APPLICATION OF MULTISERVER QUEUEING TO CALL CENTRES

A THESIS SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS
FOR MASTER OF SCIENCE IN
MATHEMATICAL STATISTICS DEGREE
IN THE FACULTY OF SCIENCE

By

JACOB MAJAKWARA

Supervisor: Professor I. Szyszkowski

RHODES UNIVERSITY
FACULTY OF SCIENCE
DEPARTMENT OF STATISTICS
JUNE 2009

.....to all who are oppressed by the ignorance of others and to everyone who is seeking himself or herself. Remember to take a chance! All life is a chance. The person who goes furthest is generally the one who is willing to do and dare. The "sure thing" boat never gets far from the shore. Also remember that the purpose of life is to live a life of purpose.....

Abstract

The simplest and most widely used queueing model in call centres is the $M/M/k$ system, sometimes referred to as Erlang-C. For many applications the model is an over-simplification. Erlang-C model ignores among other things busy signals, customer impatience and services that span multiple visits. Although the Erlang-C formula is easily implemented, it is not easy to obtain insight from its answers (for example, to find an approximate answer to questions such as "how many additional agents do I need if the arrival rate doubles?"). An approximation of the Erlang-C formula that gives structural insight into this type of question would be of use to better understand economies of scale in call centre operations. Erlang-C based predictions can also turn out highly inaccurate because of violations of underlying assumptions and these violations are not straightforward to model. For example, non-exponential service times lead one to the $M/G/k$ queue which, in stark contrast to the $M/M/k$ system, is difficult to analyse.

This thesis deals mainly with the general $M/GI/k$ model with abandonment. The arrival process conforms to a Poisson process, service durations are independent and identically distributed with a general distribution, there are k servers, and independent and identically distributed customer abandoning times with a general distribution. This thesis will endeavour to analyse call centres using $M/GI/k$ model with abandonment and the data to be used will be simulated using EZSIM-software. The paper by Brown *et al* [3] entitled "Statistical Analysis of a Telephone Call Centre: A Queueing-Science Perspective," will be the basis upon which this thesis is built.

Key Words call centre; contact centre; queueing theory; multiserver queue; Poisson process; service times; arrival times; uniform distribution; exponential distribution; multiserver queue with customer abandonment.

ACKNOWLEDGMENT

Professor Radloff as the Head of Statistics department made a meaningful contribution by allowing my registration in the Statistics department, otherwise I would not have found myself at Rhodes University.

Professor I. Szyszkowski made this research to be what it is through his professional advice and constant guidance throughout. Without him, this research would not have been undertaken. The rest of the Statistics department staff helped in one way or the other for the completion of the thesis. Also fellow mates: Opeoluwa Oyedele and Greg Webster offered their help in editing this thesis. Lastly my family deserves mentioning as they persevered in my absence. This research was made possible by the Scholarship offered by Andrew Mellon Foundation.[1]

[1] The financial assistance from ***Andrew Mellon Foundation*** towards this research is hereby acknowledged. Opinions and views expressed and conclusions arrived at are those of the author and do not necessary reflect those of Rhodes University or the Mellon Foundation.

Contents

1 INTRODUCTION 1

1.1 Background 1

1.2 Aims and Objectives 2

1.3 Significance of the Thesis 2

1.4 Layout of the Thesis 4

2 QUEUEING THEORY 5

2.0.1 Why Queueing Theory? 5

2.1 Probability Distributions 6

2.1.1 Discrete Probability Distributions 8

2.1.1.1 Bernoulli Distribution 8

2.1.1.2 Binomial Distribution 8

2.1.1.3 Poisson Distribution 9

2.1.2 Continuous Probability Distributions 10

2.1.2.1 Uniform Distribution 10

2.1.2.2 Exponential Distribution 10

2.1.2.3 Gamma Distribution 12

2.1.2.4 Erlang Distribution 13

2.1.2.5 Weibull Distribution 14

2.1.3 Poisson Process 15

2.2 General Description of a queueing system 17

2.2.1 Notation 18
2.2.2 Utilisation 19
2.2.3 Cost Equation 20
2.2.4 Steady-State Probabilities 21
2.3 Single Server (Channel) Queues 22
2.3.1 Single Server Exponential Queueing System ($M/M/1$) 22
2.3.2 The System $M/G/1$ Queue 25
2.4 Multiserver Queues 27
2.4.1 $M/M/k/k$ Queue 27
2.4.2 $M/M/k$ Queue 28

3 CALL CENTRES 29
3.1 Introduction 29
3.2 Call Centres as Queueing Systems 33
3.3 Models for Call Centres 33
3.3.1 Erlang-C Model ($M/M/k$) 34
3.3.2 Erlang-A Model ($M/M/k/r+M$) 35
3.3.3 Other Models 36
3.4 Performance Measures 37
3.5 Traffic Intensity 38
3.6 Workforce Management: Staffing 39
3.6.1 Square-root Staffing 40
3.6.2 Real-time Staffing 41
3.6.3 Short-term Staffing 42
3.6.4 Long-term Staffing 42

4 METHODOLOGY 43
4.1 Limitations of Mathematical Approach 43
4.2 Simulation Software (EZSIM) 44
4.3 Specifications for the Model 49

4.4 Parameter Estimation 51
4.5 Beyond Erlang-A 53

5 ANALYSIS **54**

5.1 Performance measures 54
5.2 Assumptions 55
5.3 Distribution of Service Times 55
5.4 Distribution of Abandoning Times 56
5.4.1 Light Loads 56
5.4.1.1 Testing for Equality of Performance Measures 63
5.4.1.2 Comparison of Exponential and Gamma Distributions with Erlang Distributions 65
5.4.1.3 Coefficient of Variation 71
5.4.2 Heavy Loads 73

6 CONCLUSIONS **76**

6.1 Limitations 76
6.2 Areas of Further Research 76
6.3 Conclusion 77

List of Figures

3.1 Stages that calls pass through 31
3.2 Process flow of calls 32

4.1 Flowchart of the main module in the queueing simulation program 45
4.2 Flowchart of the procedure for using the stand-alone EZSIM environment 47

5.1 Bar graph of estimated mean *length of queue* for different distributions of abandoning times with 4, 5 and 6 servers respectively. 62
5.2 Bar graph of estimated mean *length of queue* for different distributions with 4, 5 and 6 servers respectively. 70

List of Tables

5.1 **Comparison of steady-state performance measures** for different distributions of abandonment; exponential (M/M/4/10+M), uniform (M/M/4/10+U(0.5, 1.5)), log-normal (M/M/4/10−LN(1, 1)) and gamma (M/M/4/10−gamma(2, 2)) with 4 servers and traffic intensity of 0.926. 57

5.2 **Comparison of steady-state performance measures** for different distributions of abandonment; exponential (M/M/4/10+M), uniform (M/M/4/10+U(0.5, 1.5)), log-normal (M/M/4/10−LN(1, 1)) and gamma (M/M/4/10+gamma(2, 2)) with 5 servers and traffic intensity of 0.926. 60

5.3 **Comparison of steady-state performance measures** for different distributions of abandonment; exponential (M/M/4/10+M), uniform (M/M/4/10+U(0.5, 1.5)), log-normal (M/M/4/10−LN(1, 1)) and gamma (M/M/4/10−gamma(2, 2)) with 6 servers and traffic intensity of 0.926. 61

5.4 95% Confidence Intervals for exponential, uniform, log-normal and gamma distributions of abandoning times. 64

5.5 **Comparison of steady-state performance measures** of different Erlang distributions (M/M/5/12+E(1, 0.5), M/M/5/12+E(1, 2) and Erlang-*A*) with gamma distribution (M/M/5/12+gamma(2, 2)) with 4 servers and traffic intensity of 0.926. 66

5.6 **Comparison of steady-state performance measures** of different Erlang distributions (M/M/5/12+E(1, 0.5), M/M/5/12+E(1, 2) and Erlang-A) with gamma distribution (M/M/5/12–gamma(2,2)) with 5 servers and traffic intensity of 0.926. 67

5.7 **Comparison of steady-state performance measures** of different Erlang distributions (M/M/5/12+E(1, 0.5), M/M/5/12+E(1, 2) and Erlang-A) with gamma distribution (M/M/5/12 | gamma(2, 2)) with 6 servers and traffic intensity of 0.926. 68

5.8 95% Confidence intervals for exponential, gamma and Erlang distributions of abandoning times 69

5.9 Coefficient of variation for estimated mean *length of queue*. 71

5.10 Coefficient of variation for estimated mean *waiting times* in the system. . 72

5.11 Coefficient of variation for estimated mean *server utilisation*. 72

5.12 Validity of abandoning distribution with traffic intensity of 1.1. 74

6.1 Model specification for Erlang-A. ii

6.2 Output from model using Erlang-A. iii

6.3 Uniform abandoning times with 0.926 traffic intensity. ii

6.4 Log-normal abandoning times with 0.926 traffic intensity. iii

6.5 Gamma abandoning times with 0.926 traffic intensity. iv

6.6 Gamma abandoning times with 0.926 traffic intensity. v

6.7 Erlang abandoning times with 0.926 traffic intensity. vi

6.8 Abandoning times with traffic intensity of 1.1 for different distributions. . vii

Basic Notation and Abbreviations

ACD: automatic call distributor

CTI: computer telephone integration

IVR: interactive voice response

PBX: private branch exchange

ICT: information and communication technology

FCFS: first come first served

LIFO: last in first out

$M/M/k$: Erlang-C model

$M/M/k/k$: Erlang-B model

$M/M/k/r+M$: Erlang-A model

L: average number of customers in the system

L_Q: average number of customers in the queue

W: average amount of time a customer spends in the system

W_Q: average amount of time a customer spends in the queue

V: virtual waiting time

P_n: steady-state proportion of time that the birth and death process is in state n

ρ: traffic intensity or utilisation factor

λ: arrival rate

μ: service rate

θ: abandonment rate

1 INTRODUCTION

1.1 Background

Queueing theory was conceived by A. K. Erlang (Erlang (1917) [A]) at the beginning of the 20th century and has flourished since, to become one of the central research themes of Operations Research. The study of queues with multiple servers dates back over fifty years to the seminar paper of Kiefer and Wolfowitzs (Dvoretzky, Kiefer and Wolfowittz (1956) [B]). During the last two decades, there has been an explosive growth in the number of companies that provide services via the telephone as well as in the variety of telephone services provided. Companies that offer services via the telephone are called call centres and are mathematically modelled as queueing systems and analysed using queueing theory. Multi-server queueing plays a central role in the analysis of call centre data.

The rapid growth of telephone call centres and in general contact centres has generated interest in the performance of multiserver queueing models. The call centre industry is thus vast and rapidly expanding in terms of both workforce and economic scope. More broadly, the continued growth in both the economic importance and complexity of call centres has prompted increasingly deep investigation of their operations. This is manifested by a growing body of academic work devoted to call centres; research ranging in discipline from Mathematics and Statistics, through to Operations Research, Industrial Engineering, Information Technology and Human Resource Management, all the way to Psychology and Sociology. Call centres are generally multiserver systems and often have

a very large number of servers. Therefore it is natural to look for insight into system performance. The goal is to improve the quality and the efficiency of the service of call centres using multiserver queueing models.

A central challenge in designing and managing a service operation in general and a call centre in particular, is to achieve a desired balance between operational efficiency and service quality. Typically, call centre goals are formulated based on the provision of service at a given quality, subject to a specified budget. There is need for continuous research to achieve a balance between operational efficiency and service quality.

1.2 Aims and Objectives

The aim of this thesis is to analyse call centres using the $M/GI/k/r+GI$ system. The process conforms to Poisson arrivals, service durations are independent and identically distributed with a general distribution, abandonments are independent and identically distributed with a general distribution, there are k servers and r extra waiting spaces.

The objectives are to:

- use $M/GI/k/r+GI$ systems to model call centres;
- calculate the mean service time, mean delay and mean utilisation;
- calculate proportion of time servers are idle or busy;
- calculate the probability of abandoning.

1.3 Significance of the Thesis

The rapid growth of telephone call centres and more general customer contact centres has generated renewed interest in the performance of multiserver queueing models when the

number of servers is large. Queues in service operations are often the arena where customers, service providers (servers or agents) and managers establish contact in order to jointly create the service experience. Process-wise, queues play in services much the same role as inventories in manufacturing. But in addition, "human queues" express preferences, complaints, abandonment and even spread around negative impressions. Managers can use queues as indicators for control and opportunities improvement. Indeed, queues provide unbiased quantifiable measures, in terms of which performance is relatively easy to monitor and goals are naturally formulated.

Technological progress has significantly affected the development of the call-centre industry. Computer-Telephone Integration (CTI) provides numerous opportunities for combining telephone services with e-mail and internet services. Consequently, many call centres evolve to contact centres; a big, growing, complicated and increasingly important part of the business landscape. Lots of data is gathered in telephone switches and call tracking databases.

Research in quantitative call centre management is concerned with the development of scientifically-based design principles and tools that support and balance service quality and efficiency. Queueing models constitute a natural convenient nurturing ground for the development of such principles and tools. Therefore the need to research along this line is of paramount importance. Even though there has been distinguished research along these lines (for example, Whitt (2005) [16] and Avramidis [2]), much remains to be done because the multi-server queue presents a formidable challenge.

1.4 Layout of the Thesis

This thesis has six chapters, of which the present one, the introduction, is the first chapter. Chapter two details literature review of Queueing theory and discusses all those models that can be used in call centres and chapter 3 discusses literature review of call centres. The methodology to be used in this research is covered in chapter four. Detailed analysis of the thesis is done in chapter five and the results are presented. Recommendations and conclusion will be presented in the last chapter of this thesis, chapter six.

2 QUEUEING THEORY

Congestion is a natural phenomenon in real systems. A service facility gets congested if there are more people than the server(s) can possibly handle. Very often the congestion is caused by variability in the arrival pattern of the customers or in the service mechanism or both. Therefore, any model must be expressed in terms of random processes and should yield conclusions in probabilistic terms. Queueing is therefore the mechanism that is used to handle congestion and helps to organise the various elements of the system in a manner conducive to modelling.

Queueing theory - is a branch of applied mathematics which attempts to construct and analyse models for what might be called unpredictable congestion.

Queueing system - consists of a servicing facility, a process of arrival of customers who wish to be served by the facility and the process of service.

Many recent developments in queueing theory have been driven in large part by a great interest in applications that involve human customers (for example, in the rapidly growing call centre sector).

2.0.1 Why Queueing Theory?

Networks and computers running multiuser, and multitasking operating systems can be viewed as interconnected queueing systems.

- Uses of queueing analysis are to:
  - analyse and understand system behaviour using real life data;
  - project from an existing system to a future system;
  - develop an analytic model for use in designing a system and
  - create simulations that models a system.
- Queueing theory can be used to analyse the performance of:
  - computer systems;
  - networks;
  - medical facilities, transportation systems, etc and
  - call or contact centres.

For detailed understanding of queueing theory, there is need for complete understanding of probability distributions, since queueing theory is rooted in them.

2.1 Probability Distributions

Probability theory provides the foundation for queueing theory and probability theory itself is rooted in set theory. For the sake of completness, some of the fundamental notions of probability are briefly described in what follows.

Sample space - is a set of all possible outcomes of an experiment.

An event - is a subset of a sample space. The mathematical definition of an event involves the notions of sample space and Borel fields, but for practical purposes, the intuitive notion of an event is sufficient. Examples of the kind of events of interest in queueing theory are; an arbitrary customer finds the server(s) busy, an arbitrary customer must wait more than two minutes for service or the number of waiting customers at certain time is n.

Events are called mutually exclusive if their intersection is an empty set (that is, if the occurrence of one excludes the possibility of the other occurring). A set of events is said to be exhaustive if the union of the events is the same as the sample space.

Random variable - is a real valued function defined on the sample space. It is the outcome of the experiment that is random and not the assigning of a real valued number to each possible outcome of the experiment.

Random variables are denoted by capital letters, X, Y , etc. The expected value or mean of X is denoted by $E(X)$ and its variance by $\sigma^2(X)$, where $\sigma(X)$ is the standard deviation of X. An important quantity is the coefficient of variation of the positive random variable X, defined as

$$c_X = \frac{\sigma(X)}{E(X)}$$

The coefficient of variation is a dimensionless measure of the variability of the random variable X.

Consider a sample space S. Let A be a subset of S then, the probability of A is the function on S denoted as $P(A)$ and satisfies the following three axioms:

1. $0 \leq P(A) \leq 1$.
2. $P(S) = 1$.
3. The probability of the union of mutually exclusive events is equal to the sum of the probabilities of these events, i.e. $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ where A_i are mutually exclusive events for $i = 1, ..., \infty$.

We use the notation $P(A \mid B)$ for the conditional probability of A given B (the probability that event A occurs given that event B is known to have occurred). It is defined as

$$P(A \mid B) = \frac{P(A \bigcap B)}{P(B)}$$

If events A and B are independent (that is, if one of them occurs then the probability of the other to occur is not affected), then

$$P(A \mid B) = P(A),$$

which implies that

$$P(A \bigcap B) = P(A)P(B)$$

2.1.1 Discrete Probability Distributions

This section discusses a number of important distributions which have been found useful for describing the distribution of random variables in queueing theory.

2.1.1.1 Bernoulli Distribution

Consider an experiment which has only two possible outcomes. Let us call them "success" and "failure". These two outcomes are mutually exclusive and exhaustive events. The Bernoulli random variable assigns the value $X = 1$ to the "success" outcome and the value $X = 0$ to the "failure" outcome. Let p be the probability of the "success" outcome. Since "success" and "failure" are mutually exclusive and exhaustive events, the probability of the "failure" outcome is $1 - p$. The probability distribution function in terms of the Bernoulli random variable is:

$$P(X = 1) = p,$$

$$P(X = 0) = 1 - p$$

where p is such that, $0 \leq p \leq 1$.

2.1.1.2 Binomial Distribution

Assume that n independent Bernoulli trials are performed, let X be a random variable representing the number of successes in these n trials. Such a random variable is called

binomial random variable with parameters n and p. Its probability function is given by

$$P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$$

for $i = 0, 1, 2, ..., n$.

2.1.1.3 Poisson Distribution

Among the discrete probability distributions, the Poisson distribution is the most applicable in queueing theory. A Poisson random variable with parameter λ, where $\lambda > 0$, has the following distribution

$$P(X = i) = \frac{e^{-\lambda}\lambda^i}{i!}$$

for $i = 0, 1, 2, ..., \infty$.

The Poisson random variable accurately models the number of calls arriving at a telephone exchange or internet service provider in a short period of time; for example a few seconds or a minute. The importance of the Poisson random variable lies in its property to approximate the binomial random variable in cases when n is very large and p is very small, so that $n * p$ is neither too large nor too small. The Poisson process $N(t)$ usually represents the number of events in an interval $[0, t]$, such that:

1. the number of events occurring in an interval of length t is independent of the number of events occurring in any other non-overlapping interval of length t;

2. the distribution of $N(t)$ is the same for all intervals of length t, no matter where the interval begins;

3. two events cannot occur simultaneously and

4. no matter how small t is, there is a positive probability that an event will occur in the interval $[s, s + t]$.

The first two relate to the idea of independent and stationary increments, respectively. The Poisson distribution has been successfully used to describe such diverse phenomena as the number of busy channels in a telephone system, customer demand for service, etc.

2.1.2 Continuous Probability Distributions

There are five continuous random variables that are of particular interest in queueing theory. These are uniform, exponential, gamma, Erlang and Weibull distributions.

2.1.2.1 Uniform Distribution

The uniform distribution is a continuous type of distribution, which may assume any value on a real line segment of nonzero length. The probability density function of the uniform random variable takes non-negative values over the interval $(a\,, b)$ and is given by

$$f(x) = \begin{cases} \frac{1}{b-a}, & if\ a < x < b \\ 0, & elsewhere \end{cases}$$

A uniform random variable over the interval (0, 1) is very important in simulations. Almost all computers programs have a function which generates uniform (0, 1) random deviates.

2.1.2.2 Exponential Distribution

The most common stochastic queueing models assume that inter-arrival times and service times are exponentially distributed. It is one of the most important continuous distributions in queueing theory. The density of an exponential distribution with parameter $\lambda > 0$ is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & if\ x \geq 0 \\ 0, & otherwise \end{cases}$$

An exponential random variable can be used to model the time until the next call arrives at a switch board. Interestingly, such time does not depend on how long ago the last call was received. This property is called the memory-less property of a random variable. More precisely, a random variable is said to posses the memory-less property if, $P(X > t+s \mid X > t) = P(X > s)$. This is true for the exponential distribution and is proved as follows

$$\begin{aligned} P(X > t+s \mid X > t) &= \frac{P(X>t+s, X>t)}{P(X>t)} \\ &= \frac{P(X>t+s)}{P(X>t)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} \\ &= \frac{e^{-\lambda t}e^{-\lambda s}}{e^{-\lambda t}} \\ &= e^{-\lambda s} \\ &= P(X > s) \end{aligned}$$

The exponential distribution is the only continuous distribution that exhibits the memory-less property (Markovian property). This property of the exponential random variable makes it useful in describing inter-arrival times and service times in queueing theory. As shall be shown later, the exponential random variable is integrally related to the Poisson random variable. This relationship will prove to be of paramount importance throughout the discussion of queueing theory.

If X_1, X_2, ..., X_n are independent exponential random variables with parameters λ_1, λ_2, ..., λ_n respectively, then $\min(X_1, X_2, ..., X_n)$ is again an exponential random variable with parameter $\lambda_1 + \lambda_2 + \cdots + \lambda_n$ and the probability that X_i is the smallest one is given by $\lambda_i/(\lambda_1+\lambda_2+\cdots+\lambda_n)$, $i = 1, 2, ..., n$. This can be seen as follows, let X_1 and X_2 be exponentially distributed random variables with parameters λ_1 and λ_2. It is also of interest to know the distribution of $X = \min(X_1, X_2)$. In other words, we are interested in the distribution of the time that passes until the first one of the two random variables

X_1 and X_2 occurs. Note that,

$$\begin{aligned} P(X > t) &= P(min(X_1, X_2) > t) \\ &= P(X_1 > t, X_2 > t) \\ &= P(X_1 > t)P(X_2 > t) \quad \textit{by independence of } X_1 \textit{ and } X_2 \\ &= e^{-\lambda_1 t}e^{-\lambda_2 t} \\ &= e^{-(\lambda_1+\lambda_2)t} \end{aligned}$$

Thus, the distribution of X is exponential with parameter $\lambda_1 + \lambda_2$.

