8-2007

Approximation Methods for the Standard Deviation of Flow Times in the G/G/s Queue

Xiaofeng Zhao

University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a dissertation written by Xiaofeng Zhao entitled "Approximation Methods for the Standard Deviation of Flow Times in the G/G/s Queue." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Management Science.

Kenneth C. Gilbert, Major Professor

We have read this dissertation and recommend its acceptance:

Mandyam M. Srinivasan, Melissa R. Bowers, Funda Sahin

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Xiaofeng Zhao entitled "Approximation Methods for the Standard Deviation of Flow Times in the G/G/s Queue". I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Management Science.

Kenneth C. Gilbert
Major Professor

We have read this dissertation
and recommend its acceptance:

Mandyam M. Srinivasan

Melissa R. Bowers

Funda Sahin

Accepted for the Council:

Carolyn R. Hodges
Vice Provost and Dean of the
Graduate School

(Original signatures are on file with official student records.)

APPROXIMATION METHODS FOR THE STANDARD DEVIATION OF FLOW TIMES IN THE G/G/s QUEUE

A Dissertation
Presented for the
Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Xiaofeng Zhao
August 2007

Dedication

To my parents

Yurong Zhang and Hongyi Zhao

For their unconditional support, encouragement, and love

Acknowledgement

There are many individuals to whom I owe much gratitude. My deepest thanks go to my advisor and dissertation committee chair Dr. Kenneth Gilbert, who has endured copious discussions on how to approach this topic and waded through numerous iterations of this dissertation, aiming for the highest degree of lucidity. He has always been generous with his time, expertise and enthusiasm. He has taught me so much on how to do research. Without his guidance, insight and commitment, completing this dissertation would not have been possible.

I would also like to express my appreciation to the other members of my dissertation committee: Dr. Mandyam Srinivasan, Dr. Melissa Bowers, and Dr. Funda Sahin, who give generosity of their time and expertise to help me refine and focus on my research, greatly strengthening it.

I will always be grateful for the opportunity to study and work with all faculty, staff and fellow students at the College of Business, who teach me and demonstrate a true commitment to quality education. A special thank you is due to Dr. Frank Guess for his invaluable support and indefatigable positive attitude. He has taught me so much on how to become an excellent scholar. I would also like to thank Dr. Russell Zaretzki for his helpful discussions and comments, Ms. Amber Clay for helping me learn the simulation software. In particular, I would like to thank Ms. Jane Moser, who helps me so much to make my journey at University of Tennessee a pleasant and wonderful one.

My wife and son deserve special recognition for encouraging my education. I owe a large debt of gratitude to them for their patience, understanding, and never-ending support. Last but certainly not the least, I deeply thank my parents and my sister’s family. Although this dissertation is still very foreign to them, they are constantly stressing the importance of education and unconditionally supporting my pursuit of higher goals.

Abstract

We provide approximation methods for the standard deviation of flow time in system for a general multi-server queue with infinite waiting capacity ($G/G/s$). The approximations require only the mean and standard deviation or the coefficient of variation of the inter-arrival and service time distributions, and the number of servers.

These approximations are simple enough to be implemented in manual or spreadsheet calculations, but in comparisons to Monte Carlo simulations have proven to give good approximations (within $\pm 10\%$) for cases in which the coefficients of variation for the inter-arrival and service times are between 0 and 1. The approximations also have the desirable properties of being exact for the specific case of Markov queue model $M/M/s$, as well as some imbedded Markov queuing models ($E_k/M/1$ and $M/E_\alpha/1$).

The practical significance of this research is that (1) many real world queuing problems involve the $G/G/s$ queuing systems, and (2) predicting the range of variation of the time in the system (rather than just the average) is needed for decision making. For example, one job shop facility with which the authors have worked, guarantees its customers a nine day turnaround time and must determine the minimum number of machines of each type required to achieve nine days as a "worst case" time in the system. In many systems, the "worst case" value of flow time is very relevant because it represents the lead time that can safely be promised to customers. To estimate this we need both the average and standard deviation of the time in system.

The usefulness of our results stems from the fact that they are computationally simple and thus provide quick approximations without resorting to complex numerical techniques or Monte Carlo simulations. While many accurate approximations for the $G/G/s$ queue have been proposed previously, they often result in algebraically intractable expressions. This hinders attempts to derive closed-form solutions to the decision variables incorporated in optimization models, and inevitably leads to the use of complex numeric methods. Furthermore, actual application of many of these approximations often requires specification of the actual distributions of the inter-arrival time and the service time. Also, these results have tended to focus on delay probabilities and

average waiting time, and do not provide a means of estimating the standard deviation of the time in the system.

We also extend the approximations to computing the standard deviation of flow times of each priority class in the $G/G/s$ priority queues and compare the results to those obtained via Monte Carlo simulations. These simulation experiments reveal good approximations for all priority classes with the exception of the lowest priority class in queuing systems with high utilization. In addition, we use the approximations to estimate the average and the standard deviation of the total flow time through queuing networks and have validated these results via Monte Carlo Simulations.

The primary theoretical contributions of this work are the derivations of an original expression for the coefficient of variation of waiting time in the $G/G/s$ queue, which holds exactly for $G/M/s$ and $M/G/1$ queues. We also do some error sensitivity analysis of the formula and develop interpolation models to calculate the probability of waiting, since we need to estimate the probability of waiting for the $G/G/s$ queue to calculate the coefficient of variation of waiting time.

Technically we develop a general queuing system performance predictor, which can be used to estimate all kinds of performances for any steady state, infinite queues. We intend to make available a user friendly predictor for implementing our approximation methods. The advantages of these models are that they make no assumptions about distribution of inter-arrival time and service time. Our techniques generalize the previously developed approximations and can also be used in queuing networks and priority queues. Hopefully our approximation methods will be beneficial to those practitioners who like simple and quick practical answers to their multi-server queuing systems.

Key words and Phrases: Queuing System, Standard Deviation, Waiting Time, Stochastic Process, Heuristics, $G/G/s$, Approximation Methods, Priority Queue, and Queuing Networks.

Contents

Chapter 1

Introduction and motivation

1.1 Introduction

The goal is to develop approximation methods for estimating both average and standard deviations of flow times in the $G/G/s$ queue and networks of $G/G/s$ queues with or without priority classes.

We present approximations for the standard deviation of waiting time in system for a general multi-server queue with infinite waiting capacity ($G/G/s$). The $G/G/s$ model has a single service facility with s identical servers, unlimited waiting capacity and the first come first served queue discipline. The inter-arrival times are independent and identically distributed (i.i.d.) with a general distribution, the service times are also independent and identically distributed with a general distribution.

We also extend the approximations to computing the standard deviation of flow times of each priority class in the $G/G/s$ priority queues. In addition, we use the approximations to estimate the average and the standard deviation of the total flow time through queuing networks. We have validated these results via Monte Carlo Simulations by using Extend simulation program.

Most real world queuing problems are the $G/G/s$ systems. They do not satisfy the assumptions of Markov queuing model $M/M/s$. Inter-arrival times are not always exponential; service times are also unlikely to be exponential. To address systems with non-exponential inter-arrival and service time distributions, we must turn to the $G/G/s$ queue, which reflects the real world.

Unfortunately, without the memory-less property of the exponential distribution to facilitate analysis, we can not compute exact performance measures for the $G/G/s$ queue. When it comes to exact solutions of multi-server queuing systems, the more one departs from the assumption of exponential, the thornier the problem becomes, especially if this happens for the service time. Due to its inherent complexity, analysis of the $G/G/s$ queue in general is notoriously difficult.

However, this does not mean that we should give up on modeling queuing systems, only that we need to be concerned with finding good approximations. In contrast, an exact formula may be capable of giving the exact answers to the wrong problem or a mathematically intractable answer to the problem of interest. Consequently, approximations have been studied extensively.

The purpose of this research is to provide a simple yet good approximation for the standard deviation of a general multi-server queue with infinite waiting capacity ($G/G/s$). We provide a new method for the analysis of the $G/G/s$ queue that is based on heuristic and interpolation methods.

We develop models by means of a two-moment approximation, which makes use of only the mean and standard deviation or coefficient of variation (c) of the interarrival and service time distributions. The approximation method was motivated by the results of $M/M/s$ queue and imbedded Markov chain queues $G/M/s$ and $M/G/1$.This formula has the form of the exact variance of waiting times for these queues and hence it can be easily calculated. The quality of the approximation is tested by comparing it with simulation results or by comparing it with a few known numerical results in particular cases.

To develop the approximation of the standard deviation of waiting time, we have studied the equivalent (under the assumption that a good approximation exists for the average time in the queue) problem of finding a mathematically tractable formula of estimating the coefficient of variation of waiting time $c_q = \sigma_q / W_q$, where W_q and σ_q are respectively the average and standard deviation of the time in queue.

We first derive exact results of c_q for Markov queues and some imbedded Markov queue models $M/M/1$, $M/M/s$ $M/G/1$, $G/M/s$ queues, and then we apply heuristic and interpolation methods to approximate the $G/G/s$ queue and extend the approximation results to priority queues and queuing networks.

For $G/M/s$ queuing models, we find that the coefficient of variation of waiting time is just a function of the probability of waiting. They are not all related to the distribution of inter-arrival time and service time. We then generalize the expression to the $M/G/1$ queue. We conjecture that for the $G/G/s$ queue these relationships still hold and all queuing systems have the same relationship. Since we do not assume that G is specified, we must estimate it by assuming some known distribution for the service times, e.g. gamma, for which the third moment can be computed as a function of the average and standard deviation. We also have to estimate $P(T_q = 0)$. Therefore, we need to do error sensitivity analysis to show that our result is relatively insensitive to errors in estimating these inputs.

Similarly, we propose approximations for queuing networks and priority classes of the $G/G/s$ queue. For the $M/M/s$ queuing series, the departure time distribution from an $M/M/s$ queue is identical to the inter-arrival time distribution, namely, exponential. Hence, all stations are $M/M/s$ models. In general $G/G/s$ queues, our model can estimate all situations by using entering c_a as departure c_d of the previous queue. So we can estimate all kinds of $G/G/s$ queuing networks. When $G/G/s$ models appear as sub-models, simple closed form analytic formulas are useful. For multi-class jobs, we use the law of total variance to calculate pooled average and pooled variance.For priority queues, we conjecture the approximations still hold for each priority classes.

Since no closed-form analytical results are available for $G/G/s$ models, to evaluate the accuracy of our approximations, we conduct Monte Carlo simulation experiments by using the Extend simulation program to gain insight into the heuristic methods for calculating approximate steady-state performance measures of $G/G/s$ queuing system. The testing of our approximations has been based on extensive simulation experiments. These simulation experiments are indispensable parts of our research on the $G/G/s$ queue.

Although the approximation derivations may appear complicated, this approximation is simple enough to be implemented in manual or spreadsheet calculations, but in comparison to Monte Carlo simulations has proven to give good approximations (within $\pm$ 10%) for cases in which the coefficients of variation for the inter-arrival and service times are between 0 and 1. The approximation also has the desirable property of being consistent with the specific case of $M/M/s$ queue, as well as some imbedded Markov chain queuing models $M/G/1$, $G/M/s$.

This makes it possible to couple the single queue approximation with the multiple linking equations to create a spreadsheet tool for analyzing all kinds of performances of queuing networks. Although it is not the focus of this research, we believe that the research can be further extended to more complicated situations such as queues with balking, batching, and optimal design etc.

1.2 Motivation

We intend to provide a quick spreadsheet alternative to more elaborate simulation models for analyzing real world systems. Recent years have witnessed a growing volume of good quality approximations for average waiting time of the $G/G/s$ queue W_q (Sakasegawa 1977, Kimura 1986, Whitt 2004). While the accuracy of these approximations is usually satisfactory, they often result in algebraically intractable expressions. This hinders attempts to derive closed-form solutions to the decision variables incorporated in optimization models, and inevitably leads to the use of complex numerical methods or to recursive schemes of calculation. Further more, actual application of many of these approximations is often obstructed due to the thorough specification that is needed of inter-arrival or service time distribution.

Because of mathematical complications, closed-form solutions have been difficult to achieve. Consequently, approximations have been studied extensively. However, all existing approximations appear to be cumbersome or computationally demanding. It often turns out that it is not possible to develop analytical models for some queuing systems, such as the $G/G/s$ queue. It is the popular realization of this fact that has lead to the rush towards simulation techniques. While simulation may offer a way out for many analytical intractable models, it is not in itself a panacea. Simulation needs special training and is at a relatively high cost. There are also a considerable number of pitfalls one may encounter in using simulation. The success or failure of simulation study often lies in how it is used and how the output is interpreted. Because of this, simulation analysis has often been referred to as an art. Therefore, we should explore and propose analytical approximation models.

In addition, all current literature focuses on delay probability and average waiting time. We have seen very little literature dealing with variance of waiting time in the $G/G/s$ queue as well as its queuing networks and priority classes. Only bounds or approximations of waiting time have been found in the literature. When these bounds are used as approximations, they appear to be rather crude (Boxma 1979). Nevertheless, understanding the variance of flow times in the system is essential to understanding the performance of a queuing system.

We focus on the standard deviation of the total flow time in a system. In many systems, the "worst case" value of flow time is very relevant because it represents the lead time that can safely be promised to the customers. Predicting the range of variation of the time in the system (rather than just the average) is needed for decision making. To estimate this they need both the average and standard deviation of the time in system.

Another considerable portion of real world queuing situations contain priority considerations. Priority queues are generally more difficult to model than non-priority situations, but nevertheless, the priority models should not be oversimplified merely to permit solution. Full consideration of priorities is absolutely essential when we consider costs of a queuing system and optimal design. In current literature, tractable priority queuing formulas are limited to $M/M/s$. In this research, we only focus on the non-preemptive $G/G/s$ system with many priorities and introduce formulas for performance estimation.

Furthermore, existing methods are not designed to handle queuing networks. The characteristics of real world queuing systems are that they are often networked. The arrivals at a queue may be the output or a fraction of the output of more than one queue. Also, there may be several classes of jobs each having different service time distributions.

