

$$C(s,a)=\sum_{i=s}^{\infty}\pi(i)=\frac{a^s}{(s-1)!(s-a)}\left[\sum_{i=0}^{s-1}\frac{a^i}{i!}+\frac{a^s}{(s-1)!(s-a)}\right]^{-1} \quad (3.6)$$

C(s,a) is calculated using the recursive formula (3.7) with equation (3.2)

$$C(s,a)=\frac{sB(s,a)}{s-a(1-B(s,a))} \quad (3.7)$$

The probability that a caller has to wait in the queue for time, T seconds before being served, i.e. P($W_Q > T$) is given by

$$\mathrm{P}\ W_Q > T\ = C\ s,a\ e^{-(s\mu-\lambda)T} \tag{3.8}$$

where $C(s,a)$ is obtained from (3.6)

The expressions for the expected time spent waiting in the queue, (EW_Q) is given by

$$\mathrm{EW_Q} = \mathrm{C(s,a)} = \frac{1}{(1-\rho)}.\frac{\beta}{s} = \frac{\mathrm{C(s,a)}}{(s\mu-\lambda)} \tag{3.9}$$

From Little's Law the expected queue length: EL_Q is given by

$$\mathrm{EL_Q} = \lambda \mathrm{EW_Q}$$

$$\mathrm{EL_Q} = \frac{\rho \mathrm{C(s,a)}}{(1-\rho)} \tag{3.10}$$

$C(s,a)$ can also be obtained by putting $T = 0$ into (3.8) given that the condition that $\rho < 1$ is required for stability. If calls are arriving more quickly on average than the call centre is managing to serve them, $\lambda > s\mu \iff \rho > 1$, then the queue will continue to grow with no upper bound (Clark, 2007).

In order to calculate the required number of agents, supposing that the desired grade of service is of the form: answer α % of calls within T seconds.

Then we wish to find the smallest number of agents, s which should be an integer such that

$$\mathrm{P}\ W_Q \leq T\ > \alpha \iff \mathrm{P}\ W_Q > T\ < 1-\alpha \tag{3.11}$$

and the left hand side is precisely what is given by (3.8).

3.6 Case 3: Queue with abandonment(Erlang A)

In addition to the general assumptions, k > s, so no call is blocked; and also, customers, while waiting to be served abandon the tele-queue when their patience run out. Customers' abandonment times follow an exponential distribution with an *average patience* γ. This typically assumes an *M/M/s + M* queue.

For a given customer, the patience time t is the time that the customer is willing to wait for service. A wait that reaches *t seconds* results in abandonment. Let *V* denote the offered waiting time - the time a customer who has infinite patience, must wait in order to get the service.

The actual waiting/queueing time then equals

$$W_Q = \min\ V, t \quad .$$

Denote by L(t) the total number of callers in system at time t (served plus queued). Then L = {L(t), t =0} is a Markov birth-and-death process, with the following transition-rate diagram:

Figure 4: Transition-rate diagram of the Erlang-A model

Let d_j stand for the death-rate in state j, $0 \leq j < \infty$. Then

$$j \cdot \min(\mu, \gamma) \leq d_j \leq j \cdot \max(\mu, \gamma). \qquad (3.12)$$

The bounds on the left-hand and right-hand sides of (3.12) correspond to death-rates of an M/M/∞ queue with service rates min (μ,γ) and max(μ,γ), respectively. In some sense, which can be made precise via stochastic orders between distributions,

these two M/M/∞ queues provide lower and upper (stochastic) bounds for the Erlang-A system. In the special case of equal service and abandonment rates ($\mu = \gamma$), the Erlang-A and M/M/∞ models in fact coincide.

As customary, we define the limiting distribution of L by:

$$\pi_j = \lim_{t\to\infty} P\{L(t)=j\}, \qquad j \geq 0.$$

When it exists, the limit distribution is also a steady-state (or stationary) distribution, which is calculated via the following version of the steady-state equations (Mandelbaum and Zeltyn 2004):

$$\begin{cases} \lambda\pi_j = (j+1).\mu\pi_{j+1} & 0 \leq j \leq s-1 \\ \lambda\pi_j = (s\mu+(j+1-s)\gamma).\pi_{j+1} & j \geq s \end{cases} \tag{3.13}$$

from above the recipe solution can be derived:

$$\pi_j = \begin{cases} \frac{(\lambda/\mu)^j}{j!}\pi_0 & 0 \leq j \leq s \\ \prod_{k=s+1}^{j}\left(\frac{\lambda}{s\mu+(k-s)\gamma}\right)\frac{(\lambda/\mu)^s}{s!}\pi_0 & j \geq s+1 \end{cases} \tag{3.14}$$

$$\pi_0 = \left[\sum_{j=0}^{s}\frac{(\lambda/\mu)^j}{j!} + \sum_{j=s+1}^{\infty}\prod_{k=s+1}^{j}\left(\frac{\lambda}{s\mu+(k-s)\gamma}\right)\frac{(\lambda/\mu)^s}{s!}\right]^{-1} \tag{3.15}$$

The solution makes sense - equivalently the Markov process L is ergodic(that is positive recurrent periodic state of stochastic systems; tending in probability to a limiting form that is independent of the initial conditions) - if the infinite sum in (3.15) converges, which is a consequence of the lower bound in (3.12) (Mandelbaum and Zeltyn 2004)

$$\sum_{j=0}^{s}\frac{(\frac{\lambda}{\mu})^j}{j!} + \sum_{j=s+1}^{\infty}\prod_{k=s+1}^{j}\left(\frac{\lambda}{s\mu+(k-s)\gamma}\right)\frac{(\frac{\lambda}{\mu})^s}{s!} \leq \sum_{j=0}^{\infty}\frac{(\frac{\lambda}{\min(\mu,\lambda)})^j}{j!} = e^{\frac{\lambda}{\min(\mu,\lambda)}}$$

Formulae (3.14) and (3.15) include infinite sums that can cause numerical problems. To overcome these, Palm C. (1957) represented the Erlang-A steady-state distribution, and some of its important performance measures, in terms of special functions. We define the Gamma function

$$\Gamma(x) \triangleq \int_0^\infty t^{x-1}e^{-t}dt \qquad x > 0,$$

And the incomplete Gamma function

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

Let $A(x,y) \triangleq \frac{xe^y}{y^x} \cdot \gamma(x,y) = 1 + \sum_{j=1}^{\infty} \frac{y^j}{\prod_{k=1}^{j}(x+k)} \qquad x > 0, y \geq 0$ (3.16)

(The second equality is taken from Palm C. (1957).)

and $$E_{1,s} = \frac{\frac{a^s}{s!}}{\sum_{i=0}^{s-1}\frac{a^i}{i!}}$$ (3.17)

which denotes the probability of blocking in the M/M/s/s system above(3.1) given that $E_{1,0=0}$ so that $E_{1,s} = \frac{\rho E_{1,s-1}}{1+\rho E_{1,s-1}} \quad s \geq 1$

In which ρ is the offered load per agent already defined.

In the Appendix, it is deduced from (3.16) the following solution for the steady-state distribution:

$$\pi_j = \begin{cases} \pi_s . \frac{s!}{j!.(\frac{\lambda}{\mu})^{n-j}}, 0 \leq j \leq s \\ \pi_s \frac{(\frac{\lambda}{\gamma})^{n-j}}{\prod_{k=1}^{j-n}\left(\frac{s\mu}{\gamma}+k\right)}, j \geq s+1 \end{cases}$$ (3.18)

where

$$\pi_s = \frac{E_{1,s}}{1+\left[A\left(\frac{s\mu}{\gamma}, \frac{\lambda}{\gamma}\right) - 1 \right].E_{1,s}} \tag{3.19}$$

3.6.1 The delay probability P {W>0}

In this derivation, calculations are based on conditioning and the incomplete gamma function. The delay probability P {W > 0}, represents the fraction of customers who are forced to actually wait for service. (The others are served immediately upon calling.) Recall that this measure identifies operational regimes of performance. Following Palm (1957), it is shown in the Appendix A that the representations (3.16) and (3.18) immediately imply,

$$P\{W > 0\} = \sum_{j=s}^{\infty} \pi_j = \frac{A\left(\frac{s\mu}{\gamma}, \frac{\lambda}{\gamma}\right).E_{1,s}}{1+\left(A\left(\frac{s\mu}{\gamma}, \frac{\lambda}{\gamma}\right) - 1 \right).E_{1,s}} \tag{3.20}$$

The first equality in (3.20) follows from *Poisson Arrivals See Time Averages* (PASTA)

Where A(x,y) and $E_{1,s}$ are given by (3.16) and (3.17) respectively above (Mandelbaum and Zeltyn 2004)

3.6.2 Fraction of customers who abandon P{Ab}

To calculate the probability to abandon, which represents the fraction abandoning, define $P_j\{Sr\}$ to be the probability of ultimately getting served, for a customer that encounters all servers busy and j customers are in queue, upon arrival (implying, s + j customers are in the system). Competition among exponentials now implies that

$$P_0\ \mathrm{Sr}\ = \frac{s\mu}{s\mu+\gamma}$$

Then $$P_1\ \mathrm{Sr}\ = \frac{s\mu+\gamma}{s\mu+2\gamma}.P0\ \mathrm{Sr}\ = \frac{s\mu}{s\mu+2\gamma}$$

here we set a condition on the first event, that, on arrival encounters all servers busy and a single customer is in the queue; this event is either a service completion (with probability $\frac{s\mu+\gamma}{s\mu+2\gamma}$) or an abandonment. More generally, via induction:

$$P_j\{\mathrm{Sr}\} = \frac{s\mu+j\gamma}{s\mu+(j+1)\gamma}.P_{j-1}\{\mathrm{Sr}\} = \frac{s\mu}{s\mu+(j+1)\gamma}\ j \geq 1$$

The probability to abandon service, given all servers busy and j customers in the queue upon arrival, finally equals

$$P_j\ \mathrm{Ab}\ = 1-P_j\ \mathrm{Sr}\ = \frac{j+1\ \gamma}{s\mu+\ j+1\ \gamma}, j \geq 0 \qquad (3.21)$$