2.1.2.3 Gamma Distribution

Let Y_1, Y_2, ..., Y_n be the times between the occurrence of $n+1$ successive events. Define X as

$$X = \sum_{i=1}^{n} Y_i$$

- If Y_i is an exponential random variable with parameter $\lambda > 0$, $i = 1, 2, ..., n$ and $Y_i's$ are independent, then X is gamma distributed with density function

$$f(x) = \frac{\lambda^n}{(n-1)!}x^{n-1}e^{-\lambda x}I_{(0,\infty)}(x),$$

where

$$I_{(0,\infty)}(x) = \begin{cases} 1, & if\ x \in (0, \infty) \\ 0, & elsewhere \end{cases}$$

A more general form of the gamma density function is

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda x}I_{(0,\infty)}(x),\ \lambda > 0,\ \alpha > 0,$$

where

$$\Gamma(\alpha) = \int_0^\infty e^{-x}x^{\alpha-1}dx$$

If α is a positive integer, it is easy to show that $\Gamma(\alpha) = (\alpha - 1)!$

- If α is not an integer, the gamma random variable cannot be represented by the sum of identically distributed exponential random variables.
- If $\alpha = 1$, the gamma density function reduces to an exponential density function.

2.1.2.4 Erlang Distribution

The Erlang distribution is a continuous probability distribution developed by A. K. Erlang. It is a random variable that is closely related to exponential and gamma random variables. The Erlang distribution is a special case of the gamma distribution when the shape parameter is an integer. It represents the sum of a series of exponential distributions. A Gamma(a, b) distribution is equal to an Erlang(m, b) (where the first and second values are the scale and shape parameters respectively) distribution with $a = m$, when a is an integer. To illustrate how this random variable arises in the context of queueing theory, let us consider a service facility that services units at rate μ. Service is performed in k phases, where the distribution of service time in each phase is exponential with rate $k\mu$. If Y_i is the time the unit spends in the i^{th} phase, $i = 1, 2, ..., k$, then the density function of Y_i is given by

$$g(y_i) = k\mu e^{-k\mu y_i} I_{(0,\infty)}(y_i)$$

The total time, X, spent in service is then

$$X = \sum_{i=1}^{k} Y_i$$

Since the distribution of the sum of exponential distribution is a gamma, therefore the density function of X is

$$f(x) = \frac{(k\mu)^k}{(k-1)!} x^{k-1} e^{-k\mu x} I_{(0,\infty)}(x)$$

Therefore, an Erlang distribution is gamma distribution with $\lambda = k\mu$.

A. K. Erlang worked a lot on traffic modelling. The Erlang distribution was developed to examine the number of telephone calls which might be made at the same time to the operators of the switching stations. This work on telephone traffic engineering has been expanded to consider waiting times in queueing systems in general. Thus, there are two Erlang models, both used in modelling traffic:

- Erlang-B model; this is the easier of the two, and can be used, for example, in a call centre to calculate the number of trunks one needs to carry a certain amount of phone traffic with a certain "target service" (see Vose (2007) [20]).
- Erlang-C model; this formula is much more difficult and is often used, for example, to calculate how long callers will have to wait before being connected to an agent in a call centre.

2.1.2.5 Weibull Distribution

The Weibull distribution is one of the most commonly used distributions in reliability engineering because of the many shapes it attains for various values of the slope parameter (β). It models a great variety of data and life characteristics.

- If $\beta = 1$, the Weibull distribution is identical to the exponential distribution;
- if $\beta = 2$, the Weibull distribution is identical to the Rayleigh distribution;
- if β is between 3 and 4, the Weibull distribution approximates the normal distribution.
- The Weibull distribution approximates the log-normal distribution for several values of β for example$\beta = 0.18$. For most populations, more than fifty samples are required to differentiate between the Weibull and log-normal distributions.

The 2-parameter Weibull probability density function is given by

$$f(T) = \frac{\beta}{\eta}\left(\frac{T}{\eta}\right)^{\beta-1}\exp\left[-\left(\frac{T}{\eta}\right)^{\beta}\right] I_{(0,\infty)}(T)$$

where $T \geq 0,\ \beta > 0,\ \eta > 0$

and

- $\eta = scale\ parameter,$
- $\beta = shape\ parameter\ (or\ slope).$

The Weibull distribution is extremely flexible. It is capable of representing a wide variety of data including left-skewed, symmetrical and right-skewed distributions.

2.1.3 Poisson Process

When the Poisson random variable was discussed, properties that it possesses were defined. The relationship between exponential and Poisson random variables was pointed out. Let us examine these properties in detail.

Let us consider an arrival process $\{N(t),\ t \geq 0\}$, where $N(t)$ denotes the total number of arrivals up to time t, with $N(0) = 0$, and which satisfies the following assumptions:

1. The number of events occurring in an interval of length t is independent of the number of events occurring in any other non-overlapping interval of any length (independent increments). The distribution of $N(t)$ is the same for all intervals of the length t, no matter where the interval begins (stationary increments).

2. The probability that an arrival occurs between t and $t + h$ is equal to $\lambda h + o(h)$. This is written as $P(N(h) = 1) = \lambda h + o(h)$, where λ is a constant independent

of $N(h)$, h is an incremental element and $o(h)$ denotes a quantity that becomes negligible when compared to h as $h \to 0$; that is,

$$\lim_{h \to 0} \frac{o(h)}{h} = 0$$

3. P(more than one arrival between t and $t+h$) $= o(h)$, that is, $P(N(h) \geq 2) = o(h)$.

By definition, the Poisson process has **stationary increments**, that is, for any $t_2 > t_1$, the random variables $N(t_2) - N(t_1)$ and $N(t_2+u) - N(t_1+u)$ have the same distribution for any $u > 0$. In both cases, the distribution is Poisson with parameter, $\lambda(t_2 - t_1)$.

If $N(h)$ has a Poisson distribution, with parameter λh, then

$$\begin{aligned} P(N(h) = 0) &= e^{-\lambda h} \\ &= 1 - \lambda h + \tfrac{(-\lambda h)^2}{2!} + \tfrac{(-\lambda h)^3}{3!} + \ldots \\ &= 1 - \lambda h + o(h) \end{aligned}$$

where

$g(h) = \frac{(-\lambda h)^2}{2!} + \frac{(-\lambda h)^3}{3!} + \ldots$ is $o(h)$ and we have used the Taylor series for $e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots$ with $x = \lambda h$. Using this result

$$\begin{aligned} P(N(h) = 1) &= \lambda h P(N(h) = 0) \\ &= \lambda h(1 - \lambda h + \tfrac{(-\lambda h)^2}{2!} + \tfrac{(-\lambda h)^3}{3!} + \ldots) \\ &= \lambda h - (\lambda h)^2 + \tfrac{(\lambda h)^3}{2!} - \tfrac{(\lambda h)^4}{3!} + \ldots \\ &= \lambda h + o(h) \end{aligned}$$

where $g(h) = -(\lambda h)^2 + \frac{(\lambda h)^3}{2!} - \frac{(\lambda h)^4}{3!} + \ldots$ is an $o(h)$ since $\lim_{h \to 0} \frac{g(h)}{h} = 0$

The Poisson process is an extremely useful process for modelling purposes in many practical applications, for example, to model arrival processes for queueing models or demand processes for inventory systems. It is empirically found that in many circumstances, the arising stochastic processes can be well approximated by a Poisson process.

2.2 General Description of a queueing system

Among others, a queueing model is characterized by

- *The arrival process of customers.*

  Usually we assume that the inter-arrival times are independent and have a common distribution. In many practical situations, customers arrive according to a Poisson process (that is exponential inter-arrival times). Customers may arrive one by one or in batches. An example of batch arrivals is the customs office at the border where travel documents of bus passengers have to be checked.

- *The behaviour of customers.*

  Customers may be patient and willing to wait or customers may be impatient and leave after a while. For example, in call centres, customers will hang up when they have to wait too long before an operator is available and they may possibly try again after a while.

- *The service times.*

  Usually we assume that the service times are independent and identically distributed, and that they are independent of the inter-arrival times. For example, the service times can be deterministic, exponentially distributed or have a general distribution. It can also occur that service times are time-dependent or dependent on queue length. For example, the processing rates of the machines in a production system can be increased once the number of jobs waiting to be processed becomes too large.

- *The service discipline.*

  Customers can be served one by one or in batches (bulk servers such as buses, elevators, etc. can serve more than one customer at a time). We have many possibilities for the order in which they enter service.

These are:

- first come first served; or first in, first out (FIFO), that is in order of arrival;
- service in random order (SIRO), zero length;
- last come first served; or last in, first out (LIFO);
- priorities (for example rush orders first, shortest processing time first or customers with high priority are served first);
- processor sharing (in computers that equally divide their processing power over all jobs in the system). Customers are served equally and they all effectively experience the same delay.

- *The service capacity.*

  There may be a single server or a group of servers helping the customers.

- *The waiting room.*

  There can be limitations with respect to the number of customers in the system. For example, in a data communication network, only finitely many cells can be transmitted in a switch.

2.2.1 Notation

A commonly used shorthand notation, called Kendall notation, for queue models describes the arrival process, service distribution, the number of servers and the buffer size (waiting room) as follows

$$arrival\ process/service\ distribution/number\ of\ servers/waiting\ room$$

Commonly used characters for the first two positions in this shorthand notation are:

- D- Deterministic,
- M- Markovian (Poisson for the arrival process or Exponential for the service time), and M stands for memory-less,
- G- General distribution,
- GI- General and independent distribution.

The first position specifies the inter-arrival time distribution and the second one, the service time distribution. The third position specifies the number of servers (k) and the fourth position is used for the number of buffer places in addition to the number of servers and it is usually not used if the waiting room is unlimited.

For example, $M/M/1$ denotes a single-server queue with Poisson arrival process and exponential service time with infinite buffer places. An $M/G/k/k$ denotes a queue with k-servers and no additional waiting room except at the servers, with the arrival process being Poisson and the service time following a general distribution.

2.2.2 Utilisation

An important measure for queueing systems' performance is the utilisation.

Utilisation - is the proportion of time that a server is busy on average.

If we have multi-server queues then the system utilisation is the average of individual server utilisation.

Major characteristics that need to be studied to understand the behaviour of a queueing system are the queue length (number of customers waiting at time t), the waiting time (the time a new arrival will have to wait till his service commences) and the length of

the busy period (the length of time when the server will be continuously busy). These factors are dependent on the input process, service mechanism and the queue discipline, which are subject to uncertainties and hence, are better described as random variables. The queue length and the waiting time are stochastic processes (family of random variables indexed by a time parameter) whose behaviour is given by transition distributions, whereas the busy period is a random variable whose distribution is of particular interest. Expected values and other moments of these distributions need to be obtained for a greater understanding of the processes involved.

2.2.3 Cost Equation

Consider a system in equilibrium in which customers arrive, remain in the system for a length of time, and then depart. Let λ_a be the arrival rate of customers who actually enter the system, W and W_Q be the mean waiting time in the system and queue respectively, and L and L_Q be the mean number of customers in the system and queue respectively. If entering customers are required to pay money according to some rule to the system, then we have the following basic cost identity:

$$\textit{average rate at which the system earns} = \lambda_a * \textit{average amount an entering customer pays} \tag{2.1}$$

Equation 2.1 is an important and simple queueing theory result that applies to $G/G/1$ queues (and to other systems). It is known as Little's formula. Suppose that in the basic cost identity, each customer pays \$1 per unit time while in the system, then equation 2.1 yields

$$L = \lambda_a W \tag{2.2}$$

The well known Little's formula embodied in equation 2.2 is one of the general and useful results in queueing theory. Little's formula applies to any system in equilibrium in which customers arrive, spend a certain amount of time and depart. Its applicability is not limited to single-server queues, or single queue systems, or systems with infinite buffer.

It was first proved by John D. C. Little (Ross (2000) [12]) in the context of a steady-state queueing theory system. Little's theorem can be applied to the queue itself, that is:

$$L_Q = \lambda_a W_Q \tag{2.3}$$

If the cost rule is applied to service, then we obtain:

$$average\ number\ of\ customers\ in\ service = \lambda_a E[S] \tag{2.4}$$

where $E[S]$ is defined as the average amount of time a customer spends in service. Little's formula gives a very important relation among the mean number of customers in the system, the mean service time and the average number of customers entering the system per unit time. It assumes that the capacity of the system is sufficient to deal with the customers (that is, the number of customers in the system does not grow to infinity).

2.2.4 Steady-State Probabilities

This thesis will only deal with stationary type of continuous Markov chains. Let $X(t)$ denote the number of customers in the system at time t and define P_n, $n \geq 0$, by

$$P_n = \lim_{t \to \infty} P\{X(t) = n\}$$

If the limit exists, then P_n is the limiting or long-run probability that there will be n customers in the system. In most of the cases, it turns out to be the long-run proportion of time that the system contains exactly n customers. There are two other sets of limiting probabilities $\{a_n, n \geq 0\}$ and $\{d_n, n \geq 0\}$, where

$a_n = proportion\ of\ customers\ that\ find\ n\ customers\ in\ the\ system\ when\ they\ arrive$

$d_n = proportion\ of\ customers\ leaving\ behind\ n\ customers\ in\ the\ system\ when\ they\ depart$

In a system in which customers arrive one at a time and are served one at a time, these two proportions will be equal (that is $a_n = d_n$), since in the long-run, the rate of transitions from n to $n+1$ equals the rate from $n+1$ to n. This means that the rate at which arrivals find n equals the rate at which departures leave n. Hence, on average, arrivals and departures always see the same number of customers (Ross (2000) [12]).

2.3 Single Server (Channel) Queues

2.3.1 Single Server Exponential Queueing System ($M/M/1$)

In the $M/M/1$ process, the queue-size increases by only one, decreases by only one and stays an exponential amount of time at each state. It is equivalent to a birth-and-death process. In this queue the first two Ms refer to the fact that both the inter-arrivals and service distributions are exponential, that is, they are Markovian or memory-less and 1 means that there is a single server. Using the idea that in the long run, the rate at which transitions into state j occur must equal the rate at which transitions out of state j occur, we can determine the limiting probabilities P_n, $n = 0, 1, 2, \ldots$. Now when there are n, $n \geq 0$, customers in the system, using the rate equality principle (that is, the rate at which the process enters state n equals the rate at which it leaves state n) we can determine the limiting probabilities as follows: for state 0, the process can leave only through an arrival. Since the arrival rate is λ and the proportion of time that the process is in state 0 is P_0, then the rate at which the process leaves state 0 is λP_0. On the other hand, state 0 can be reached from state 1 via a departure (that is, if there is one person and the person completes service, then the system becomes empty). Since the service rate is μ and the proportion of time that the system has exactly one customer is P_1, it follows that the rate at which the process enters state 0 is μP_1. Hence we get our first equation

$$\lambda P_0 = \mu P_1$$

Let us consider state 1. The process can leave this state either by an arrival or a departure.

Hence, when the process is in state 1, it will leave at rate $(\lambda + \mu)P_1$. State 1 can be entered either from state 0 via an arrival or from state 2 via a departure. Hence, the rate at which the process enters state 1 is $\lambda P_0 + \mu P_2$. Similarly other balance equations can be obtained by the same reasoning, to obtain

$$\begin{array}{lrcl} \textit{State} & \textit{rate at which the process leaves} & = & \textit{rate at which it enters} \\ 0 & \lambda P_0 & = & \mu P_1 \\ 1 & (\lambda+\mu)P_1 & = & \lambda P_0 + \mu P_2 \\ 2 & (\lambda+\mu)P_2 & = & \lambda P_1 + \mu P_3 \\ n,\ n \geq 1 & (\lambda+\mu)P_n & = & \lambda P_{n-1} + \mu P_{n+1} \end{array} \tag{2.5}$$

In order to solve these equations, we re-write them in this format

$$\begin{aligned} P_1 &= \tfrac{\lambda}{\mu}P_0, \\ P_{n+1} &= \tfrac{\lambda}{\mu}P_n + \left(P_n - \tfrac{\lambda}{\mu}P_{n-1}\right), \quad n \geq 1 \end{aligned}$$

Solving in terms of P_0 yields

$$\begin{array}{rcccccl} P_0 &=& P_0, & & & & \\ P_1 &=& \tfrac{\lambda}{\mu}P_0 & & & & \\ P_2 &=& \tfrac{\lambda}{\mu}P_1 + \left(P_1 - \tfrac{\lambda}{\mu}P_0\right) & = & \tfrac{\lambda}{\mu}P_1 & = & \left(\tfrac{\lambda}{\mu}\right)^2 P_0, \\ P_3 &=& \tfrac{\lambda}{\mu}P_2 + \left(P_2 - \tfrac{\lambda}{\mu}P_1\right) & = & \tfrac{\lambda}{\mu}P_2 & = & \left(\tfrac{\lambda}{\mu}\right)^3 P_0, \\ P_4 &=& \tfrac{\lambda}{\mu}P_3 + \left(P_3 - \tfrac{\lambda}{\mu}P_2\right) & = & \tfrac{\lambda}{\mu}P_3 & = & \left(\tfrac{\lambda}{\mu}\right)^4 P_0, \\ \vdots & & \vdots & & \vdots & & \\ P_{n+1} &=& \tfrac{\lambda}{\mu}P_n + \left(P_n - \tfrac{\lambda}{\mu}P_{n-1}\right) & = & \tfrac{\lambda}{\mu}P_n & = & \left(\tfrac{\lambda}{\mu}\right)^{n+1} P_0, \end{array}$$

Using the fact that $\sum_{n=0}^{\infty} P_n = 1$, thus we get

$$\begin{aligned} 1 &= \sum_{n=0}^{\infty} P_n \\ &= \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n P_0 \\ &= \frac{P_0}{1-\frac{\lambda}{\mu}} \end{aligned}$$

from which it follows that

$$P_n = \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right), \quad n \geq 1. \tag{2.6}$$

The above is only valid when $\frac{\lambda}{\mu} < 1$, otherwise the sum would be infinite. Now the rest of the other values can be expressed as follows

$$\begin{aligned} L &= \sum_{n=0}^{\infty} nP_n \\ &= \sum_{n=0}^{\infty} n \left(\frac{\lambda}{\mu}\right)^n \left(1 - \frac{\lambda}{\mu}\right) \\ &= \left(1 - \frac{\lambda}{\mu}\right) \frac{\lambda}{\mu} \sum_{n=0}^{\infty} \left(\left(\frac{\lambda}{\mu}\right)^n\right)' \\ &= \left(1 - \frac{\lambda}{\mu}\right) \frac{\lambda}{\mu} \left(\sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n\right)' \\ &= \left(1 - \frac{\lambda}{\mu}\right) \frac{\lambda}{\mu} \left(\frac{1}{1-\frac{\lambda}{\mu}}\right)' \\ &= \left(1 - \frac{\lambda}{\mu}\right) \frac{\lambda}{\mu} \left(\frac{1}{1-\frac{\lambda}{\mu}}\right)^2 \\ &= \frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}} \\ &= \frac{\lambda}{\mu-\lambda} \end{aligned} \tag{2.7}$$

Since $\lambda_a = \lambda$, it follows that

$$\begin{aligned} W &= \frac{L}{\lambda} \\ &= \frac{1}{\mu-\lambda} \end{aligned}$$

$$\begin{aligned} W_Q &= W - E[S] \\ &= W - \frac{1}{\mu} \\ &= \frac{\lambda}{\mu(\mu-\lambda)} \end{aligned}$$

$$\begin{aligned} L_Q &= \lambda W_Q \\ &= \frac{\lambda^2}{\mu(\mu-\lambda)} \end{aligned} \tag{2.8}$$

2.3.2 The System $M/G/1$ Queue

Let us define work for an arbitrary queueing system as the sum of the remaining service times of all customers in the system at time t. Let us consider the cost rule that each customer pays at rate of y per unit time when the remaining service time is y, irrespective of being in queue or service. Let V denote the average work in the system, then

$$V = \lambda_a E\left[\textit{amount paid by a customer}\right]$$

Now, let S and W_Q^* denote the service time and the time an arbitrary customer spends waiting in queue, respectively. Therefore, the customer pays at a constant rate of S per unit while he waits in queue and at rate of $S - x$ after spending an amount of time x in service, that is

$$E[\textit{amount paid by a customer}] = E\left[SW_Q^* + \int_0^S (S-x)dx\right]$$

and thus

$$V = \lambda_a E\left[SW_Q^*\right] + \frac{\lambda_a E\left[S^2\right]}{2} \tag{2.9}$$

If customer's service time is independent of his waiting time in queue, then we have

$$V = \lambda_a E\left[S\right] W_Q + \frac{\lambda_a E\left[S^2\right]}{2} \tag{2.10}$$

This model assumes Poisson arrivals (that is, inter-arrivals follow exponential distribution), a general distribution for service times and a single server. Generally, we assume

that customers are served in the order they come.

$$Customer's\ wait\ in\ queue = work\ in\ the\ system\ when\ he\ arrives \tag{2.11}$$

Taking expectation of equation 2.11 yields

$$W_Q = average\ work\ as\ seen\ by\ an\ arrival$$

Since these are Poisson arrivals, the average work as seen by an arrival will equal average work in the system

$$W_Q = V.$$

Hence

$$V = \lambda E[S] W_Q + \frac{\lambda E\left[S^2\right]}{2}$$

upon simplifying yields the **Pollaczek** formula (Ross (2000) [12]),

$$W_Q = \frac{\lambda E\left[S^2\right]}{2\left(1 - \lambda E[S]\right)} \tag{2.12}$$

where $E[S]$ and $E\left[S^2\right]$ are the first two moments of the service distribution. The quantities L, L_Q and W can be obtained from equation 2.12 as

$$\begin{array}{rcccl} L_Q & = & \lambda W_Q & = & \frac{\lambda^2 E\left[S^2\right]}{2(1-\lambda E[S])} \\ W & = & W_Q + E[S] & = & \frac{\lambda E\left[S^2\right]}{2(1-\lambda E[S])} + E[S] \\ L & = & \lambda W & = & \frac{\lambda^2 E\left[S^2\right]}{2(1-\lambda E[S])} + \lambda E[S] \end{array} \tag{2.13}$$

2.4 Multiserver Queues

2.4.1 $M/M/k/k$ Queue

The $M/M/k/k$ Queue is a queueing system in which arrivals that find all servers busy do not enter but are lost to the system. This system has Poisson arrivals, service times are exponentially distributed and there are k servers. The balance equations are

$$\begin{array}{ccrcl} \textit{State} & & \textit{Rate leave} & = & \textit{rate enter} \\ 0 & & \lambda P_0 & = & \mu P_1 \\ 1 & & (\lambda+\mu)\,P_1 & = & 2\mu P_2 + \lambda P_0 \\ 2 & & (\lambda+2\mu)\,P_2 & = & 3\mu P_3 + \lambda P_1 \\ i,\ 0<i<k & & (\lambda+i\mu)\,P_i & = & (i+1)\,P_{i+1} + \lambda P_{i-1} \\ k & & k\mu P_k & = & \lambda P_{k-1} \end{array}$$

Rewriting gives

$$\begin{aligned} \lambda P_0 &= \mu P_1 \\ \lambda P_1 &= 2\mu P_2 \\ \lambda P_2 &= 3\mu P_3 \\ &\vdots \\ \lambda P_{k-1} &= k\mu P_k \end{aligned}$$

which can be re-written as

$$\begin{array}{rcccl} P_1 &=& \frac{\lambda}{\mu}P_0 & & \\ P_2 &=& \frac{\lambda}{2\mu}P_1 &=& \frac{\left(\frac{\lambda}{\mu}\right)^2}{2}P_0 \\ P_3 &=& \frac{\lambda}{3\mu}P_2 &=& \frac{\left(\frac{\lambda}{\mu}\right)^3}{3!}P_0 \\ & \vdots & & & \\ P_k &=& \frac{\lambda}{k\mu}P_{k-1} &=& \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!}P_0 \end{array}$$

and also using $\sum_{i=0}^{k} P_i = 1$, we obtain

$$P_i = \frac{\left(\frac{\lambda}{\mu}\right)^i / i!}{\sum_{j=0}^{k} \left(\lambda E[S]\right)^j / j!}, \quad i = 0, 1, ..., k$$

Since $E[S] = \frac{1}{\mu}$, where $E[S]$ is the mean service time, the preceding can be written as

$$P_i = \frac{\frac{(\lambda E[S])^i}{i!}}{\frac{\sum_{j=0}^{k}(\lambda E[S])^j}{j!}}, \qquad i = 0,\ 1,\ ...,\ k \tag{2.14}$$

This is the same system as the previous model except that the service distribution is general (that is, $M/G/k/k$), with no queue allowed. This model is sometimes called the **Erlang loss** (Erlang-B) system. The equation 2.14 remains valid for more general systems, but the proof is more advanced (Ross (1996) [11]).