Queuing networks can be described as a group of nodes where each node represents a service facility of some kind with servers at nodes. In most general cases, customers may arrive from outside the system to any node and may depart from the system from any nodes. Thus, customers may enter the system at some node, traverse from node to node in the system, and depart from some node, where not all customers necessarily enter and leave at the same nodes or taking the same path once having entered the system. In our research we consider tandem queue models in which there is a series of service stations through which each service unit must progress prior to leaving the system.

The advantages of these models are that they make no assumptions about distribution of inter-arrival time and service time. Therefore, they are more general than other infinite queuing models. Our techniques generalize the previously developed approximations and can be used in all kinds of real world queue situations including queuing networks and priority queues.

Chapter 2

Literature review

Queuing theory has been studied thoroughly throughout the past decades, but many problems still remain unsolved, in spite of the effort and intelligence devoted to them. Among these problems, the analysis of the $G/G/s$ queuing system has survived the attacks of many excellent mathematician and management scientists, due to its inherent complexity.

Queuing systems have provided many models for different kinds of queues. There are many queuing systems of practical interest for which exact analysis is difficult due to the generality in their stochastic structures. Most real world queuing systems are $G/G/s$ queuing systems. They don't satisfy the assumptions of the $M/M/s$ queuing model.

Unfortunately, without the memory-less property of the exponential to facilitate analysis, we can't compute exact performance measures for the $G/G/s$ queue. To deal with the difficulty, we often need approximation. Therefore, $G/G/s$ approximation models are still subject to active research (Whitt 2004 etc). The following is a brief literature review on the approximations of $G/G/s$ queue over the last 30 years.

Sakasegawa (1977) provided a closed-form approximation formula for the $G/G/s$ queue. He suggested the following closed-form expression for approximating the mean waiting time in the $G/G/s$ queue.

$$W_q(G/G/s) = \left(\frac{c_a{}^2 + c_s{}^2}{2}\right)\left(\frac{\rho^{\sqrt{2(s+1)}-1}}{s(1-\rho)}\right)\frac{1}{\mu}$$

This approximation has several nice properties. First, it is exact for the $M/M/s$ queue. It neatly separates into three terms: a dimensionless variability term V, utilization term U and a time term T (service time). Whitt (1983) discussed this formula. Although it may appear complicated, it does not require any type of iterative algorithm to solve and is therefore easily implemented in a spreadsheet program. This also makes it possible to couple the single queue approximation with multiple-server to create a spreadsheet tool for analyzing the performance of a series of queues.

Kimura (1986) provided a simple two-moment approximation formula for the mean waiting time in a $G/G/s$ queue. This formula has the form of a combination of the exact mean waiting times for $D/M/s$, $M/D/s$ and $M/M/s$ queues, and hence it can be easily calculated. It depends only on the first two moments $M/M/s$ of inter-arrival times and service times.

Bertsimas (1987) discussed an analytic approach to a general class of a $G/G/s$ queuing system, but he assumed G is the class of Coxian probability density functions, which is a subset of the PDF that have rational Laplace transforms. Although the method of stages he presented is not immediately extendable to distributions which do not have rational Laplace transform, he believed that this separable property holds for the more general model. He used conjecture methods. Whitt (1983) also made some conjectures about the equilibrium waiting time distribution in the $M/G/s$ queue. He presented several conjectures about the qualitative behavior of multi-servers queues and some supporting evidence based on light-traffic limits and heavy-traffic limits and a special family of service time distributions.

Whitt (1988) developed a closed form approximation for the mean steady-state workload or virtual waiting time in a $G/G/1$ queue, using the first two moments of the service-time distribution. Girish and Hu proposed an approximation technique which combines the light and heavy traffic characteristics. They showed how this can be applied for estimating the waiting time moments of the $G/G/1$ queue.

Whitt (1993) briefly mentioned an alternative approach for approximating the variance of waiting time. It is to approximate the tail probability by a simple exponential distribution $P(W > x) \approx \alpha e^{-\eta x}$, where η and α are obtained from the limit $e^{\eta x} P(W > x) \to \alpha \quad as \quad x \to \infty$. Since the asymptotic is known to hold in considerable generality. Whitt conjecture that for $G/G/s$ it still holds and analogs could be established.

Kimura (1994) developed a diffusion-approximation model for stable $G/G/s$ queues. He considered the standard $G/G/s$ queuing system with s homogenous servers in parallel, unlimited waiting spaces, the FIFO discipline, and IID service times which are independent of a renewal arrival process. For the $G/G/s$ case, possible approaches are quite limited and essentially heuristic by nature. The queuing length in $G/G/s$ queue is approximated by a diffusion process on the non-negative real line. Some heuristics on the state space and the infinitesimal parameters of the approximating diffusion process are introduced to obtain an approximation formula for the steady-state queuing length distribution. It is shown that the formula is consistent with the exact results for the $M/M/s$ and $G/M/s$ queues.

An alternative approach to approximating steady-state distributions is simple exponential approximation using an asymptotic method: approximate the steady-state waiting-time tail probability $P(W > x)$ by $\alpha e^{-\eta x}$, where η and α are called determined from the limit $e^{\eta \alpha} P(W > x) \to \alpha \ as \, x \to \infty$. The parameters η and α are called the asymptotic decay rate and asymptotic constant, respectively. Abate and Whitt (1994) discussed exponential approximations for steady-state distributions in the $G/G/s$ model based on asymptotic method. The key quantity is the asymptotic decay rate η, which in general depends on more than basic queuing parameters.

Gross & Harris (2002) and Kleinrock, L. (1975 & 1976) systematically summarized all queuing concepts and theories in their book “Fundamentals of Queuing Theory” and “Queuing Systems”. We develop our formula and approximations mostly based on the basic theories and extensive discussions of the concepts and theory of the steady state queues in the books.

The above discussions focused on delay probability or mean waiting time or queuing length. We have seen very few discussions about the standard deviation of waiting time in the $G/G/s$ queue in the literature. Only Whitt (1993) mentioned the approximation for the variance of steady state waiting time. However, no further details were provided. He just suggested using the formula for M/G/1 as approximations for M/G/s and $G/G/s$. The idea is that the conditional delay should depend much more on the service time distribution than the inter-arrival time distribution. Seelen and Tijms (1984) provided additional support for this approximation principle.

Whitt (2004) summarized the diffusion approximation for the $G/G/s$ queue. He developed diffusion approximation for the queue length stochastic process in the $G/G/s$ queuing model.

He pointed out that because the asymptotic delay probability function has proven to be so important for the Markovian $M / M / s$ queue, he found analogs for the non-Poisson arrival process and non-exponential service-time distribution. In his research he primarily focused on an approximation for the steady-state delay probability and the steady-state probability that all servers are busy in the $G / G / s$ model.

A serious defect in the previous diffusion approximation models is that they are not consistent with exact results available for particular cases (Kimura 1986 and Whitt 2004). For instance, none of the previous diffusion approximations for the queuing length distribution in the $G / G / s$ queue are consistent with any exact results, even with the $M / M / s$ queue. It is obvious that the lack of consistency makes diffusion models less reliable.

In our research, we estimate the average and standard deviation of waiting time by means of two moment approximation, which makes use of only the mean, and standard deviation or coefficient of variation (c) of inter-arrival and service time distribution. We consider the standard $G / G / s$ queuing system with s homogeneous servers in parallel, unlimited waiting capacity, the first come first served discipline and independent sequence of independent and identically distributed (i.i.d.) inter-arrival times and service times.

Chapter 3

Queuing theory basics and assumptions

In this chapter, we first introduce basic queuing concepts and notations, as well as the theory basics behind them. We then outline the research design and methodology.

Variability is the quality of non-uniformity of a class of entities. Variability exists in all operations systems and can have an enormous impact on performance. Worst cases represent systems where performance is degraded by variability. From a management point of view, it is clear that the ability to deal effectively with variability and uncertainty will be an important skill for the foreseeable future. For this reason, the ability to measure, understand, and manage variability is crucial to effective operations management (Hopp and Spearman 2001).

To effectively analyze variability, we must be able to quantify it. We do this by using standard measures from statistics and stochastic models to define a set of variability classes. Variance is a measure of absolutely variability, as is the standard deviation, defined by the square root of the variance. Often, however, absolute variability is less important than relative variability. A reasonable relative measure of variability of a random variable is the standard deviation divided by the mean, which is the coefficient of variation. Using this unit-less ratio, we can make consistent comparisons of the level of variability in both process and flows. We use the coefficient of variation for representing and analyzing variability in operations systems.

The subject of queuing systems is not directly concerned with optimization. Rather it attempts to explore, understand, and compare various queuing situations and thus indirectly achieve optimization approximately. In general, unlike optimization theory in which the main concern is to maximize or minimize an objective function subject to constraints, queuing theory is mostly a mathematically descriptive theory. It attempts to formulate, interpret, and predict for the purpose of better understanding of queues and for the sake of introducing remedies.

3.1 Queuing systems basics

Queuing systems represent an example of a much broader class of interesting dynamic systems, which, for convenience, we refer to as “systems of flow”. Flow systems are one in which some customers or items flow, move, or are transferred through one or more channels in order to go from one point to another. In this research, we merely consider steady flow.

Queuing theory is a branch of applied mathematics utilizing concepts from the field of stochastic processes. It has been developed in an attempt to predict fluctuating demands from observational data and to enable an organization to provide adequate service for its customers with tolerable waiting. However, the theory also basically improves understanding of a queuing situation, enabling better control. The predictions help the management to anticipate situations and to take appropriate measures to alleviate congestion.

In practice, we observe that actual process time typically represents only a small fraction (5 to 10 percent) of the total cycle time in a plant (Hopp and Spearman 2001). This has been documented in numerous published surveys (e.g. Bradt 1983). The majority of the extra time is spent waiting for various resources (e.g. workstations transport devices, machine operations, etc).So it is important to estimate the variance of waiting time.

A queuing system combines the components that have been considered so far: an interarrival process, a service process, and a queue. Arrivals can consist of individual customers or batches. Customers can be identical or have different characteristics. Interarrival times can be constant or random. The work station can have a single server or several servers in parallel, which can have constant or random process times. The queuing discipline can be first come first served (FIFO), last come first serve (LIFO), and a variety of priority schemes. The variety of queuing systems is almost endless.

Regardless of the queuing system under consideration, the primary job of queuing theory is to characterize performance measures in terms of descriptive parameters.

Queuing Notations and Measures

To use queuing theory to describe the performance of a single queue, we assume the following basic parameters are known:

λ : Arrival rate of entering customers

μ : Service rate of each servers

ρ : Average utilization of servers ($\rho = \lambda / s\mu$)

s : Number of parallel servers at station

c_a : Coefficient of variation of inter-arrival time

c_s : Coefficient of variation of service time

The performance measures we will focus on are:

L : Average number of customers in system

L_q : Average number of customers waiting in queue

W : Average time a customer spends in system

W_q : Average time a customer spends waiting in queue

σ_q : Standard deviation of waiting time

c_q : Coefficient of variation of waiting time

$P_n(t)$ =probability of n customers in system at time t

In addition to the above parameters, a queuing system is characterized by a host of specific assumptions, including the type of arrival and process time distributions, dispatching rules, balking protocols, batch arrivals or processing, whether it consists of a networking of queues, whether it has single or multiple customer classes and many others. Following convention, we use Kendall's notation, which characterizes a queuing station by means of four parameters: $A / B / s / b$

Where A describes the distribution of inter-arrival times, B describes the distribution of service times, s is the number of servers at the stations, and b is the maximum number of customers that can be in the system. For instance:

D: constant (deterministic) distribution

M: exponential (Markov) distribution

E_k : Erlang distribution

E_α : Gamma distribution

G: general distribution

In many situations, queue size is not explicitly restricted (e.g. the buffer is very large).
We indicate this case as $A/B/s/\infty$ or simply as $A/B/s$. In our research, we focus on $G/G/s$ queue.

Some fundamental relations

Before considering specific queuing systems, we note that some important relationships hold for all single queue systems (i.e. regardless of the assumptions about inter-arrival and process time distributions, number of servers, etc.)

(1) Utilization, which is the measure of traffic intensity, is given by $\rho = \lambda / \mu s$

(2) Relation between total mean time spent in the system and mean time spent in queue W_q .

Since means are additive, we have E (time in system) =E (time in queue) +E (time in service), i.e. $W = W_q + t_s$

(3) Applying Little's rule to any queue yields a relation among W, L, W_q , L_q and the arrival rate $\lambda : L = \lambda W$; $L_q = \lambda W_q$ .Using these relations and knowledge of any one of the four performance measures $W_q, W, L_q, and\, L$, we can complete the other three.

All fundamental relations are exact, even if the independence assumptions of the $G/G/s$ model are dropped. As a consequence, in a complicated open queuing network model, these relations are valid without any assumption.

3.2. Stochastic process and Markov chains

Queuing theory is also a branch of management science utilizing concepts from the field of stochastic processes. A stochastic process is the mathematical abstraction of empirical process whose development is governed by probability laws. From the point of view of the mathematical theory of probability, a stochastic process is best defined as a family of random variables, $\{X(t), t \in T\}$ defined over some index set or parameter space T. The set T is sometimes also called the time range and $X(t)$ denotes the state of the process at time t. Depending upon the nature of the time range, the process is either a discrete parameters or continuous Markov chain as follows:

(i) If T is a countable sequence, for example, $T = \{0, \pm 1, \pm 2, ...\}$

Then the stochastic process $\{X(t), t \in T\}$ is said to be a discrete time process defined on the index set T.

(ii) If T is an interval or an algebraic combination of intervals, for example

$T = \{t : -\infty < T < +\infty\}$ *or* $T = \{t : 0 < T < +\infty\}$

Then the stochastic process $\{X(t), t \in T\}$ is called a continuous time process defined on the index set T.