It follows that

$$P\ [\mathrm{Ab}\mid W>0] = \sum_{j=s}^{\infty}\pi_j P_{j-s}\ \mathrm{Ab}\ / P\ \ W>0\ = \frac{1}{\rho A\left(\frac{s\mu}{\gamma},\frac{\lambda}{\gamma}\right)}+1-\frac{1}{\rho} \qquad 3.22)$$

So that the fraction abandoning, P{Ab}, is simply the product

P[Ab|W > 0] x P{W > 0}

3.6.3 Relations between E(W), P{Ab} and E(Q)

A remarkable property of Erlang-A, which generalizes to other models with customer patience γ that is exponentially distributed exp(γ), (Mandelbaum and Zeltyn 2004) is the linear relation between the fraction abandoning P{Ab} and average wait E[W]: $$P\ \mathrm{Ab}\ = \gamma.\mathrm{E}\ W \qquad (3.23)$$

Proof: The proof is based on the balance equation

$$\gamma.\mathrm{E}\ \mathrm{Q} \quad = \quad \lambda.\mathrm{P}\ \mathrm{Ab} \quad , \tag{3.24}$$

and on the Little's formula

$$\mathrm{E}\ \mathrm{Q} \quad = \lambda.\mathrm{E}\ \mathrm{W} \quad , \tag{3.25}$$

where Q is the steady-state queue length. The balance equation (3.24) is steady-state equality between the rate that customers abandon the queue (left hand side) and the rate that abandoning customers (customers who eventually abandon) enter the system. Substituting the Little's formula (3.25) into (3.24) yields formula (3.23).

Observe that (3.23) is equivalent to

$$\mathrm{P}\ \mathrm{Ab} \mid \mathrm{W} > 0 \quad = \quad \gamma.\mathrm{E}\ \mathrm{W} \mid \mathrm{W} > 0 \tag{3.26}$$

Then, the average waiting time of delayed customers is computed via (3.22) and (3.26):

$$\mathrm{E}[\mathrm{W} \mid \mathrm{W} > 0] = \frac{1}{\gamma}\left[\frac{1}{\rho \mathrm{A}\left(\frac{s\mu}{\gamma}, \frac{\lambda}{\gamma}\right)} + 1 - \frac{1}{\rho}\right] \tag{3.27}$$

The unconditional average wait E[W] equals the product of (3.20) with (3.27).

3.6.4 How to dimensioning the Call Centre Queue with abandonment.

An economical optimal staffing level can be obtained by a trade-off between staffing cost, cost of customers waiting and cost of abandonment.

Using the trade-off between staffing cost and customers' waiting cost

Let the average operational cost (per unit of time) be equal to

$$U(s,\lambda) = c.s + \lambda w.\,P\ \ Ab\ , \tag{3.28}$$

where **c** is the staffing cost, and **w** is the customers waiting cost. The objective is to find the staffing level **s*** that minimizes operational cost, U (s, λ). Note that instead of customers waiting cost, abandonment cost could suffice for the analysis.

We define the customers waiting /staffing cost ratio by $r \triangleq w/c$, and let $v \triangleq \sqrt{\frac{\mu}{\gamma}}$.

Assuming that $w > c/\mu$ (Otherwise, the asymptotic optimal policy is $s^* = 0$: not to provide service at all.) Then the asymptotic optimal staffing level, (Mandelbaum and Zeltyn 2004), is equal to

$$s^* = [R + y^*(r;v).\sqrt{R}] \tag{3.29}$$

where the square brackets in (3.25) denote the nearest integer value, the function y*(.) is defined by

$$y^*(r;v) \triangleq \arg min_{-\infty<y<\infty}\ \ c.y + w.\gamma v.[1+(h(yv))/(vh(-y))]^{(-1)}.[h(yv)-yv]$$

and $h(x) = \Phi(x)/(1 - \Phi(x))$ is the hazard rate of the standard normal distribution, ($\phi(x)$ is its density function and $\Phi(x)$ is the cumulative distribution function and their expressions are given below).

$$\phi\ x = \frac{1}{\sqrt{2\pi}} e^{-x/2} \tag{3.30}$$

$$\Phi(x) = \int_{-\infty}^{x} \phi(y)dy \tag{3.31}$$

3.7 Simulation Approach

3.7.1 Introduction

We analyzed the data gathered to derive the inter arrival time between subsequent callers using the beginning (**beg)** value in data collected (sample data is in **Table C-1)** and tabulating them against number of callers that shared the same inter arrival time. We did the same with the **total time value** from the data collected to get tabulation for the service distribution.

Table 3.1 Source data format

Cust #	beg(hr:min)	end (hr:min)	waite_time (hr:min)	total time (hr:min)
1	7:07	7:12	0:00	0:05
2	7:12	7:21	0:00	0:09
.	.	.	.	.

From the tabulations we then derived a forecast of arrivals and service distributions. With the distributions derived, we created two lookup tables that were used to simulate the actual occurrence of arrivals to the call centre and service provided to the callers. Operation time of the call centre is bounded by the start time and closing time of the call centre and this serve d as a lookup table determining the start and end time for the simulation. Frequency of arrivals at the call centre is determined by a lookup on the arrival distribution table whilst the service time is derived from lookup on the service distribution table.

Each call that arrives is assigned a customer number in sequence according to time of arrival at the call centre. Each caller's service start time is determined by the time an agent becomes available. A call that does not get a free agent on arrival waits in a queue until an agent becomes available. The call's service time is determined by the

looked up value from the service distribution table. The simulations are run varying the number of agents attending to the calls starting with two agents. The format of the simulation is of the form in table 3.1 below.

Table 3.2 Simulation of arrival and service received

Cust.	Interarrival	Arrival	Service	Server #1		Server #2		Server #3		Server #4		Server #5		Wait
#	Time	Time	Time	Start	End	Start	End	Start	End	Start	End	Start	End	Time
	(min)	(hr: min)	(min)	(hr: min)	(hr: min)	(hr: min)	(hr: min)	(hr: min)	(hr: min)	(hr: min)	(hr: min)	(hr: min)	(hr: min)	(hr: min)
Start		7:00												

Table 3.3 formulas used in Simulation	
Cust#	=IF(ISTEXT(F12),"Closed",A11+1)
Interarrival Time(min)	=VLOOKUP(RAND(),arrival,2)
Service Time(min)	=VLOOKUP(RAND(),service,2)
Arrival Time (hr:min)	=IF(ISTEXT(F11),"",IF(C12+F11>closed_time,"Closed",C12+F11))
#1 Start Time (hr:min)	=IF(ISTEXT(F12),"",MAX(F$11:F11,$F12,L$11:L11,Agent_St_time))
#2 Start Time (hr:min)	=IF(ISTEXT(F12),"",MAX(F$11:F11,$F12,N$11:N11,Agent_St_time))
Actual Time (hr:min)	=IF(ISTEXT(F12),"",MIN(G12:H12))
Next Server	=IF(ISTEXT(F12),"",MATCH(I12,G12:H12,0))
Server #1 Start (hr:min)	=IF(J12=1,I12,"")
Server #1 End (hr:min)	=IF(ISTEXT(K12),"",K12+E12)
Server #2 Start (hr:min)	=IF(J12=2,I12,"")
Server #2 End (hr:min)	=IF(ISTEXT(M12),"",M12+E12)
Wait	=IF(ISTEXT(A12),"",I12-F12)
Total	=IF(ISTEXT(A12),"",O12+E12)

For each call that arrived at the call centre, the call was assigned to the first agent (server) that was available, starting the check for availability from agent number one through to the agent number five depending on the number of agents attending to the calls.

From two agents to five agents, hundred simulations were run. The daily longest wait and the average wait of each run were recorded. To portray the agent utilization, twenty (20) simulations each were run for the four agent number situations (2-5agents). The number of calls each agent attended to was also recorded for further analysis.

3.7.2 Excel functions used for data analysis and simulations

The **VLookup** (figure 3) function was used to search for a value in the left-most column of a *table_array* and returned the value in the same row based on specified *index_number*.

The **IsText**(figure 3) function was used to check if the value in cell was is a text value.

The **Match**(figure 3) function was used in searching for a value in an array and returned the relative position of that item.

The **Rand**(figure 3) function return a random number that is greater than or equal to 0 and less than 1. It returned a new random number each time the spreadsheet recalculated.

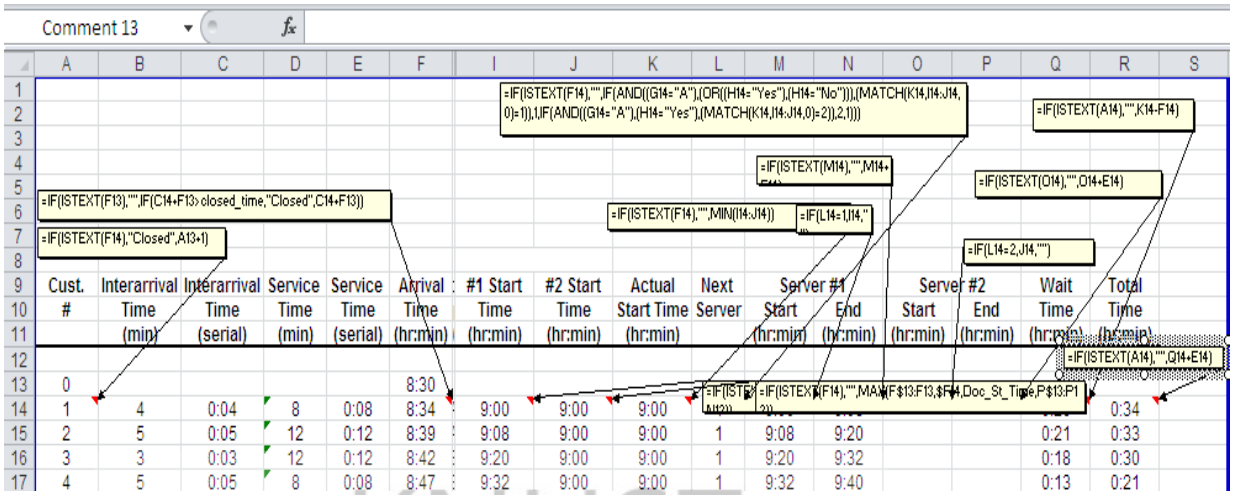

Figure 3 is interface of the simulation software for 2 agent module, showing the formulas used in the various columns.