2.4.2 $M/M/k$ Queue

The balance equations are

State	*Rate leave*	=	*rate enter*
0	λP_0	=	μP_1
1	$(\lambda + \mu) P_1$	=	$2\mu P_2 + \lambda P_0$
2	$(\lambda + 2\mu) P_2$	=	$3\mu P_3 + \lambda P_1$
$i,\ 0, i < k$	$(\lambda + i\mu) P_i$	=	$(i+1) P_{i+1} + \lambda P_{i-1}$
$n,\ n \geq k$	$(\lambda + k\mu) P_n$	=	$k\mu P_{n+1} + \lambda P_{n-1}$

and also using the fact that $\sum_{i=0}^{\infty} P_i = 1$, we obtain

$$P_i = \begin{cases} \dfrac{\frac{\left(\frac{\lambda}{\mu}\right)^i}{i!}}{\sum_{i=0}^{k-1} \frac{\left(\frac{\lambda}{\mu}\right)^i}{i!} + \frac{\left(\frac{\lambda}{\mu}\right)^k k\mu}{k!(k\mu - \lambda)}}, & i \leq k \\ \dfrac{\left(\frac{\lambda}{k\mu}\right)^i k^k}{k!} P_0, & i > k \end{cases}$$

3 CALL CENTRES

3.1 Introduction

Service engineering is a newly emerging discipline that seeks to develop scientifically-based engineering principles and tools, often culminating in software, which support the design and management of service operations. A contact centre is a collection of resources providing an interface between the service provider and its remote customers. The classical contact centre is the telephone call centre, containing a collection of customer service representatives who talk to customers over the telephone. Due to advances in Information and Communication Technology, the number, size and scope of contact centres, as well as the number of people who are employed there or use them as customers, is growing explosively. Call centres are locations where calls are placed or received in high volume for the purpose of sales, marketing, customer service, telemarketing, technical support, or other specialised business activity. In a call centre, the service representatives are supported by quite elaborate information-and-communication-technology (ICT) equipment, such as a private branch exchange (PBX), an automatic call distributor (ACD), a personal computer (PC) and assorted databases. There are different kinds of call centres namely:

- call centres with only inbound traffic (customer-generated calls);
- call centres with only outbound traffic (agent-generated calls like tele-marketing);
- or a combination of these.

Inbound call centres are usually supported by interactive voice response (IVR) units, which serve as elaborate answering machines. Through a selection of menus, IVR units attempt to respond to the customer's needs and if necessary, help route the call to an appropriate service representative (agent).

This research focuses only on inbound traffic and models of those calls that are passed to agents by the IVR. If the utilisation of ACD signifies the basic paradigm of the first generation of call centres, the adoption of interactive voice response, introduction of call blending and development of web-enabled multimedia contact centres might be seen as significant steps in their late evolution. A call centre remains defined fundamentally by the integration of telephone and computer technologies.

In analysing call centres, it is necessary to take account of differences in relation to a number of important variables like size, industrial sector, market conditions, complexity and call cycles times, the nature of operations (inbound or outbound), the precise manner of technological integration, the effectiveness of representative organisations and management styles, priorities, and human resource practices. The opposing goals of efficiency and excellent service are both central to call centres. High levels of service are important since the number of completely satisfied customers is one of the few predictors of long-term profitability. Efficiency and service are more salient than in most service organisations. To achieve efficiency, call centre management focuses on the selection, implementation and use of technology. The technology is used to facilitate the physical concentration of staff, labour scheduling, staff monitoring and high productivity rates.

Telephone call centre agents provide tele-services as they speak with customers over the phone. They interact with a computer terminal, inputing and retrieving information related to customers and their requests. Customers are either being served or are waiting in what is called a tele-queue, a phantom queue which they share, invisible to each other and to agents who serve them. Customers wait in this queue until either

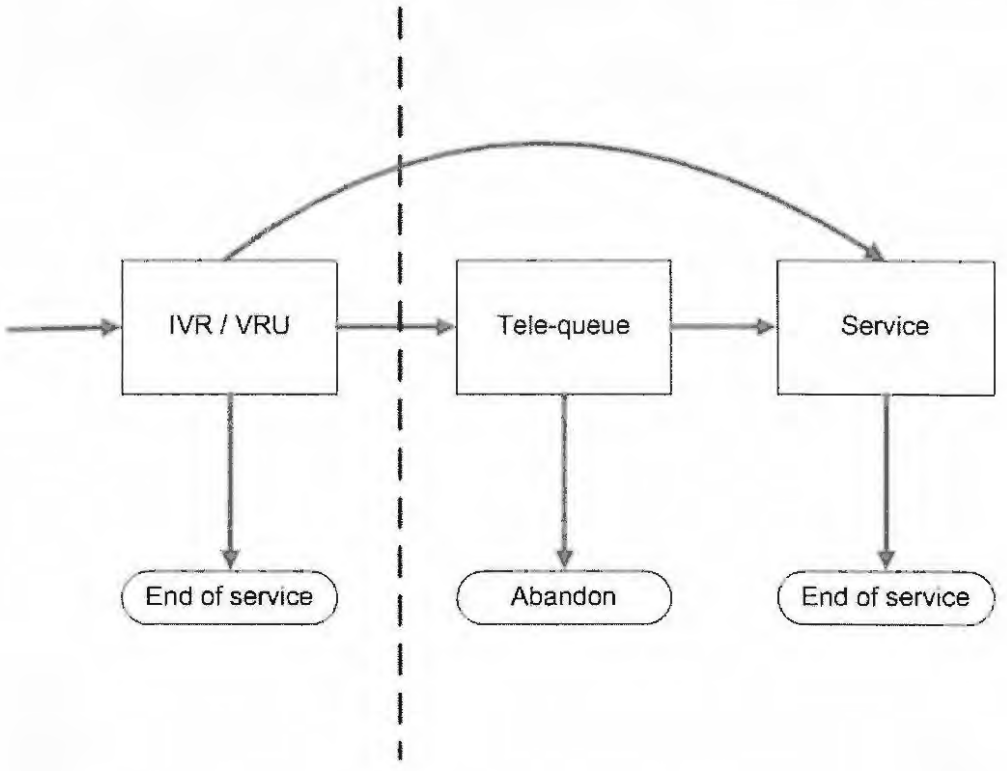

Figure 3.1: Stages that calls pass through

an agent is allocated to serve them (through supporting software) or they become impatient and abandon the tele-queue. Customers in the tele-queue are nominally served on a first-come-first-serve (FCFS) basis and customer's place in queue are distinguished by the time at which they arrive to the queue. In a queueing model of a call centre the customers are callers, the servers are communication equipment (IVR) or telephone agents and queues are populated by callers that await service. Figure 3.1 illustrate the stages that calls go through.

Callers are first served by IVR (interactive voice response) unit and some of the calls are terminated at this stage, and those callers who request to talk to agents pass the dashed line in the diagram (Figure 3.1). Each call that crosses the dashed line can be

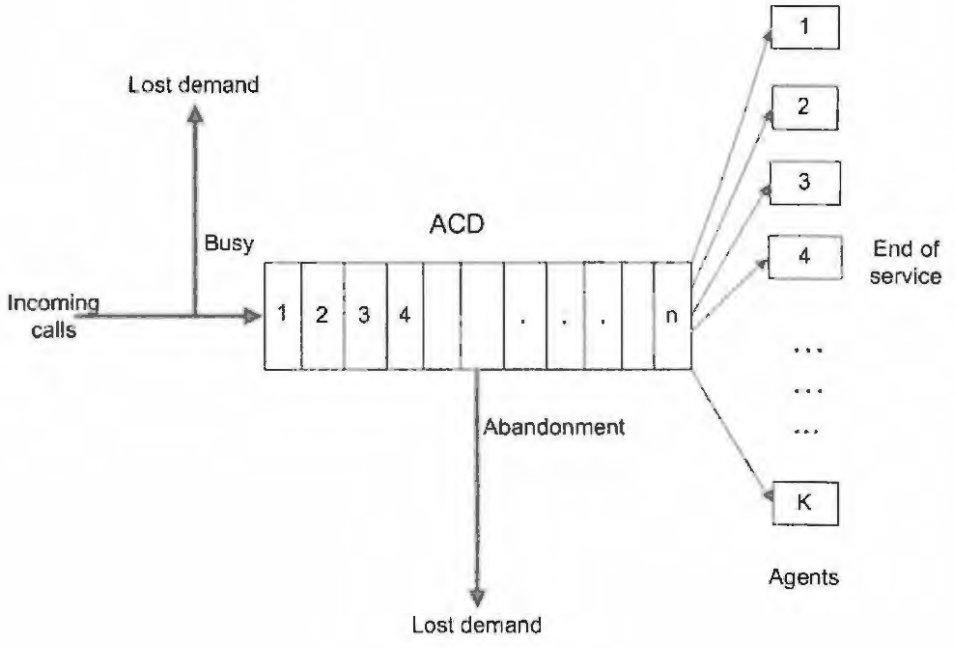

Figure 3.2: Process flow of calls

thought as passing through up to three stages. The first stage is the arrival stage, which is triggered by calls exit from the IVR. If no server or appropriate server is available, then the call enters the second stage, the queueing stage. The last stage is service. Calls that are served immediately skip the queueing stage and calls that abandon never enter the service stage. Figure 3.2 illustrate what happens to the callers who pass through the IVR stage.

Figure 3.2 shows a simplified representative of traffic flows in a call centre. Incoming calls form a single queue, waiting for service from one of k statistically identical agents. There are $k + r$ telephone trunk-lines. These are connected to an Automatic Call Distributor (ACD) which manages the queue, connects customers to available agents and also archives

operational data. Customers who arrive when all lines are occupied encounter a busy signal. The model in figure 3.2 ignores multiple service types and skilled-based routing that are present in many call centres. However, a lot of questions still remain open even for models with homogeneous servers or customers.

3.2 Call Centres as Queueing Systems

Call centres can be viewed as queueing systems. Figure 3.2, which is an operational scheme of a simple call centre shows clearly that a call centre is a queueing system. In a queueing model of a call centre, the customers are callers, servers (resources) are telephone agents or communication equipment and tele-queues consists of callers that await service by the system resource. A Modern call centre is often a complicated queueing network. The mere incorporation of an IVR, prior to joining the agent's tele-queue, already creates two stations in tandem, not mentioning multiple teams of specialised or cross-trained agents.

3.3 Models for Call Centres

Call centres are a growing part of the economy and they are complicated because they involve multiple sites with multiple groups of agents having different skills, serving multiple classes of customers with different needs. Another reason why call centres are complicated is that waiting customers may abandon. Moreover, the probability distributions of both service times and abandonment times often are not nearly exponential, making it inappropriate to directly apply a simple Markovian model (Whitt (2005) [16]), a model with memory-less property. Assuming that waiting customers cannot see the queue, it is natural to assume that the customers abandoning times are identically and independent distributed (i.i.d) with a general distribution. Several models have been and are still been developed for call centres, and a lot of research is still required in this engineering service.

Scientific models are prerequisites for climbing the performance ladder and the Erlang models constitute the starting point. The most commonly used models are the Erlang-C $(M/M/k)$ and Erlang-A $(M/M/k/r+M)$ models.

3.3.1 Erlang-C Model $(M/M/k)$

This model was introduced by Erlang (Erlang (1917) [A]), the founder of queueing theory. It has been prevalent in call centre applications for many years, being the mathematical engine of workforce management. This is the most simplified and easy to use model. Erlang-C assumes Poisson arrivals at a constant rate λ, exponentially distributed service times with rate μ, and k independent statistically-identical agents. The model assumes infinite patience of customers, that is custo, 1911; Erlang, 1917mers who are delayed in queue keep on waiting until they are served, there is no abandonment.

This model does not acknowledge customer's heterogeneity, server's skill levels, or time-dependent parameters. But models which ignore abandonment either distort or fail to provide information that is important to call centre managers. This is so because of the following:

1. Abandonment statistics constitute the only ACD (automatic call distributor) measurement that is customer-subjective. Those who abandon declare that the service offered is not worth its wait.

2. Some call centres focus only on the average *waiting time* of only those who get served. This does not acknowledge abandoning customers. But under such circumstances, the service order that optimises performance is last-in-first-out (LIFO), which clearly suggests that a distorted focus has been chosen.

3. Ignoring abandonment can cause either under- or over-staffing: if service level is measured only for those customers who reach service, the result is unjustly optimistic. The effect of an abandonment is less delay for those further back in line as

well as for future arrivals. This would lead to under-staffing. On the other hand, using workforce management tools that ignore abandonment would result in over-staffing as actually fewer agents are needed in order to meet abandonment-ignorant service goals.

3.3.2 Erlang-A Model ($M/M/k/r+M$)

The classical $M/M/k$ queueing model, called Erlang-C, is the model most frequently used in workforce management of call centres. Customer abandonment is not a minor, let alone a negligible aspect of a call centre operations. The Erlang-A model is an Erlang-C model onto which exponentially distributed customer patience is added, hence the "$+M$" notation.

Mandelbaum and Zeltyn (2004) [8] introduced a simple way to model abandonment. They suggested to enrich Erlang-C by associating each arrival (caller) with an exponentially distributed *patience time* with mean θ^{-1}. An arrival encounters an offered *waiting time*, which is defined as the time that this customer would have to wait given that his or her patience is infinite. If the offered *waiting time* exceeds the customer's *patience time*, then the call is abandoned, otherwise the customer awaits service. The patience parameter, θ, will be referred to as the individual abandonment rate. We denote this model by

$$M/M/k/r+M$$

and refer to it as Palm/Erlang-A. Here the A stands for abandonment and r is the extra waiting space. The model interpolates between Erlang-C and Erlang-B. The latter is the $M/M/k/k$ model, in which there are k trunk lines. Hence, customers that cannot be served immediately are blocked.

The Erlang-A model is characterised by four parameters, which are

- λ - arrival rate or calls per unit of time.
- μ - service rate.
- k - number of servers or agents.
- θ - individual abandonment rate ($1/\theta$ is the average patience time).

These parameters are needed for performance measures that are necessary for efficiency purposes. One of the performance measure that is important and is rarely used in practice is the fraction of customers who encounter a delay. This is a useful measure of congestion. General performance measures considered are

- probability of abandonment,
- average waiting time and
- probability of waiting.

In this model, the processes of arrivals, patience and service are mutually independent. For a customer, the *patience time* θ is the time that the customer is willing to wait for service, a wait that reaches θ results in an abandonment.

3.3.3 Other Models

Whitt (2005) [16] approximated $M/GI/k/r + GI$ model by $M/M/k/r + M(n)$ model, where $M(n)$ denotes state-dependent Markovian abandonment, n being position of the n^{th} caller in the queue. He made this approximation because it produced a Markovian model that can be analysed. Some use inhomogeneous Poisson arrival and consider service time as log-normally distributed (see Brown *et al* (2005) [3]). Generally a lot of models are still being proposed.

3.4 Performance Measures

The performance level of a call centre is usually measured in terms of the waiting time of calls and the productivity of the call centre employees (often called agents). One of the main problems in managing a call centre is the uncertainty in call volume and the fact that calls need to be answered quickly (on average between 10 - 20 seconds). The most popular measure of operational performance is the fraction of served customers that have been waiting less than some given time. For example, in a call centre that caters to emergency calls, waiting times should be very small (if not zero). A common rule of thumb is the goal that at least 80% of the customers be served within 20 seconds. Another important measure that is rarely used in practice is the fraction of customers who encounter a delay. This is a useful measure of congestion. General performance measures considered are:

1. **Blockage** - what percentage of customers will not be able to access the centre at a given time due to insufficient network facilities in place.

2. **Abandon Rate** - call centres measure the number of abandons as well as the rate since both correlate with retention and revenues. While abandonment is affected by the average waiting time in queue, there are a number of other factors that influence this number, such as individual caller tolerance, time of the day and availability of service alternatives.

3. **Self-Service Availability** - increasingly, a call centres activities are being off-loaded today from call centre agents to self-service alternatives.

4. **Agent Occupancy** - it is a measure of time an agent is busy on calls compared to available or idle time, calculated by dividing workloads hours by staff hours. Occupancy is an important measure of how well the call centre has scheduled its staff and how efficiently it is using its resources. If occupancy is too low, agents are sitting around idle with not enough to do.

These performance measures should be able to answer the following questions:

- What is the probability of abandonment?
- How long does a customer expect to wait in the queue before they are served? Also how long will they have to wait before the service is complete?
- What is the probability of a customer having to wait longer than a given time interval before they are served?
- What is the average length of the queue?
- What is the probability that the queue will exceed a certain length?
- What is the expected utilisation of the server and the expected time period during which he will be fully occupied? In fact if we can assign costs to factors such as customer waiting time and server idle time then we can investigate how to design a system at minimum total cost.

For a call centre system, the mean service time is one essential quantity for calculating several basic performance measures, such as average waiting time in the system or average delay in the queue. When combined with a prediction of future arrival rates, it can also be used to predict the future workload that will arrive to the system, which can be used for agent staffing and capacity planning.

3.5 Traffic Intensity

One factor that is of importance is traffic intensity $\rho = \frac{arrival\ rate}{departure\ rate}$, where arrival rate is the number of arrivals per unit time and departure rate is the number of departures per unit time. Traffic intensity is a measure of congestion of the system. If it is near to zero, then there is very little queueing and as the traffic intensity increases (to near 1 or even greater than 1) the amount of queueing increases. Traffic intensity is determined by the

arrival rate (λ), service rate (μ) and the number of servers (k), that is

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

There are two types of traffic in telecommunications

- offered traffic,
- carried traffic.

Offered traffic is the mean number of arrivals per mean service time. Namely, it is equal to the ratio $\frac{\lambda}{\mu}$. This ratio is thetraffic intensity ρ for $M/M/1$ queue. In $M/M/1$, we must have that ρ cannot exceed unity for stability and it also represents the server utilisation which cannot exceed unity in this queue. Offered traffic is measured in *Erlangs* named after the Danish mathematician A. K. Erlang who was the originator of queueing theory and tele-traffic. One *Erlang* represents traffic load of one arrival, on average, per mean service time.

Carried traffic is defined as the mean number of customers or calls leaving the system after completing service during a time period and is equal to the mean service time. It is also measured in Erlangs and it is equal to the mean number of busy servers which is equal to the mean queue size.

In practice the number of servers is limited and the offered traffic is higher than the carried traffic because some of the calls are blocked due to call congestion when all circuits are busy.

3.6 Workforce Management: Staffing

As the technology has become more sophisticated, product and process knowledge as well as customer information have been embedded in the system, reducing training costs.

Continual control can be maintained over the call times, call volumes and virtually every activity the employee performs. Additionally, technology allows monitoring of the quality of the agents' interactions. Supervisors have the ability to assess agent performance by randomly checking their calls or computer screens. Thus agents can be monitored closely for performance and burnout, and appropriate interventions made.

It is important for a call centre's manager to be able to anticipate the impact of changes on the service level. Examples of such changes are an increase in the call arrival rate due to a marketing campaign, or a change in the number of agents on shift. The classical formula such as Erlang-C or Erlang-A, if fed by point forecasts of the arrival, service and time to abandonment rate for the target period, can be used to find the minimal staffing that meets all targets for performance constraints. Brown *et al* (2005) [3] find Erlang-A to work well against empirical data.

3.6.1 Square-root Staffing

- The offered load (intensity) parameter R represents the amount of work (measured in time units of service) that arrives to the system per unit time. It is significant to the staffing problem since R and its neighbourhood provide nominal staffing levels, deviations from which could result in extreme performance (staffing high above R would result in a very high quality of service and staffing far below R would result in a very high utilisation).