Markov Process

A discrete stochastic process $\{X(t), t = 0,1,2,...\}$ or a continuous-parameter stochastic process $\{X(t), t > 0\}$ is said to be a Markov process if, for any set of n time points $t_1 < t_2 < ... < t_n$ in the index set or time range of the process, the conditional distribution of $X(t_n)$, given the values of $X(t_1), X(t_2), X(t_3), ..., X(t_{n-1})$, depends only on $X(t_{n-1})$, the immediately preceding value; more precisely, for any real numbers $x_1, x_2, ..., x_n$,

$$\Pr\{X(t_n) \le x_n \mid X(t_1) = x_1, ... X(t_{n-1}) = x_{n-1}\}$$
$$= \Pr\{X(t_n) \le x_n \mid X(t_{n-1}) = x_{n-1}\}$$

Markov processes are classified according to:

(i) The nature of the index set of the process(whether discrete or continuous);

(ii) The nature of state space of the process.

A real number x is said to be a state of a stochastic process $\{X(t), t \in T\}$ if there exists a time point t in T such that the $\Pr\{x - h < X(t) < x + h\}$ if possible for every h>0. The set of possible states constitutes the state space of the process. If the state space is discrete, the Markov process is a Markov chain.

A discrete Markov process with discrete state space is a discrete Markov chain. A Markov chain is finite if the space is finite; otherwise, it is s denumerable or infinite Markov chain. Since the system is observed at a discrete set of time points, let the successive observations be denoted by $X_0, X_1, X_2, \ldots, X_n, \ldots$. It is assumed that X_n is a random variable whose value represents the state of the system at the nth time point. The sequence $\{X_n\}$ is called a chain if it is assumed that there are only a finite number of states in which the system may be found at any point within the given time range. The sequence $\{X_n\}$ is thus a Markov chain if each random variable X_n is discrete and the following holds: for any integer m>2 and any set of m points $n_1 < n_2 < \ldots < n_m$, the conditional distribution of X_{n_m}, given values of $X_{n_1}, X_{n_2}, \ldots X_{n_{m-1}}$, Depend only on $X_{n_{m-1}}$, the immediately preceding value; that is,

$$\Pr\left\{X_{n_m} = x_{n_m} \middle| X_1 = x_{n_1}, \ldots X_{n_{m-1}} = x_{n_{m-1}}\right\} = \Pr\left\{X_{n_m} = x_{n_m} \middle| X_{n_{m-1}} = x_{n_{m-1}}\right\}$$

A continuous Markov process with discrete state space is called a continuous Markov chain, while for continuous state space and discrete parameter space, the process is called a discrete parameter of Markov process. If both the state spaces and parameters spaces are continuous, it is called a continuous parameter Markov process.

3.3 Research design and methodology

In general, models can be classified into two types: descriptive and prescriptive. Descriptive models which describe some current real world situation, while prescriptive models are models which prescribe what the real world situation should be, that is optimal behavior at which to aim. Most of the queuing models are descriptive in that for given types of arrivals and service patterns, and specified queuing discipline and configuration, the state probabilities, expected value measures of effectiveness, and variations which describe the system are obtained.

The subject of queuing is not directly concerned with optimization. Rather it attempts to explore, understand, and compare various queuing situations and thus indirectly achieve optimization approximately. In general, unlike optimization theory in which the main concern is to maximize or minimize an objective function subject to constraints, queuing theory is mostly a mathematically descriptive theory. It attempts to formulate, interpret, and predict for the purpose of better understanding of queues and for the sake of introducing remedies.

For simplicity, we restrict our consideration to systems with a single job class (i.e. single customer, no batching). Of course, most operations systems have multiple products. But we can develop the key insights into the role of variability in systems with single job class models. More ever, these models can be extended to approximate the behavior of multiple job classes and batching systems.

In this research, we develop descriptive models. We consider initially the $M/M/1$ and $M/M/s$ queuing systems because they yield important intuition and serve as building blocks for more general systems. Then we analyze imbedded Markov chain queuing models. We present the exact formula for $M/M/1$, $M/M/s$ and the imbedded Markov chain queuing models $G/M/s$, $M/G/1$ and show how we apply approximation methods to extend it to general $G/G/s$ queue. We validate our results via Monte Carlo simulation by using the Extend simulation program. The following is the outline of the next two chapters.

I. An exact method for estimating the standard deviation of waiting time, applicable to $M/M/s$, $G/M/s$ and $M/G/1$.

a. Formula for coefficient of variation of waiting time, applicable to $M/M/s$, $G/M/s$ and $M/G/1$.

b. Formula for $P(T_q > 0)$ for $M/M/s$, $M/G/1$ and $E_k/M/1$.

II. A heuristic approximation method for the $G/G/s$ queues

a. First approximate $G/G/s$ queue using the $M/G/1$ having the same λ and μ.

b. Then approximate the $M/M/s$ queue with $M/M/1$ queue having the same arrival rate and same probability of waiting $P(T_q > 0)$.

c. Do error sensitivity analysis to show the formula is relatively insensitive to errors in estimating inputs.

d. Then use the interpolation method to estimate $P(T_q > 0)$ for $G/G/s$ queue having the same c_a and c_s and the same arrival rate and service rate as $G/G/1$.

Chapter 4

Exact Methods for M/M/s, G/M/s and M/G/1

4.1. Formula for the coefficient of variation of waiting time

(1)Exact coefficient of variation of waiting time for $M/M/1$ queue

One of the simplest queuing systems to analyze is $M/M/1$. This model assumes exponential inter-arrival times, a single server with exponential process times, a first come first served discipline, and unlimited space for customers waiting in queue. While not an accurate representation of most systems, the $M/M/1$ queue is tractable and offers valuable insights into more complex and realistic systems.

The key to analyzing the $M/M/1$ queue is the memory-less property of the exponential distribution. To begin, we require information about the inter-arrival and service times. Since both are assumed to be exponential, all we need to know are the means (because the standard deviation is equal to the mean for the exponential distribution). Beyond that, the only other information we need is how many customers are currently in the system. Because the inter-arrival and process time distributions are memory-less, the time since the last arrival and the time the current customer has been in process are irrelevant to the future behavior of the system. Because of this, the state of the system can be expressed as a single number n, representing the number of costumers currently in the system. By computing the long-run probability of being in each state, we can characterize all the long-term (steady state) performance measures, including $L_q, L, W, and\ W_q$.

Performance measures

The various steady state performance measures for $M/M/1$ queue can be computed from the results derived in many literatures (Gross and Harris 2002, Kleinrock 1975).
Some basic notations and fundamental relations and performance measures for $M/M/1$:
$Var(x) = E(X^2) - [E(X)]^2$ (For all queues)

$L = \frac{1}{1-\rho}$, $L_q = \frac{\rho^2}{1-\rho}$

$W_q = \frac{\rho}{1-\rho} \cdot \frac{1}{\mu}$,$W = W_q + \frac{1}{\mu}$

$P_n(t)$ =probability of n customers in system at time t

$P_n = \rho^n (1-\rho)$ where $(\rho = \lambda / \mu)$

P_0 =probability of no customers in system $P_0 = (1-\rho)$

$\rho = 1 - P_0$: Probability of a customer waiting

Assumptions for $M/M/1$ and $M/M/s$:

- Infinite model assuming that there is no limit to the waiting capacity.
- Identical servers and infinite waiting capacity
- Interarrival times and service times are exponentially distributed
- First come first served discipline
- $\rho = \lambda/(s\mu) < 1$ (steady state)

For the $M/M/1$ model, we first parallel Gross & Harris (1985) .The density function for the inter-arrival times and services times are given respectively, as

$$a(t) = \lambda e^{-\lambda t} \qquad \text{b}(t) = \mu e^{-\mu t}$$

Where $1/\lambda$ is the mean inter-arrival time; $1/\mu$ is the mean service time. We define:

T_q =time spent waiting in queue

$W_q(t)$ = the probability of a customer waiting a time less than or equal to t for service.

$W_q(0) = \Pr\{$ system empty at an arrival $\}$=$\Pr\{\mathrm{T_q} \le 0\} = \Pr\{\mathrm{T_q} = 0\}$

q_n =conditional probability of n customers in the system given arrival is about to occur

$$W_q(0) = P_0 = 1 - \rho$$

Since the service distribution is memory-less, the distribution of the time required for n completions is independent of the time of the current arrival and is the convolution of n exponential random variables.

In addition, since the input is Poisson, the arrival points are uniformly spaced and hence the probability that an arrival finds n in the system as identical to the stationary distribution of system size.

Therefore, we may write that:

$$
\begin{aligned}
&W_q(t) = \Pr\{T_q \le t\} \\
&= \sum_{n=1}^{\infty} [\Pr\{n \text{ completions in } \le \text{ t } \mid \text{arrival found n in system }\} \cdot \mathrm{P_n} + W_q(0) \\
&= (1-\rho) \sum_{n=1}^{\infty} \mathrm{p^n} \int_0^t \frac{\mu(\mu \mathrm{x})^{\mathrm{n-1}}}{\mathrm{(n-1)!}} \, \mathrm{e}^{\mu \mathrm{x}} dx + (1\text{-}\rho) \\
&= (1-\rho)\, \rho \int_0^t \mu \, \mathrm{e}^{-\mu \mathrm{x}} \sum_{n=1}^{\infty} \frac{\mu(\mu \mathrm{x})^{\mathrm{n-1}}}{\mathrm{(n-1)!}} dx + (1\text{-}\rho) \\
&= (1-\rho)\, \rho \int_0^t \mu \, \mathrm{e}^{-\mu \mathrm{x}(1\text{-}\rho)} \mathrm{dx} + (1\text{-}\rho) \\
&= 1\text{-}\rho \mathrm{e}^{-\mu \mathrm{x}(1\text{-}\rho)\mathrm{t}} \qquad\qquad (\mathrm{t}>0)
\end{aligned}
$$

So the distribution of waiting time in queue is

$$
W_q(t) = \begin{cases} 1-\rho & (\mathrm{t} = 0) \\ 1-\rho e^{-\mu(1-\rho)t} & (\mathrm{t} > 0) \end{cases}
$$

With the probability distribution of T_q , we can calculate the expected waiting time, which is denoted by W_q .

$$W_q = E[T_q] = \int_0^\infty t dW_q(t) \qquad \text{(Riemann-Stieltjes)}$$

$$=0(1-\frac{\lambda}{\mu}) + \int_0^\infty t\frac{\lambda}{\mu}(\mu-\lambda)\mathrm{e}^{-(\mu-\lambda)t}dt$$

$$=\frac{\lambda}{\mu}\int_0^\infty t(\mu-\lambda)\mathrm{e}^{-(\mu-\lambda)t}dt$$

$$=\frac{\lambda}{\mu(\mu-\lambda)}$$

In order to calculate σ_q^2 , we use definition $\sigma_q{}^2 = E[T_q^2] - (E[T_q])^2$

So we first need to know $E[T_q^2]$.

$$E[T_q{}^2] = \int_0^\infty t^2 dW_q(t)$$

$$= \int_0^\infty t^2\frac{\lambda}{\mu}(\mu-\lambda)\mathrm{e}^{-(\mu-\lambda)t}dt$$

$$=\frac{\lambda}{\mu}\int_0^\infty t^2(\mu-\lambda)\mathrm{e}^{-(\mu-\lambda)t}dt$$

In order to calculate above integration, we first look at integration $\int_0^\infty t^2\lambda\mathrm{e}^{-\lambda t}dt$.

Using integration by parts: $\int$ u(x)dx=u(x)v(x)-$\int$ v(x)du(x)

We can obtain:

$$\int_0^\infty t^2\lambda\mathrm{e}^{-\lambda t}dt = \int_0^\infty t^2 d(-\mathrm{e}^{\lambda t})$$

$$=-\mathrm{t}^2\cdot e^{-\lambda t}\Big|_0^\infty + \int_0^\infty \mathrm{e}^{-\lambda t}dt^2$$

$$=\frac{2}{\lambda^2}$$

$$E[T^2] = \frac{\lambda}{\mu}\int_0^\infty t^2(\mu-\lambda)e^{-(\mu-\lambda)t}dt$$

$$= \frac{\lambda}{\mu}\cdot\frac{2}{(\mu-\lambda)^2} = \frac{2\lambda}{\mu(\mu-\lambda)^2}$$

Therefore,

$$\sigma_q^2 = E[T_q^{\,2}] - (E[T_q])^2$$

$$= \frac{\lambda}{\mu}\cdot\frac{2}{(\mu-\lambda)^2} - \frac{\lambda^2}{\mu^2(\mu-\lambda)}$$

$$= \frac{\rho(2-\rho)}{\mu^2(1-\rho)^2}$$

We obtain:

$$c_q = \frac{\sigma_q}{t_q} = \sqrt{\frac{\rho(2-\rho)}{\mu^2(1-\rho^2)}}\Bigg/\frac{\rho}{(1-\rho)\mu}$$

$$= \sqrt{\frac{2-\rho}{\rho}}$$

In the previous section, we defined $\mathrm{W_q}(t)$ the probability of a customer waiting a time less than or equal to t for service. $W_q(t) = P\{T_q \le t\}$.

$$W_q(t) = \begin{cases} 1-\rho & (\mathrm{t} = 0) \\ 1-\rho e^{-\mu(1-\rho)t} & (\mathrm{t} > 0) \end{cases}$$

We know $P\{T_q > t\} = 1 - P\{T_q \le t\}$. So $P\{T_q > t\}$ is the probability of a customer waiting a time greater than t for service.

Hence,

$$P\{T_q > t\} = 1 - W_q(t) = \begin{cases} \rho & (\mathrm{t} = 0) \\ \rho e^{-\mu(1-\rho)t} & (\mathrm{t} > 0) \end{cases}$$

The probability of a customer waiting $P(T_q > 0) = \rho$

Therefore, we have

$$c_q = \sqrt{\frac{2 - P(T_q > 0)}{P(T_q > 0)}}$$
$$\Rightarrow \sigma_q = \sqrt{\frac{2 - P(T_q > 0)}{P(T_q > 0)}} \cdot W_q$$
$$\Rightarrow \sigma_q^2 = \frac{2 - P(T_q > 0)}{P(T_q > 0)} \cdot W_q.$$

Exact coefficient of variation of the waiting time for $M / M / s$ queue

We now derive the exact coefficient of variation c_q of waiting time for $M / M / s$ queue.