CHAPTER 4

DATA ANALYSIS

4.1 Introduction

In Chapter three, three queue modules were formulated. The blocking probability, the probability of wait, the probability of abandonment and the queue length are some expressions derived. A simulation method for optimizing human resource scheduling at our chosen case study was also discussed. In this chapter the data collected and the simulations results are analyzed.

4.2 Expressions for the minimum number required for each queue module

From the 1st queue module (Erlang B) the expression for the fraction of arriving customers who find all the agents busy is given by equation 3.1. The minimum number of agents required to achieve this is calculated using equation 3.2.

From the 2nd queue module (Erlang C) the expression for the fraction of arriving customers who find all the agents busy and have to wait in the queue is given by equation 3.8

The minimum number of agents required to achieve this is calculated using equation 3.7 together with equation 3.2

From the 3rd queue module (Erlang A) the expression for minimum number of agents required to minimize staff and waiting cost is given by 3.29

4.3 Data

From the data collected, the arrival and service probability distributions, **Table 4-0 and Table 4-2** respectively were determined, out of which the summary of arrival and service distributions, **Table 4-1** and **Table 4-3** respectively were extracted and are used as lookup tables for inter arrival between calls and the service distribution

in the simulations. From the data collected, the call centre operation start time is 7am (07:00) and closing time is 7pm (19:00); from this we have another lookup table **Table 4-4** that determined the start and end time of daily operations. Our goal was to derive an arrival and service forecast and to use a simulation method to derive the minimum number of agents that minimized staff and waiting costs.

Table 4-0 Inter Arrival Time Probability Distribution

time between arrivals	frequency	frequency density to the nearest whole number
1	923	2
2	17003	32
3	498	1
4	652	1
5	14000	26
6	588	1
7	12218	23
8	1026	2
9	6042	11

Source: Field Data, 2011

Table 4-1 Inter Arrival Time Forecast

Inter arrival Time(minutes)	Probability
2	0.35
5	0.28
7	0.25
9	0.12

Source: Field Data, 2011

Table 4-2 Service Time Probability Distribution		
Service Time (minutes)	frequency	frequency density to the nearest whole number
0	0	0
1	12	0
3	112	0
4	873	2
5	22010	42
6	821	2
7	16033	30
8	1028	2
9	12061	23

Source: Field Data, 2011

Table 4-3 Service Time Forecast	
Service Time (minutes)	Probability
5	0.45
7	0.31
9	0.24

Source: Field Data, 2011

Table 4-4 Start time and end time	
Start Time	Close Time
7:00	19:00

Sample simulations for each agent (server) situation are presented in appendix D. The result of the hundred simulation runs are presented in **Table 4-5**(Appendix B). The maximum longest wait, maximum average wait is also presented in **Table 4-6**(appendix B). The number of calls attended to by the servers (agents) in each server situations is presented in **Table 4-**7(Appendix B). Finally the maximum and minimum number of calls served by each server (agent) in the four server situations is presented in **Table 4-8**(Appendix B).

4.4 Discussion of Simulation Results

In the simulation with 2, 3, 4 and 5 agents (servers) it became clear that a smaller number of agents could handle the volume of calls that arrive at the ITHelpDesk.

With a two agents (servers) the load on the system is too heavy for the system to handle. With about 150 calls, each agent would have to handle over 70 calls each with virtually no breaks between calls. For hundred (100) simulations of about one hundred and fifty (150) calls each, the average waiting time was zero, but some of the callers had to wait for over five (5) minutes to be served, with a maximum waiting time reaching as much as nineteen (19) minutes for the hundred (100) simulations, (Table 4-4). These results are clearly unacceptable for the call centre, because callers would normally abandon the call after waiting in the queue for more than three minutes.

With a three (3) server situation the load on the system is still heavy for system to handle. With about 150 calls, each agent would have to handle about 50 calls each with small breaks between calls. In these case also, for the hundred (100) simulations the average waiting time was zero, but some of the callers had to wait for over three (3) minutes to be served, with the maximum waiting time being six (6) minutes, (Table 4-4). This result is also not perfect for the call centre, because of the caller patience of about three (3) minutes, but, only a few would have abandoned their calls as compared to the two server situation.

With a four (4) server situation, the load on the system is light enough for system to handle smoothly. With about 150 calls, each agent would have to handle about 40 calls each with breaks between calls. In these case also, the average waiting time for hundred (100) simulations was zero, with the maximum time a caller had to wait being one (1) minute , (Table 4-4), and even this maximum waiting time of (1) minute is experienced by a very small percentage of the callers. This result is good for a call centre, because none of the callers were likely to abandon their calls due to waiting.

With a five (5) server the load on the system is very light on the system. With about 150 calls, each agent would have to handle about 30 calls each with so many break intervals between calls. The simulations, (Table 4-3), actually show that two agents would virtually be idle as most of the traffic would be handled by the first three

agents. Like the previous cases the average waiting time was zero for the hundred (100) simulations and no caller had to wait to be served, (Table 4-4). This result is very efficient for a call centre by way of its performance, because none of the callers would abandon their calls due waiting. Though the call centre performance metrics, like answer X% of calls that arrive within seconds or minutes of arrival; or number of abandoned calls should be less than 3% of all calls that arrive, would be met, the problem, however, is that the cost of the agents employed to handle the calls now comes into play, as the a high percentage of the cost of operating the call centre is tied in agent cost. Five agents would do the job with none of the callers having to wait at all, however the extra agent needed to achieve this level performance makes the five server situation a huge expenditure as compared to the four server situation, with only a small fraction of customers with the possibility of waiting for about a minute before being served.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

At the IThelpdesk six agents attend to the calls but the outcome of simulation results show that four agents could attend to the calls. With four agents there would be only a few callers who would wait for about a minute to be served and no caller is likely to abandon his calls.

With the number of agents down by two, the agent cost is also reduced. Assuming an hourly rate of ¢20.00 per agent, then the daily staff cost is also down by ¢480.00 from ¢1,440.00 for six agents to ¢960.00 for four agents.

We have been able to forecast about 35% of the calls arriving within two minutes after each other, 28% within five minutes, 25% within 7 minutes and 12% within 7 minutes.

For the handling time about 45% are handled within five minutes, 31% within seven minutes and 24% within nine minutes.

Expressions for the minimum number of agents required to minimize the staffing cost for the various queue modules have been derived.

5.2 Recommendations

We recommend a maximum of four agents to handle the calls at the IThepldesk. The number could further be reduced to two during a low peak period.

We recommend the use of the simulation method to complement other tools used to help determine the approximate number of agents required to serve the volume of calls at the other call centers and at other queue related service centres like hospitals and banks.

We recommend the use of Erlang A workforce management tools instead of the Erlang C workforce management tools which is prevalent on the market.

References

1. Chokshi, R. (1999), "Decision support for call center management using simulation", Winter Simulation Conference, Phoenix ,AZ, pp. 1634-1639.
2. Cooper, R.B. (2000), Queueing Theory, Published in Encyclopedia Of Computer Science, Fourth Edition (Anthony Ralston, Edwin D. Reilly, David Hemmendinger, eds.), Groves Dictionaries, Inc.,pp. 1496-1498
3. Feinberg, M.A. (1990). Performance characteristics of automated call distribution systems. GLOBECOM '90., IEEE. San Diego, CA, pp. 415-419.
4. Feinberg, R.A., Kim, I.-S., Hokama, L., Ruyter, K. de, and Keen, C.(2000). Operational determinants of caller satisfaction in the call center. International Journal of Service Industry Management, Vol. 11, pp. 131-141
5. Green, L., and Kolesar, P. (1989). Testing the validity of a queueing model of police patrol. Management Science, Vol. 37 pp. 84-97.
6. Green, L., and Kolesar, P. (1991). The pointwise stationary approximation for queues with nonstationary arrivals. Management Science, Vol. 37 pp. 84-97.
7. Gulati, S. and Malcolm, S. (2001), "Call center scheduling technology evaluation using simulation", Winter Simulation Conference, Arlington, VA pp. 1841-1846.
8. Halfin, S., and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. Operations Research, Vol. 29 pp. 567-587
9. Hall, B. and Anton, J. (1998), "Optimizing your call center through simulation", Call Center Solutions Magazine, pp. 1-10.

10. Hertz, D. (1980), "Análise de risco em investimentos de capital", Biblioteca Harvard de Administração de Empresas, Vol. 8, N. 3, pp. 1-14.
11. Hillier, F. and Lieberman, G. (1995), Introduction to Operations Research, New York: McGraw-Hill.
12. Jennings, O.B., Mandelbaum, A., Massey, W.A. and Whitt, W. (1996). Server staffing to meet time-varying demand. Management Science, Vol. 42, No 10, pp. 1383-1394.
13. Jongbloed, G. & Koole, G. (2001). Managing Uncertainty in Call Centers using Poisson Mixtures. Applied Stochastic Models in Business and Industry, Vol.17 pp. 307-318,
14. Klungle, R. and Maluchnik, J. (1997), "The role of simulation in call center management", MSUG Conference, pp. 1-10.
15. Klungle, R. (1999), "Simulation of a claims call center: a success and a failure", Winter Simulation Conference, Phoenix ,AZ, pp. 1648-1653.
16. Mandelbaum, A., Massey, W.A., Reiman, M.I. and Rider, R. (1999). Time varying multiserver queues with abandonments and retrials. In Key, P. and Smith, D. (Ed), Proceedings of the 16th International Teletraffic Conference, Edinburg, Scotland, pages 355-364
17. Mehrotra, V., Profozich, D. and Bapat, V. (1997), "Simulation: the best way to design your call center", Telemarketing & Call Center Solutions, pp. 1-5.
18. Mehrotra, V. (1997), "Ringing Up Big Business", OR/MS Today, Vol. 24, N. 4, pp.18-2
19. Mehrotra, V. and Fama, J. (2003), "Call Center Simulation Modeling: Methods, Challenges and Opportunities", Winter Simulation Conference, New Orleans, LA, pp. 135-143.