  The square root safety staffing (Avramidis (2005)[2]) $k = R + \delta$ (where $\delta = \beta R^{\frac{1}{2}}$) is for achieving a given delay probability α under an offered load (traffic). $R = \frac{\lambda}{\mu}$ denote the average offered load (where λ is the average call arrival rate and $\frac{1}{\mu}$ is the mean call duration), is the safety staffing above the load to account for stochastic variability. In this formula k is the number of servers and β is a quality of service parameter; the larger it is, the better is the operational service level. The approximation has been extended to more general queues, and is very robust. Large k

ensures simultaneously high quality of service and high server utilisation, which characterise a quality and efficiency driven call centre. Of course, in practice the value of k derived from this formula must be rounded to an integer.

The square-root staffing rule has a conceptual dimension that clearly shows the economies of scale in running a large call centre. It also has an economic dimension, which allows one to determine actual values for the constant β by trading off service level and agents costs. Indeed, for large call centres

- $\beta = 1$ or larger would give rise to negligible abandonment (quality-driven call centre).
- $\beta = -1$ or lower would give about $8 - 12\%$ abandonment (efficiency-driven call centre).
- β around 0 (preferably positive) would result in about $2 - 3\%$ abandonment (Whitt (1999) [15]).

Staffing recommendations depend on the measure of performance to be controlled as well as on the patience distribution beyond the mean. A common naive "deterministic" approach to staffing can yield good to very good results, in the presence of abandonment.

3.6.2 Real-time Staffing

This is a dynamic staffing in the real time scale of length of a call, done in response to observed system state, including information about the history of the call centre on that day and information about the calls currently in process. The whole purpose is to have sufficient flexibility to be able to add agents when they are needed and to pull them off to do alternative work when they are not needed. Of course, call centre managers routinely do some form of real-time staffing, but systematic real-time staffing based on substantial data and analysis so far is only a dream. Real-time staffing places great challenges upon queueing theory, because it requires that we consider the time-dependent behaviour of the

queueing system. Usually, performance analysis is confined to a description of the steady-state behaviour. However, recent research has begun seeking algorithms to describe the time-dependent behaviour of queueing systems (see Whitt (2005) [16]).

3.6.3 Short-term Staffing

This is the daily staffing done in response to forecasted demand and knowledge of the available agents. A significant challenge in short-term staffing is that the call arrival rate varies significantly over the day. In some cases, the call holding times are sufficiently short that the time-dependence can be safely ignored. Then, it is appropriate to use a dynamic steady state, using the parameters that are appropriate at any given instant (a short-term average), rather than the long-run average parameters over an entire day. A significant component of short-term staffing is scheduling work shifts for the agents, including breaks for coffee and lunch.

3.6.4 Long-term Staffing

This staffing is done in the time scale of the length of time required to hire and train agents. There are different challenges in the long-term staffing when it takes a relatively long time to train new agents. Over the longer time scale, it is also important to address agents atrition and agents career paths. The purpose is to have agents and customers both satisfied.

4 METHODOLOGY

4.1 Limitations of Mathematical Approach

Classic queueing theory is often too mathematically restrictive to be able to model all real-world situations exactly. This restriction arises because the underlying assumptions of the theory do not always hold in the real world. For example, the mathematical models often assume infinite numbers of customers, infinite queue capacity or no bounds on inter-arrival or service times, when it is quite apparent that these bounds must exist in reality. Often, although the bounds do exist, they can be safely ignored because the differences between the real-world and theory is not that different. In other cases the theoretical solution may either prove intractable or insufficiently informative to be useful.

Alternative means of analysis have thus been devised in order to provide some insight into problems which do not fall under the mathematical scope of queueing theory, though they are often scenario-specific since they generally consist of computer simulations and/or of analysis of experimental data. Therefore in-order to analyse $M/GI/k/\tau + GI$ model, we have to resort to simulation sinc a theoretical solution is intractable. A free software, EZSIM (downloaded at http://www-rcf.usc.edu/~khoshnev/software.html) was used to simulate results presented for the $M/GI/k/\tau + GI$ model.

4.2 Simulation Software (EZSIM)

EZSIM is the simulation software used in the results presented in this thesis. It is a general purpose process-oriented simulation modelling tool for discrete systems involving entity flow. Models in EZSIM are presented in network form. Each node in the model network represents a process and branches show the entity path from one node to another. The major issues in simulation program development consists of the initialisation routine, the input routine, the event timing routine, the arrival event routine, the departure event routine, the statistics routine and the output routine (see Khoshmnevis (1994) [4]). Figure 4.1 shows the main modules in a queueing simulation program.

General event-based simulation programs have the following modules:

- *Main routine* - transfers control between the major modules of the program.
- *Initialisation routine* - initialises all variables and clear the statistical data that may been gathered in a previous run.
- *Events timing routine* - locates the most imminent future event, advances the simulation clock to the time of the event and calls the corresponding event-processing routine.
- *Future events list (calendar)* - contains the list of the unprocessed future events. Other information such as the attributes of entities causing the event, may be stored in this structure.
- *Event processing routines* - are individual modules each representing an event in the system.
- *Library routines* - include a module for pseudo-random number generation and several modules for random variates with various distribution types.

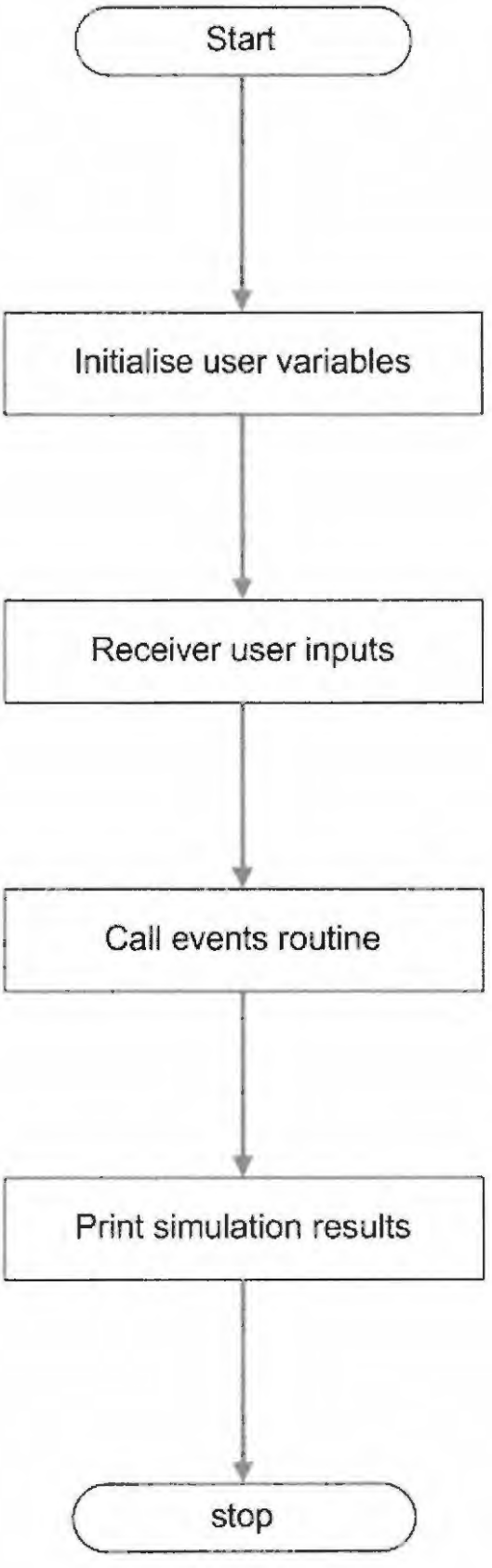

Figure 4.1: Flowchart of the main module in the queueing simulation program

- *Statistics routine* - collects and processes certain statistics that are specified by the user and general quantities desired in statistics reports such as mean, standard deviation, minimum, maximum and last value observed at the end of a simulation.
- *Output routine* - gathers the values collected by the statistics routine and may perform some operations on these values to create measures such as overall averages.

EZSIM allows its user to quickly build a model of the system under study, run the model in either the batch mode or animation mode, verify the model and observe the desired statistics. The user can quickly change model parameters or configuration and run the model several times in a single session.

Numerous windows (menus and context-sensitive help prompts) are available throughout the above stages. EZSIM enables the user to concentrate on the system structure and high-level dependencies while the system checks the integrity of the model structure as it is being constructed by the user. The following stages are involved in a complete stand-alone EZSIM session and are shown in figure 4.2.

- model network construction,
- nodal parameter specification,
- model initialisation,
- desired statistics specification,
- execution in batch or in real-time with animation,
- output observation on screen,
- model and output disk file and/or hard-copy generation,
- possible model modification and re-execution.

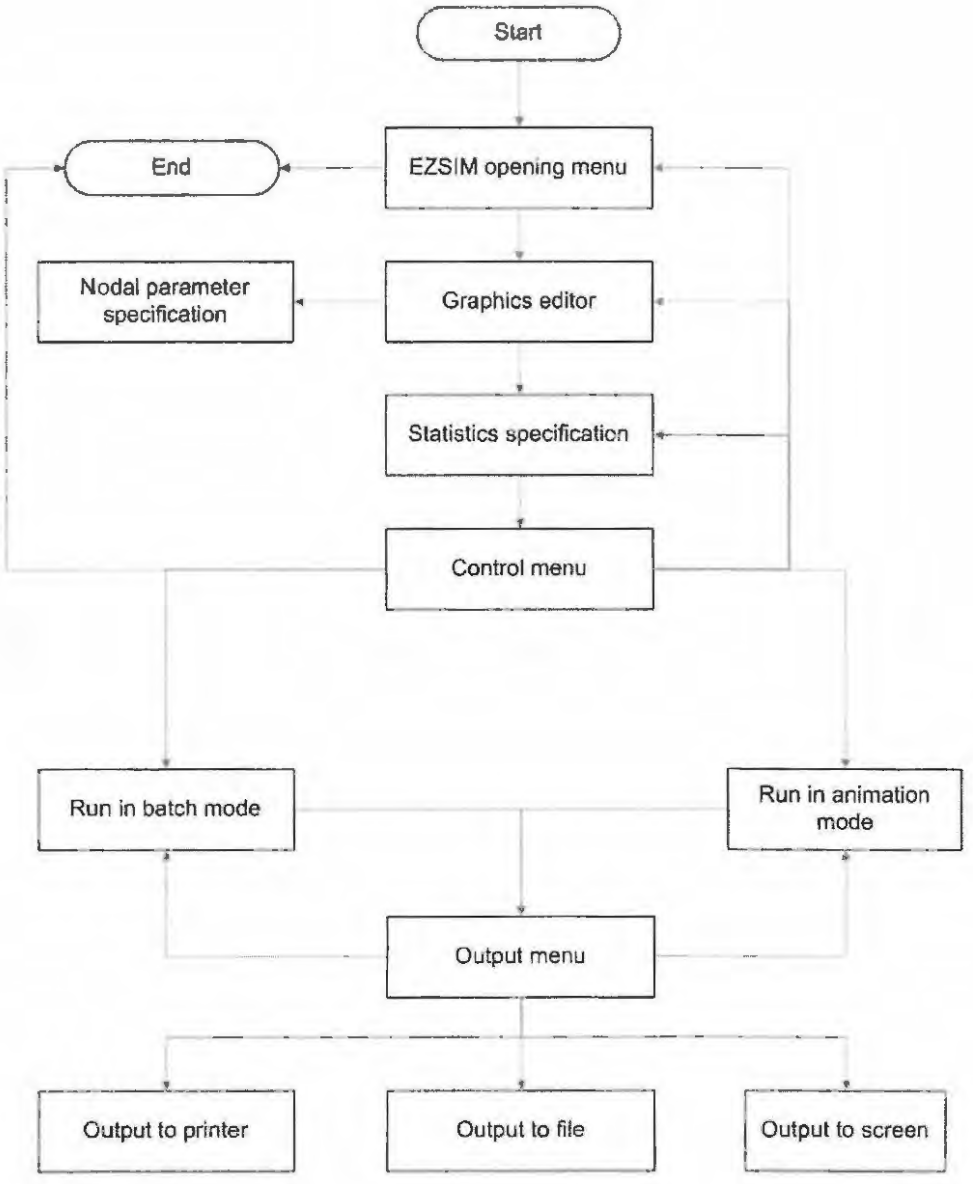

Figure 4.2: Flowchart of the procedure for using the stand-alone EZSIM environment

EZSIM starts with its opening menu, which allows for file creation, retrieval, listing, renaming and deletion. After this specification the graphics screen is invoked. It provides an environment for construction or modification of a model network. The first stage under graphics screen is network construction activity. Using the help key (H) while in the graphics mode, all graphics control parameters and node definitions can be reviewed on the screen

A node is a common process that some entities go through. After node selection a dedicated window that guides the user in identifying each parameter relevant to the node in question is shown. After completing the answers to the questions regarding a node, the control key returns to the graphics screen for selection of the next node or further network construction and editing. Hitting the Esc key after the completion of the nodal parameter identification stage starts the system initialisation stage. This initialisation stage occurs only if there are quantities (such as user variables and resources) that need to be specified.
Nodes commonly used in queueing problems are Source, Queue, Facility, Delay and Terminate.

Source node creates entities. Each entity that the Source node creates may have a name, which becomes an attribute for the entity (called NAME). It is possible to specify the total number of entities to be created by the Source node.

Facility node acts as a server. Entities remain in the node for the duration of their service. When the node is occupied, the arriving entities have to wait until the node is free. A Facility node must be preceded by a Queue node. Multiple parallel or series servers may be specified for a given Facility.

Queue node represents buffers before Facility nodes and is always succeeded by a facility node. A queue may have capacity limitation, various priority disciplines such as first-in-fist-out (FIFO), last-in-first-out (LIFO) and so on.

Delay node is used for creating a delay that corresponds to the traversal time of the

entity from one node to another, or may also be used for collecting several incoming branches and for creating several outgoing branches.

Terminate node ends the path of entities. When the entity enters a Terminate node it is considered to be out of the network.

EZSIM has a unique capability of provision for prevention of coding as well as logical errors in simulation. This is so since all names of nodes, variables, attributes, and resources are entered only once when building an EZSIM model. All subsequent references to names are made through selection windows. By incorporating the generic logical rules governing discrete systems, EZSIM also offers some provision for preventing logical errors. For example, connection between certain nodes are not allowed, system variables cannot be overwritten, the nature of a considerable number of statistics types are automatically distinguished, the user is aided in building conditional and other forms of expressions with minimal chance for errors, and an event animation module for effective checking of entity flow and possible blockages is provided.

4.3 Specifications for the Model

In this thesis, a single call centre with a single group of agents and serving a single group of callers will be considered. The model to be used assumes Poisson arrivals, exponential service times, and general and independently distributed abandonment. EZSIM will be used to simulate performance measures using the following specifications

Source node: the node is called ARRIVE and the entity name is CUST; first creation time is zero; time between creations is exponential with different inter-arrival times depending on the number of servers and traffic intensity; default maximum number of creation is used and time to stop creation is 2200.

Queue node: the node is called LINE; queue capacity is 10 if servers are 4, 12 if servers are 5 and 15 if servers are 6. For full queue situation BALK is selected; when queue

is full, entities balk to LOST (a Terminate node); for balking traversal time, zero is selected and the same for initial number in queue. FIFO is selected as the queue discipline; for maximum waiting time for entities in the queue, different distribution are selected each at a time and when waiting time is up, entities go to ABANDON (another Terminate node).

Facility node: the name for the node is SERVER. Number of parallel servers used are 4, 5 and 6. For service duration, exponential is used with mean 1. For other specifications, their defaults are selected.

Terminate node: three nodes are used and are labelled LOST, ABANDON and LEAVE and only LEAVE is connected to the Facility node. Their defaults are selected in all the nodes.

For the Source node, two traffic intensities are used which are 0.926 and 1.1. For traffic intensity of 0.926, three inter-arrival times are used 0.27, 0.22 and 0.18 which corresponds respectively to 4, 5 and 6 servers. Likewise for traffic intensity of 1.1, inter-arrival times used are 0.2083, 0.167 and 0.1389. The Source node is connected to the Queue node and the Queue node to the Facility node then the Facility node to the Terminate node labelled LEAVE.

In the Desired Statistics, LOST, ABANDON, SERVED and TOTAL were used as Statistics names and Counting of entities at a node was selected for each of the statistics name. Finally length of simulation run was 2300 and length of transient period was 1100 (measurements in minutes). Initially the simulation was run in animation mode and when everything was observed to be in-order, the batch mode was used to get the statistics for a sample of ten using different seeds for each of the nodes.

4.4 Parameter Estimation

In order to apply the model, it is necessary to input values for its four parameters: λ (arrival rate), μ (service rate), θ (abandonment rate) and k (number of servers). In call centres, a model should be used to support solutions of the staffing problem, namely: how many agents should be answering calls during a specified time period. Typically, the goal is to provide a satisfactory service level (for example, fraction abandoning less than 5%), but sometimes one optimises an economic measure - minimise cost or maximise revenues.

Arrivals: An arrival is an event that indicates a need for a service. The stream of customers that demand service at a particular facility will be called the arrival pattern or process. Typically such a stream must be thought of as stochastic, for the exact instants at which customers appear tend to vary haphazardly and unpredictably. Common call-centre practice assumes that the arrival process is Poisson with a rate that remains constant for blocks of time, often individual 30 minutes or 60 minutes, then a queueing model is fit for each block of time. The intended model $M/GI/k/r+GI$ queue ignores the time dependence almost found in call arrival processes, but the time dependence often tends to be not too important over short time intervals, such as $15-60$ minutes.

The goal is to predict these arrival rates, over short time-intervals (15, 30 minutes or 1 hour), chosen so that the rates are approximately constant during an interval. Then the time homogeneous model is applied separately over each interval. The goal can be achieved in two stages. First, time series algorithms are used to predict daily volumes, taking into account trends and special days. Second, one uses parametric regression techniques for predicting the fraction of arrivals per time-interval, out of the daily-total. This fraction, combined with the daily total, yields actual arrival rates per each time-interval.

Service: The service facility may be described as the element of the service system that

actually satisfies the demand customer. Service durations are assumed general and independent. Average service times tend to be relatively stable from day to day. However, they often change depending on the time of day. Some studies (Whitt (2005) [15]) find that exponential distribution provides an adequate fit to empirical data. Also other parametric families that arose in applications includes gamma and log-normal. Brown *et al* (2005) [3] find that the log-normal distribution provides an excellent fit to the data, especially after excluding the short time.

Idle-time, namely the time that an agent is immediately accessible for service, is normally used to estimate the average service time during any time interval. When the agents are busy, arriving calls often queue. Often when people find themselves in circumstances involving delay, they react to minimise or avoid it. Others balk, refuse to join the line or queue, when they find the line to be too long and others renege (abandon), leaving a line after a period spent waiting for service.

Waiting time: All customers who abandon the tele-queue would have waited. Also, the times at which customers who are served would have abandoned, had they not been served, are not observed. Therefore, the characterisation of patience and time to abandon is based on censored data (Miller (1981) [9]). The maximum waiting time a customer is willing to wait in queue is his *patience time*, A, also know as time to abandonment. The time a customer must wait before beginning service is his *virtual queue time*, V. The *actual waiting time* is $T = min(A, V)$, terminated by either abandonment (when $V > T$), or beginning of service ($V = T$). With respect to parametric models of patience, the Weibull distribution is a possible candidate because of its wide use in survival analysis which involve data that is censored.

The ACD (automatic call distributor) collects data on T and the abandonment indicator, I $(V > T)$; The *patience time*, A, cannot be observed. Real data will encounter the classical problem of censoring, and therefore requires techniques from the field of survival analysis (Miller (1981) [9]). This procedure is used to estimate

the distribution of patience.

4.5 Beyond Erlang-A

A natural first approximation to try for the $M/GI/k/r+GI$ queueing model is the more elementary Erlang-A model, $M/M/k/r+GI$, where we obtain both the exponential time to abandon and the exponential service time distribution by using exponential distributions with the same means as given in the distributions. This research will concentrate more on the distribution of abandoning times, hence the $M/GI/k/r+GI$ will be approximated with $M/M/k/r+GI$, where we have assumed that the service times are exponential.

There is a vast literature on statistical inference and forecasting, but surprisingly little has been devoted to stochastic processes and much less to queueing models in general and call centres in particular. Indeed, the practice of statistics and time series in the world of call centres is still at its infancy, and serious research is required to bring it to par with its needs.

5 ANALYSIS

All simulation experiments reported in this thesis were based on 10 independent replications of 2300 runs observed using different seeds from one to twenty. The statistics were collected after a transient period of 1100 runs as it was observed that statistics were now reliable. Time was in minutes. Two different traffic intensities of 0.926 and 1.1 were used for all the considered distributions. This is in line with other studies (Nikolic (2006) [10] and Whitt (2005) [15]) where similar intensities were used. Experiments were repeated with 3, 4 and 5 servers.

Independent replicates made it possible to reliably estimate confidence intervals using the t-statistic and for all estimates, half confidence interval width at 95% confidence interval were calculated.

In trying to understand the behaviour of the $M/GI/k/r+GI$ model, an important initial insight is that, in contrast to single-server queues, the *waiting times* in multi-server queues (with large number of servers) tend to be quite small relative to the mean service times. This phenomenon is well established in call centres and is reflected by the classical 80/20 rule (Whitt (1999) [15]).

5.1 Performance measures

Performance measure is the specific representation of a call centre capacity, process or outcome deemed to be relevant to the assessment of perfomance.

The following perfomance measures are considered:

- L_Q - *length of queue*,
- W - *waiting time*,
- U - *server utilisation*,
- B - *proportion of busy periods*,
- I - *proprortion of idle periods*,
- $P(A)$ - *probability of abandonment*, and
- $P(Balk)$ - *probability of balking*.

5.2 Assumptions

The following assumptions are made in the analysis of the simulated data:

- replication of experiments are independent for each seed;
- service times are exponentially distributed;
- Erlang-A (model with abandoning time that is exponentially distributed) is considered as the standard for comparison with other distributions; and
- callers who abandon do not retry.