For $M / M / s$, we first still parallel Gross and Harris (1985). From the general birth-death model, we have,

$$P_{n+1} = \frac{\lambda_n + \mu_n}{\mu_{n+1}} P_n - \frac{\lambda_{n-1}}{\mu_{n+1}} P_{n-1} \qquad (\mathrm{n} \geq 1)$$
$$\mathrm{P}_1 = \frac{\lambda_0}{\mu_1} P_0$$
$$\Rightarrow P_n = \frac{\lambda_{n-1}\lambda_{n-2}\ldots\lambda_0}{\mu_n\mu_{n-1}\ldots\mu_1} P_0 \ = \ P_0 \prod_{i=1}^{n} \frac{\lambda_{i-1}}{\mu_\mathrm{i}} \qquad (\mathrm{n} \geq 1).$$

For the multi-server model, since the input is Poisson and the service exponential, we have a birth-death process. Hence, $\lambda_n = \lambda$ for all n and

$$\mu_n = \begin{cases} n\mu & (1 \leq \mathrm{n} \leq \mathrm{s}) \\ s\mu & (\mathrm{n} \geq \mathrm{s}) \end{cases}$$

$$\Rightarrow P_n = \begin{cases} \dfrac{\lambda^n}{n!\,\mu^n} P_0 & (1 \leq \mathrm{n} \leq \mathrm{s}) \\ \dfrac{\lambda^n}{s^{n-s} s!\,\mu^n} P_0 & (\mathrm{n} \geq \mathrm{s}) \end{cases}$$

To determine P_0 , use the boundary condition, $\sum_{n=0}^{\infty} P_n = 1$.

This gives,

$$P_0\left[\sum_{n=0}^{s-1}\frac{\lambda^n}{n!\mu^n}+\sum_{n=s}^{\infty}\frac{\lambda^n}{s^{n-s}s!\mu^n}\right]=1.$$

Define : $\quad \mathrm{r}=\lambda/\mu, \quad \rho=r/s=\lambda/s\mu$

$$\Rightarrow P_0\left[\sum_{n=0}^{s-1}\frac{r^n}{n!}+\sum_{n=s}^{\infty}\frac{r^n}{s^{n-s}s!}\right]=1.$$

Consider series,

$$\sum_{n=s}^{\infty}\frac{r^n}{s^{n-s}s!}=\frac{r^s}{s!}\sum_{n=s}^{\infty}\left(\frac{r}{s}\right)^{n-s}$$

$$=\frac{r^s}{s!}\sum_{m=0}^{\infty}\left(\frac{r}{s}\right)^m=\frac{r^s}{s!}\frac{1}{(1-r/s)} \quad . \qquad (\mathrm{r/s}=\rho<1)$$

Therefore , we can write,

$$P_0=\left[\sum_{n=0}^{s-1}\frac{r^n}{n!}+\frac{r^s}{s!}\frac{s}{(s-r)}\right]^{-1}$$

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

When $\mathrm{s}=1$,

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

This reduces to $M/M/1$ when s=1.

We consider measures of effectiveness for $M/M/s$ utilizing the steady-state probabilities in a manner similar to that used for $M/M/1$ model.

By definition,

$$L_q=\sum_{n=s}^{\infty}(n-s)P_n$$

$$=\sum_{n=s}^{\infty}(n-s)\frac{r^n}{s^{n-s}s!}P_0=\frac{r^s}{s!}P_0\sum_{m=1}^{\infty}m\rho^m$$

$$\Rightarrow L_q=\left[\frac{(r^{s+1}/s)}{s!(1-r/s)^2}\right]P_0=\left[\frac{(\lambda/\mu)^s\lambda\mu}{(s-1)!(s\mu-\lambda)^2}\right]P_0.$$

Using Little's rule, we can also obtain,

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

$$W = \frac{1}{\mu} + [\frac{(\lambda/\mu)^s \mu}{(s-1)!(s\mu-\lambda)^2}]P_0,$$

$$L = \lambda W = \frac{\lambda}{\mu} + [\frac{(\lambda/\mu)^s \lambda\mu}{(s-1)!(s\mu-\lambda)^2}]P_0.$$

When s=1, they all reduce to $M/M/1$.

For our interest, we want to know $W_q(0)$ *and* $W_q(t)$ .We proceed in a manner similar to that of $M/M/1$.

Let T_q represent the random variable "time spent waiting in queue".

$W_q(t)$: The distribution of waiting time in queue.

Hence,

$$W_q(0) = \Pr\{s-1 \text{ or less in the system}\}$$

$$= \sum_{n=0}^{s-1} P_n$$

$$= P_0 \sum_{n=0}^{s-1} \frac{r^n}{n!}$$

$$\because \sum_{n=0}^{s-1} \frac{r^n}{n!} = \frac{1}{P_0} - \frac{sr^s}{s!(s-r)}$$

$$\Rightarrow W_q(0) = P_0[\frac{1}{P_0} - \frac{sr^s}{s!(s-r)}]$$

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

For $T_q > 0$, FIFO is assumed, hence,

$$W_q(t) = \Pr\{T_q \le t\}$$

$$= \sum_{n=s}^{\infty} [\Pr\{\text{ n-s } +1 \text{ completions in } \le t \mid \text{arrival found n in system }\} \cdot P_n] + W_q(0). \quad (t>0)$$

When $n \geq s$, the system service rate is Poisson with mean $s\mu$, so that the time between successive completions is exponential with mean $1/(s\mu)$ and

$$W_q(t) = P_0 \sum_{n=s}^{\infty} \frac{r^n}{s^{n-s} s!} \int_0^t \frac{\mu s(\mu s x)^{n-s}}{(n-s)!} e^{-\mu s x} dx + W_q(0) \qquad (\mathrm{t} > 0)$$

$$= P_0 \frac{r^s}{(s-1)!} \int_0^t \mu e^{-\mu s x} \sum_{n=s}^{\infty} \frac{(\mu r x)^{n-s}}{(n-s)!} dx + W_q(0)$$

$$= P_0 \frac{r^s}{(s-1)!} \int_0^t \mu e^{-\mu s x} dx + W_q(0)$$

$$= P_0 \frac{r^s}{(s-1)!} \int_0^t \mu e^{-\mu x(s-r)} dx + W_q(0)$$

$$= \frac{r^s (1 - e^{-\mu(s-r)t})}{(s-1)!(s-\lambda/\mu)} P_0 + W_q(0)$$

$$= \frac{(\lambda/\mu)^s (1 - e^{-(\mu s - \lambda)t})}{(s-1)!(s-\lambda/\mu)} P_0 + W_q(0). \qquad (\mathrm{t} > 0)$$

Hence the distribution of waiting time in queue is then

$$W_q(t) = \begin{cases} 1 - \dfrac{\mathrm{s}(\lambda/\mu)^{\mathrm{s}}}{s!(s-\lambda/\mu)} \mathrm{P}_0 & (\mathrm{t} = 0) \\[2ex] \dfrac{(\lambda/\mu)^s (1 - e^{-(\mu s - \lambda)t})}{(s-1)!(s-\lambda/\mu)} P_0 + W_q(0) & (\mathrm{t} > 0). \end{cases}$$

When s=1, it reduces to M/M/1.

Now we can derive $\sigma_q^{\ 2}$.By definition, $\sigma_q^{\ 2} = E[T_q^2] - (E[T_q])^2$.

By using Little's rule we already obtained,

$$E[T_q] = W_q = [\frac{(\lambda/\mu)^s \mu}{(s-1)!(s\mu - \lambda)^2}] P_0.$$

Here we derive the above formula in an alternative way.

$$E[T_q]=\int_0^\infty t dW_q(t)=\int_0^\infty P_0[\frac{(\lambda\mu)^s\mu}{(s-1)!(s-\lambda/\mu)}]e^{-(\mu s-\lambda)t}(\mu s-\lambda)tdt$$

$$=[\frac{(\lambda\mu)^s\mu}{(s-1)!(s-\lambda/\mu)}]P_0\int_0^\infty e^{-(\mu s-\lambda)t}t(\mu s-\lambda)dt$$

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

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

This is the same as derived by using Little's rule.

Similarly,

$$E[T_q^2]=\int_0^\infty t^2 dW_q(t)=\int_0^\infty P_0[\frac{(\lambda/\mu)^s}{(s-1)!(s-\lambda/\mu)}]e^{-(\mu s-\lambda)t}(\mu s-\lambda)t^2dt$$

$$=[\frac{(\lambda/\mu)^s}{(s-1)!(s-\lambda/\mu)}]P_0\int_0^\infty e^{-(\mu s-\lambda)t}\ (\mu s-\lambda)t^2dt$$

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

$$=[\frac{(\lambda/\mu)^s 2\mu}{(s-1)!(s\mu-\lambda)^3}]P_0.$$

Both $E(T_q)$ and $E(T_q^2)$ reduces to the $M/M/1$ when s=1 and $P_0=1-\rho$.

Now we can calculate σ_q^2 by definition:

$$\sigma_q^2=E[T_q^2]-(E[T_q])^2$$

$$=[\frac{(\lambda/\mu)^s 2\mu}{(s-1)!(s\mu-\lambda)^3}]P_0-[\frac{(\lambda/\mu)^s\mu}{(s-1)!(s\mu-\lambda)^2}]^2P_0{}^2.$$

We can verify this as follows, when s=1, $P_0=1-\rho$ it reduces to M/M/1.

$$\sigma_q^2=E[T_q^2]-(E[T_q])^2=[\frac{(\lambda/\mu)^s 2\mu}{(s-1)!(s\mu-\lambda)^3}]P_0-[\frac{(\lambda/\mu)^s\mu}{(s-1)!(s\mu-\lambda)^2}]^2P_0^2$$

$$=[\frac{(\lambda/\mu)^1 2\mu}{(1-1)!(s\mu-\lambda)^3}](1-\rho)-[\frac{(\lambda/\mu)^2\mu^2}{(1-1)!(\mu-\lambda)^4}](1-\rho)^2$$

$$=[\frac{2\lambda}{(\mu-\lambda)^3}](1-\frac{\lambda}{\mu})-[\frac{\lambda^2}{(\mu-\lambda)^4}](1-\rho)^2$$

$$=\frac{2\mu\lambda-\lambda^2}{\mu^4(1-\lambda/\mu)^2}=\frac{2\mu\rho\mu-(\rho\mu)^2}{\mu^4(1-\rho)^2}$$

$$=\frac{2\rho-\rho^2}{\mu^2(1-\rho)^2}.$$

Which is the same as $M/M/1$ when s=1.

Similar to the $M/M/1$, we know $P\{T_q > t\} = 1 - P\{T_q \le t\}$

Since $P\{T_q > t\}$ is the probability of a customer waiting a time greater than t for service, we have,

$$P\{T_q > t\}=1-W_q(t)=\begin{cases}\dfrac{s(\lambda/\mu)^s}{s!(s-\lambda/\mu)}P_0 & (t=0)\\ 1-[\dfrac{(\lambda/\mu)^s(1-e^{-(\mu s-\lambda)t})}{(s-1)!(s-\lambda/\mu)}P_0+W_q(0)] & (t>0)\end{cases}$$

The probability that a customer has to wait is,

$$P(T_q > 0)=\frac{s(\lambda/\mu)^s}{s!(s-\lambda/\mu)}P_0 \quad .$$

By combining above formula of $P(T_q > 0)$, and $W_q = [\frac{(\lambda/\mu)^s\mu}{(s-1)!(s\mu-\lambda)^2}]P_0$, we can obtain,

$$P(T_q > 0) = (s\mu-\lambda)W_q \quad . \tag{4.1}$$

Now we can calculate c_q .By definition, $c_q^2 = \frac{\sigma_q^{\ 2}}{W_q^2}$.

We have already derived,

$$\sigma_q^2 = E[T_q^2]-(E[T_q])^2=[\frac{(\lambda/\mu)^s 2\mu}{(s-1)!(s\mu-\lambda)^3}]P_0-[\frac{(\lambda/\mu)^s\mu}{(s-1)!(s\mu-\lambda)^2}]^2 P_0^{\ 2}.$$

Also we know

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

$$W_q = [\frac{(\lambda / \mu)^s \mu}{(s-1)!(s\mu - \lambda)^2}] P_0 .$$

Hence,

$$\sigma_q^2 = [\frac{2W_q}{s\mu - \lambda} - W_q{}^2] .$$

Therefore, by definition,

$$c_q^2 = \frac{\sigma_q^2}{W_q^2} = \frac{2W_q/(s\mu - \lambda) - W_q{}^2}{W_q^2}$$

$$= \frac{2}{(\mu s - \lambda) W_q} - 1$$

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

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

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

$$= \frac{2}{P(T_q > 0)} - 1$$

$$= \frac{2 - P(T_q > 0)}{P(T_q > 0)}.$$

So we obtain,

$$\frac{\sigma_q^2}{W_q^2} = \frac{2 - P(T_q > 0)}{P(T_q > 0)}$$

$$\Rightarrow \sigma_q^2 = \frac{2 - P(T_q > 0)}{P(T_q > 0)} \cdot W_q^2$$

$$\Rightarrow \sigma_q = \sqrt{\frac{2 - P(T_q > 0)}{P(T_q > 0)}} \cdot W_q$$

$$\Rightarrow c_q = \sqrt{\frac{2 - P(T_q > 0)}{P(T_q > 0)}}.$$

(2). Exact coefficient of variation of waiting time for $G/M/1$ and $G/M/s$ queues

In this section, we investigate some important queuing models that cannot be studied in the framework of the birth-death process. They are imbedded Markov chain queuing models: $G/M/1$ and $G/M/s$.

In the previous analysis, we concentrated mainly on queues with Poisson input and exponential service times. These assumptions imply that the future evolution of the system from some time t depends only on the state of the system at time t, and is independent of the history of the system prior to time t. In these models, the state of the system could always be specified in terms of the number of customers present.

Now we analyze queues for which knowledge of the number of customers present at any time t is not sufficient information to permit complete analysis of the model. For example, consider the case in which the service times are assumed exponential, but the customers' arrival epochs are separated by a constant time interval. Then the future evolution of the system from time t would dependent not only on the number of customers present at time t, but also on the elapsed time since the last customer arrival epoch (because the arrival epoch of the next customer is strictly determined by the arrival epoch of the last customer).