20. Palm, C. (1957) Research on telephone traffic by Full Availability Groups. Tele, vol.1, pp. 107
21. Palm, C. (1953). Methods of judging the annoyance caused by congestion. Tele, Vol 4. pp. 189-208.
22. Puhalskii, A.A., and Reiman, M.I. (2000). The multiclass GI/PH/N queue in the Halfin-Whitt regime. Advances in Applied Probability, Vol. 32, pp. 564–595
23. Pidd, M. (1998), Computer Simulation in Management Science, New York: Willey.
24. Ragsdale, C. (2001), Spreadsheet Modeling and Decision Analysis, Tennessee: South-Western
25. Riley, D. (2005), "Simulating a Virtual Customer Service Center", Winter Simulation Conference, Orlando FL, pp. 56-61.
26. Robbins, T.R., Medeiros, D.J., and Dum, P. (2006). Evaluating Arrival Rate Uncertainty in Call Centers. Proc. 38th Winter Simulation Conf., pp. 2180-2187
27. Sze, D.Y. (1984). A queueing model for telephone operator staffing. Operations Research, Vol. 32, pp. 229-249.
28. Steckley, S., Henderson, S. and Mehrotra, V. (2005), "Performance Measures for Service Systems with a Random Arrival Rate", Winter Simulation Conference, Orlando FL, pp. 566-575.
29. Saliby, E. (1989), Repensando a Simulação: a Amostragem Descritiva, São Paulo: Atlas. Worthington, D. and Wall, A. (1999), "Using the discrete time modeling approach to evaluate the time-dependent behavior of queueing

systems", Journal of the Operational Research Society, Vol. 50, pp. 777-788.

30. Tanir, O. and Booth, R. J.(1999). Call Center Simulation in Bell Canada. Proc. Winter Simulation Conference, Phoenix, AZ, pp. 1640-1647
31. Whitt, W. (2005). Engineering Solution of a Basic Call Center Model. Management Sci., vol. 51, no. 2, pp. 221-235.
32. Whitt, W. (2006). Staffing a Call Center with Uncertain Arrival Rate and Absenteeism. Production And Operations Management, Vol. 15, No. 1. pp. 88-102
33. Yonamine, J. (2006), O Setor de Call Centers e Métodos Quantitativos: uma Aplicação da Simulação, Dissertation (M. Sc. in Business Administration), Rio de Janeiro: UFRJ/COPPEAD
34. Avramidis, A.N., Deslauriers, A. and L'Ecuyer, P. (2004). Modelling Daily Arrivals to a Telephone Call Center. Management Science, Vol. 50, No. 7, pp. 896-908. http://mansci.journal.informs.org/cgi/content/abstract/50/7/896(accessed 20th, Jan 2010)
35. Altman, E., Jiménez, T. and Koole, G.M. (2001). On the comparison of queueing systems with their fluid limits. Probability in the Engineering and Informational Sciences, Vol. 15 pp. 165-178. Sophia Antipolis, France.http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.638&rep=rep1&type=pdf (accessed 20th, Jan 2009)
36. Baccelli, F. and Hebuterne, G. (1981). On queues with impatient customers. International symposium on computer performance, North-Holland. pp 159-

179. *citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.638&rep...* (accessed 20th, Jan 2009)

37. Bapat, V. and Pruitte Jr, E. (1998), "Using simulation in call centers", Winter Simulation Conference, Washington, DC, pp. 1395-1399. http://portal.acm.org/citation.cfm?id=293496(accessed 20th, Jan 2010)
38. Borst, S.C., Mandelbaum, A., and Reiman, M.I., (2000). Dimensioning large call centers. Working paper. Operations Research, Vol. 52, pp. 17–34. http://iew3.technion.ac.il/serveng/References/references.html. (accessed 20th, Jan 2010)
39. Borst, S.C. and Seri, P. (2000). Robust algorithms for sharing agents with multiple skills. Working paper. (Brandt, A., and M. Brandt. 1999. **Ed)** *www.scribd.com/doc/.../Overview-of-Routing-and-Staffing-**Algorithms***. (accessed 20th, Jan 2010)
40. Bouzada, M. (2006), The use of quantitative tools in call centers: a case-Contax, Thesis (Rio de Janeiro: UFRJ/COPPEAD.) UFRJ / ADMINISTRATION, http://www.joscm.com.br/previous/22/download/04%20JOSCM_VOL2_NUMBER%202_4.pdf
41. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. (2005). Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective. J. Amer. Statist. Assoc., 100, pp 36-50 Technical Report, Department of Statistics, The Wharton School University of Pennsylvania. http://iew3.technion.ac.il/serveng/References/references.html(accessed 14th, March 2010)

42. Bouzada, M. (2009) Dimensioning a Call Center: Simulation or Queue Theory? The Flagship Research Journal of International Conference of the Production and Operations Management Society, Vol. 2 Numbers 2 http://www.joscm.com.br/previous/22/download/04%20JOSCM_VOL2_NUMBER%202_4.pdf(accessed 14th, March 2010)
43. Clark, S.(2007) Robust Staff Level Optimisation in Call Centres, Thesis, Jesus College University of Oxford http://eprints.maths.ox.ac.uk/660/1/clarke.pdf (accessed 20th December, 2009)
44. Gans, N., Koole, G. and Mandelbaum, A. (2003). Telephone Call Centers: Tutorial, Review, and Research Prospects. Invited review paper by, Manufacturing and Service Operations Management, (M&SOM), vol 5 (2) pp. 79–141, ,http://iew3.technion.ac.il/serveng/References/references.html(accessed 20th, April 2010)
45. Garnett, O. and Mandelbaum, A. (2004) An introduction to skills-based routing and its operational complexities. Teaching note. Technion, Israel. http://iew3.technion.ac.il/serveng2004/Lectures/SBR.pdf(accessed 24th, April 2010)
46. Garnett, O., Mandelbaum, A. and Reiman, M. I. (2002) "Designing a call center with impatient customers," Manufacturing and Service Operations Management, M & SOM, vol. 4, pp. 208–227, http://ect.bell-labs.com/who/marty/pub/gmr.pdf(accessed 24th, April 2010)
47. Mandelbaum, A. (2002). Call centers. Research bibliography with abstracts.

Technical report, Technion, Israel Institute of Technology Version 3, 137 pages ie.technion.ac.il/serveng /References/ccbib.pdf. (accessed 24th, April 2010)

48. Mandelbaum, A., Massey, W.A., Reiman, M.I., Rider, R., and Stolyar, A. (2000). Queue lengths and waiting times for multiserver queues with abandonment and retrials. Proceedings of the Fifth IN-FORMS Telecommunications Conference http://iew3.technion.ac.il/serveng/References/references.html(accessed 24th, April 2010)
49. http://iew3.technion.ac.il/serveng/References/references.html(accessed 24th, April 2010)
50. Mandelbaum, A., Sakov, A. and Zeltyn, S. (2000). Empirical analysis of a call center. Technical Report, Technion, Israel Institute of Technology, http://iew3.technion.ac.il/serveng/References/references.html. (accessed 24th, April 2010)
51. Mandelbaum A. and Zeltyn S. (2004) The Palm/Erlang-A Queue, with Applications to Call Centers. Teaching note to Service Engineering course. Faculty of Industrial Engineering & Management Technion, Haifa 32000, ISRAEL http://iew3.technion.ac.il/serveng/References/references.html. (accessed 24th, April 2010)
52. Koole, G. (2002). Queueing Models of Call Centers: An Introduction, Technion, Israel Institute of Technology www.cs.vu.nl/obp/callcenters(accessed 24th, April 2010)
53. Paragon (2005), What is Simulation of processes / systems? Arena Contact Center, *www.erlang.com.br/simulacao.asp*(accessed 24th, April 2010)

Bibliography

1. Avramidis, A. and L'ecuyer, P. (2005), "Modeling and Simulation of Call Centers", Winter Simulation Conference, Orlando FL pp. 144-152.
2. Baccelli, F.and P. Bremaud, P. (2003). Elements of Queueing Theory. Springer-Verlag, Berlin, 2nd edition,
3. Borst, S., Mandelbaum, A., and Reiman, M. (2000), "Dimensioning of large call centers", Preprint. http://iew3.technion.ac.il/serveng/References/references.html.
4. Brown, L. D., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. (2002), "Multifactor Poisson and gamma-poisson models for call center arrival times," Technical Report, University of Pennsylvania.
5. Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyin, S. and Zhao, L. (2002), "Statistical analysis of a telephone call center: a queueing-science perspective" (working paper 03-12), Wharton Financial Institutions Center.
6. Foh, C. H., Zukerman, M. and NANYANG(2003) Poisson Arrivals See Time Averages, by TECHNOLOGICAL UNIVERSITY, tutorial series
7. AT&T. Call Center Statistics (2002) Call Center News Service Web Site, callcenternews.com/resources/statistics.shtml. (accessed 14th, Feb 2010)
8. Call Center Data (2002), Technion, Israel Institute of Technology. http://iew3.technion.ac.il/serveng/callcenterdata/index.html. (accessed 14th, Feb 2010)
9. Datamonitor. As reported on resources.talisma.com/ver call statistics.asp. (accessed 20th, March, 2010)

10. Heckley, G. (2010) Offshoring and the Labour Market: IT and Call Centres considered. Available from: http://www.statistics.gov.uk. (accessed 24th, April 2010)
11. Koole, G. (2010) Optimization of Business Processes: An Introduction to Applied Stochastic Modeling. Available from: http://obp.math.vu.nl/callcenters/ (accessed 22nd, May, 2010)
12. Mandelbaum, A., and Schwartz, R. (2002), "Simulation Experiments with M/G/100 Queues in the Halfin-Whitt (Q.E.D) Regime", Technical Report, Technion http://iew3.technion.ac.il/serveng/References/references.html. (accessed 24th, April 2010)
13. 4CallCenters Software (2002), http://iew3.technion.ac.il/serveng/4CallCenters/Downloads.htm. (accessed 25th, June 2010)
14. Koole, G.(2007) Call Center Mathematics A scientific method for understanding and improving contact Department of Mathematics, Vrije Universiteit Amsterdam, and CCmath consulting and software centers www.math.vu.nl/~koole/ccmath(accessed 29th, April 2010)
15. Reynolds, P.(2003) call centre staffing maths www.connectionsmagazine.com/articles/3/020.html (accessed 24th, April 2011)
16. Mandelbaum, A. and Zeltyn, S. (2007). The M/M/n+G queue: Summary of performance http://iew3.technion.ac.il/serveng/References/, (accessed 27th, June 2010)