5.3 Distribution of Service Times

Already, some work has been done in the distribution of service times (Brown *et al* (2005) [3]) and some distributions such as log-normal and gamma have provided a good fit to certain types of call centre data. This research assumes that services times are exponentially distributed and hence we focused more on the distribution of abandonment.

5.4 Distribution of Abandoning Times

To define performance measures we examine, S (the event that a typical customer who enters the system is eventually served) and A (the event that a typical customer who enters the system abandons before starting service). The mean delay (system *waiting time*) of a customer is the time from the moment a customer arrives until his or her service is completed.

5.4.1 Light Loads

Table 5.1 shows the simulations using Erlang-A model and three other distributions of abandonment. The distributions considered are uniform, log-normal and gamma with parameters (0.5, 1.5); (1, 1) and (2, 2) respectively (where the first and the second values are the scale and shape parameters respectively). All the models have common inter-arrival time of 0.27, mean service time $\mu^{-1} = 1.0$ and mean time to abandon 1.0 (all measurements are in minutes). Using the Erlang-A model as the standard, we noticed that the estimate of mean *length of queue* was bigger for the other distributions of abandonment and this caused the *waiting times* to be larger as well. The estimate of mean *length of queue* seems almost the same for exponential and gamma distributions of abandonment and their standard errors are also similar (with gamma having 1.8% while exponential has 2.4%).

The estimate of variance for the *length of queue* was small for exponential distribution while for the rest of other distributions for abandonment are very high (as high as 1.87 for log-normal while for exponential it is 0.87). The estimate for mean *waiting time* in queue was high for uniform and log-normal and almost doubled that of exponential and gamma with the exponential distribution having a bigger standard error of 6% compared to 5% for gamma. Similarly, exponential and gamma distributions have small estimates of variance for *waiting time* in queue but the exponential distribution had double standard error as compared to the gamma distribution (0.1 for exponential and

Table 5.1: **Comparison of steady-state performance measures** for different distributions of abandonment; exponential (M/M/4/10+M), uniform (M/M/4/10+U(0.5, 1.5)), log-normal (M/M/4/10+LN(1, 1)) and gamma (M/M/4/10+gamma(2, 2)) with 4 servers and traffic intensity of 0.926.

Performance Measure	M/M/4/10+M	M/M/4/10+U(0.5,1.5)	M/M/4/10+LN(1,1)	M/M/4/10+GAMA(2,2)
$E(L_Q)$	0.318 ± 0.017	0.584 ± 0.029	0.882 ± 0.063	0.353 ± 0.013
$Var(L_Q)$	0.869 ± 0.037	1.327 ± 0.041	1.871 ± 0.077	0.935 ± 0.022
E(W)	0.077 ± 0.046	0.141 ± 0.006	0.219 ± 0.014	0.083 ± 0.0035
Var(W)	0.18 ± 0.0072	0.275 ± 0.006	0.433 ± 0.015	0.187 ± 0.0035
E(U)	1.549 ± 0.0137	1.626 ± 0.025	1.679 ± 0.021	1.564 ± 0.015
Var(U)	1.736 ± 0.0086	1.795 ± 0.013	1.828 ± 0.013	1.751 ± 0.010
E(I)	0.146 ± 0.0047	0.116 ± 0.008	0.096 ± 0.006	0.14 ± 0.0067
E(B)	0.503 ± 0.0064	0.501 ± 0.0098	0.501 ± 0.005	0.5 ± 0.0034
P(A)	0.17 ± 0.008	0.121 ± 0.008	0.082 ± 0.006	0.158 ± 0.0045
P(Balk)	0	0	0.0064 ± 0.0013	0

0.05 for gamma).

There seems to be not much difference in the estimates of mean *server utilisation* for all the distributions, although uniform and gamma distributions of abandonment shows high standard errors, 3.5% for the uniform distribution which was almost double that of the exponential distribution (which was 1.9%). The estimates of variance for *server utilisation* was almost the same for all the considered distributions. The estimate of mean for *proportion of busy periods* was almost the same, although the log-normal distribution of abandonment had a smaller standard error of 0.7% (exponential had 0.9% and 0.87% for gamma). Proportion of time that servers are idle was smallest for the log-normal distribution. This is because this distribution has the longest *waiting time* as well as the longest queue as compared to other distributions for abandonment. Hence servers are busier for the log-normal distribution than for the other distributions. The exponential distribution for abandonment had the biggest percentage of estimate of mean for *server idleness*.

It can be concluded that estimates of the mean and variance for *server utilisation* and estimates of means for *proportion of idleness* as well as *proportion of busy periods* seems not that different for the four distributions of abandonment. But if we consider estimates of mean *length of queue* and mean *waiting time* in queue, exponential and gamma distributions seems to be the best.

The estimate for the *probability of abandoning* was larger for the Erlang-A model (with exponential abandonment distribution), doubling that of log-normal with similar standard errors. But log-normal had high *probability of balking* ($P(Balk)$), that is, more people do balk and therefore reduce the number of those who abandon. The gamma distribution had a low estimate for *probability of abandonment* as compared to the exponential distribution (exponential had 0.17 and 0.16 for gamma). Generally gamma and exponential distributions, although having some differences, seems to show similar

performance measures.

Table 5.2 shows the same distributions for abandonment as in table 5.1, but this time with 5 servers. The inter-arrival time was 0.22 and the rest of the other parameters remained the same as in the last section. The same trend observed in table 5.1 is also reflected in this table 5.2, although the estimate of mean *length of queue* had slightly increased and the estimate of mean *waiting time* in the system had slightly decreased. The estimate of mean *proportion of idle periods* is shorter for uniform and log-normal distributions, and high for the exponential distribution. The estimate of mean *proportion of busy periods* changed slightly for all the considered distributions.

The estimates of mean and variance of *server utilisation* as well as mean for *idle* and *busy periods* changed slightly for all different distributions of abandonment. The estimate of mean *server utilisation* was 2 for all the distributions, and the estimate of mean *proportion of busy periods* estimate was 0.5 for all the distributions. As observed before (table 5.1), the log-normal distribution had the smallest estimate for *probability of abandoning* with 0.004 being the *probability of balking*, while other distributions do not have customers who balk. Gamma and exponential distributions have estimates of 0.128 and 0.144 respectively for *probability of abandonment*. These two tables (tables 5.1 and 5.2) confirm that gamma distribution of abandonment seems to be the best amongst the four, followed by exponential distribution (although it had a high probability of abandoning). This is because good performance measures should have smaller values for mean *length of queue*, mean *waiting time*, mean *proportion of idle periods* and *probability of abandoning* as well as smaller variances and *standard errors*. With different sets of data from call centres, one would expect different distributions to suit different situations.

The same trend (as shown in tables 5.1 and 5.2) was observed with 6 servers as shown in table 5.3. The estimate of mean *length of queue* increased as the number of servers increased for all the distributions (figure 5.1). The variance estimates also increased

Table 5.2: **Comparison of steady-state performance measures** for different distributions of abandonment; exponential (M/M/4/10+M), uniform (M/M/4/10+U(0.5, 1.5)), log-normal (M/M/4/10+LN(1, 1)) and gamma (M/M/4/10+gamma(2, 2)) with 5 servers and traffic intensity of 0.926.

Performance r Measure	M/M/5/12+M	M/M/5/12+U(0.5,1.5)	M/M/5/12+LN(1,1)	M/M/5/12+GAMA(2,2)
$E(L_Q)$	0.335 ± 0.0158	0.614 ± 0.03	1.032 ± 0.036	0.374 ± 0.023
$Var(L_Q)$	0.945 ± 0.0315	1.441 ± 0.05	2.187 ± 0.047	1.011 ± 0.031
E(W)	0.066 ± 0.0035	0.124 ± 0.005	0.208 ± 0.0066	0.075 ± 0.0038
Var(W)	0.162 ± 0.0054	0.251 ± 0.0079	0.413 ± 0.009	0.167 ± 0.0048
E(U)	1.964 ± 0.022	2.059 ± 0.021	2.129 ± 0.03	1.976 ± 0.026
Var(U)	2.155 ± 0.013	2.235 ± 0.012	2.284 ± 0.017	2.176 ± 0.016
E(I)	0.137 ± 0.0056	0.108 ± 0.0056	0.089 ± 0.0079	0.132 ± 0.007
E(B)	0.502 ± 0.054	0.501 ± 0.0063	0.505 ± 0.007	0.5 ± 0.0067
P(A)	0.144 ± 0.0066	0.089 ± 0.0077	0.073 ± 0.004	0.128 ± 0.006
P(Balk)	0	0	0.004 ± 0.0011	0

Table 5.3: **Comparison of steady-state performance measures** for different distributions of abandonment; exponential (M/M/4/10+M), uniform (M/M/4/10+U(0.5, 1.5)), log-normal (M/M/4/10+LN(1, 1)) and gamma (M/M/4/10+gamma(2, 2)) with 6 servers and traffic intensity of 0.926.

Performance measure	M/M/6/15+M		M/M/6/15+U(0.5,1.5		M/M/6/15+LG(1,1)		M/M/6/15+GAMA(2,2)	
$E(L_Q)$	0.366	± 0.02	0.77	± 0.047	1.102	± 0.072	0.439	± 0.017
$Var(L_Q)$	1.018	± 0.043	1.69	± 0.048	2.339	± 0.094	1.141	± 0.0235
E(W)	0.06	± 0.003	0.125	± 0.006	0.185	± 0.011	0.071	± 0.0023
Var(W)	0.147	± 0.005	0.246	± 0.006	0.367	± 0.014	0.159	± 0.0023
E(U)	2.404	± 0.014	2.535	± 0.023	2.59	± 0.022	2.445	± 0.0173
Var(U)	2.607	± 0.01	2.708	± 0.017	2.747	± 0.014	2.635	± 0.0103
E(I)	0.126	± 0.004	0.092	± 0.0056	0.081	± 0.004	0.114	± 0.005
E(B)	0.502	± 0.0045	0.497	± 0.0068	0.498	± 0.0045	0.5	± 0.0048
P(A)	0.133	± 0.0053	0.083	± 0.006	0.0068	± 0.0049	0.12	± 0.0053
P(Balk)		0		0	0.001	± 0.0003		0

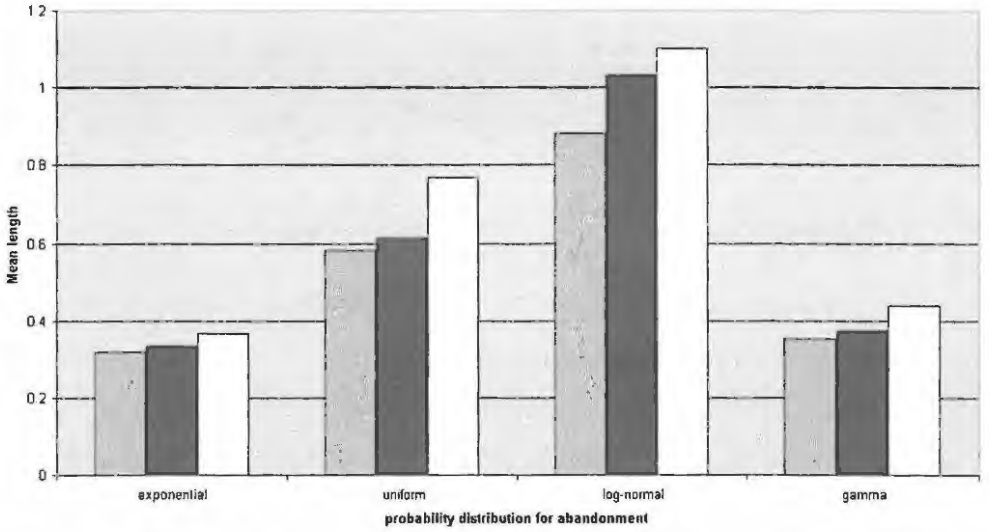

Figure 5.1: Bar graph of estimated mean *length of queue* for different distributions of abandoning times with 4, 5 and 6 servers respectively.

across the different distributions of abandonment. The estimated mean *waiting time* in system and its variance slightly decreased for all the distribution although the traffic intensity was kept at 0.926. The estimate of mean *server utilisation* and its variance also increased across the different distributions. The mean *proportion of idle* or *busy periods* estimates changed slightly, but the *probability of abandonment* decreased as the number of servers increased, with exponential decreasing from 0.17 to 0.13 as servers increased 4 four to 6. The estimate for *probability of abandoning* for the gamma distribution decreased from 0.16 to 0.12 as servers increased from four to six. Generally, as the servers are increased whilst maintaining the same traffic intensity, *the probability of abandoning* and mean *waiting times* estimates decreased.

5.4.1.1 Testing for Equality of Performance Measures

We compared the perfomance measures for all other distributions of abandoning time with that of exponential. Table 5.4 revealed that the other two distributions for abandonment (that is, uniform and log-normal) did not have overlapping confidence intervals with that of exponential for all the perfomance measures. This means that the perfomance measures for exponential distribution are not the same as for the uniform and log-normal distributions. The confidence intervals for gamma and exponential distributions for abandonment were overlapping and the majority of gamma's perfomance measures were contained in the confidence intervals of exponential (for example, estimates of mean *waiting times*, mean of idle as well as busy periods), therefore statistically, the perfomance measures for gamma and exponential are not necessarily different. The gamma distribution had a smaller confidence width when compared to exponential (see tables 5.1 to 5.3) distribution of abandonment and this implies that gamma distribution has better performance measures than exponential.

Table 5.4: 95% Confidence Intervals for exponential, uniform, log-normal and gamma distributions of abandoning times.

95% Confidence Interval

Using four servers with exponential, uniform, log-normal and gamma

Perfprmanc Measure	M/M/4/10+M	M/M/4/10+U(0.5,1.5)	M/M/4/10+LN(1,1)	M/M/10+GAMA(2,2)
$E(L_Q)$	(0.30; 0.335)	(0.555; 0.613)	(0.819; 0.945)	(0.340; 0.366)
$Var(L_Q)$	(0.832; 0.906)	(1.286; 1.368)	(1.794; 1.948)	(0.913; 0.957)
E(W)	(0.031; 0.123)	(0.135; 0.147)	(0.105; 0.233)	(0.08; 0.087)
Var(W)	(0.173; 0.187)	0.135; 0.147)	0.205; 0.233)	(0.078; 0.087)
E(U)	(1.535; 1.563)	(1.601; 1.651)	1.658; 1.7)	(1.549; 1.579)
Var(U)	(1.729; 1.745)	(1.782; 1.808)	(1.815; 1.841)	(1.75; 1.752)
E(I)	0.141; 0.1507	(0.108; 0.124)	(0.09; 0.102)	(0.133 0.147)
E(B)	(0.496; 0.509)	(0.491; 0.511)	(0.496; 0.506)	(0.497; 0.503)

Using five servers with exponential, uniform, log-normal and gamma

$E(L_Q)$	(0.319; 0.351)	(0.611; 0.617)	(0.996; 1.068)	(0.351; 0.397)
$Var(L_Q)$	0.914; 0.977)	(1.436; 1.446)	(2.14; 2.234)	(0.98; 1.042)
E(W)	(0.064; 0.07)	(0.119; 0.129)	(0.201; 0.215)	(0.071; 0.079)
Var(W)	0.157; 0.167)	(0.243; 0.259)	(0.404; 0.422)	(0.162; 0.172)
E(U)	(1.942; 1.986)	(2.038; 2.08)	(2.1; 2.132)	(1.95; 2.002)
Var(U)	(2.142; 2.168)	(2.223; 2.247)	(2.267; 2.301)	(2.76; 2.192)
E(I)	(0.131; 0.143)	(0.102; 0.114)	(0.081; 0.097)	(0.125; 0.139)
E(B)	0.448; 0.556)	(0.496; 0.507)	(0.498; 0.512)	(0.493; 0.507)

5.4.1.2 Comparison of Exponential and Gamma Distributions with Erlang Distributions

Since we have observed that exponential and gamma distributions have the same performance measures, we compared them with the Erlang distribution with parameters (1, 2) and (1, 0.5) (where the first and second values are the scale and shape parameters respectively). Again comparison were made with 4, 5 and 6 servers. Table 5.5 shows these performance measures for 4 servers. The two Erlang distributions (depicted as E_2 (Erlang(1, 2)) and $E_{0.5}$ (Erlang(1, 0.5))) showed that $E_{0.5}$ had better performance measures when compared to E_2 (see figure 5.2). E_2 has estimates of mean *length of queue* and *waiting time* in system that is three times bigger when compared to $E_{0.5}$ and the same relates to the standard errors, but $E_{0.5}$ had a smaller variance of *length of queue* estimate with high standard error of 13.2% compared to 5% of E_2, which is lower. There is not much difference in the estimates of mean *server utilisation* for the two distributions and the same applies to the estimates of mean *proportion of idle* and *busy periods*. E_2 had a smaller *probability of abandoning* of 0.14 compared to 0.20 of $E_{0.5}$. Generally $E_{0.5}$ has better performance measures when compared to E_2. The same trend was maintained with 5 and 6 servers (tables 5.5 and 5.6 respectively) .

On comparing $E_{0.5}$ with exponential and gamma distributions of abandonment, it was observed that $E_{0.5}$ has much better performance measures than the other two distributions (gamma and exponential). The estimates of mean *server utilisation* and mean *proportion of idle periods* were slightly higher for $E_{0.5}$ when compared with the other two distributions (gamma and exponential). Also the estimate for *probability of abandoning* was high for $E_{0.5}$. The trend was the same when servers were increased from 4 to 6 (tables 5.6 and 5.7).

Table 5.5: **Comparison of steady-state performance measures** of different Erlang distributions (M/M/5/12+E(1, 0.5), M/M/5/12+E(1, 2) and Erlang-A) with gamma distribution (M/M/5/12+gamma(2, 2)) with 4 servers and traffic intensity of 0.926.

Performance Measure	M/M/4/10+M	M/M/4/10+GAMA(2,2)	M/M/4/10+E(1,2)	M/M/4/10+E(1, 0.5)
$E(L_Q)$	0.318 ± 0.017	0.353 ± 0.013	0.511 ± 0.033	0.181 ± 0.0078
$Var(L_Q)$	0.869 ± 0.037	0.935 ± 0.022	1.232 ± 0.049	0.585 ± 0.1320
E(W)	0.077 ± 0.046	0.083 ± 0.0035	0.125 ± 0.0077	0.045 ± 0.004
Var(W)	0.18 ± 0.0072	0.187 ± 0.0035	0.273 ± 0.011	0.114 ± 0.004
E(U)	1.549 ± 0.0137	1.564 ± 0.015	1.614 ± 0.024	1.488 ± 0.0125
Var(U)	1.736 ± 0.0086	1.751 ± 0.010	1.783 ± 0.012	1.689 ± 0.009
E(I)	0.146 ± 0.0047	0.14 ± 0.0067	0.122 ± 0.008	0.172 ± 0.0045
E(B)	0.503 ± 0.0064	0.5 ± 0.0034	0.505 ± 0.009	0.501 ± 0.0063
P(A)	0.17 ± 0.008	0.158 ± 0.0045	0.139 ± 0.005	0.197 ± 0.0055
P(Balk)	0	0	0.0006 ± 0.0003	0

Table 5.6: **Comparison of steady-state performance measures** of different Erlang distributions (M/M/5/12+E(1, 0.5), M/M/5/12+E(1, 2) and Erlang-A) with gamma distribution (M/M/5/12 + gamma(2,2)) with 5 servers and traffic intensity of 0.926.

Performance Measure	M/M/5/12+M	M/M/5/12+GAMA(2,2)	M/M/5/12+E(1,2)	M/M/5/12+E(1, 0.5)
$E(L_Q)$	0.335 ± 0.0158	0.374 ± 0.023	0.517 ± 0.023	0.195 ± 0.005
$Var(L_Q)$	0.945 ± 0.0315	1.011 ± 0.031	1.326 ± 0.0503	0.636 ± 0.0136
E(W)	0.066 ± 0.0035	0.075 ± 0.0038	0.104 ± 0.006	0.04 ± 0
Var(W)	0.162 ± 0.0054	0.167 ± 0.0048	0.24 ± 0.0101	0.103 ± 0.03460
E(U)	1.964 ± 0.022	1.976 ± 0.026	2.013 ± 0.0135	1.88 ± 0.0169
Var(U)	2.155 ± 0.013	2.176 ± 0.016	2.203 ± 0.0083	2.095 ± 0.0097
E(I)	0.137 ± 0.0056	0.132 ± 0.007	0.122 ± 0.003	0.164 ± 0.005
E(B)	0.502 ± 0.054	0.5 ± 0.0067	0.501 ± 0.004	0.499 ± 0.004
P(A)	0.144 ± 0.0066	0.128 ± 0.006	0.113 ± 0.004	0.1646 ± 0.0049
P(Balk)	0	0	0.0003 ± 0.0003	0

Table 5.7: **Comparison of steady-state performance measures** of different Erlang distributions (M/M/5/12+E(1, 0.5), M/M/5/12+E(1, 2) and Erlang-A) with gamma distribution (M/M/5/12+gamma(2, 2)) with 6 servers and traffic intensity of 0.926.