Clearly, we need a new method of analysis. A powerful method for the analysis of certain queuing models, such as the method in the above example, is that of the imbedded Markov chain, introduced by Kendall (1951). This is a brilliant expository research, in which for the first time Kendall hinted at his concept of the Imbedded Markov chain, subsequently developed by him and other researchers.

As with the birth-and-death process, there is a vast theory of Markov chains. We will aim at as direct an approach to the analysis of our queuing models as possible, without extended excursions into the surrounding theory structure. Thus, we introduce the main ideas behind the theory of the imbedded Markov chain, and show how these ideas facilitate the analysis of certain important queuing models.

That is, we will study queuing models in which the input process and service time distribution function are such that the imbedded Markov chain analysis is applicable.

Consider the following input process: customers arrives at epoch $T_1, T_2, ...T_k, ...$. The inter-arrival times $T_{k+1} - T_k (k = 0,1,...; T_0 = 0)$ are identically distributed, mutually independent, positive random variables with distribution function

$$G(x) = P\{T_{k+1} - T_k \le x\} .$$

Independent of the index k, the input process is then said to be recurrent. Queues with recurrent input can sometimes be studied by the imbedded Markov chain analysis.

In short, for the $G/M/1$, $G/M/s$ queuing models studied in this section, a Chapman-Kolmogorov analysis is not possible, since we no longer have a Markov process because of the relaxation of the exponential assumption on the inter-arrival times and/or service times. However, for many of the models considered here, while we no longer have a Markov process, there is nevertheless, imbedded within this non-Markov stochastic process a Markov chain (i.e. imbedded Markov chain). For these types of models, we can employ some of the theory of Markov chains.

We assume that service times are exponential and no assumption is made concerning the arrival pattern other than that successive inter-arrival times are independent and identical distributed. For these cases, results can be obtained for s parallel servers using an analysis similar to that for the $s = 1$ case with a slight increase in complexity in certain probability calculations. So we first consider $s = 1$ and then generalize to s servers.

We use the Imbedded Markov Chain approach.

$G/M/1$ Queue

For $G/M/1$, by using Laplace transforms (Kleinrock 1975), we have

$W_q(t) = 1 - re^{-\mu(1-r)t} \qquad (t \ge 0)$.

By definition (Kleinrock 1975),

r =E [number of times state E_{k+1} is reached between two successive visits to state E_k]

The conditional PDF for $G/M/1$ queue waiting time is the exponential distribution. (Kleinrock 1975, Gross & Harris 2002).

Compared with $M/M/1$, they have exactly the same form (replace ρ with r).

By straight forward calculation, we also have that the mean waiting time in $G/M/1$ is

$$W_q = \frac{r}{\mu(1-r)} \quad .$$

Here we need to know r (0< r <1).

For $M/M/1$, r reduces to ρ, which the probability of a customer is waiting $P(T_q > 0)$.

From $W_q(t) = 1 - re^{-\mu(1-r)t} \qquad (t \geq 0)$,

we have $P(T_q > 0) = 1-(1-r) = r$, which has the same form as $P(T_q > 0) = 1-(1-\rho) = \rho$ for $M/M/1$.

Hence similar to $M/M/1$, for $G/M/1$, we have

$$c_q = \sqrt{\frac{2-P(T_q > 0)}{P(T_q > 0)}}$$

$$\Rightarrow \sigma_q = \sqrt{\frac{2-P(T_q > 0)}{P(T_q > 0)}} \cdot W_q$$

$$\Rightarrow \sigma_q^2 = \frac{2-P(T_q > 0)}{P(T_q > 0)} \cdot W_q$$

$G/M/s$ Queue

For $G/M/s$, $W_q(t) = 1 - \frac{Cr^s}{1-r} e^{-s\mu(1-r)t} \quad (t \geq 0)$ where C is a constant (Gross & Harris 1985)

$$W_q = E[T_q] = \int_0^\infty t dW_q(t)$$
$$= \int_0^\infty t \frac{Cr^s}{1-r} e^{-s\mu(1-r)t} s\mu(1-r)dt$$
$$= \frac{Cr^s}{1-r} \int_0^\infty te^{-s\mu(1-r)t} s\mu(1-r)dt$$
$$= \frac{Cr^s}{1-r} \frac{1}{s\mu\left(1-r\right)}$$
$$= \frac{Cr^s}{s\mu\left(1-r\right)^2}.$$

In order to calculate variation: $\sigma_q^2 = E[T_q{}^2] - (E[T_q])^2$.

First we need to know $E[T_q{}^2]$.

$$E[T_q{}^2] = \int_0^\infty t^2 \frac{Cr^s}{1-r} e^{-s\mu(1-r)t} s\mu(1-r)dt$$
$$= \frac{Cr^s}{1-r} \int_0^\infty t^2 e^{-s\mu(1-r)t} s\mu(1-r)dt$$
$$= \frac{Cr^s}{1-r} \frac{2}{[s\mu(1-r)]^2}$$
$$= \frac{2Cr^s}{s^2\mu^2(1-r)^3}.$$

Hence by definition,

$$\sigma_q^2 = E[T_q{}^2] - (E[T_q])^2$$
$$= \frac{2Cr^s}{s^2\mu^2(1-r)^3} - [\frac{Cr^s}{s\mu(1-r)^2}]^2$$
$$= \frac{2Cr^s}{s^2\mu^2(1-r)^3} - \frac{C^2r^{2s}}{s^2\mu^2(1-r)^4}$$
$$= \frac{2Cr^s(1-r) - C^2r^{2s}}{s^2\mu^2(1-r)^4}.$$

For $G/M/s$ we want to verify that

$$c_q = \sqrt{\frac{2 - P(T_q > 0)}{P(T_q > 0)}}$$

$$\Rightarrow \sigma_q = \sqrt{\frac{2 - P(T_q > 0)}{P(T_q > 0)}} \cdot W_q$$

$$\Rightarrow \sigma_q^2 = \frac{2 - P(T_q > 0)}{P(T_q > 0)} \cdot W_q$$

i.e. We want to verify $\dfrac{\sigma_q^2}{W_q^{\ 2}} = \dfrac{2 - P(T_q > 0)}{P(T_q > 0)}$.

$$\text{LHS}=\frac{\sigma_q^2}{W_q^{\ 2}} = \frac{2Cr^n(1-r) - C^2 r^{2n}}{n^2\mu^2(1-r)^4} \Bigg/ \frac{Cr^n}{n\mu(1-r)^2}$$

$$= \frac{2}{Cr^n/(1-r)} - 1 = \frac{2}{P(T_q > 0)} - 1.$$

$$\text{RHS}=\frac{2 - P(T_q > 0)}{P(T_q > 0)} = \frac{2}{P(T_q > 0)} - 1.$$

Therefore, LHS=RHS, $\dfrac{\sigma_q^{\ 2}}{W_q^{\ 2}} = \dfrac{2 - P(T_q > 0)}{P(T_q > 0)}$

$$\Rightarrow \sigma_q^{\ 2} = \frac{2 - P(T_q > 0)}{P(T_q > 0)} \cdot W_q^{\ 2}$$

$$\Rightarrow \sigma_q = \sqrt{\frac{2 - P(T_q > 0)}{P(T_q > 0)}} \cdot W_q$$

$$\Rightarrow c_q = \sqrt{\frac{2 - P(T_q > 0)}{P(T_q > 0)}}.$$

The above relationship does not depend at all on the interarrival time distribution or the number of servers s. This implies that for $G/M/s$ queues, all of the needed information about the interarrival time distribution and the number of servers is contained in the probability of waiting $P(T_q > 0)$.

(3) Exact coefficient of variation of waiting time for $M/G/1$ queue

For $M/G/1$, we can also use the imbedded Markov chain approach. The imbedded process is Markovian. This allows the utilization of Markov chain theory in the analysis of the $M/G/1$ model (Gross & Harris 2002, Kleinrock 1975).

For the expected number of customers in system, we have Pollaczek-Klintchine formula

$$L = \rho + \frac{\rho^2 + \lambda^2 \sigma_s^2}{2(1-\rho)}.$$

By Little's rule, we can obtain $L_q, W\ and\ W_q$ easily.

For waiting time PDF, we have (Gross &Harris 1985)

$$W_q(t) = (1-\rho)\sum_{n=0}^{\infty} \rho^n [R^{(n)}(t)].$$

We know the variance of waiting time is $\sigma_q^{\ 2} = W_q^{\ 2} + \frac{\lambda E[s^3]}{3(1-\rho)}$ and average waiting time is

$W_q = \frac{\lambda E[s^2]}{2(1-\rho)}$ (Kleinrock 1976), where $E[s^2],\ E[s^3]$ are the second and the third moment of service time distribution. For the $M/G/1$ queue, we know $P(T_q > 0) = \rho$ and $P(T_q = 0) = 1 - P(T_q > 0)$. Therefore,

$$c_q = \frac{\sigma_q}{W_q} = \sqrt{1 + \frac{\lambda E[s^3]}{3(1-\rho)W_q^{\ 2}}}$$

$$= \sqrt{1 + \frac{E[s^3]}{3\lambda}\frac{4(1-\rho)}{(E[s^2])^2}}$$

$$= \sqrt{1 + \frac{E[s^3]}{3\lambda}\frac{4(1-P(T_q>0))}{(E[s^2])^2}}$$

$$= \sqrt{1 + \frac{4E[s^3]}{3\lambda}\frac{P(T_q=0))}{(E[s^2])^2}}.$$

The above formula can be used if there are ways of estimating $P(T_q > 0)$ and the second and third moment of the distribution of the service times $E[s^2],\ E[s^3]$ or the third moment can be computed when the first and second moments are known.

When $G = M$ it reduces to $M / M / 1$, where the service time is exponentially distributed, $f(x,\mu) = \mu e^{-\mu x}$

We know that $E[x^n] = M_x^{(n)}(0) = \frac{d^n M_x(t)}{dt^n}\Big|_{t=0}$, so we use Moment Generating Function $M_x(t)$ to calculate $E[s^2]$ *and* $E[s^3]$ .For the exponential distribution, $M_x(t) = (1 - \frac{t}{\mu})^{-1}$,

$$E[s^2] = M_s''(t)\Big|_{t=0} = \frac{2}{\mu^2}$$

$$E[s^3] = M_s'''(t)\Big|_{t=0} = \frac{6}{\mu^3}.$$

Hence

$$c_q = \sqrt{1 + \frac{E[s^3]}{3} \frac{4(1-\rho)}{\lambda (E[s^2])^2}}$$

$$= \sqrt{1 + \frac{2(1-\rho)}{\rho}} = \sqrt{\frac{2-\rho}{\rho}}$$

$$= \sqrt{\frac{2 - P(T_q > 0)}{P(T_q > 0)}}.$$

It has the same expression as $G / M / s$. So this formula is a generalization of the formula for $G / M / s$ and we have a more general form that applies to $G / M / s$ and $M / G / 1$.

$M / G / s$ Queue

$M / G / s$ queue does not possess the imbedded Markov chain property (Gross and Harris 1985).But $M / M / s$ and $M / D / s$ queues are Markovian (Saaty 1961). Whitt (1993) conjectured that we can use the exact formula for an $M / G / 1$ approximation for the $M / G / s$ model. The idea is that the conditional delay should depend more on the service time distribution than the interarrival time distribution. Seelan and Tijms (1984) provided additional support for this approximation. This supports our conjectures that for $M / G / s$ our results still provide a good approximation for the coefficient of variance.

In summary, for Markov queues $M/M/s$ and imbedded Markov queues $G/M/s$, $M/G/1$ the exact formula for the coefficient of variation of waiting is

$$c_q = \sqrt{1+\frac{E[s^3]}{3\lambda}\frac{4P(T_q=0)}{(E[s^2])^2}}. \qquad (4.2)$$

The above formula can be used if there are ways of estimating $P(T_q > 0)$ and the third moment of the distribution of the service times. The later can be accomplished if we make an assumption that the service time distribution can be approximated if we could have assumed any distribution for which the third moment can be computed as a function of the first and second. We conjecture that the same assumption is justified in an approximation method for general $G/G/s$ queues.

4.2 Exact Formulas for the probability of waiting for $M/M/s$, $G/M/s$, $M/G/1$ and $E_k/M/1$ queues

In order to estimate the variance of waiting time, we know from formula (4.2) that the key point is to calculate the probability of waiting if $E[s^2]$ and $E[s^3]$ are known. In this section, we derive the exact formula for $M/M/s$, $M/G/1$ and $E_k/M/1$ queues.

(1) $M/M/s$ Queue

From the previous discussion, we know the probability of waiting for $M/M/s$ is

$$P(T_q > 0) = (s\mu - \lambda)W_q\ . \qquad (4.1)$$

However, this relation holds only for the $M/M/s$ queue. We disprove it for different queues as follows.

For the $G/M/1$ queue,

we know from Kleinrock $W_q(t) = 1 - re^{-\mu(1-r)t}$ $\quad (t > 0)$.

So $P(T_q > 0) = r$ and $W_q = \dfrac{r}{\mu(1-r)}$, only when $r = \rho = \lambda/\mu$ (1) holds.

For $G/M/s$ queue, $W_q(t) = 1 - \dfrac{Cr^s}{1-r}e^{-\mu s(1-r)t}$. $\quad (t > 0)$

So $P(T_q > 0) = \frac{Cr^s}{1-r}$ and $W_q = \frac{Cr^s}{s\mu(1-r)^2}$. Only when $r = \rho = \lambda/(s\mu)$ (4.1) holds.

For $M/G/1$ queue, we have $W_q(t) = (1-\rho)\sum_{n=0}^{\infty}\rho^n[R^n(t)]$, $P_n = (1-\rho)\rho^n$.

So $P(T_q > 0) = \rho = \lambda/\mu$.

From Pollacczek-Klintchhine formula, we know

$L = \rho + \frac{\rho^2 + \lambda^2\sigma_s^2}{2(1-\rho)}$, where σ_s^2 is the variance of service time.