17. Mandelbaum, A., Massey, W.A. and Reiman, M.I.(1998), Strong approximations forMarkovian service networks", Queueng Systems, Vol.30, , pp. 149-201, http://iew3.technion.ac.il/serveng/References/references.html (accessed 27th, April 2010)
18. Zeltyn, S. and Mandelbaum, A. (2005), "Call centers with impatient customers: Many server asymptotics of the M/M/n+G queue," Queueing Systems, vol. 51, pp. 361–402, Faculty of Industrial Engineering & Management Technion, Haifa 32000, ISRAEL, http://iew3.technion.ac.il/serveng/References/references.html(accessed 24th, April 2010)

Appendix A

Derivation of some Erlang-C performance measures

Derivation of Equations (3.6) and (3.9)

Denote $p_i(t)$ = P(Xt =i), the probability that there are i customers in the system at time t. Suppose a customer arrives into the system at time τ. Applying the law of total probability:

$$\mathrm{P}\,(W_Q \le T \setminus \tau) = \sum_{i=0}^{N} \mathrm{P}\,(W_Q \le T \setminus \mathrm{X}\tau = i)\mathrm{P}\,(\mathrm{X}\tau = i)$$

$$= \sum_{i=0}^{N} \mathrm{P}\,(W_Q \le T \setminus \mathrm{X}\tau = i)\mathrm{P}i\,(\tau) \qquad \text{(A.1)}$$

If $i < s$, then an arriving customer will be served immediately, so

$$\mathrm{P}\,(W_Q \le T \setminus \mathrm{X}\tau = i) = 1 \text{ for } i < s \qquad \text{(A. 2)}$$

If $i \ge s$, then the probability that the customer will be served within time T from now is equal to the probability that more than i-s customers are served and leave the system within time T. We have assumed that service times are exponential with rate 1/μ and so the number of customers leaving the system after being served follows a Poisson process. Hence, we have

$$\mathrm{P}\,(W_Q \le T \setminus \mathrm{X}\tau = i) = \mathrm{P}\,(Y > i - s)$$

$$= \sum_{k=i-s+1}^{\infty} \frac{(T\mu s)^k}{k!} e^{-T\mu s} \text{ for } i \ge s \qquad \text{(A. 3)}$$

where $Y \sim Po(T\mu s)$, is a Poisson distribution with mean $T\mu s$. Putting (A.2) and (A.3) into (A.1) yields

$$P(WQ \leq T \setminus \tau) = \sum_{i=0}^{i-s} Pi(\tau) + \sum_{i=s}^{N} \left(e^{-T\mu s} \sum_{k=i-s+1}^{\infty} \frac{(T\mu s)^k}{k!} \right) Pi(\tau) \qquad \text{(A. 4)}$$

This equation effectively gives an *instantaneous* grade of service; if a customer arrives into the system time τ, then this is the probability that they will be served within time *T*. However, as described in Section 1.2.2, grades of service are often calculated over periods of time such as an hour. Suppose we wish to calculate the grade of service over the interval $I = [t_0, t_1]$ where the time-dependent arrival rate, $\lambda(t)$, is defined on *I*. Then the probability that a call arrives into the system at time τ, given that it arrives in *I*, is given by

$$f(\tau) = \frac{\lambda(\tau)}{\int_{to}^{t1} \lambda(t)dt}$$

Now, using the continuous version of the law of total probability:

$$P(WQ \leq T) = \int_{to}^{t1} P(WQ \leq T \setminus \tau) f(\tau) d\tau$$

$$= \frac{\int_{to}^{t1} P(WQ \leq T \setminus \tau) f(\tau) d\tau}{\int_{to}^{t1} \lambda(t)dt} \qquad \text{(A. 5)}$$

where $P(WQ \leq T \setminus \tau)$ is given by (A.4).

We begin by deriving (3.6), using (3.4), (3.5) and (A.4). Indeed, assuming a constant arrival rate and putting (3.4) into (A.4) gives:

$$P(W_Q \leq T) = \sum_{i=0}^{i-s} \frac{a^i}{i!} \pi(0) + \sum_{i=s}^{\infty} \left(e^{-T\mu s} \sum_{k=i-s+1}^{\infty} \frac{(T\mu s)^k}{k!} \right) \frac{a^i}{s! s^{i-s}} \pi(0)$$

Then, re-arranging the above and using (3.5):

$$P(W_Q > T) = 1 - \sum_{i=0}^{i-s} \frac{a^i}{i!}\,\pi(0) - \sum_{i=s}^{\infty} \left(e^{-T\mu s} \sum_{k=i-s+1}^{\infty} \frac{(T\mu s)^k}{k!} \right) \frac{a^i}{s!s^{i-s}}\,\pi(0)$$

$$= (\pi(0)^{-1} - \sum_{i=0}^{i-s} \frac{a^i}{i!} - \sum_{i=s}^{\infty} \left(e^{-T\mu s} \sum_{k=i-s+1}^{\infty} \frac{(T\mu s)^k}{k!} \right) \frac{a^i}{s!s^{i-s}})\pi(0)$$

$$= (\frac{a^s}{(s-1)!(s-a)} - \sum_{i=s}^{\infty} \left(e^{-T\mu s} \sum_{k=i-s+1}^{\infty} \frac{(T\mu s)^k}{k!} \right) \frac{a^i}{s!s^{i-s}})\pi(0)$$

$$= \frac{a^s}{(s-1)!(s-a)} (1 - \sum_{i=s}^{\infty} \left(e^{-T\mu s} \sum_{k=i-s+1}^{\infty} \frac{(T\mu s)^k}{k!} \right) \frac{a^{i-s}(s-a)}{ss^{i-s}})\pi(0)$$

Now, defining

$$C(s,a) = \frac{a^s}{(s-1)!(s-a)}\,\pi(0)$$

Gives (3.7) nothing that

$$1 = (1-\rho) \sum_{j=0}^{\infty} \rho^j$$

$$= (1-\rho)\, e^{-T\mu s} \sum_{j=0}^{\infty} \rho^j \sum_{k=0}^{\infty} \frac{(T\mu s)^k}{k!}$$

letting j= i - s

$$P(W > T) = C(s,a)\, e^{-T\mu s}((1-\rho) \sum_{j=0}^{\infty} \rho^j \sum_{k=0}^{\infty} \frac{(T\mu s)^k}{k!} - (1-\rho) \sum_{j=0}^{\infty} \rho^j \sum_{k=j+1}^{\infty} \frac{(T\mu s)^k}{k!})$$

$$= C(s,a)\, e^{-T\mu s}((1-\rho) \sum_{j=0}^{\infty} \rho^j \sum_{k=0}^{j} \frac{(T\mu s)^k}{k!})$$

Changing the order of summation:

$$P(W_Q > T) = C(s,a)\, e^{-T\mu s}((1-\rho) \sum_{k=0}^{\infty} \frac{(T\mu s)^k}{k!} \sum_{j=k}^{\infty} \rho^j)$$

$$=C(s,a)\, e^{-T\mu s}((1-\rho)\sum_{k=0}^{\infty}\frac{(T\mu s)^k}{k!}\frac{\rho^k}{(1-\rho)})$$

$$=C(s,a)\, e^{-T\mu s}(\sum_{k=0}^{\infty}\frac{(\rho T\mu s)^k}{k!})$$

$$=C(s,a)\, e^{-T\mu s}e^{\rho T\mu s}$$

However, $\rho T\mu s = \lambda T$ and so:

$$P(W_Q > T) = C(s,a)\, e^{-T\mu s}e^{\lambda T}$$

$$= C(s,a)\, e^{-(\mu s-\lambda)T} \qquad \text{(A. 5i)}$$

which is precisely (3.6). To derive (3.8), denote the probability density of the waiting time distribution by

$$f(T) = \frac{d}{dT}P(W_Q \le T)$$

$$=(\mu s-\lambda)C(s,a)\, e^{-(\mu s-\lambda)T}$$

Then

$$E\, W_Q = \int_0^{\infty} Tf(T)dT$$

$$=(\mu s-\lambda)C(s,a)\int_0^{\infty} Te^{-(\mu s-\lambda)T}dT$$

$$=(\mu s-\lambda)C(s,a(\,[T\frac{e^{-(\mu s-\lambda)T}}{-(\mu s-\lambda)}]_0^{\infty}-\int_0^{\infty}\frac{e^{-(\mu s-\lambda)T}}{-(\mu s-\lambda)}dT)$$

but the first term is zero and so

$$EW_Q = -(\mu s-\lambda)C(s,a)[\frac{e^{-(\mu s-\lambda)T}}{(\mu s-\lambda)^2}]_0^{\infty}$$

$$=\frac{C(s,a)}{\mu s-\lambda} \qquad \text{(A. 5ii)}$$

Finally, (3.9) follows from Little's Law from (Optimization of Business Process, Koole,200). This tells us that

$$\mathrm{E}L_Q \quad = \quad \lambda\, \mathrm{E}W_Q$$

$$= \frac{\rho C(s,a)}{1-\rho} \qquad \text{(A. 5iii)}$$

Derivation of some Erlang-A performance measures

Steady-state distribution. Using formulae (3.15), (3.16) and definition (3.17) one gets

$$\pi_0{}^{-1} = \sum_{j=0}^{s} \frac{(\lambda/\mu)^j}{j!} + \frac{(\lambda/\mu)^s}{s!} \cdot \sum_{j=s+1}^{\infty} \prod_{k=s+1}^{j} \left(\frac{\lambda}{s\mu + (k-s)\gamma}\right)$$

$$= \frac{(\lambda/\mu)^s}{s!} \cdot \left[\frac{1}{E_{1,s}} + \sum_{j=1}^{\infty} \frac{(\lambda/\gamma)^j}{\prod_{k=1}^{j} \left(\frac{s\mu}{\gamma} + k\right)}\right] = \frac{(\lambda/\mu)^s}{s!} \cdot \left[\frac{1}{E_{1,s}} + A\left(\frac{s\mu}{\gamma}, \frac{\lambda}{\gamma}\right) - 1\right]$$