Performance measure	M/M/6/15+M		M/M/6/15+GAMA(2,2)		M/M/6/15+E(1,2)		M/M/6/15+E(1, 0.5)	
$E(L_Q)$	0.366	± 0.02	0.439	± 0.017	0.583	± 0.028	0.213	± 0.0076
$Var(L_Q)$	1.018	± 0.043	1.141	± 0.0235	1.464	± 0.0496	0.689	± 0.016
E(W)	0.06	± 0.003	0.071	± 0.0023	0.095	± 0.005	0.033	± 0.00346
Var(W)	0.147	± 0.005	0.159	± 0.0023	0.221	± 0.007	0.094	± 0.0037
E(U)	2.404	± 0.014	2.445	± 0.0173	2.225	± 0.019	2.33	± 0.0151
Var(U)	2.607	± 0.01	2.635	± 0.0103	2.662	± 0.014	2.54	± 0.0111
E(I)	0.126	± 0.004	0.114	± 0.005	0.107	± 0.0048	0.146	± 0.0037
E(B)	0.502	± 0.0045	0.5	± 0.0048	0.497	± 0.0048	0.5	± 0.0053
P(A)	0.133	± 0.0053	0.12	± 0.0053	0.104	± 0.0052	0.157	± 0.005
P(Balk)		0		0		0		0

Table 5.8: 95% Confidence intervals for exponential, gamma and Erlang distributions of abandoning times

95% Confidence Interval

Using four servers with exponetial, gamma and Erlang abandonment

Perfprmanc Measure	M/M/4/10+M	M/M/10+GAMA(2,2)	M/M/4/10+E(1,2)	M/M/4/10+E(1,0.5)
$E(L_Q)$	(0.30; 0.335)	(0.340; 0.366)	0.478; 0.544)	(0.175; 0.189)
$Var(L_Q)$	(0.832; 0.906)	(0.913; 0.957)	(1.183; 1.281)	(0.473; 0.717)
E(W)	(0.031; 0.123)	(0.08; 0.087)	(0.117; 0.133)	(0.041; 0.049)
Var(W)	(0.173; 0.187)	(0.078; 0.087)	(0.262; 0.284)	(0.11; 0.118)
E(U)	(1.535; 1.563)	(1.549; 1.579)	(1.59; 1.638)	(1.476; 1.501)
Var(U)	(1.729; 1.745)	(1.75; 1.752)	(1.771; 1.795)	(1.68; 1.699)
E(I)	0.141; 0.1507	(0.133 0.147)	(0.114; 0.13)	(0.168; 0.177)
E(B)	(0.496; 0.509)	(0.497; 0.503)	(0.496; 0.514)	(0.495; 0.507)

Using five servers with exponential, gamma and Erlang abandonment

$E(L_Q)$	(0.319; 0.351)	(0.351; 0.397)	(0.494; 0.54)	(0.19; 0.2)
$Var(L_Q)$	0.914; 0.977)	(0.98; 1.042)	(1.276; 1.376)	(0.622; 0.65)
E(W)	(0.064; 0.07)	(0.071; 0.079)	(0.098; 0.11	
Var(W)	0.157; 0.167)	(0.162; 0.172)	(0.23; 0.25)	(0.1; 0.1065)
E(U)	(1.942; 1.986)	(1.95; 2.002)	(2.0; 2.027)	(1.863; 1.897)
Var(U)	(2.142; 2.168)	(2.76; 2.192)	(2.195; 2.211)	(2.085; 2.105)
E(I)	(0.131; 0.143)	(0.125; 0.139)	(0.019; 0.125)	(0.159; 0.169)
E(B)	0.448; 0.556)	(0.493; 0.507)	(0.497; 0.505)	(0.495; 0.503)

Using six servers with exponential, gamma and Erlang abandonment

$E(L_Q)$	(0.364; 0.368)	(0.422; 0.456)	(0.555; 0.611)	(0.206; 0.221)
$Var(L_Q)$	(0.935; 1.061)	(1.119; 1.168)	(1.414; 1.514)	(0.673; 0.675)
E(W)	(0.053; 0.063)	(0.069; 0.073)	(0.09; 0.1)	(0.03; 0.036)
Var(W)	(0.142; 0.152)	(0.157; 0.161)	(0.214; 0.228)	(0.09; 0.0944)
E(U)	(2.39; 2.418)	(2.428; 2.462)	(2.206; 2.244)	(2.315; 2.345)
Var(U)	(2.59; 2.608)	(2.625; 2.645)	(2.648; 2.676)	(2.529; 2.551)
E(I)	(0.124; 0.13)	(0.109; 0.119)	(0.102; 0.112)	(0.143; 0.15)
E(B)	(0.498; 0.507)	(0.495; 0.505)	(0.495; 0.502)	(0.495; 0.505)

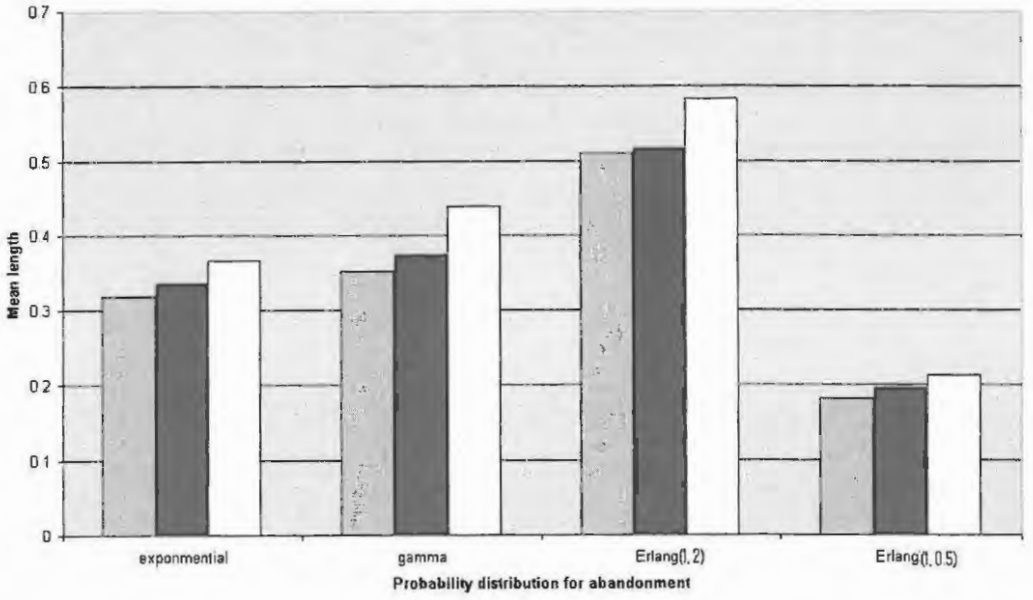

Figure 5.2: Bar graph of estimated mean *length of queue* for different distributions with 4, 5 and 6 servers respectively.

In testing equality of performance measures for exponential and Erlang distributions of abandoning time, we noted that none of the Erlang distributions have confidence intervals that overlap with that of the exponential distribution (as shown in table 5.8). Good performance measures should have smaller values for mean *length of queue*, mean *waiting time*, mean *proportion of idle periods* and *probability of abandoning* as well as smaller *standard errors*. Although the confidence intervals of exponential distribution do not overlap with the confidence intervals of E_2 and $E_{0.5}$, $E_{0.5}$ had the best perfomance measures. This is because it has the smallest values for mean *length of queue*, mean *waiting time*, mean *proportion of idle periods* and *probability of abandoning* as well as smaller *standard errors* when compared to the rest of distributions of abandonment.

Table 5.9: Coefficient of variation for estimated mean *length of queue*.

Distribution	4 Servers	5 Servers	6 Servers
Exponential	2.931	2.902	2.757
Gamma	2.739	2.688	2.433
Log-normal	1.551	1.433	1.388
Uniform	1.973	1.955	1.688
Erlang(1, 2)	2.172	2.227	2.075
Erlang(1, 0.5)	4.226	4.090	3.897

5.4.1.3 Coefficient of Variation

The coefficient of variation is a dimensionless measure of variation and we checked the performance measures' variability for the different distributions of abandonment. A comparison was made for all the distributions that have been discussed so far.

Scrutiny of these distributions of abandonment was done by considering variability of performance measures. Generally, from table 5.9, it was observed that the relative variability for *length of queue* was different for all the distributions and tended to decrease as the number of servers increased. This is because with more calls, there is need to have more servers and this reduces variability as calls are answered quickly. Those distributions that had poor performance measures (log-normal and uniform) had low coefficient of variation and the opposite was true for those distributions that have good performance measures. E_2 and $E_{0.5}$ have marked differences in variations with $E_{0.5}$ doubling the variability of E_2. The trend is, the better the distribution in performance measures the higher the coefficient of variation. Although this seems strange, this variability is expected since we are dealing with callers.

Table 5.10 shows that as the calls increased resulting in increase in the number of servers, the relative variability of amount of time in the system also increased. This is because generally with more calls one also expects the amount of *waiting time* to increase. The relative variability in this case was even greater than the relative variability of *length of queue*. This is because this is total delay, which includes waiting time in the

Table 5.10: Coefficient of variation for estimated mean *waiting times* in the system.

Distribution	4 Servers	5 Servers	6 Servers
Exponential	5.510	6.098	6.390
Gamma	5.210	5.449	5.616
Log-normal	3.005	3.090	3.275
Uniform	3.719	4.040	3.968
Erlang(1, 2)	4.180	4.711	4.948
Erlang(1, 0.5)	7.503	8.023	9.291

Table 5.11: Coefficient of variation for estimated mean *server utilisation*.

Distribution	4 Servers	5 Servers	6 Servers
Exponential	0.851	0.747	0.672
Gamma	0.846	0.747	0.664
Log-normal	0.805	0.710	0.640
Uniform	0.824	0.726	0.649
Erlang(1, 2)	0.827	0.737	0.733
Erlang(1, 0.5)	0.873	0.770	0.684

queue as well as service time. The same trend of the different distribution for abandonment was maintained, that is, distributions with good performance measures have poor relative variability.

The coefficient of variation for *server utilisation* (table 5.11) was the same for all the distributions considered for abandoning time. The relative variability decreased as the number of servers increased, despite the fact that the traffic intensity remained the same. Despite the type of distribution used for abandoning, the relative variability was not different for *server utilisation* unlike the variation in *length of queue* and *waiting time* in the system.

Erlang(1, 0.5) has the best performance measures although it has the highest relative variability for *length of queue* and amount of *waiting time*. Exponential and gamma distributions have similar performance measures as well as the coefficient of variation. Although the other distributions have a generally low coefficient of variation, their performance measures are not good enough. Variation on its own is not of concern in call

centres because generally length of calls vary a lot and so to specify a good distribution using solely this measure is not good enough.

5.4.2 Heavy Loads

Having used traffic intensity of 0.926 in the previous subsections, and having seen that the exponential, gamma and Erlang ($E_{0.5}$) distributions are more or less the same, we investigated if there was any difference when traffic intensity changes. Using traffic intensity of 1.1, which is relatively heavy loads, table 5.12 was analysed. The estimated mean *length of queue* was shorter for the Erlang distribution ($E_{0.5}$) of abandonment than the gamma and exponential distributions and had smaller standard error as well. This also applied to the variance of *length of queue*. The mean delay (system *waiting time*) was similar for all the distributions of abandonment as well as estimate of variance for *waiting times* for the three distributions. The estimates of mean and variance for *server utilisation* was almost the same across the different distributions and all the distributions had almost the same percentage standard errors. The exponential distribution had a high proportion of mean idle estimate due to the fact that its estimates for mean *length of queue* and *waiting time* are slightly smaller than the other distributions.

The mean proportion of busy periods estimate was exactly the same across all the distributions whether light or heavy loads. The Erlang distribution ($E_{0.5}$) had also a high *probability of abandoning* of 0.26. compared to 0.22 and 0.24 for gamma and exponential distributions respectively. Overall, although there are slight differences in these distributions, they gave us similar performance measures. The same pattern that was shown in lighter loads was also observed for heavy loads. Therefore changes in traffic intensity does not alter the perfomance of these distributions for abandonment. The performance measures have shown that Erlang distribution had the best performance measures compared to the rest of other distributions, although it had greater relative variability.

Table 5.12: Validity of abandoning distribution with traffic intensity of 1.1.

Traffic intensity = 1.1

Performance measure	M/M/4/10+M	M/M/4/10+GAMA(2, 2)	M/M/4/10+E(1, 0.5)
$E(L_Q)$	0.511 ± 0.0149	0.582 ± 0.0286	0.285 ± 0.008
$Var(L_Q)$	1.156 ± 0.0298	1.254 ± 0.041	0.76 ± 0.013
E(W)	0.104 ± 0.0037	0.114 ± 0.006	0.059 ± 0.0023
Var(W)	0.209 ± 0.0053	0.215 ± 0.007	0.13 ± 0.0034
E(U)	1.693 ± 0.0068	1.713 ± 0.0122	1.63 ± 0.0123
Var(U)	1.824 ± 0.005	1.836 ± 0.0077	1.773 ± 0.0048
E(I)	0.09 ± 0	0.084 ± 0.0037	0.116 ± 0.0037
E(B)	0.5 ± 0.067	0.5 ± 0.00477	0.499 ± 0.0063
P(A)	0.235 ± 0.0057	0.224 ± 0.0089	0.261 ± 0.0067
P(Balk)	0	0.00012 ± 0.0001	0

Comparing lighter loads and heavy loads, it was noticed that the quality of approximation using lighter loads were better than in heavy loads because in heavy loads the arrival rate is bigger than service rate so, to maintain stability, there will be a lot of callers that will abandon (leading to high *probability of abandoning*). Because of more calls arriving, there will be longer *lengths of queue* and *waiting times* in system. That should be expected because in lighter loads where the arrival rate was smaller than the service rate, shorter *length of queue* and *waiting times* in system as well as few abandonment were observed.

Normally, the time that the facility is not in use is the time during which money is spent but no revenue is collected. It is therefore important to design systems that will maintain high *server utilization*. Most of the distribution that were considered for abandonment had a mean of 1. This was done so as to compare them using the same mean. The results showed that the distributions for abandonment must not necessarily be exponential (as some research claim, Brown *et al* (2002) [3] and Madelbaum (2002) [7]). Other distributions (Erlang and gamma) are as similar or sometimes better for example Erlang$(1, 0.5)$. This shows that the distribution for abandonment is a general one and not necessarily exponential. With slight modification of parameters most of the distributions may give better performance measurements than the exponential distribution (compare E_2 and $E_{0.5}$).

6 CONCLUSIONS

6.1 Limitations

Every software irrespective of how good it is has its own limitations and EZSIM is no exception. The software is difficult to implement with a lot of servers as it was not designed to cater for a large number of servers and has limited simulation time, which makes it difficult to loop and had to resort to the use of different seeds and then find averages. Financial constrains made it difficult to use licensed software, but the results obtained were tested for their validity by changing loads. In this work we used 4 to 6 servers and the obtained results were consistent for light and heavy loads.

6.2 Areas of Further Research

This model did not consider retrials, that is when a customer re-dials into the centre after having encountered a busy signal or having abandoned. It has been observed that in most call centres, the majority of retrials is due to customer abandonment, because the bottleneck resource are the agents, not the number of telephone lines. There is need for some research that focuses on retrials since the retrial volume can be of the order of first-time calls. There is need for testing these results using a large number of servers as well to incorporate multi-skills in the call centre. Also with advent of contact centres, call centres that includes e-mail and internet, more research is needed in understanding their operations. However, results obtained in this research can still be extended to contact centres with some modification.

6.3 Conclusion

Modern call centres operate under many uncertainties and complexities, notably, uncertain and/or time-varying primitives and complex daily control and routing control actions. These realities stretch the limits of existing analytical models from queueing theory, optimal queueing control and stochastic programming. The high operational complexity and the prevalent uncertainty suggest that simulation modelling and simulation-based decision-making should have a central role in the management of call centres. Simulation appears to be the most viable option for accurate performance measures and subsequent decision support, hence the use of EZSIM in this research.

From the various distributions that were considered for abandoning times, not much difference could be inferred from them, with the exception of uniform and log-normal. The Erlang(1, 0.5) distribution seems to be better f than exponential and gamma distributions by analysing the performance measures. In general it can be concluded that depending on the nature of data, the abandoning time is general and independently distributed, different distributions may be used for different sets of call centre data. Service times have been observed to follow a general distribution, though the exponential distribution was used in this research. The $M/GI/k/r + GI$ model is the best model for call centres although it is generally difficult to analyse and in this research EZSIM software made it possible to analyse.

Appendix

The simulation was done by EZSIM which is a free software that can be downloaded at http://www-rcf.usc.edu/~khoshnev/software.html. Excel was used to create the tables for the different distributions for the abandoning times. The tables on the next pages show the performance measures for the different distributions used for the abandoning time. All statistics obtained from the software are measured in minutes and traffic intensity of 0.926 and 1.1 were used. The distributions that were considered were exponential, uniform, log-normal, gamma and Erlang.

The model used in this thesis had six nodes, that is the Source, Queue, Facility and three Terminates nodes. The three Terminate nodes were labelled as Lost, Abandon and Leave. Leave was connected from the Facility node that had parallel servers. Those who were balked were sent to the Lost node and those who were abandoned were sent to the Abandon node. The model used is shown below and an example of results obtained from the model with an exponential abandonment distribution are shown overleaf.

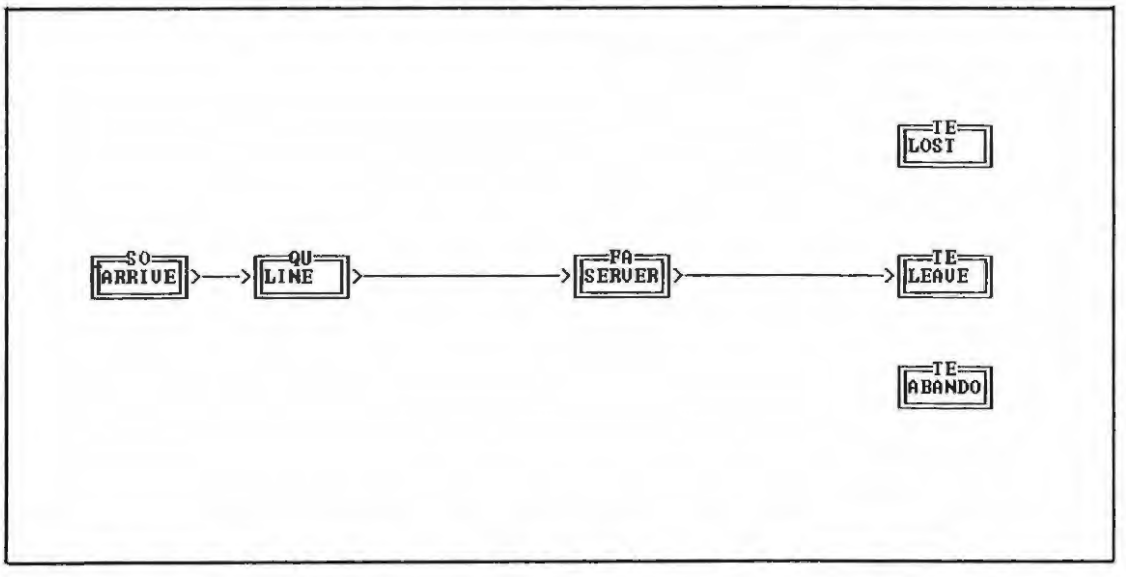

Table 6.1: Model specification for Erlang-*A*.

```
Disk file name: G:\TM-M-M.EZ
Project name:  ERLANG-A DISTRIBUTIO
Date:  01/08/09
Analyst:  JACOB

Node name: ARRIVE
Node type: SO
Entity name: CUST
First creation time: 0
Time between creations: EXPON(0.2083,11)
Maximum no. of creations<INF>:
Time to stop creation<INF>: 2200

Node name: LINE
Node type: QU
Queue capacity <INF>: 10
For full queue situation:  BALK
When queue is full, entities balk to: LOST
Balking traversal time: 0
Initial number in the queue <0>: 0
Queue discipline: FIFO

Node name: LEAVE
Node type: TE
Termination count <INF>:

Node name: LOST
Node type: TE
Termination count <INF>:

Node name: ABANDO
Node type: TE
Termination count <INF>:

Node name: SERVER
Node type: FA
Number of parallel servers <1>: 4
Service duration: EXPON(1,11)
Schedule breakdowns <N>: N
Does the facility use other resources <N>: N

Information on desired statistics
Name: LOST
Statistics type: COUNT
Variable type: OBS.BASED
Collect at node: LOST
Name: ABANDON
Statistics type: COUNT
Variable type: OBS.BASED
Collect at node: ABANDO
Name: SERVED
Statistics type: COUNT
Variable type: OBS.BASED
Collect at node: LEAVE
```

Table 6.2: Output from model using Erlang-*A*.

```
*** E Z S I M STATISTICAL REPORT ***

Simulation Project: ERLANG-A DISTRIBUTIO
Analyst: JACOB
Date: 01/08/09
Disk file name: G:\TM-M-M.OUT

Current Time: 2200.78    Transient Period: 1100.00

Q U E U E S:
NAME      MIN/MAX/LAST      MEAN      STD       MEAN      STD
          LENGTH            LENGTH    LENGTH    DELAY     DELAY
-------------------------------------------------------------------
LINE       0/10/ 0          0.71      1.41      0.13      0.24


F A C I L I T I E S:
NAME      NBR     MIN/MAX/LAST   MEAN      STD       MEAN      MEAN
          SRVRS   UTILIZATION    UTLZ      UTLZ      IDLE      BUSY
-------------------------------------------------------------------
SERVER    4        0/ 4/ 0       1.76      1.87      0.07      0.51


V A R I A B L E S:
NAME      MEAN        STD         MIN         MAX         No.OBSRVD
-------------------------------------------------------------------
LOST      4.00E+00    0.00E+00    4.00E+00    4.00E+00    4
ABANDON   1.50E+03    0.00E+00    1.50E+03    1.50E+03    1502
SERVED    3.80E+03    0.00E+00    3.80E+03    3.80E+03    3797
```

Exponential abandoning times with 0.926 traffic intensity.