Using Little's rule $L_q = L - \frac{\lambda}{\mu} = \rho + \frac{\rho^2 + \lambda^2\sigma_s^2}{2(1-\rho)} - \rho = \frac{\rho^2 + \lambda^2\sigma_s^2}{2(1-\rho)}$.

$W_q = \frac{L_q}{\lambda} = \frac{\rho^2 + \lambda^2\sigma_s^2}{2\lambda(1-\rho)}$.

It is not hard to see $P(T_q > 0) \neq (s\mu - \lambda)W_q$. For instance, for M/D/1 queue,

$P(T_q > 0) = \rho = \frac{\lambda}{\mu} \neq (\mu - \lambda)\frac{\rho^2}{2\lambda(1-\rho)}$.

except for the case when $\lambda = \mu$.

Also we know for the M/D/1 queue, $W_q = \frac{\lambda}{2(\mu - \lambda)}$.

For $M/E_k/1$ queue, $W_q = \frac{k+1}{2k}\cdot\frac{\lambda}{\mu(\mu-\lambda)}$.

Since $\frac{\lambda}{\mu} \neq (\mu-\lambda)\cdot\frac{k+1}{2k}\cdot\frac{\lambda}{\mu(\mu-\lambda)}$ except for k=1, which is M/M/1 queue.

We can also see $P(T_q > 0) \neq (s\mu - \lambda)W_q$. Therefore, we have to explore the formula of the probability of waiting for $G/G/s$ queues.

(2) $M/G/1$ Queue

For the $M/G/1$ queue, we already see from the above that $W_q(t) = (1-\rho)\sum_{n=0}^{\infty}\rho^n[R^n(t)]$, $P_n = (1-\rho)\rho^n$.

So we have $P(T_q > 0) = \rho = \lambda/\mu$.

(3) $E_k / M / 1$ Queue

For $E_k / M / 1$**,** we can also calculate the exact result for the probability of waiting.

We assume that the interarrival times are Erlang type k distributed, with a mean of $1/\lambda$. We can look therefore at an arrival having passed through k phases, each with a mean time of $1/(k\lambda)$, prior to actually entering the system.

For the Erlang distribution, $f(x,k,\lambda) = \dfrac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!}$, $\quad$ *for* $x > 0$.

Mean $E(x) = k/\lambda$ and variance $\sigma^2 = k/\lambda^2$, so $c_a = \dfrac{\sigma_a}{E(x)} = \dfrac{\sqrt{k}}{\lambda} \Big/ \dfrac{k}{\lambda} = \dfrac{1}{\sqrt{k}}$.

Therefore, we can compute $k = 1/c_a^{\;2}$ to get Erlang (actually gamma) distribution parameter k.

The probability of no wait for service upon arrival is given by $q_0 = 1 - r^k$ (Gross and Harris (2002)), so the probability of waiting is $1 - q_0 = r^k$. r is the root of the characteristic equation

$[\mu D^{K+1} - (\lambda + \mu)D + \lambda]p_n = 0 \quad (n \geq 0)$

There is one and only one root in (0, 1) and $p_n = (1-r)r^n \;\; (n \geq 0, 0 < r < 1)$.

To obtain r , repeating the above characteristic equation with λ replaced by $k\lambda$, we have

$$\mu r^{k+1} - (k\lambda + \mu)r + k\lambda = 0 .$$

Given λ, μ, k , we can use Newton's method with initial value of $(\lambda/\mu)^{\frac{1}{k}}$ to obtain a unique $r\,(0 < r < 1)$ so that we can calculate the probability of waiting $P(T_q > 0) = r^k$.

Chapter 5
Heuristic Approximation Methods for the $G/G/s$ queue

5.1 Basic assumptions for the $G/G/s$ queue

There is no question that in principle, both from a scientific and an aesthetic viewpoint the most desirable way of resolving problems arising from any queue process is to formulate a precise math model and derive solutions by mathematical analysis.

However, such an ideal approach---the traditional analytical procedure is not usually possible. Without the memory-less property of the exponential to facilitate analysis, we can not compute exact performance measures for the $G/G/s$ queue. When it comes to exact solutions of multi-server queuing systems, the more one departs from the assumption of exponential, the thornier the problem becomes, especially if this happens for the service time. Due to its inherent complexity, analysis of the $G/G/s$ queue in general is difficult.

In this research, we consider the standard steady state $G/G/s$ queuing system with homogeneous servers in parallel, unlimited waiting room, the first come first served discipline and independent sequence of independent and identically distributed (i.i.d) interarrival times and service times. We assume that the general interarrival time and service time distributions are each partially specified by their first two moments. Equivalently, we assume that the arrival process is partially specified by the arrival rate λ (the mean interarrival time is $1/\lambda$). Similarly, we assume that the service-time distribution is partially specified by its process rate μ (the mean process time is $1/\mu$). All descriptions of these models thus depend only on the basic 5 parameter $s, \lambda, \mu, c_a, and\ c_s$. To apply the approximations, the above 5 queue specifications are assumed to be known.

Each customer arrives according to an arrival process and is served once at each queue, with the order of the queues being the same for all customers. Each queue has unlimited waiting space, the FIFO discipline, and i.i.d service times that are independent of the other random quantities in the model. The problem is to determine, for a given fixed external arrival process, the standard deviation of waiting time in system per customer. More generally, the object is to determine if variability, utilization, and server numbers actually matters.

Given that the external arrival process (i.e., the interarrival times are i.i.d.), which we also assumes here, the model is specified by the distribution of the service times at each queue and the distribution of external inter-arrival times. This problem is difficult for general distributions because exact expressions for the expected steady-state waiting time typically are unavailable; primarily because the arrival processes to all queues after the first typically are not renewal processes.

With this approximation procedure, each distribution is partially characterized by its first two moments, or equivalently, by its mean and squared coefficient of variation. The closed-form formulas give an approximate squared coefficient for the arrival process to each queue and an approximate expected steady-state waiting time. The expressed steady-state waiting time for queues in series actually depends on the distributions beyond their first two moments, but experience indicates that a fairly good approximation can often be obtained given this partial information (Kimura 1986).

A list of assumptions for the $G/G/s$ queue is as follows:

1. Identical s homogeneous servers. Infinite steady state queue.
2. Interarrival and service times are independent and identically distributed with general distributions.
3. One piece flow. No delays due to batching.
4. Utilization $\rho = (\lambda / s\mu)$ <1 so that the system is stable.
5. First come first served queue discipline

For the $G/G/s$ queue, W *and* W_q are not directly accessible. Except for the $G/M/s$ queue, where L is given by the well known Pollaczek-Khintchine formula, measures possess explicit general formulas in terms of known inputs to the queue. However, numerous highly accurate approximations for either the mean queue length or the mean waiting time (the latter two being related via Little's rule) have been developed for the $G/G/s$ queue, and intensively explored.

Since the accuracy of these approximations is commonly very high, we use these approximations of W *and* W_q in our models.

5.2 Average waiting time of the $G/G/s$ queue

As we commented previously, without the memory-less property of the exponential distribution to facilitate analysis, we can not compute exact performance measures for the $G/G/s$ queue. This does not mean that we should give up researching on the $G/G/s$ queue, only that we need to be concerned with finding good approximations. We can estimate the average waiting time of $G/G/s$ queue by means of "a two moment" approximation, which makes use of only the mean and standard deviation of the inter-arrival and process time distributions. Because it works well, this approximation is the basis of several commercially available queuing analysis packages (Hopp and Spearman 2001).

Much of the following development parallels Hopp and Spearman (2001). We proceed by introducing an expression for the waiting time in queue W_q and then computing the other performance measures. The approximation for W_q, which was first investigated by Kingman, is given by

$$W_q(G/G/1) = \left(\frac{c_a^{\,2} + c_s^{\,2}}{2}\right)\left(\frac{\rho}{1+\rho}\right)\frac{1}{\mu}$$

This approximation has several nice properties. First, it is exact for the $M/M/1$ queue. It also happens to be exact for the $G/G/1$ queue, although this is not evident from our discussion here. Finally, it neatly separates into three terms: a dimensionless variability term V, a utilization term U and a time term T, as $W_q(G/G/1) = VUT$. We refer to this as Kingman's equation or as the VUT equation. From it, we see that if the V factor is less than one, then the waiting time, and hence other congestion measures, for the $G/G/1$ queue will be smaller than those for the $M/M/1$ queue. Thus, the VUT equation shows that the $M/M/1$ case represents an intermediate case for a single server analogous to that represented by the worst case for waiting.

The VUT equation gives us a tool for analyzing a queue consisting of single server. However, in real-world systems, queuing systems often consists of multiple servers. The reason is that often more than a single server is required to achieve the desired workstation capacity. To analyze and understand the behavior of multi-server queues, we need a more general model.

To develop an approximation for this situation, note that for $G/G/1$, the approximation can be rewritten as

$$W_q(G/G/1) = \left(\frac{c_a^{\ 2} + c_s^{\ 2}}{2}\right) \cdot W_q(M/M/1). \qquad (5.1)$$

Where $W_q = \frac{\rho}{1-\rho} \cdot \frac{1}{\mu}$ is the waiting time in queue for the $M/M/1$ queue. This suggests the following approximation for the $G/G/s$ queue.

$$W_q(G/G/s) = \left(\frac{c_a^{\ 2} + c_s^{\ 2}}{2}\right) \cdot W_q(M/M/s). \qquad (5.2)$$

Sakasegawa (1977) presented the following closed-form expression for the mean waiting time in the $G/G/s$ queue:

$$W_q(G/G/s) = \left(\frac{c_a^{\ 2} + c_s^{\ 2}}{2}\right)\left(\frac{\rho^{\sqrt{2(s+1)}-1}}{s(1-\rho)}\right)\frac{1}{\mu}\ . \qquad (5.3)$$

Expression (5.3) is the multi-server version of the VUT equation. The V and T terms are identical to the single server version given in expression (5.1), but the U term is different.

Whitt (1983) discussed this formula in more detail. Although it may appear complicated, it does not require any type of iterative algorithm to solve and is therefore easily implemented in a spreadsheet program. This also makes it possible to couple the single-station approximation with the multiple-server to create a spreadsheet tool for analyzing the performance of a series of queues. This formula is used in our research when calculating mean waiting time for the $G/G/s$ queue.

5.3 Coefficient of variation of the deviation of waiting time for the $G/G/s$ queue

For the standard deviation of a general multi-server queue with infinite waiting capacity ($G/G/s$), we conjecture it has the properties of $G/M/s$ and $M/G/1$ queues as follows.

$$c_q = \sqrt{1 + \frac{4E[s^3]}{3\lambda} \frac{P(T_q = 0)}{(E[s^2])^2}} \qquad (5.4)$$

We conjecture that formula (5.4) can be used as an approximation for the $G/G/s$ queue since it applies to $G/M/s$ and $M/G/1$**,** and can be used as an approximation for $M/G/s$ queue. To estimate c_q using formula (1), it is necessary to estimate $P(T_q = 0)$ and $E[s^3]$. Since we do not assume that $E[s^3]$ is specified, we must estimate it by assuming some known distribution for the service times, e.g. gamma, for which the third moment can be computed as a function of the average and standard deviation. We also have to estimate $P(T_q = 0)$. Therefore, we need to do error sensitivity analysis to show that formula (5.4) is relatively insensitive to errors in estimating these inputs.

(1) Sensitivity to Errors in Estimating of the Input Parameters

We first examine the sensitivity of the formula (5.4) to errors in estimating $E[s^3]$, given that the other parameters $P(T_q > 0)$ and $E[s^2]$ are known. We find that the formula is relatively insensitive to the errors in estimating $E[s^3]$.

Theorem1: Suppose a small change in $E[s^3]$, expressed as a proportion P, is $\Delta E[s^3] = P \cdot E[s^3]$, the resulting change in c_q, also expressed as a proportion is at most $P/2$, $\frac{\Delta c_q}{c_q} \le 0.5P$.

Proof: The partial derivative of c_q with respective to $E[s^3]$ is:

$$\frac{\partial c_q}{\partial E[s^3]} = \frac{4P(T_q = 0)}{3\lambda (E[s^2])^2} \frac{1}{2c_q}.$$

Note that the derivative approaches 0 as $P(T_q = 0)$ approaches 0.

Also, we observe that $c_q \geq 1$, since $\frac{4E[s^3]}{3\lambda}\frac{P(T_q = 0)}{(E[s^2])^2} \geq 0$.

When $E[s^3]$ changes by a small amount $\Delta E[s^3] = P \cdot \Delta E[s^3]$, the corresponding change in c_q is

$$\Delta c_q = P \cdot E[s^3] \cdot \frac{\partial c_q}{\partial E[s^3]}$$
$$= P \cdot E[s^3] \cdot \frac{4P(T_q = 0)}{3\lambda(E[s^2])^2} \cdot \frac{1}{2c_q}$$
$$= P \cdot \frac{c_q^{\ 2} - 1}{2c_q}.$$

Expressing Δc_q as a proportion gives

$$\frac{\Delta c_q}{c_q} = P \cdot \frac{c_q^{\ 2} - 1}{2c_2^{\ 2}}.$$

Since $c_q \geq 1$, it can be observed that $0 \leq \frac{\Delta c_q}{c_q} \leq \frac{P}{2}$.

So the formula of c_q is not sensitive to $E[s^3]$ .We also observe that the above expression approaches 0 when $c_q \to 1$.

We can do the same sensitivity analysis on $P(T_q = 0)$ and draw the same conclusion that c_q is not sensitive to errors of $P(T_q = 0)$.

Theorem 2: Suppose a small change in $P(T_q = 0)$, expressed as a proportion P, is $\Delta P(T_q = 0) = P \cdot P(T_q = 0)$, the resulting change in c_q, also expressed as a proportion is at most $P/2$, that is $\frac{\Delta c_q}{c_q} \leq 0.5P$.

Proof: similar to the proof of theorem (1), we can do the similar sensitivity analysis on $P(T_q = 0)$ and draw the same conclusion that c_q is not sensitive to errors of $P(T_q = 0)$.