Hence

$$\pi_0 = \frac{E_{1,s}}{1 + \left[A\left(\frac{s\mu}{\gamma}, \frac{\lambda}{\gamma}\right) - 1\right] . E_{1,s}} \cdot \frac{s!}{(\lambda/\mu)^s}$$

For $1 \leq j \leq s$

$$\pi_j = \pi_0 . \frac{(\lambda/\mu)^j}{j!} = \frac{E_{1,s}}{1 + \left[A\left(\frac{s\mu}{\gamma}, \frac{\lambda}{\gamma}\right) - 1\right] . E_{1,s}} \cdot \frac{s!}{j! . (\lambda/\mu)^{s-j}}$$

Specifically,

$$\pi_s = \frac{E_{1,s}}{1 + \left[A\left(\frac{s\mu}{\gamma}, \frac{\lambda}{\gamma}\right) - 1\right] . E_{1,s}} \qquad \text{(A. 6)}$$

Finally, for j>s

$$\pi_j=\pi_s.\frac{\lambda^{j-s}}{\prod_{k=1}^{j-s}(s\mu+k\gamma)}=\frac{E_{1,s}}{1+\left[A\left(\frac{s\mu}{\gamma},\frac{\lambda}{\gamma}\right)-1\right].E_{1,s}}.\frac{(\frac{\lambda}{\gamma})^{j-s}}{\prod_{k=1}^{j-s}(\frac{s\mu}{\gamma}+k)} \quad (A.7)$$

Probability of wait. From *Poisson Arrivals See Time Averages* (PASTA) , (A.6) and (A.7), the delay probability is equal to

$$P\{W>0\}=\sum_{j=s}^{\infty}\pi_j=\frac{E_{1,s}}{1+\left[A\left(\frac{s\mu}{\gamma},\frac{\lambda}{\gamma}\right)-1\right].E_{1,s}}.\left[1+\sum_{j=s+1}^{\infty}\frac{(\lambda/\gamma)^{j-s}}{\prod_{k=1}^{j-s}(\frac{s\mu}{\gamma}+k)}\right]$$

$$=\frac{A\left(\frac{s\mu}{\gamma},\frac{\lambda}{\gamma}\right).E_{1,s}}{1+\left[A\left(\frac{s\mu}{\gamma},\frac{\lambda}{\gamma}\right)-1\right].E_{1,s}} \quad (A.8)$$

Probability to abandon. Preliminary calculations required. Differentiating (3.16), gives

$$\frac{\delta}{\delta y}A(x,y)=\frac{\delta}{\delta y}\left[\frac{xe^y}{y^x}\gamma(x,y)\right]=\frac{x}{y}+\left(1-\frac{x}{y}\right).A(x,y).$$

Then, for x > 0, y > 0,

$$\sum_{j=0}^{\infty}\frac{(j+1)y^j}{\prod_{k=1}^{j+1}(x+k)}=\frac{\delta}{\delta y}\left[\sum_{j=1}^{\infty}\frac{y^j}{\prod_{k=1}^{j}(x+k)}\right]$$

$$=\frac{\delta}{\delta y}[A(x,y)-1]=\frac{\delta}{\delta y}A(x,y)=\frac{x}{y}+\left(1-\frac{x}{y}\right).A(x,y) \quad (A.9)$$

Using (A.8) and (3.21), the conditional probability to abandon is equal to

$$P\{Ab|W>0\}=\frac{\sum_{j=s}^{\infty}\pi_j.P_{j-s\{Ab\}}}{P\{W>0\}}$$

$$=\frac{1}{A\left(\frac{s\mu}{\gamma},\frac{\lambda}{\gamma}\right)}.\sum_{j=s}^{\infty}\frac{(\lambda/\gamma)^{j-s}}{\prod_{k=1}^{j-s}(\frac{s\mu}{\gamma}+k)}.\frac{\gamma(j+1-s)}{s\mu+\gamma(j+1-s)}$$

(by convention, $\prod_{k=1}^{0}(\frac{s\mu}{\gamma}+k) \triangleq 1$)

$$= \frac{1}{A\left(\frac{s\mu}{\gamma},\frac{\lambda}{\gamma}\right)} \cdot \sum_{j=0}^{\infty} \frac{\left(\frac{\lambda}{\gamma}\right)^{j} \cdot (j+1)}{\prod_{k=1}^{j+1}(\frac{s\mu}{\gamma}+k)} = \frac{1}{A\left(\frac{s\mu}{\gamma},\frac{\lambda}{\gamma}\right)} \cdot \frac{\delta}{\delta y}\left[A\left(\frac{s\mu}{\gamma}, y\right)\right]_{y=\lambda/\gamma}$$

$$= \frac{1}{A\left(\frac{s\mu}{\gamma},\frac{\lambda}{\gamma}\right)} \cdot [\frac{s\mu}{\lambda} + (1-\frac{s\mu}{\lambda})A\left(\frac{s\mu}{\gamma},\frac{\lambda}{\gamma}\right)] = \frac{1}{\rho A\left(\frac{s\mu}{\gamma},\frac{\lambda}{\gamma}\right)} + 1 - \frac{1}{\rho}$$

where the last line follows from (A.9).

Appendix B

Simulation Results

Table 4-5 Summary of 100 simulation for 2, 3,4 and 5 servers

	2 server			3 server			4 server			5 server		
Day	No of Cust	Longest wait	Avg wait	No of Cust	Longest wait	Avg wait	No of Cust	Longest wait	Avg wait	No of Cust	Longest wait	Avg wait
1	129	5	0	130	1	0	133	0	0	130	0	0
2	130	5	0	131	3	0	135	0	0	132	0	0
3	131	6	0	134	3	0	135	0	0	135	0	0
4	133	5	0	134	3	0	136	0	0	135	0	0
5	134	6	0	135	0	0	136	0	0	135	0	0
6	135	8	0	136	3	0	136	0	0	136	0	0
7	135	8	0	137	3	0	136	0	0	136	0	0
8	135	12	0	137	3	0	137	0	0	137	0	0
9	137	6	0	138	0	0	137	1	0	137	0	0
10	137	8	0	138	3	0	138	0	0	137	0	0
11	137	5	0	139	3	0	138	0	0	138	0	0
12	138	8	0	139	1	0	139	0	0	138	0	0
13	139	8	1	140	3	0	139	0	0	138	0	0
14	139	7	0	140	3	0	139	0	0	138	0	0
15	139	7	0	140	1	0	140	1	0	138	0	0
16	139	7	0	140	3	0	140	1	0	138	0	0
17	140	5	0	141	3	0	140	0	0	138	0	0
18	140	7	0	141	1	0	140	1	0	138	0	0
19	140	6	0	141	1	0	140	1	0	139	0	0
20	140	10	0	141	1	0	140	0	0	139	0	0
21	140	8	0	141	3	0	140	0	0	139	0	0
22	141	9	0	141	3	0	141	0	0	140	0	0
23	141	15	1	141	3	0	141	1	0	141	0	0
24	141	6	0	141	3	0	141	1	0	141	0	0
25	141	8	0	142	3	0	141	0	0	141	0	0
26	141	11	1	142	3	0	141	0	0	141	0	0
27	141	8	0	142	3	0	142	0	0	141	0	0
28	141	9	0	142	3	0	142	0	0	141	0	0
29	141	11	1	142	3	0	142	0	0	141	0	0
30	142	10	0	142	3	0	142	0	0	141	0	0
31	142	5	0	142	3	0	142	1	0	141	0	0
32	142	7	0	142	3	0	142	0	0	142	0	0
33	142	13	1	143	3	0	143	0	0	142	0	0
34	142	7	0	143	3	0	143	0	0	142	0	0

35	142	6	0	143	3	0	143	1	0	142	0	0
36	143	9	1	143	3	0	143	1	0	142	0	0
37	143	8	0	143	3	0	144	1	0	142	0	0
38	143	12	1	144	3	0	144	1	0	142	0	0
39	143	11	0	144	0	0	144	1	0	143	0	0
40	144	10	1	144	3	0	144	1	0	143	0	0
41	144	12	1	144	3	0	144	0	0	144	0	0
42	144	5	0	144	3	0	144	1	0	144	0	0
43	144	7	0	145	4	0	144	1	0	144	0	0
44	144	8	1	145	3	0	144	0	0	144	0	0
45	145	5	0	145	3	0	144	1	0	144	0	0
46	145	7	0	145	3	0	144	1	0	144	0	0
47	145	11	0	145	3	0	145	0	0	144	0	0
48	145	9	1	145	1	0	145	0	0	144	0	0
49	146	8	0	145	3	0	145	1	0	144	0	0
50	146	12	1	146	3	0	145	1	0	145	0	0
51	146	8	0	146	3	0	145	0	0	145	0	0
52	146	8	0	146	3	0	145	0	0	145	0	0
53	146	6	0	146	3	0	145	0	0	145	0	0
54	147	6	0	146	1	0	145	1	0	145	0	0
55	147	13	1	147	3	0	145	0	0	145	0	0
56	147	5	0	147	3	0	146	0	0	146	0	0
57	147	6	0	147	3	0	146	1	0	146	0	0
58	147	9	0	147	3	0	146	0	0	146	0	0
59	148	12	1	147	3	0	147	1	0	146	0	0
60	148	8	0	147	3	0	147	1	0	146	0	0
61	148	12	1	147	3	0	147	1	0	146	0	0
62	148	10	0	148	3	0	147	0	0	146	0	0
63	148	19	0	148	3	0	147	0	0	146	0	0
64	148	5	0	148	1	0	147	0	0	146	0	0
65	148	8	1	148	3	0	148	0	0	147	0	0
66	148	10	0	148	1	0	148	1	0	147	0	0
67	148	8	0	148	3	0	148	1	0	147	0	0
68	148	12	1	148	3	0	148	0	0	147	0	0
69	148	6	0	149	3	0	148	1	0	148	0	0
70	148	10	0	149	3	0	149	0	0	148	0	0
71	148	10	1	149	3	0	149	1	0	148	0	0
72	148	10	1	149	3	0	149	1	0	149	0	0
73	149	6	0	149	3	0	149	0	0	149	0	0
74	149	8	1	149	3	0	149	0	0	149	0	0
75	150	16	2	149	3	0	150	1	0	149	0	0
76	150	8	0	150	1	0	150	0	0	149	0	0
77	150	7	1	150	3	0	150	1	0	150	0	0
78	150	14	1	150	3	0	151	0	0	150	0	0
79	150	9	1	150	3	0	151	1	0	150	0	0