TRAFFIC INTENSITY = 0.926

M/M/4/10+M interarrival time=0.27, mean service time=1min, mean time abandon= 1min

seed	mean leng	std length	mean dela	std delay	mean utlz	std utlz	mean idle	mean busy	balk	abandon	served	total	P(A)
1	0.31	0.87	0.08	0.18	1.54	1.73	0.15	0.5	0	681	3477	4158	0.163781
2	0.35	0.93	0.08	0.2	1.57	1.75	0.14	0.51	0	692	3455	4147	0.166868
3	0.36	0.98	0.09	0.19	1.57	1.74	0.14	0.51	0	796	3419	4215	0.188849
4	0.32	0.85	0.08	0.18	1.57	1.75	0.14	0.51	0	748	3349	4097	0.182573
5	0.3	0.82	0.07	0.17	1.53	1.73	0.15	0.51	0	674	3365	4039	0.166873
7	0.34	0.9	0.08	0.19	1.56	1.75	0.14	0.51	0	781	3423	4204	0.185775
8	0.28	0.81	0.07	0.17	1.51	1.71	0.16	0.48	0	631	3416	4047	0.155918
9	0.31	0.83	0.07	0.17	1.56	1.74	0.14	0.5	0	693	3501	4194	0.165236
10	0.29	0.83	0.07	0.17	1.54	1.73	0.15	0.5	0	645	3442	4087	0.157817
11	0.32	0.87	0.08	0.18	1.54	1.73	0.15	0.5	0	723	3402	4125	0.175273
Mean	0.318	0.869	0.077	0.18	1.549	1.736	0.146	0.503	0	706.4	3424.9	4131.3	0.170896
sd	0.024413	0.051274	0.006403	0.01	0.019209	0.012	0.006633	0.009	0	52.09261	44.37894	59.98842	0.010999
error	0.017463	0.036677	0.00458	0.007153	0.013741	0.008584	0.004745	0.006438	0	37.26222	31.74457	42.91015	0.007867
95% C.I	0.32±0.017	0.869±0.03	0.077±0.0	0.18 ±0.00	1.549 ±0.0	1.736 ±0.0	0.146±0.0	0.503 ±.0064					0.17 ±0.00

M/M/5/12+M

1	0.33	0.86	0.06	0.15	1.98	2.18	0.13	0.5	0	735	4304	5039	0.145862
2	0.35	0.98	0.07	0.17	1.96	2.16	0.14	0.5	0	723	4364	5087	0.142127
3	0.35	1	0.07	0.17	1.96	2.16	0.14	0.5	2	775	4279	5056	0.153283
4	0.33	0.95	0.06	0.16	1.95	2.16	0.14	0.5	0	715	4289	5004	0.142886
5	0.37	0.99	0.07	0.17	2.03	2.12	0.12	0.52	0	813	4269	5082	0.159976
6	0.33	0.93	0.07	0.16	1.94	2.16	0.14	0.5	0	723	4248	4971	0.145444
7	0.34	0.98	0.07	0.17	1.97	2.16	0.14	0.5	0	738	4424	5162	0.142968
8	0.34	0.95	0.07	0.16	1.99	2.17	0.13	0.51	0	736	4324	5060	0.145455
9	0.33	0.93	0.06	0.16	1.95	2.16	0.14	0.49	0	716	4420	5136	0.139408
10	0.28	0.86	0.06	0.15	1.91	2.12	0.15	0.5	0	597	4273	4870	0.122587
Mean	0.335	0.945	0.066	0.162	1.964	2.155	0.137	0.502	0.2	727.1	4319.4	5046.7	0.144
sd	0.022023	0.044102	0.004899	0.007483	0.030397	0.018574	0.00781	0.007483	0.6	52.11804	59.71298	79.52239	0.009135
error	0.015753	0.031547	0.003504	0.005353	0.021743	0.013286	0.005587	0.005353	0.43	37.28041	42.71313	56.88294	0.006534
95% C.I	.335±0.015	0.945±0.03	0.066±0.00	0.162±0.00	1.964±0.02	2.155±0.010	0.137±0.00	0.502±0.054					0.144±0.00

M/M/6/15+M

1	0.32	0.93	0.05	0.14	2.37	2.58	0.13	0.49	0	729	5205	5934	0.122851
2	0.35	0.96	0.06	0.14	2.42	2.62	0.12	0.51	0	778	5370	6148	0.126545
3	0.37	1.04	0.06	0.15	2.39	2.6	0.13	0.5	0	791	5259	6050	0.130744
4	0.36	1.01	0.06	0.14	2.39	2.6	0.13	0.5	0	836	5263	6099	0.137072
5	0.38	1.06	0.06	0.15	2.42	2.62	0.12	0.51	0	853	5257	6110	0.139607
6	0.36	1.02	0.06	0.15	2.4	2.61	0.13	0.5	0	771	5302	6073	0.126955
7	0.37	1.03	0.06	0.15	2.4	2.61	0.13	0.51	0	831	5282	6113	0.13594
9	0.43	1.15	0.07	0.16	2.43	2.62	0.12	0.5	0	918	5314	6232	0.147304
10	0.37	0.99	0.06	0.15	2.43	2.62	0.12	0.5	0	783	5374	6157	0.127172
11	0.35	0.99	0.06	0.14	2.39	2.59	0.13	0.5	0	794	5265	6059	0.131045
mean	0.366	1.018	0.06	0.147	2.404	2.607	0.126	0.502	0	808.4	5289.1	6097.5	0.132524
sd	0.027968	0.060148	0.004714	0.006749	0.020111	0.014181	0.005164	0.006325	0	52.76404	52.5578	78.63453	0.007418
error	0.020006	0.043024	0.003372	0.004828	0.014385	0.010144	0.003694	0.004524	0	37.7425	37.59497	56.24784	0.005306
95% C. I	0.366±0.02	1.018±0.04	0.06±0.003	0.147±0.00	2.404±0.012	2.607±0.01	0.126±0.004	0.502±0.00	0				0.133±0.00

Table 6.3: Uniform abandoning times with 0.926 traffic intensity.

Traffic Intesity 0.926

M/M/4/10+U(0.5;1.5)				interarrival time=0.27,	service time= 1								
seed	mean leng	std length	mean dela	std delay	mean utlz	std utlz	mean idle	mean busy	balk	abandon	served	total	P(A)
1	0.57	1.29	0.14	0.27	1.64	1.8	0.11	0.5	2	472	3589	4063	0.11617
2	0.66	1.42	0.16	0.29	1.66	1.82	0.11	0.52	2	581	3565	4148	0.140068
3	0.61	1.35	0.15	0.28	1.68	1.82	0.1	0.52	2	514	3629	4145	0.124005
4	0.57	1.32	0.14	0.27	1.68	1.78	0.13	0.49	0	468	3558	4026	0.116244
5	0.58	1.34	0.14	0.28	1.63	1.79	0.12	0.5	4	479	3602	4085	0.117258
6	0.54	1.26	0.13	0.27	1.57	1.77	0.13	0.49	3	455	3539	3997	0.113836
7	0.61	1.38	0.14	0.28	1.64	1.8	0.11	0.51	3	554	3583	4140	0.133816
8	0.61	1.38	0.14	0.28	1.64	1.8	0.11	0.51	3	554	3583	4140	0.133816
9	0.57	1.29	0.14	0.27	1.63	1.8	0.11	0.49	0	457	3703	4160	0.109856
10	0.52	1.24	0.13	0.26	1.59	1.77	0.13	0.48	0	436	3627	4063	0.10731
mean	0.584	1.327	0.141	0.275	1.626	1.795	0.116	0.501	1.9	497	3597.8	4096.7	0.121238
sd	0.040056	0.057552	0.008756	0.006498	0.03534	0.017795	0.01075	0.013703	1.45	50.24164	46.59948	57.80052	0.011173
error	0.028652	0.041167	0.006263	0.006079	0.025279	0.012729	0.007689	0.009802	1.04	35.93821	33.33294	41.34513	0.007992
95% C I	0.584±0.02	1.327±0.04	0.141±0.00	0.275±0.00	1.626±0.02	1.795±0.010	0.116±0.00	0.501±0.0098					0.121±0.00

M/M/5/12+U(0.5;1.5)				interarrival time=0.22, service time =1								
1	0.54	1.32	0.11	0.23	2.04	2.22	0.11	0.49	0	384	4599	4983	0.077062
2	0.66	1.49	0.13	0.26	2.09	2.25	0.1	0.51	0	502	4542	5044	0.099524
3	0.63	1.49	0.13	0.26	2.04	2.23	0.11	0.51	1	468	4529	4998	0.093637
4	0.59	1.42	0.12	0.25	2.02	2.21	0.12	0.49	0	431	4550	4981	0.086529
5	0.64	1.46	0.13	0.26	2.06	2.24	0.11	0.51	1	454	4493	4948	0.091754
6	0.62	1.46	0.12	0.25	2.03	2.22	0.12	0.49	0	463	4476	4939	0.093744
7	0.6	1.4	0.12	0.24	2.08	2.25	0.1	0.5	1	396	4725	5122	0.077314
8	0.67	1.53	0.13	0.26	2.09	2.25	0.1	0.5	0	540	4555	5095	0.105986
9	0.63	1.5	0.13	0.26	2.1	2.26	0.1	0.51	0	464	4645	5109	0.09082
10	0.56	1.34	0.12	0.24	2.04	2.22	0.11	0.5	0	356	4624	4980	0.071486
mean	0.614	1.441	0.124	0.251	2.059	2.235	0.108	0.501	0.3	445.8	4573.8	5019.9	0.088786
sd	0.041687	0.069833	0.006992	0.011005	0.028848	0.017159	0.007888	0.008756	0.48	55.57937	75.36401	67.69613	0.010779
error	0.029819	0.049952	0.005001	0.007872	0.020635	0.012274	0.005642	0.006263	0.35	39.75633	53.90842	48.42353	0.00771
955 C. I	0.614±0.03	1.441±0.05	0.124±0.00	0.251±0.00	2.059±0.02	2.235±0.010	0.108±0.00	0.501±0.0063					0.089±0.00

M/M/6/15+U(0.5;1.5)				interarrival time= 0.18,	service time = 1								
1	0.73	1.68	0.12	0.24	2.49	2.68	0.1	0.48	0	497	5538	6035	0.082353
2	0.69	1.61	0.12	0.24	2.52	2.69	0.1	0.5	0	444	5582	6026	0.073681
3	0.87	1.8	0.14	0.26	2.58	2.74	0.08	0.51	0	608	5651	6259	0.09714
4	0.71	1.66	0.12	0.24	2.5	2.68	0.1	0.49	0	461	5588	6049	0.076211
5	0.84	1.8	0.14	0.26	2.59	2.75	0.08	0.51	0	596	5591	6187	0.096331
6	0.79	1.74	0.13	0.25	2.56	2.73	0.09	0.5	0	542	5578	6120	0.088562
7	0.74	1.65	0.12	0.24	2.52	2.7	0.09	0.49	0	485	5717	6202	0.078201
8	0.75	1.67	0.12	0.24	2.54	2.71	0.09	0.5	0	492	5645	6137	0.080169
9	0.86	1.68	0.12	0.25	2.53	2.7	0.09	0.49	0	523	5615	6138	0.085207
10	0.72	1.62	0.12	0.24	2.52	2.7	0.1	0.5	0	457	5618	6075	0.075226
mean	0.77	1.691	0.125	0.246	2.535	2.708	0.092	0.497	0	510.5	5612.3	6122.8	0.083308
sd	0.065659	0.067569	0.008498	0.008433	0.032745	0.024404	0.007888	0.009487	0	56.64166	49.69697	77.59983	0.008415
error	0.046966	0.048333	0.006079	0.006032	0.023423	0.017456	0.005642	0.006786	0	40.51619	35.5486	55.50772	0.00602
95% C. I	0.77±0.047	1.69±0.048	0.125±0.00	0.246±0.00	2.535±0.02	2.708±0.010	0.092±0.00	0.497±0.0068					0.083±0.00

Table 6.4: Log-normal abandoning times with 0.926 traffic intensity.

Traffic inetsity 0.926

M/M/4/10+LN(1,1)

seed	mean leng	std length	mean dela	std delay	mean utlz	std utlz	mean idle	mean busy	balk	abandon	served	total	P(A)	P(Balk)
1	0.95	2	0.23	0.45	1.67	1.83	0.1	0.5	35	366	3692	4093	0.08942	0.0086
2	0.95	1.93	0.24	0.45	1.69	1.83	0.09	0.5	18	342	3691	4051	0.08442	0.0044
3	0.84	1.81	0.21	0.41	1.69	1.84	0.09	0.5	24	316	3737	4077	0.07751	0.0059
4	0.79	1.78	0.2	0.43	1.62	1.8	0.11	0.49	29	286	3643	3958	0.07226	0.0073
5	0.85	1.85	0.21	0.44	1.68	1.82	0.1	0.51	19	319	3635	3973	0.08029	0.0048
6	0.77	1.74	0.19	0.41	1.64	1.8	0.11	0.49	15	285	3667	3967	0.07184	0.0038
7	0.83	1.79	0.21	0.41	1.69	1.83	0.09	0.5	27	310	3775	4112	0.07539	0.0066
8	0.83	1.79	0.21	0.41	1.69	1.83	0.09	0.5	27	310	3775	4112	0.07539	0.0066
9	0.99	1.98	0.24	0.46	1.71	1.85	0.09	0.51	27	380	3746	4153	0.0915	0.0065
10	1.02	2.04	0.25	0.46	1.71	1.85	0.09	0.51	41	400	3712	4153	0.09632	0.0099
mean	0.862	1.871	0.219	0.433	1.679	1.828	0.096	0.501	26.2	331.4	3707.3	4065	0.08143	0.0064
sd	0.067661	0.107129	0.019692	0.021628	0.028848	0.017512	0.008433	0.007379	7.86	39.2972	50.612	74.94	0.00857	0.0019
error	0.062705	0.07663	0.014086	0.015471	0.020635	0.012526	0.006032	0.005278	5.62	28.1095	36.203	53.61	0.00613	0.0013
95% C.	0.862±0.0	1.871±0.07	0.219±0.010	0.433±0.011	1.679±0.0	1.828±0.010	0.096±0.00	0.501±0.005					0.082±0	0.006±0

M/M/5/12+LN(1,1)

1	0.95	2.11	0.19	0.4	2.09	2.26	0.1	0.49	27	325	4614	4966	0.06545	0.0054
2	1.11	2.25	0.22	0.42	2.16	2.3	0.08	0.51	29	431	4696	5156	0.08359	0.0056
3	0.99	2.07	0.2	0.39	2.11	2.28	0.09	0.5	7	338	4732	5077	0.06657	0.0014
4	1.06	2.21	0.21	0.41	2.18	2.31	0.08	0.51	18	369	4794	5181	0.07122	0.0035
5	1.09	2.28	0.22	0.43	2.14	2.29	0.09	0.51	33	403	4594	5030	0.08012	0.0066
6	1.04	2.26	0.21	0.43	2.08	2.26	0.1	0.49	16	355	4673	5044	0.07038	0.0032
7	1.03	2.16	0.21	0.41	2.16	2.3	0.08	0.51	18	363	4638	5019	0.07233	0.0036
8	1.03	2.16	0.21	0.41	2.16	2.3	0.08	0.52	20	355	4634	5009	0.07087	0.004
9	1.05	2.19	0.21	0.42	2.15	2.3	0.08	0.51	18	363	4677	5058	0.07177	0.0036
10	0.97	2.18	0.2	0.41	2.06	2.24	0.11	0.5	19	372	4533	4924	0.07555	0.0039
mean	1.032	2.187	0.208	0.413	2.129	2.284	0.089	0.505	20.5	367.4	4658.5	5046	0.07278	0.0041
sd	0.050509	0.066341	0.009189	0.012517	0.040947	0.02319	0.011005	0.009718	7.41	30.4492	73.509	78.22	0.00562	0.0015
error	0.036129	0.047454	0.006573	0.008953	0.02929	0.016588	0.007872	0.006952	5.3	21.7806	52.582	55.95	0.00402	0.0011
95% C.	1.032±0.0	2.187±0.04	0.208±0.00	0.413±0.00	2.129±0.0	2.284±0.010	0.089±0.00	0.505±0.007					0.073±0.0	0.0041±

M/M/6/15/LN(1,1)

1	1.3	2.54	0.21	0.39	2.66	2.79	0.07	0.5	5	460	5760	6225	0.0739	0.0008
2	1.08	2.31	0.18	0.37	2.59	2.74	0.08	0.5	10	353	5716	6079	0.05807	0.0016
3	1.03	2.24	0.17	0.35	2.56	2.73	0.09	0.49	7	356	5672	6035	0.05899	0.0012
4	1.12	2.42	0.19	0.38	2.59	2.75	0.08	0.5	8	411	5631	6050	0.06793	0.0013
5	1.12	2.36	0.19	0.38	2.61	2.76	0.08	0.51	8	347	5767	6122	0.05668	0.0013
6	1.06	2.23	0.18	0.35	2.6	2.75	0.08	0.5	2	353	5809	6164	0.05727	0.0003
7	1.17	2.45	0.2	0.38	2.59	2.75	0.08	0.5	8	386	5777	6171	0.06255	0.0013
8	1.17	2.45	0.2	0.38	2.59	2.75	0.08	0.5	8	386	5777	6171	0.06255	0.0013
9	1.04	2.29	0.17	0.36	2.56	2.73	0.08	0.49	3	340	5807	6150	0.05528	0.0005
10	0.93	2.1	0.16	0.33	2.55	2.72	0.09	0.49	3	301	5735	6039	0.04984	0.0005
mean	1.102	2.339	0.185	0.367	2.59	2.747	0.081	0.498	6.2	369.3	5745.1	6121	0.06031	0.001
sd	0.099976	0.130848	0.015811	0.018886	0.031269	0.019465	0.005676	0.006325	2.74	43.7824	57.699	66.19	0.0068	0.0005
error	0.071515	0.093596	0.01131	0.013509	0.022367	0.013924	0.00406	0.004524	1.96	31.3179	41.273	47.35	0.00487	0.0003
96% C.	1.102±0.07	2.339±0.09	0.185±0.010	0.367±0.012	2.59±0.022	2.747±0.010	0.081±0.00	0.498±0.0045					0.0068±0	0.001±0

Table 6.5: Gamma abandoning times with 0.926 traffic intensity.

Traffic Intensity 0.926

M/M/4/10+GAM(2,2)

seed	mean leng	std length	mean dela	std delay	mean utlz	std utlz	mean idle	mean busy	balk	abandor	served	total	P(A)
1	0.39	0.97	0.09	0.19	1.6	1.77	0.13	0.51	0	717	3443	4160	0.172356
2	0.37	0.98	0.09	0.19	1.56	1.74	0.14	0.5	0	669	3410	4079	0.164011
3	0.33	0.88	0.08	0.18	1.57	1.75	0.14	0.5	0	626	3485	4111	0.152274
4	0.36	0.95	0.09	0.19	1.56	1.75	0.14	0.5	0	638	3469	4107	0.155345
5	0.34	0.92	0.08	0.19	1.55	1.773	0.15	0.5	0	624	3400	4024	0.15507
6	0.34	0.91	0.08	0.18	1.56	1.74	0.14	0.5	0	630	3440	4070	0.154791
7	0.36	0.95	0.08	0.19	1.58	1.76	0.13	0.5	0	641	3507	4148	0.154532
8	0.36	0.95	0.08	0.19	1.58	1.76	0.13	0.5	0	641	3507	4148	0.154532
9	0.34	0.92	0.08	0.19	1.56	1.74	0.14	0.5	0	618	3438	4056	0.152367
10	0.34	0.92	0.08	0.18	1.52	1.73	0.16	0.49	0	652	3438	4090	0.159413
mean	0.353	0.935	0.083	0.187	1.564	1.7513	0.14	0.5	0	645.6	3454	4099	0.157469
sd	0.018288	0.030277	0.00483	0.00483	0.021187	0.01419	0.009428	0.004714	0	29.14	37.24	44.13	0.006288
95%	0.013081	0.021657	0.003455	0.003455	0.015155	0.01015	0.006744	0.003372	0	20.844	26.64	31.57	0.004498
95% C.I	0.353±0.010	0.935±0.02	0.083±0.00	0.187±0.00	1.564±0.011	1.751±0.010	0.14±0.006	0.5±0.0034					0.158±0.00

M/M/5/12+GAMA(2,2)

1	0.39	1.01	0.08	0.17	2	2.19	0.13	0.5	0	637	4373	5010	0.127146
2	0.38	1.03	0.07	0.17	1.98	2.17	0.13	0.5	0	634	4386	5020	0.126295
3	0.4	1.04	0.08	0.17	1.99	2.19	0.13	0.51	0	700	4370	5070	0.138067
4	0.38	1	0.08	0.17	2	2.19	0.13	0.5	0	638	4442	5080	0.125591
5	0.33	0.96	0.07	0.16	1.93	2.15	0.15	0.49	0	570	4346	4916	0.115948
6	0.42	1.09	0.08	0.18	2.01	2.2	0.12	0.51	0	729	4419	5148	0.141608
7	0.35	0.98	0.07	0.16	1.97	2.17	0.13	0.5	0	642	4399	5041	0.127356
8	0.35	0.98	0.07	0.16	1.97	2.17	0.13	0.5	0	642	4399	5041	0.127356
9	0.41	1.06	0.08	0.17	2.01	2.2	0.12	0.51	0	693	4366	5059	0.136984
10	0.33	0.96	0.07	0.16	1.9	2.13	0.15	0.48	0	572	4313	4885	0.117093
mean	0.374	1.011	0.075	0.167	1.976	2.176	0.132	0.5	0	645.7	4381	5027	0.128344
sd	0.032387	0.043576	0.00527	0.006749	0.035963	0.022706	0.010328	0.009428	0	51.193	36.66	77.13	0.008436
error	0.023166	0.03117	0.00377	0.004828	0.025725	0.016242	0.007388	0.006744	0	36.618	26.22	55.17	0.006035
95% C	0.374±0.02	1.011±0.03	0.075±0.00	0.167±0.00	1.976±0.02	2.176±0.010	0.132±0.00	0.5±0.0067					0.128±0.00

M/M/6/15+GAMA(2,2)

1	0.42	1.1	0.07	0.16	2.46	2.64	0.11	0.5	0	706	5421	6127	0.115228
2	0.42	1.13	0.07	0.16	2.43	2.62	0.12	0.49	0	663	5458	6121	0.108316
3	0.45	1.14	0.07	0.16	2.47	2.65	0.11	0.51	0	782	5368	6150	0.127154
4	0.46	1.17	0.07	0.16	2.45	2.64	0.11	0.5	0	761	5464	6225	0.122249
5	0.43	1.13	0.07	0.16	2.41	2.62	0.12	0.5	0	685	5349	6034	0.113523
6	0.47	1.17	0.08	0.16	2.46	2.65	0.11	0.51	0	791	5357	6148	0.12866
7	0.42	1.12	0.07	0.16	2.45	2.63	0.11	0.5	0	693	5372	6065	0.114262
8	0.46	1.18	0.07	0.16	2.47	2.65	0.11	0.5	0	813	5444	6257	0.129934
9	0.46	1.18	0.07	0.16	2.45	2.64	0.11	0.5	0	751	5343	6094	0.123236
10	0.4	1.09	0.07	0.15	2.4	2.61	0.13	0.49	0	694	5317	6011	0.115455
mean	0.439	1.141	0.071	0.159	2.445	2.635	0.114	0.5	0	733.9	5389	6123	0.119802
sd	0.023781	0.032813	0.003162	0.003162	0.024152	0.014337	0.006992	0.006667	0	51.982	52.8	77.75	0.007415
error	0.017011	0.023471	0.002262	0.002262	0.017276	0.010256	0.005001	0.004769	0	37.183	37.77	55.61	0.005304
955 C. I	0.439±0.011	1.141±0.02	0.071±0.00	0.159±0.00	2.445±0.012	2.635±0.010	0.114±0.00	0.5±0.0048					0.12±0.005

Table 6.6: Gamma abandoning times with 0.926 traffic intensity.