Heuristically, we can analyze the sensitivity of the method to errors in estimating $P(T_q > 0)$.

Figure 5.1 below shows the coefficient of variation as a function $P(T_q > 0)$. To illustrate the effect of the service time distribution, we show curves for exponential, Erlang (with k =4) and deterministic service time distributions.

Figure 5.1 shows that for $P(T_q > 0) \geq 0.4$ and $0 \leq c_s \leq 1$, the curve is relatively flat, with the value of c_q ranging between 1 and 2. Over this same region the curve is relatively insensitive to changes in the distribution of the service times or to small changes in $P(T_q > 0)$.
(When $P(T_q > 0)$ is small the estimate becomes very sensitive to errors in estimating these parameters. However, when $P(T_q > 0)$ is small, W_q is also small, as is $\sigma_q = c_q \cdot W_q$. In this

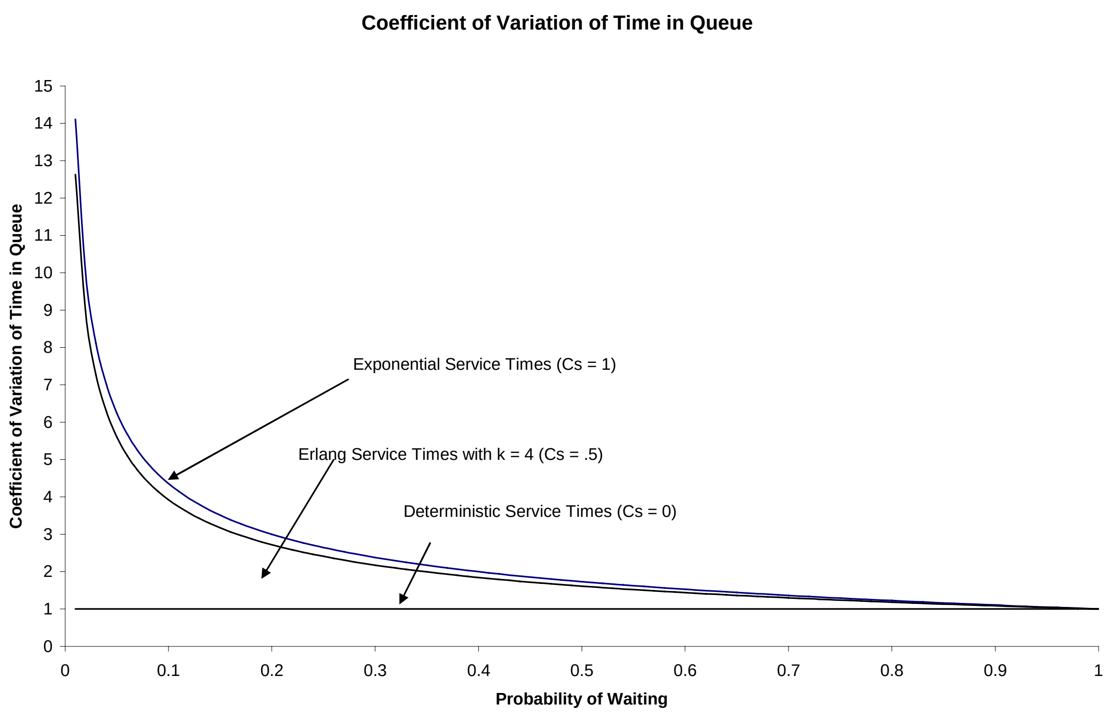


Figure 5.1Coefficient of variation of waiting time as a function of the probability of waiting.

case errors in estimating c_q would not have much impact on the estimated distribution of the total time in the system.)

The curve becomes steeper if $c_s > 1$. In our simulations we find the method does not necessarily give a good approximation in these cases.

In general, it can be shown that a given percentage error in estimating the probability of waiting will give a smaller percentage error in estimating the coefficient of variation of waiting time. For example, if the true probability of waiting is 0.8, the coefficient of variation is 1.22. If the estimate obtained for the probability of waiting is anywhere in the $\pm$ 10% range (0.72, 0.88), the resulting estimate of the coefficient of variation will be in range (1.13, 1.33) an error range of (-7.8%, 8.8%).

Hence, from equation (5.4) we can show that c_q has a fairly small range of variation (and should therefore be easy to estimate) when the probability of waiting is not small[1] (see footnote) and α is greater than 1.

(2) Implementation in spreadsheet for practical use

In order to implement the approximations in spreadsheet for practitioners, we analyze the specific case of $M/E_\alpha/1$ queue, where E_α is gamma distribution and α is the shape parameter in the gamma service time distribution. The resulting expression is

$$c_q = \sqrt{1 + \frac{4(1 - P(T_q > 0)(\alpha + 2)}{3P(T_q > 0)(\alpha + 1)}} \qquad (5.5)$$

When $\alpha = 1$, formula (5.5) reduces to $M/M/1$ queue. We show that the formula can provide a good approximation for $G/G/s$ queues using a gamma distribution approximation to the service time distribution with $\alpha = [E(service\ \ time)/\sigma(service\,time]^2$.

---

[1] When $P(T_q > 0)$ is small, W_q the average waiting time is also small as is $\sigma_q = c_q \cdot W_q$. In this case errors in Estimating c_q would not have much impact on the estimated distribution of the total time in the system.

When the service time distribution is gamma distribution, we know $E(s) = \alpha / \mu$,

and $\sigma_s^{\ 2} = \alpha / \mu^2$, so $c_s = \frac{\sigma_s}{E(s)} = \frac{\sqrt{\alpha}}{\lambda} \Big/ \frac{\alpha}{\lambda} = \frac{1}{\sqrt{\alpha}}$ and $\alpha = (E(s)/\sigma_s)^2$.

Proof: We know that $E[x^n] = M_x^{(n)}(0) = \frac{d^n M_x(t)}{dt^n}\big|_{t=0}$, so we use the Moment Generating Function $M_x(t)$ to calculate $E[s^2]$, $E[s^3]$. For the gamma distribution, we know the mean $1/\mu = \alpha\beta$, hence $\beta = 1/\alpha\mu$, $M_x(t) = (1-\beta t)^{-\alpha}$ for $t < 1/\beta$.

$$E[s] = M_s^{'}(t)\big|_{t=0} = \alpha\beta .$$
$$E[s^2] = M_s^{''}(t)\Big|_{t=0} = \alpha(\alpha+1)\beta^2 .$$
$$E[s^3] = M_s^{'''}(t)\Big|_{t=0} = \alpha(\alpha+1)(\alpha+2)\beta^3 .$$

Hence,

$$c_q = \sqrt{1 + \frac{E[s^3]4(1-\rho)}{3\lambda(E[s^2])^2}}$$
$$= \sqrt{1 + \frac{4(1-\rho)(\alpha+2)}{3P(T_q > 0)(\alpha+1)}}$$

For $M/G/1$ queue, we know $P(T_q > 0) = \rho$.

Hence, $c_q = \sqrt{1 + \frac{4(1-P(T_q > 0)(\alpha+2)}{3P(T_q > 0)(\alpha+1)}}$.

The sensitivity of the method to α

Numerical analysis shows that, when $\alpha \geq 1$ and $P(T_q > 0)$ is not small, c_q is not sensitive to α. In other words, α has little impact on c_q when $P(T_q > 0)$ is large. So we conjecture that the formula can provide a good approximation for $G/G/s$ queues using a gamma distribution approximation to the service time distribution with $\alpha = [E(service\ \ time)/\sigma(service\, time]^2$ under the assumption that the coefficient of variation of the service times is less than 1, given α values greater than 1.

Therefore, we conjecture that formula (5.5) works well for the $G/G/s$ queue. We don't know the accuracy of the approximations for $G/G/s$ queue. Since no closed-form analytical results are available for $G/G/s$ models, to evaluate the accuracy of the $G/G/s$ approximations, we resort to Monte Carlo simulation experiments using the Extend simulation program. We conduct simulation experiments to gain insight into the analog methods for calculating approximate steady-state performance measures of $G/G/s$ queuing system. We compare our results to simulation experiments and a few numerical results.

5.4 Interpolation methods to estimate the probability of waiting in the $G/G/s$ queue

In this section we analyze the probability of waiting for the $G/G/s$ queue. From 4.2, we know the probability of waiting formula $P(T_q > 0) = (s\mu - \lambda)W_q$ holds only for the $M/M/s$ queue. So we need to consider other methods to estimate the probability of waiting for the $G/G/1$ and $G/G/s$ queues. Since we already know exact results of the probability of waiting for some queues ($D/D/1, D/M/1$, $M/G/1$, $E_k/M/1$, $M/D/1, M/M/1$), we use an interpolation method to approximate the probability of waiting for $G/G/1$ queue.

Before using the interpolation method, we first approximate the $M/M/s$ queue with an $M/M/1$ queue having the same arrival rate and same probability of waiting $P(T_q > 0)$.The reason is that when we use the interpolation method, we only know exact results for single server queues ($D/D/1$, $M/G/1$, $E_k/M/1$).

Step1. We first approximate via $M/M/s$: use the formula for $M/M/s$ queue to get initial estimate Pw = $\text{Pw}_{\text{m/m/s}} = P(T_q > 0) = (s\mu - \lambda)W_q$ using only λ, μ , and s.

Step2. Find an approximating $M/M/1$ queue. Find the service rate μ' of the $M/M/1$ queue that has arrival rate λ and has $\text{Pw}_{\text{m/m/1}}$ equal to $\text{Pw}_{\text{m/m/s}}$:

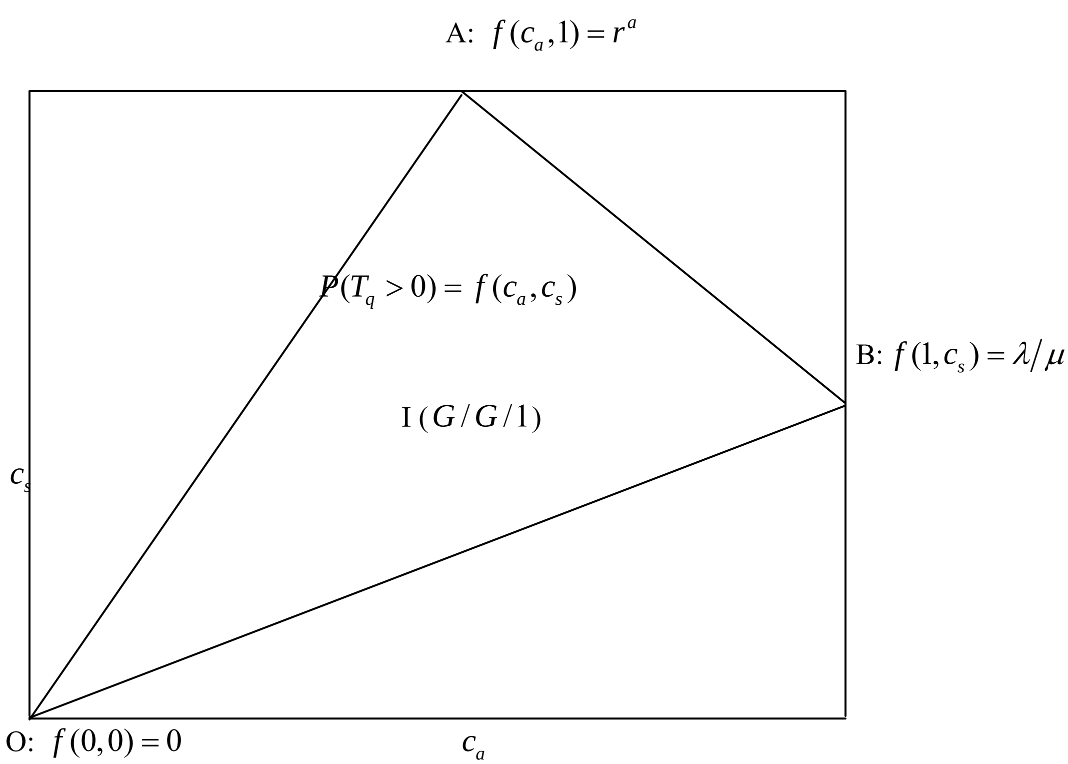
$$\mu' = \frac{\lambda}{(s\mu - \lambda)W_q}.$$

So in this way we can approximate the performance of multiple server queues as a single server queue.

Interpolation approximation methods were studied by Boxma (1979), and Reiman and Simon (1984). They consider a queuing system with a Poisson arrival process and evaluate the light traffic derivatives and the heavy traffic limit. But the selection of the function is quite arbitrary and there is no systematic way of selecting the correct function to be used, except for some special cases. Another interpolation approximation was proposed for the average workload in the $G/G/1$ queue. This approximation works well, but the evaluation of the approximation parameters is not straightforward.

Our approach considers queues with general independent arrivals and service distributions and we give a very easy procedure to calculate the probability of waiting. We use a point based interpolation method to approximate the probability of waiting for the $G/G/1$ queue. The result is easily implemented in the spreadsheet.

Point based interpolation



We then approximate $P\ (T_q > 0)$ for the resulting $G/G/1$ queue. Given λ and μ′, we wish to estimate $P\ (T_q > 0)$ as a function of the coefficient of variation of the arrival time c_a and the coefficient of variation of the service time c_s : $P(T_q > 0) = f(c_a, c_s)$

We compute $P(T_q > 0)$ for three points surrounding (c_a, c_s).

$f(0,0) = 0$

$f(1, c_s) = \lambda/\mu^{'}$

$f(c_a, 1) = r^k$, where $k = 1/c_s^2$ and r is the root of the equation:

$\mu' r^{k+1} - (k\lambda + \mu')r + k\lambda = 0$

In the above equation we are assuming that the interarrival time distribution can be approximated using an Erlang distribution and apply the formula for $E_k/M/1$ queues (Gross and Harris 2002). Then we estimate $f(c_a, c_s)$ from the plane passing through the three points $(0,0,0)$, $(c_a, 1, f(c_a, 1))$ and $(1, c_s, f(1, c_s))$.