80	150	7	1	151	3	0	151	0	0	151	0	0
81	150	6	0	151	3	0	151	1	0	151	0	0
82	151	12	1	151	3	0	151	0	0	151	0	0
83	151	8	1	152	3	0	151	0	0	151	0	0
84	151	16	0	152	3	0	151	1	0	151	0	0
85	151	7	0	152	3	0	152	1	0	151	0	0
86	151	8	0	152	3	0	152	1	0	152	0	0
87	151	10	0	152	3	0	153	1	0	152	0	0
88	152	10	1	153	4	0	154	0	0	152	0	0
89	152	12	1	153	3	0	154	0	0	153	0	0
90	152	14	1	153	3	0	154	1	0	154	0	0
91	152	8	2	153	3	0	154	1	0	155	0	0
92	153	11	1	154	3	0	155	0	0	155	0	0
93	155	14	2	154	3	0	155	1	0	155	0	0
94	155	10	1	155	3	0	156	1	0	157	0	0
95	155	10	1	156	4	0	156	1	0	157	0	0
96	156	8	1	156	6	0	156	1	0	158	0	0
97	157	8	1	156	3	0	156	1	0	159	0	0
98	159	10	0	156	6	0	159	1	0	159	0	0
99	159	12	2	157	3	0	159	0	0	160	0	0
100	160	12	1	163	3	0	160	1	0	160	0	0

Table 4-6	Highest Maximum & Average Waiting time			
	Number of Servers			
	2	**3**	**4**	**5**
max longest wait(mins)	19	6	1	0
max average wait(mins)	2	0	0	0

Table 4-7 No of customers served by each CRS per server type for 20 days

Day	2 server			3 server				4 server					5 server					
	No of Cust	CRS1	CRS2	**No of Cust**	CRS1	CRS2	CRS3	**No of Cust**	CRS1	CRS2	CRS3	CRS4	**No of Cust**	CRS1	CRS2	CRS3	CRS4	CRS5
1	139	82	57	134	75	45	14	136	74	45	15	2	136	77	44	13	2	0
2	140	81	59	138	77	46	15	138	78	41	13	6	137	75	44	15	3	0
3	140	82	58	138	79	49	10	140	80	44	13	3	137	73	48	14	2	0
4	140	80	60	140	76	47	17	140	80	43	16	1	138	75	48	11	3	1
5	142	85	57	142	81	47	14	141	80	46	13	2	138	75	48	13	2	0
6	142	86	56	142	74	52	16	141	74	42	20	5	138	79	42	14	3	0
7	143	83	60	142	78	46	18	142	78	50	12	2	141	72	47	16	5	1
8	144	79	65	142	76	48	18	144	75	46	19	4	142	78	44	16	3	1
9	144	82	62	143	79	45	19	144	80	46	18	0	144	76	51	13	4	0
10	144	85	59	143	78	47	18	148	77	53	17	1	144	78	49	16	1	0
11	144	85	59	143	82	47	14	148	77	54	14	3	146	80	44	17	4	1
12	145	85	60	143	75	50	18	149	74	51	20	4	147	79	44	18	5	1
13	146	87	59	147	80	49	18	149	76	48	18	7	147	71	53	19	3	1
14	147	86	61	147	83	47	17	149	78	51	17	3	147	78	45	21	3	0
15	147	80	67	148	78	52	18	150	74	48	22	6	148	76	49	17	6	0
16	148	84	64	149	80	50	19	150	76	51	22	1	150	73	48	19	9	1
17	152	84	68	152	74	55	23	151	78	49	20	4	150	78	47	22	3	0
18	152	82	70	152	78	52	22	152	80	46	21	5	154	75	53	22	3	1
19	155	84	71	153	76	53	24	154	80	51	19	4	154	78	49	19	6	2
20	156	86	70	157	81	56	20	155	82	46	21	6	158	77	55	21	4	1

Table 4-5 Maximum & Minimum No of customers served by each CRS per server type for 20 days																		
	2			3				4					5					
	No of Cust	CRS1	CRS2	No of Cust	CRS1	CRS2	CRS3	No of Cust	CRS1	CRS2	CRS3	CRS4	No of Cust	CRS1	CRS2	CRS3	CRS4	CRS5
max	158	87	72	161	83	57	24	158	82	54	25	7	161	81	55	23	9	2
min	139	79	56	134	74	45	10	136	74	41	12	0	136	71	42	11	1	0

Appendix C

Sample call data for a day

Table C1 Sample call data for a day

Cust #	Beg (hr:min)	end (hr:min)	waite_time (hr:min)	total time (hr:min)	Cust #	Beg (hr:min)	end (hr:min)	waite_time (hr:min)	total time (hr:min)
1	7:07	7:09	0:00	0:02	51	10:58	11:05	0:00	0:07
2	7:12	7:21	0:00	0:09	52	11:05	11:08	0:00	0:03
3	7:14	7:23	0:00	0:09	53	11:07	11:14	0:00	0:07
4	7:19	7:26	0:00	0:07	54	11:09	11:16	0:00	0:07
5	7:24	7:27	0:00	0:03	55	11:14	11:23	0:00	0:09
6	7:33	7:42	0:00	0:09	56	11:19	11:24	0:00	0:05
7	7:35	7:42	0:00	0:07	57	11:26	11:35	0:00	0:09
8	7:37	7:46	0:00	0:09	58	11:35	11:40	0:00	0:05
9	7:44	7:49	0:00	0:05	59	11:37	11:47	0:00	0:10
10	7:49	7:58	0:00	0:09	60	11:46	11:55	0:00	0:09
11	7:54	8:03	0:00	0:09	61	11:51	11:56	0:00	0:05
12	8:01	8:03	0:00	0:02	62	11:58	12:03	0:00	0:05
13	8:06	8:15	0:00	0:09	63	12:07	12:14	0:00	0:07
14	8:08	8:15	0:00	0:07	64	12:09	12:14	0:00	0:05
15	8:10	8:19	0:00	0:09	65	12:14	12:19	0:00	0:05
16	8:15	8:24	0:00	0:09	66	12:16	12:25	0:00	0:09
17	8:22	8:27	0:00	0:05	67	12:18	12:26	0:00	0:08
18	8:29	8:34	0:00	0:05	68	12:25	12:34	0:00	0:09
19	8:34	8:39	0:00	0:05	69	12:32	12:41	0:00	0:09
20	8:39	8:48	0:00	0:09	70	12:34	12:41	0:00	0:07
21	8:41	8:50	0:00	0:09	71	12:43	12:48	0:00	0:05
22	8:43	8:52	0:00	0:09	72	12:45	12:50	0:00	0:05
23	8:48	8:55	0:00	0:07	73	12:47	12:54	0:00	0:07
24	8:50	8:55	0:00	0:05	74	12:56	13:05	0:00	0:09
25	8:52	8:57	0:00	0:05	75	13:01	13:06	0:00	0:05
26	8:59	9:08	0:00	0:09	76	13:08	13:15	0:00	0:07
27	9:08	9:15	0:00	0:07	77	13:13	13:20	0:00	0:07
28	9:13	9:22	0:00	0:09	78	13:15	13:26	0:00	0:11
29	9:20	9:25	0:00	0:05	79	13:17	13:22	0:00	0:05
30	9:27	9:32	0:00	0:05	80	13:24	13:29	0:00	0:05
31	9:29	9:33	0:00	0:04	81	13:26	13:35	0:00	0:09
32	9:34	9:44	0:00	0:10	82	13:31	13:40	0:00	0:09

33	9:36	9:41	0:00	0:05	83	13:33	13:42	0:00	0:09
34	9:38	9:47	0:00	0:09	84	13:38	13:47	0:00	0:09
35	9:40	9:49	0:00	0:09	85	13:43	13:52	0:00	0:09
36	9:42	9:51	0:00	0:09	86	13:48	13:55	0:00	0:07
37	9:47	9:52	0:00	0:05	87	13:53	13:58	0:00	0:05
38	9:52	9:59	0:00	0:07	88	14:00	14:05	0:00	0:05
39	9:54	9:57	0:00	0:04	89	14:07	14:12	0:00	0:05
40	9:56	10:03	0:00	0:07	90	14:09	14:18	0:00	0:09
41	9:58	10:07	0:00	0:09	91	14:11	14:16	0:00	0:05
42	10:03	10:08	0:00	0:05	92	14:18	14:23	0:00	0:05
43	10:08	10:13	0:00	0:05	93	14:27	14:36	0:00	0:09
44	10:13	10:18	0:00	0:05	94	14:29	14:38	0:00	0:09
45	10:20	10:26	0:00	0:06	95	14:34	14:43	0:00	0:09
46	10:27	10:32	0:00	0:05	96	14:36	14:43	0:00	0:07
47	10:34	10:41	0:00	0:07	97	14:41	14:50	0:00	0:09
48	10:39	10:42	0:00	0:03	98	14:46	14:51	0:00	0:05
49	10:46	10:55	0:00	0:09	99	14:51	15:00	0:00	0:09
50	10:51	10:56	0:00	0:05	100	14:58	15:05	0:00	0:07