Traffic intensity = 0.926

M/M4/10+GAMA(0.5,2)														
seed	mean leng	std length	mean dela	std delay	mean utlz	std utlz	mean idle	mean busy	balk	abandor	served	total	P(A)	P(Balk)
1	1.04	2.05	0.26	0.48	1.71	1.85	0.09	0.5	37	339	3781	4157	0.081549	0.008901
2	1.05	2.1	0.26	0.5	1.7	1.84	0.09	0.5	36	320	3712	4068	0.078663	0.00885
3	0.92	1.87	0.24	0.45	1.69	1.84	0.09	0.5	29	293	3720	4042	0.072489	0.007175
4	0.92	1.89	0.23	0.46	1.71	1.84	0.09	0.5	22	283	3791	4096	0.069092	0.005371
5	0.91	1.97	0.24	0.47	1.63	1.81	0.11	0.49	41	268	3663	3992	0.072144	0.010271
6	1.03	2.09	0.25	0.49	1.72	1.85	0.08	0.51	40	338	3753	4131	0.08182	0.009683
7	1.02	2.04	0.26	0.48	1.72	1.86	0.08	0.5	30	306	3823	4159	0.073575	0.007213
8	1.02	2.04	0.26	0.48	1.72	1.86	0.08	0.5	30	306	3823	4159	0.073575	0.007213
9	0.98	2.04	0.24	0.48	1.68	1.82	0.09	0.49	43	328	3719	4090	0.080196	0.010513
10	0.93	2.02	0.24	0.49	1.63	1.81	0.11	0.49	42	268	3699	4009	0.06685	0.010476
mean	0.982	2.011	0.248	0.478	1.691	1.838	0.091	0.498	35	306.9	3748	4090	0.074995	0.008567
sd	0.056529	0.07781	0.011353	0.014757	0.034785	0.018738	0.011005	0.006325	6.9	24.154	54.53	62.08	0.005267	0.001751
error	0.040436	0.055658	0.008121	0.010556	0.024882	0.013403	0.007872	0.004524	5	17.278	39	44.4	0.003768	0.001252
95% C.	0.982±0.04	2.011±0.05	0.248±0.00	0.478±0.011	1.691±0.02	1.838±0.010	0.091±0.00	0.498±0.0045					0.075±0.00	0.0086±0.0

M/M/5/12+GAMA(0.5,2)														
1	0.89	1.99	0.19	0.38	2.11	2.28	0.09	0.49	13	261	4701	4975	0.052462	0.002613
2	0.96	2.04	0.2	0.4	2.12	2.28	0.09	0.5	12	249	4758	5019	0.049611	0.002391
3	1.01	2.22	0.21	0.43	2.1	2.27	0.09	0.5	32	308	4670	5010	0.061477	0.006387
4	1.1	2.27	0.22	0.43	2.13	2.29	0.09	0.5	21	332	4668	5021	0.066122	0.004182
5	1.08	2.31	0.22	0.44	2.12	2.28	0.09	0.5	30	315	4692	5037	0.062537	0.005956
6	0.99	2.11	0.21	0.41	2.13	2.29	0.09	0.5	21	246	4755	5022	0.048984	0.004182
7	1.07	2.24	0.21	0.43	2.13	2.28	0.09	0.49	12	323	4709	5044	0.064036	0.002379
8	1.03	2.15	0.21	0.41	2.15	2.3	0.08	0.5	13	281	4761	5055	0.055589	0.002572
9	0.97	2.15	0.2	0.42	2.1	2.27	0.09	0.49	21	283	4681	4985	0.05677	0.004213
10	1.03	2.15	0.21	0.41	2.12	2.28	0.09	0.5	10	278	4689	4977	0.056857	0.002009
mean	1.013	2.163	0.208	0.416	2.121	2.282	0.089	0.497	19	287.6	4708	5015	0.057345	0.003688
sd	0.06343	0.100338	0.009189	0.017764	0.015239	0.009189	0.003162	0.00483	7.8	30.689	36.45	27.87	0.006009	0.001557
error	0.045372	0.071773	0.006573	0.012707	0.0109	0.006573	0.002262	0.003455	5.6	21.952	26.07	19.94	0.004299	0.001113
955 C. I	1.013±0.04	2.163±0.07	0.21±0.006	0.416±0.012	2.121±0.012	2.282±0.00	0.089±0.00	0.497±0.0035					0.057±0.00	0.0037±0.0

M/m/5/15+GAMA(0.5,2)														
1	1.28	2.67	0.21	0.42	2.61	2.76	0.07	0.5	19	394	5715	6128	0.064295	0.003101
2	1.08	2.32	0.18	0.37	2.57	2.73	0.08	0.49	8	312	5767	6087	0.051257	0.001314
3	1.33	2.74	0.22	0.43	2.64	2.78	0.07	0.5	31	373	5764	6168	0.060473	0.005026
4	1.18	2.47	0.2	0.39	2.6	2.76	0.08	0.49	7	335	5845	6187	0.054146	0.001131
5	1.31	2.63	0.22	0.42	2.62	2.77	0.07	0.5	16	357	5807	6180	0.057767	0.002589
6	1.32	2.68	0.22	0.42	2.64	2.78	0.07	0.5	4	372	5824	6200	0.06	0.000645
7	0.98	2.27	0.17	0.36	2.55	2.71	0.09	0.49	9	270	5705	5984	0.04512	0.001504
8	1.2	2.57	0.2	0.4	2.59	2.75	0.08	0.5	13	383	5724	6120	0.062582	0.002124
9	1.07	2.33	0.18	0.37	2.57	2.73	0.08	0.49	2	291	5814	6107	0.04765	0.000327
10	1.38	2.78	0.23	0.44	2.63	2.78	0.07	0.51	24	412	5675	6111	0.067419	0.003927
mean	1.213	2.546	0.203	0.402	2.602	2.755	0.076	0.497	13	349.9	5764	6127	0.057071	0.002169
sd	0.133587	0.1865	0.020575	0.028206	0.031552	0.024608	0.006992	0.006749	9.2	46.525	57.66	63.4	0.007335	0.0015
error	0.095556	0.133405	0.014717	0.020176	0.02257	0.017602	0.005001	0.004828	6.6	33.279	41.24	45.35	0.005247	0.001073
95% C.I	1.213±0.09	2.546±0.13	0.203±0.010	0.402±0.02	2.602±0.02	2.755±0.010	0.076±0.00	0.497±0.0048					0.057±0.00	0.0022±0.0

Table 6.7: Erlang abandoning times with 0.926 traffic intensity.

Traffic intensity = 0.926

M/M/4/10+E(1,2)

seed	mean leng	std length	mean dela	std delay	mean utlz	std utlz	mean idle	mean busy	balk	abandon	served	total	P(A)	P(Balk)
1	0.49	1.22	0.12	0.27	1.59	1.77	0.13	0.49	0	524	3578	4102	0.12774	0
2	0.52	1.21	0.13	0.26	1.64	1.8	0.11	0.52	5	602	3565	4172	0.1443	0.0012
3	0.49	1.2	0.12	0.27	1.6	1.78	0.13	0.51	3	576	3537	4116	0.13994	0.0007
4	0.5	1.22	0.12	0.27	1.59	1.77	0.13	0.5	0	561	3501	4062	0.13811	0
5	0.53	1.26	0.13	0.28	1.63	1.79	0.12	0.51	2	596	3557	4155	0.14344	0.0005
6	0.46	1.18	0.11	0.27	1.57	1.76	0.14	0.5	1	537	3493	4031	0.13322	0.0002
7	0.57	1.32	0.14	0.29	1.65	1.8	0.11	0.51	5	620	3648	4273	0.1451	0.0012
8	0.57	1.32	0.14	0.29	1.65	1.8	0.11	0.51	5	620	3648	4273	0.1451	0.0012
9	0.43	1.1	0.11	0.24	1.57	1.76	0.13	0.48	1	523	3612	4136	0.12645	0.0002
10	0.55	1.29	0.13	0.29	1.65	1.8	0.11	0.52	2	596	3547	4145	0.14379	0.0005
mean	0.511	1.232	0.125	0.273	1.614	1.783	0.122	0.505	2.4	575.5	3569	4147	0.13872	0.0006
sd	0.046056	0.067954	0.010801	0.01567	0.0334	0.017029	0.011353	0.012693	2.01	37.423	54.24	79.04	0.00718	0.0005
error	0.032944	0.048608	0.007726	0.01121	0.023891	0.012181	0.008121	0.009079	1.44	26.769	38.8	56.54	0.00513	0.0003
95% C.I	0.511±0.03	1.232±0.04	0.125±0.00	0.273±0.0	1.614±0.02	1.783±0.010	0.122±0.00	0.505±0.009					0.139±0.0	0.0006±

M/M/5/12+E(1,2)

1	0.48	1.22	0.1	0.22	2.01	2.2	0.12	0.5	0	571	4430	5001	0.11418	0
2	0.52	1.35	0.1	0.25	2.02	2.21	0.12	0.5	1	538	4506	5045	0.10664	0.0002
3	0.52	1.34	0.1	0.24	2.02	2.2	0.12	0.5	7	558	4479	5044	0.11063	0.0014
4	0.49	1.24	0.1	0.23	2.01	2.21	0.12	0.5	0	570	4489	5059	0.11267	0
5	0.55	1.36	0.11	0.25	2.03	2.22	0.12	0.51	0	608	4377	4985	0.12197	0
6	0.52	1.36	0.11	0.25	1.98	2.19	0.13	0.5	0	550	4396	4946	0.1112	0
7	0.47	1.27	0.09	0.22	1.98	2.18	0.13	0.49	3	526	4567	5096	0.10322	0.0006
8	0.5	1.28	0.1	0.23	2.03	2.21	0.12	0.5	1	592	4412	5005	0.11828	0.0002
9	0.55	1.42	0.11	0.25	2.02	2.2	0.12	0.5	2	598	4467	5067	0.11802	0.0004
10	0.57	1.42	0.12	0.26	2.03	2.21	0.12	0.51	0	582	4480	5062	0.11497	0
mean	0.517	1.326	0.104	0.24	2.013	2.203	0.122	0.501	1.4	569.3	4460	5031	0.11318	0.0003
sd	0.032677	0.070427	0.008433	0.01414	0.018886	0.011595	0.004216	0.005676	2.22	26.575	57.16	45.4	0.00561	0.0004
error	0.023374	0.050377	0.006032	0.01012	0.013509	0.008294	0.003016	0.00406	1.59	19.009	40.89	32.47	0.00401	0.0003
95% C.I	0.517±0.02	1.326±0.05	0.104±0.00	0.24±0.0	2.013±0.01	2.203±0.00	0.122±0.00	0.501±0.004					0.113±0.0	0.0003±

M/M/6/15+E(1,2)

1	0.55	1.38	0.09	0.21	2.46	2.66	0.11	0.5	0	622	5363	5985	0.10393	
2	0.5	1.39	0.08	0.21	2.42	2.62	0.12	0.5	0	542	5452	5994	0.09042	
3	0.61	1.54	0.1	0.23	2.46	2.65	0.11	0.5	0	649	5406	6055	0.10718	
4	0.6	1.47	0.1	0.22	2.49	2.68	0.1	0.5	0	694	5483	6177	0.11235	
5	0.62	1.56	0.1	0.24	2.49	2.68	0.1	0.51	0	683	5409	6092	0.11211	
6	0.59	1.43	0.1	0.22	2.47	2.66	0.11	0.49	0	622	5491	6113	0.10175	
7	0.62	1.51	0.1	0.22	2.52	2.69	0.1	0.5	0	693	5572	6265	0.11061	
8	0.61	1.54	0.1	0.23	2.48	2.66	0.1	0.49	0	647	5520	6167	0.10491	
9	0.55	1.39	0.09	0.21	2.46	2.66	0.11	0.49	0	578	5500	6078	0.0951	
10	0.58	1.43	0.09	0.22	2.46	2.66	0.11	0.49	0	611	5272	5883	0.10386	
mean	0.583	1.464	0.095	0.221	2.471	2.662	0.107	0.497	0	634.1	5447	6081	0.10422	
sd	0.038887	0.069314	0.007071	0.00994	0.026437	0.019322	0.006749	0.006749	0	49.769	86.81	109.7	0.00714	
error	0.027816	0.049581	0.005058	0.00711	0.01891	0.013821	0.004828	0.004828	0	35.6	62.1	78.46	0.00511	
95% C.I	0.583±0.02	1.464±0.04	0.095±0.00	0.221±0.0	2.225±0.01	2.662±0.010	0.107±0.00	0.497±0.0048					0.104±0.0052	

Table 6.8: Abandoning times with traffic intensity of 1.1 for different distributions.

Traffic intensity=1.1

intaearrival time = 0.2083, service time = 1

seed	mean leng	std length	mean dela	std delay	mean utlz	std utlz	mean idle	mean busy	balk	abandon	served	total	P(A)	P(Balk)
M/M/4/10+GAMA(2, 2)														
1	0.63	1.32	0.12	0.23	1.72	1.84	0.08	0.5	1	1177	3737	4915	0.239471	0.000203
2	0.58	1.27	0.11	0.22	1.7	1.83	0.09	0.5	1	1074	3786	4861	0.220942	0.000206
3	0.59	1.24	0.12	0.21	1.73	1.85	0.08	0.51	1	1117	3821	4939	0.226159	0.000202
4	0.62	1.29	0.12	0.22	1.72	1.84	0.08	0.5	0	1184	3733	4917	0.240797	0
5	0.59	1.28	0.12	0.22	1.72	1.84	0.08	0.5	0	1116	3807	4923	0.226691	0
6	0.6	1.28	0.12	0.22	1.72	1.84	0.08	0.5	2	1095	3802	4899	0.223515	0.000408
7	0.52	1.17	0.1	0.2	1.69	1.82	0.09	0.49	0	992	3820	4812	0.206151	0
8	0.52	1.17	0.1	0.2	1.69	1.82	0.09	0.49	0	992	3820	4812	0.206151	0
9	0.62	1.32	0.12	0.22	1.74	1.85	0.08	0.51	1	1164	3783	4948	0.235247	0.000202
10	0.55	1.2	0.11	0.21	1.7	1.83	0.09	0.5	0	1023	3744	4767	0.2146	0
mean	0.582	1.254	0.114	0.215	1.713	1.836	0.084	0.5	0.6	1093.4	3785.3	4879	0.223973	0.000122
sd	0.039944	0.056608	0.008433	0.009718	0.017029	0.0107	0.005164	0.006667	0.7	72.379	35.296	62.61	0.01245	0.000143
error	0.028573	0.040492	0.006032	0.006952	0.012181	0.0077	0.003694	0.004769	0.5	51.773	25.247	44.79	0.008905	0.000102
95% C	0.582±0.02	1.254±0.04	0.114±0.00	0.215±0.00	1.713±0.011	.836±(	0.084±0.00	0.5±0.00477					0.224±0.00	0.00012±0
M/M/4/10+E(1, 0.5)														
1	0.26	0.72	0.05	0.13	1.61	1.76	0.12	0.49	0	1175	3595	4770	0.246331	0
3	0.29	0.76	0.06	0.13	1.64	1.78	0.11	0.51	0	1310	3525	4835	0.270941	0
4	0.29	0.75	0.06	0.13	1.61	1.77	0.12	0.49	0	1281	3626	4907	0.261056	0
5	0.3	0.79	0.06	0.14	1.64	1.78	0.11	0.51	0	1285	3528	4813	0.266985	0
6	0.28	0.77	0.06	0.13	1.61	1.77	0.12	0.5	0	1302	3466	4768	0.27307	0
7	0.29	0.77	0.06	0.13	1.65	1.78	0.11	0.5	0	1309	3677	4986	0.262535	0
8	0.29	0.77	0.06	0.13	1.65	1.78	0.11	0.5	0	1309	3677	4986	0.262535	0
9	0.28	0.75	0.06	0.12	1.61	1.77	0.12	0.49	0	1208	3653	4861	0.248509	0
10	0.29	0.77	0.06	0.13	1.62	1.77	0.12	0.51	0	1286	3509	4795	0.268196	0
11	0.28	0.75	0.06	0.13	1.62	1.77	0.12	0.49	0	1229	3633	4862	0.252777	0
mean	0.285	0.76	0.059	0.13	1.626	1.773	0.116	0.499	0	1269.4	3588.9	4858	0.261294	0
sd	0.010801	0.018856	0.003162	0.004714	0.017127	0.0067	0.005164	0.006756	0	48.091	76.209	79.97	0.009278	0
error	0.007726	0.013488	0.002262	0.003372	0.012251	0.0048	0.003694	0.006263	0	34.4	54.513	57.2	0.006637	0
95% C	0.285±0.00	0.76±0.013	0.059±0.00	0.13±0.003	1.63±0.012	1.773±(	0.116±0.00	0.499±0.0063					0.261±0.00	0
M/M/4/10+M														
1	0.54	1.24	0.11	0.22	1.69	1.82	0.09	0.5	1	1204	3710	4915	0.244964	0.000203
2	0.54	1.19	0.11	0.22	1.7	1.84	0.09	0.51	0	1193	3689	4882	0.244367	0
3	0.52	1.15	0.11	0.21	1.69	1.83	0.09	0.51	0	1153	3698	4851	0.237683	0
4	0.49	1.11	0.1	0.2	1.68	1.82	0.09	0.49	0	1100	3784	4884	0.225225	0
5	0.52	1.18	0.11	0.21	1.71	1.83	0.09	0.51	0	1155	3712	4867	0.237313	0
6	0.51	1.16	0.1	0.21	1.7	1.82	0.09	0.51	2	1142	3583	4727	0.241591	0.000423
7	0.49	1.14	0.1	0.21	1.68	1.82	0.09	0.49	0	1125	3828	4953	0.227135	0
8	0.5	1.11	0.1	0.2	1.7	1.82	0.09	0.49	0	1131	3822	4953	0.228346	0
9	0.48	1.11	0.1	0.2	1.69	1.82	0.09	0.49	0	1097	3800	4897	0.224015	0
10	0.52	1.17	0.1	0.21	1.69	1.82	0.09	0.5	0	1145	3700	4845	0.236326	0
mean	0.511	1.156	0.104	0.209	1.693	1.824	0.09	0.5	0.3	1144.5	3732.6	4877	0.234697	6.27E-05
sd	0.02079	0.041687	0.005164	0.007379	0.009487	0.007	1.46E-17	0.009428	0.67	34.77	75.832	64.78	0.007932	0.000142
error	0.014871	0.029819	0.003694	0.005278	0.006786	0.005	1.05E-17	0.006744	0.48	24.871	54.243	46.34	0.005674	0.000101
95% C	0.511±0.01	1.156±0.02	0.104±0.00	0.209±0.00	1.693±0.00	1.824±(	0.09±0	0.0094±0.0067					0.235±0.00	0

Bibliography

[1] Adans, I. and R. Jacques (2002), ***Queueing Theory,*** *Eindhoven*, Netherlands.

[2] Avramidis, A. and P. L'Ecuyer, **Modelling and Simulation of Call Centres**, *Proceedings of 2005 Winter conference.*

[3] Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn and L. Zhao, **Statistical Analysis of a Telephone Call Centre : A Queueing Science-Perspective**, *JASA*, 100(469), 2005, 35-50.

[4] Khoshmnevis, B. (1994), ***Discrete System Simulation,*** *McGraw-Hill,* New York.

[5] Koole, G. and A. Mandelbaum, ***Queueing Models of Call Centre : An Introduction***, *Annals of Operations Research*, 113, 2002, 41-59.

[6] Mandelbaum, A., **Call Centres: *Research bibliography with abstracts***, *Technical Report*, Haifa, 2002.

[7] Mandelbaum, A., A. Sakov and S. Zeltyn, ***Empirical Analysis of a Call Centre***. *Technical Report*, Haifa, 2001.

[8] Mandelbaum A and Zeltyn S (2004), **The Palm / Erlang-A Queue, with Application to Call Centres**, *Springer Berlin Heidelberg*, New York.

[9] Miler, R., G. (1981), **Survival Analysis**, *John Wiley*, New York.

[10] Nikolic, N., **Statistical Integration of Erlang's Equation**, *European Journal of Operational research*, 187, 2008, 1487-1493.

[11] Ross, S. M.(1996), **Stochastic Processes**, *Wiley*, New York.

[12] Ross, S. M. (2000), **Introduction to Probability Models** (Seventh Edition), *Academic Press*, New York.

[13] Ross, S. M. (2002), **Simulation**, *Academic Press*, New York.

[14] White, J. A., J. W. Schmodt and G. K. Bennett (1975), **Analysis of Queueing Systems**, *Academic Press*, New York.

[15] Whitt, W., **Dynamic Staffing in a Telephone Call Centre Aiming at to Immediately Answer all Calls**, *Operations Research letters*, 24(5), 1999, 205-212.

[16] Whitt, W., **Engineering Solution of a Basic Call-centre Model**, *Management Science*, 51(2), February 2005, 221-235.

[17] Whitt, W., **Stochastic Models for the Design and Management of Customer Contact Centres: Some Research Directions**, *working paper*, March 2002.

[18] Whitt, W., **Understanding the Efficiency of Multi-server Service Systems**. *Management Science*, 38 , 1992, 708–723.

[19] Zukerman, M. (2000), **Introduction to Queueing theory and Stochastic Teletraffic Models**, *Academic Press*, New York.

[20] http://www.vosesoftware.com/ModelRiskHelp/index.htm#Distributions /Continuous_distributions/Erlang_distribution.htm

[A] Erlang, A. K., **Solution of some Problems in the Theory of Probabilities of Significance in Automatic Telephone Exchanges**, *Elektrotkeknikeren*, vol 13, 1917.

[B] Dvoretzky, A., Kiefer, J. and Wolfowitz, J. , **Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator**, *Annals of Mathematical Statistics*, 27 (3), 1956, 642–669.