Appendix D

Table D-4 Sample simulation results for a five server(agents)														
Cust.	Interarrival	Arrival	Service	Server #1		Server #2		Server #3		Server #4		Server #5		Wait
#	Time	Time	Time	Start	End	Start	End	Start	End	Start	End	Start	End	Time
	(min)	(hr:min)	(min)	(hr:min)	(hr:min)	(hr:min)	(hr:min)	(hr:min)	(hr:min)	(hr:min)	(hr:min)	(hr:min)	(hr:min)	(hr:min)
Start		7:00												
1	2	7:02	7	7:02	7:09						0:00			0:00
2	5	7:07	9			7:07	7:16							0:00
3	2	7:09	7	7:09	7:16									0:00
4	2	7:11	9					7:11	7:20					0:00
5	7	7:18	5	7:18	7:23									0:00
6	2	7:20	9			7:20	7:29							0:00
7	2	7:22	7					7:22	7:29					0:00
8	7	7:29	5	7:29	7:34									0:00
9	5	7:34	9	7:34	7:43									0:00
10	2	7:36	9			7:36	7:45							0:00
11	2	7:38	7					7:38	7:45					0:00
12	7	7:45	9	7:45	7:54									0:00
13	7	7:52	9			7:52	8:01							0:00
14	2	7:54	9	7:54	8:03									0:00
15	7	8:01	5			8:01	8:06							0:00
16	2	8:03	7	8:03	8:10									0:00
17	5	8:08	5			8:08	8:13							0:00
18	2	8:10	9	8:10	8:19									0:00
19	5	8:15	9			8:15	8:24							0:00
20	2	8:17	9					8:17	8:26					0:00
21	2	8:19	9	8:19	8:28									0:00
22	5	8:24	9			8:24	8:33							0:00

23	5	8:29	5	8:29	8:34									0:00
24	5	8:34	5	8:34	8:39									0:00
25	2	8:36	9			8:36	8:45							0:00
26	7	8:43	9	8:43	8:52									0:00
27	2	8:45	9			8:45	8:54							0:00
28	9	8:54	9	8:54	9:03									0:00
29	5	8:59	9			8:59	9:08							0:00
30	7	9:06	7	9:06	9:13									0:00
31	2	9:08	9			9:08	9:17							0:00
32	7	9:15	9	9:15	9:24									0:00
33	7	9:22	9			9:22	9:31							0:00
34	2	9:24	5	9:24	9:29									0:00
35	2	9:26	7					9:26	9:33					0:00
36	7	9:33	9	9:33	9:42									0:00
37	2	9:35	9			9:35	9:44							0:00
38	5	9:40	5					9:40	9:45					0:00
39	2	9:42	5	9:42	9:47									0:00
40	5	9:47	7	9:47	9:54									0:00
41	7	9:54	5	9:54	9:59									0:00
42	7	10:01	9	10:01	10:10									0:00
43	7	10:08	9			10:08	10:17							0:00
44	5	10:13	5	10:13	10:18									0:00
45	9	10:22	9	10:22	10:31									0:00
46	7	10:29	5			10:29	10:34							0:00
47	5	10:34	9	10:34	10:43									0:00
48	7	10:41	9			10:41	10:50							0:00
49	5	10:46	9	10:46	10:55									0:00
50	5	10:51	9			10:51	11:0							0:00

							0							
51	9	11:00	5	11:00	11:05									0:00
52	7	11:07	9	11:07	11:16									0:00
53	9	11:16	5	11:16	11:21									0:00
54	2	11:18	9			11:18	11:27							0:00
55	5	11:23	9	11:23	11:32									0:00
56	7	11:30	7			11:30	11:37							0:00
57	7	11:37	9	11:37	11:46									0:00
58	5	11:42	9			11:42	11:51							0:00
59	2	11:44	9					11:44	11:53					0:00
60	2	11:46	7	11:46	11:53									0:00
61	7	11:53	7	11:53	12:00									0:00
62	9	12:02	9	12:02	12:11									0:00
63	7	12:09	9			12:09	12:18							0:00
64	5	12:14	9	12:14	12:23									0:00
65	7	12:21	7			12:21	12:28							0:00
66	2	12:23	7	12:23	12:30									0:00
67	9	12:32	5	12:32	12:37									0:00
68	9	12:41	5	12:41	12:46									0:00
69	2	12:43	9			12:43	12:52							0:00
70	5	12:48	7	12:48	12:55									0:00
71	7	12:55	7	12:55	13:02									0:00
72	7	13:02	7	13:02	13:09									0:00
73	5	13:07	7			13:07	13:14							0:00
74	2	13:09	9	13:09	13:18									0:00
75	2	13:11	5					13:11	13:16					0:00

76	5	13:16	7			13:16	13:23							0:00
77	5	13:21	9	13:21	13:30									0:00
78	7	13:28	9			13:28	13:37							0:00
79	2	13:30	5	13:30	13:35									0:00
80	9	13:39	9	13:39	13:48									0:00
81	7	13:46	7			13:46	13:53							0:00
82	2	13:48	9	13:48	13:57									0:00
83	7	13:55	5			13:55	14:00							0:00
84	7	14:02	9	14:02	14:11									0:00
85	7	14:09	7			14:09	14:16							0:00
86	5	14:14	5	14:14	14:19									0:00
87	5	14:19	7	14:19	14:26									0:00
88	7	14:26	7	14:26	14:33									0:00
89	5	14:31	5			14:31	14:36							0:00
90	7	14:38	5	14:38	14:43									0:00
91	2	14:40	9			14:40	14:49							0:00
92	5	14:45	9	14:45	14:54									0:00
93	5	14:50	5			14:50	14:55							0:00
94	5	14:55	7	14:55	15:02									0:00
95	5	15:00	9			15:00	15:09							0:00
96	7	15:07	9	15:07	15:16									0:00
97	2	15:09	9			15:09	15:18							0:00
98	7	15:16	9	15:16	15:25									0:00
99	7	15:23	9			15:23	15:32							0:00

100	2	15:25	9	15:25	15:34									0:00
101	5	15:30	5					15:30	15:35					0:00
102	7	15:37	5	15:37	15:42									0:00
103	7	15:44	5	15:44	15:49									0:00
104	5	15:49	5	15:49	15:54									0:00
105	5	15:54	9	15:54	16:03									0:00
106	9	16:03	5	16:03	16:08									0:00
107	2	16:05	9			16:05	16:14							0:00
108	2	16:07	9					16:07	16:16					0:00
109	5	16:12	9	16:12	16:21									0:00
110	5	16:17	7			16:17	16:24							0:00
111	5	16:22	5	16:22	16:27									0:00
112	5	16:27	9	16:27	16:36									0:00
113	2	16:29	7			16:29	16:36							0:00
114	7	16:36	9	16:36	16:45									0:00
115	7	16:43	7			16:43	16:50							0:00
116	2	16:45	5	16:45	16:50									0:00
117	2	16:47	9					16:47	16:56					0:00
118	2	16:49	5							16:49	16:54			0:00
119	5	16:54	5	16:54	16:59									0:00
120	9	17:03	7	17:03	17:10									0:00
121	5	17:08	9			17:08	17:17							0:00
122	2	17:10	9	17:10	17:19									0:00
123	7	17:17	9			17:17	17:26							0:00
124	7	17:24	7	17:24	17:31									0:00
125	5	17:29	5			17:29	17:34							0:00

126	2	17:31	5	17:31	17:36									0:00
127	5	17:36	9	17:36	17:45									0:00
128	7	17:43	7			17:43	17:50							0:00
129	7	17:50	5	17:50	17:55									0:00
130	2	17:52	9			17:52	18:01							0:00
131	7	17:59	7	17:59	18:06									0:00
132	7	18:06	9	18:06	18:15									0:00
133	7	18:13	7			18:13	18:20							0:00
134	5	18:18	7	18:18	18:25									0:00
135	5	18:23	5			18:23	18:28							0:00
136	7	18:30	9	18:30	18:39									0:00
137	7	18:37	9			18:37	18:46							0:00
138	5	18:42	7	18:42	18:49									0:00
139	7	18:49	9	18:49	18:58									0:00
140	7	18:56	5			18:56	19:01							0:00
141	2	18:58	5	18:58	19:03									0:00

Appendix E

Kendall's Notation

Kendall's notation is used in queueing theory to classify different queueing systems.

Its general form is

A/B/C/k/N/D + E:

Here, **A** refers to the arrival process, whilst **B** refers to the service time distribution.

If a homogeneous Poisson arrival process and exponential service times are assumed, then **A** and **B** are both written as M, corresponding to the Markovian.

Several other codes are common, including G, corresponding to a general distribution. Some authors write this as GI in order to emphasize the fact that arrivals/service times are independent.

If a process is an inhomogeneous Poisson process (a Poisson process with a time-varying rate) then this is written M_t.

C corresponds to the number of servers or, in the case of call centres, agents.

k refers to the capacity of the system, which is the number of agents plus the number of places in the queue for the call centre. Once this capacity is filled, further arrivals are blocked and are prevented from entering the system. Note that, if k = C then there is never a queue; arrivals can only enter the system when there is a free server/agent.

N refers to the calling population. This is often assumed to be infinite when the calling population is large compared to the number of agents.

D is the queueing discipline. This is typically first in, first out (FIFO) but others are possible, such as last in, first out (LIFO).

Finally, if the queueing system includes abandonments, then the patience distribution is represented as **+E**.

When **k** = ∞, **N** = ∞, **D** = FIFO and no abandonments are assumed, these are often omitted and Kendall's notation becomes **A/B/C**.

A Glossary of Call-Center Acronyms

Acronym	Description	Definition
ACD	automatic call distributor	p. 15
ANI	automatic number identification	p. 15
ASA	average speed of answer	p. 18
CRM	customer relationship management	p. 15
CSR	customer service representative	p. 7
CTI	computer-telephony integration	p. 15
DNIS	dialed number identification service	p. 15
PABX	private automatic branch exchange (also called PBX)	p. 15
PBX	private automatic branch exchange (also called PABX)	p. 15
PSTN	public switched telephone network	p. 15
TSF	telephone service factor (also called the 'service level')	p. 10
VRU	interactive voice response unit (also called IVR)	p. 15
WFM	workforce management	
IVR	interactive voice response unit (also called VRU)	