To estimate $P(T_q > 0) = f(c_a, c_s)$ for $0 < c_s < 1,\ 0 < c_a < 1$, we use point based interpolation method: given a number of points whose locations and values are known, determine the values of other points at predetermined locations. f value at any point (c_a, c_s) on the surface is given by an equation in terms of c_s and c_a. Output data structure is a polynomial function which can be used to estimate values on the surface. A linear equation can describe a tilted plane surface function

$$z = a + bx + cy\ . \tag{5.6}$$

For our research, let $z = P(T_q > 0), x = c_s, y = c_a$.

For function $z = a + bx + cy$, by plugging in the value of three known points, we have equations:

$$
\begin{aligned}
0 &= a + b \cdot 0 + c \cdot 0 \\
r^k &= a + b + c \cdot y \\
\frac{\lambda}{\mu} &= a + b \cdot x + c.
\end{aligned}
$$

Solving the equations, we have:

$$a = 0,\ b = \frac{r^k - y \cdot (\lambda/\mu)}{1 - x \cdot y}, c = \frac{(\lambda/\mu) - r^k \cdot x}{1 - x \cdot y}.$$

Hence substituting a, b, and c into function (5.6), we have:

$$z = a + bx + cy = bx + cy.$$

$$
\begin{aligned}
P(T_q > 0) &= f(c_a, c_s) = bx + cy \\
&= \frac{r^k - y \cdot (\lambda/\mu)}{1 - x \cdot y} \cdot x + \frac{(\lambda/\mu) - r^k \cdot x}{1 - x \cdot y} \cdot y \\
&= \frac{r^k - c_a \cdot (\lambda/\mu)}{1 - c_s \cdot c_a} \cdot c_s + \frac{(\lambda/\mu) - r^k \cdot c_s}{1 - c_s \cdot c_a} \cdot c_a \\
&= \frac{r^k \cdot c_s \cdot (1 - c_a) + (\lambda/\mu) \cdot c_a \cdot (1 - c_s)}{1 - c_s \cdot c_a}.
\end{aligned}
$$

Therefore, for $G/G/s$ queue, we have the probability of waiting:

$$P(T_q > 0) = \frac{r^\alpha c_s (1 - c_a) + (\lambda/\mu')(1 - c_s) c_a}{1 - c_s c_a}. \tag{5.7}$$

This approximation is consistent with the probability of waiting for the $D/D/1$, $M/G/1$ and $E_k/M/1$. For the $M/M/1$, the plane shrinks to a line, so we no longer have the plane defined. Therefore, this formula doesn't apply to the $M/M/1$ queue, for which we have an exact formula.

To estimate the interpolation method, we use $\lambda = 4$, and $\mu = 5$ as an example to calculate the probabilities of waiting for different c_a and c_s. We conclude that the method provides good approximations for the probability of waiting for the $G/G/1$ queue. In our queuing performance prediction model, we implement the approximation method. This point based interpolation is more direct and logical than other interpolation method since it uses all known information to calculate the probability of waiting. Numerical results show that it gives a good approximation.

In this interpolation research, our method is actually designed to work for the coefficients of variation less than or equal to 1. We restrict our discussion to the cases that $c_a \le 1$ and $c_s \le 1$. In other words, the interpolation approximation methods are used in these queuing systems only when $c_a \le 1$ and $c_s \le 1$.When $c_a \ge 1$ or $c_s \ge 1$, we still use the formula of probability of waiting for $M/M/s$ rather than the interpolation value to approximate that of $G/G/s$, which is $P(T_q > 0) = (s\mu - \lambda)W_q$.

In summary, our modeling assumptions are that the first and second moments of the inter-arrival and service time distributions are known. Equation (5.4) is exact for the $G/M/s$ and $M/G/1$ queues. Thus, the method for computing the coefficient of variation of waiting time in the queue is exact for any subset of these queues for which the exact probability of waiting and the second and third moments of the service time distribution is known.

We conjecture that for the $G/G/s$ queue, these relationships still hold and all queuing systems follow these rules. In other words, we use the relationship among the properties of $G/M/s$ and $M/G/1$ queues to estimate the variance of the waiting time for $G/G/s$. Based on the error sensitivity analysis, we know the formula is relatively insensitive to the errors in estimating $P(T_q = 0)$ and $E[s^3]$.

In our implementation, we assumed the service time distribution was gamma. Under these assumptions, the method gives the exact coefficient of variation of waiting times for $M/M/n$, $E_\alpha/M/1$ and $M/E_\alpha/1$ queues. In computational tests with $0 \le c_a \le 1$ and $0 \le c_s \le 1$ (Zhao 2007), we have found the method to give approximations of the standard deviation of the time in system to within (± 10%).

Chapter 6

Priority queue and queuing networks

6.1 Priority queue

Up to this point, all the models considered have the property of a first come first served discipline. This is obviously not the only manner of service, and there are many alternatives, such as last come, first served, selection in random order, and selection by priority. A very considerable portion of real life queuing situations contain priority considerations.

In priority schemes customers with the highest priorities are selected for services ahead of those with lower priorities, independent of their time of arrival into the system. Priority queues are generally more difficult to model than non-priority situations. The determination of stationary probabilities in a non-preemptive Markov system is an extremely difficult matter, well near impossible when the number of priorities exceeds two. Nevertheless, the priority models should not be oversimplified merely to permit solution. Full consideration of priorities is absolutely essential when considering the costs of queuing systems and optimal design.

There are two further refinements possible in priority situations, namely, preemption and non-preemption. In preemptive cases, a customer with the highest priority is allowed to enter service immediately even if another with lower priority is already present in service when the higher customer arrives. That is the lower priority customer in service is preempted, his service stopped, to be resumed again after the higher priority customer is served. In addition, a decision has to be made whether to continue the preempted customer's service from the point of preemption when resumed or to start anew. On the other hand, a priority discipline is defined to be non-preemptive if there is no interruption and the highest-priority customer just goes to the head of the queue to wait its turn. He can't get into service until the customer presently in services is completed, even though this customer has a lower priority.

The number of priority classes can be any number greater than one, and if there can be more than a single customer in any given priority class in the system simultaneously, then the discipline of selecting customers within the same priority class must also be specified.

In this research, we focus on the non-preemptive $G/G/s$ system with many priorities. Within each priority class the FIFO discipline holds. The determination of stationary priorities of $G/G/s$ is well near impossible when the number of priorities exceeds two. In light of this and the difficulty of handling multi-index generating functions when there are more than two priority classes, we use the similar approximation method analogous to the $M/M/s$ priority queue.

For non-preemptive Markovian systems with many priorities, we use the result of Gross and Harris (2002) to derive the formula we used in our spreadsheet.

Suppose that items of the kth priority (the smaller the number, the higher the priority) arrive before a single channel according to a Poisson distribution with parameter $\lambda_k \, (k = 1,2,...r)$ and that these customers wait on a FIFO basis within their respective priorities. Let the service distribution for the kth priority be exponential with mean $1/\mu_k$ .Whatever the priority of a unit in service, it completes its service before another item is admitted.

We begin by defining

$$\rho_k = \frac{\lambda_k}{\mu_k} \qquad (1 \le k \le r) \quad \text{and} \quad \sigma_k = \sum_{i=1}^{k} \rho_i \qquad (\sigma_0 \equiv 0, \sigma_r \equiv \rho)$$

The system is stationary for $\sigma_r = \rho = \sum_{k=1}^{r} \rho_k < 1$. We have

$$W_q^{(i)} = \frac{\sum_{k=1}^{r} (\rho_k / \mu_k)}{(1-\sigma_{i-1})(1-\sigma_i)} .$$

The analysis for the multiple-channel case is very similar to that of the proceeding model except that it must now be assumed that service is governed by identical exponential distributions for each priority at each of s channels. For multiple channels we must assume no service time distinction between priorities or else the mathematics becomes quite intractable.

Define

$$\rho_k = \frac{\lambda_k}{s\mu_k} \qquad (1 \le k \le r) \text{ and } \sigma_k = \sum_{i=1}^{k} \rho_i \qquad (\sigma_r \equiv \rho = \lambda / c\mu)$$

Again the system is completely stationary for $\sigma_r = \rho = \sum_{k=1}^{r} \rho_k < 1$. We have

$$W_q^{(i)} = \frac{E[S_0]}{(1-\sigma_{i-1})(1-\sigma_i)} = \frac{\left[s!(1-\rho)(s\mu)\sum_{n=0}^{s-1}(s\rho)^{n-s} / n! + s\mu\right]^{-1}}{(1-\sigma_{i-1})(1-\sigma_i)}$$

and the expected waiting time taken over all priorities is thus

$$W_q = \sum_{i-1}^{r} \frac{\lambda_i}{\lambda} W_q^{(i)}.$$

Hillier and Lieberman (1986) derived similar formulas as follows:

$$W_q = \frac{1}{A \cdot B_{k-1} \cdot B_k} + \frac{1}{\mu} \qquad for \quad k = 1,2,......N$$

$$A = s!(\frac{s\mu - \lambda}{\rho^s})\sum_{j=0}^{s-1}\frac{\rho^j}{j!} + s\mu$$

$$B_0 = 1$$

$$B_k = 1 - \frac{\sum_{i=1}^{k} \lambda_i}{s\mu} \qquad for \quad k = 1,2,......N \text{ I}$$

Define:

$$Fract1 = \frac{\lambda_1}{\lambda}, \quad Fract2 = \frac{\lambda_2}{\lambda}, \quad Fract3 = \frac{\lambda_3}{\lambda}, \quad Fract4 = \frac{\lambda_4}{\lambda}$$

$$\lambda = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$$

In spreadsheet, we use

$$L_{q1} = \frac{L_q \cdot Fract1 \cdot (1-\rho)}{(1 - Fract1 \cdot \rho)}$$

$$L_{q2} = \frac{L_q \cdot Fract2 \cdot (1-\rho)}{(1 - Fract1 \cdot \rho) \cdot (1 - Fract1 \cdot \rho - Fract2 \cdot \rho)}$$

$$L_{q3} = \frac{L_q \cdot Fract3 \cdot (1-\rho)}{(1 - Fract1 \cdot \rho - Fract2 \cdot \rho) \cdot (1 - Fract1 \cdot \rho - Fract2 \cdot \rho - Fract3 \cdot \rho)}$$

$$L_{q4} = \frac{L_q \cdot Fract4 \cdot (1-\rho)}{(1-Fract1\cdot\rho - Fract2\cdot\rho - Fract3\cdot\rho)\cdot(1-Fract1\cdot\rho - Fract2\cdot\rho - Fract3\cdot\rho - Fract4\cdot\rho)}.$$

The derivation follows.

From Hillier and Lieberman (1986), we have

$$W_q = \frac{1}{A\cdot B_{k-1}\cdot B_k} + \frac{1}{\mu} \qquad for \quad k = 1,2,......N$$

$$A = s!(\frac{s\mu-\lambda}{\rho^s})\sum_{j=0}^{s-1}\frac{\rho^j}{j!} + s\mu$$

$$B_0 = 1$$

$$B_k = 1 - \frac{\sum_{i=1}^{k}\lambda_i}{s\mu} \qquad for \quad k = 1,2,......N \text{ I}$$

λ_i : mean arrival rate for priority class i, for i=1, 2, .N

$$\lambda = \sum_{i=1}^{N}\lambda_i \ .$$

Little's formula still applies to individual priority class, so

$$L_k = \lambda_k W_k \qquad for\ k = 1,2,.......N\ .$$

Hence, $\quad L_q = \dfrac{\lambda_k}{A\cdot B_{k-1}\cdot B_k}.$

For k=1 and s=1

$$L_{q1} = \frac{\lambda_1}{A\cdot B_0\cdot B_1} = \frac{\lambda_1}{\left[s!(\frac{s\mu-\lambda}{(\lambda/\mu)^s})\cdot\frac{\lambda}{\mu} + s\mu\right]\cdot(1-\frac{\lambda_1}{\mu})}$$

$$= \frac{\lambda_1}{\left(\frac{\mu^2}{\lambda}\right)(1-\frac{\lambda_1}{\mu})} = \frac{\lambda^2}{\mu(\mu-\lambda)}\frac{\lambda_1}{\lambda}\left(1-\frac{\lambda}{\mu}\right)\Big/(1-\frac{\lambda_1}{\lambda}\frac{\lambda}{\mu})$$

$$= \frac{L_q\cdot Fract1\cdot(1-\rho)}{(1-Fract1\cdot\rho)}.$$

$$L_{q2} = \frac{\lambda_2}{A \cdot B_1 \cdot B_2} = \frac{\lambda_2}{\left(\frac{\mu^2}{\lambda}\right) \cdot (1-\frac{\lambda_1}{\mu}-\frac{\lambda_2}{\mu}) \cdot (1-\frac{\lambda_1}{\mu})}$$ ��

$$= \frac{\lambda \cdot \frac{\lambda_2}{\mu^2}}{(1-\frac{\lambda_1}{\mu})((1-\frac{\lambda_1}{\mu}-\frac{\lambda_2}{\mu})}$$

$$= \frac{\frac{\lambda^2}{\mu(\mu-\lambda)} \frac{\lambda_2}{\lambda} \left(1-\frac{\lambda}{\mu}\right)}{(1-\frac{\lambda_1}{\lambda}\frac{\lambda}{\mu})((1-\frac{\lambda_1}{\lambda}\frac{\lambda}{\mu}-\frac{\lambda_2}{\lambda}\frac{\lambda}{\mu})}$$

$$= \frac{L_q \cdot Fract2 \cdot (1-\rho)}{(1-\frac{\lambda_1}{\lambda}\rho)((1-\frac{\lambda_1}{\lambda}\rho-\frac{\lambda_2}{\lambda}\rho)}$$

$$= \frac{L_q \cdot Fract2 \cdot (1-\rho)}{(1-Fract1 \cdot \rho)(1-Fract1 \cdot \rho - Fract2 \cdot \rho)}.$$

Similarly, we can derive L_{q3} and L_{q4} etc.

For multi-servers, $s \neq 1$

$$A = s!(\frac{s\mu-\lambda}{\rho^s})\sum_{j=0}^{s-1}\frac{\rho^j}{j!} + s\mu \qquad \text{(Note here } \rho = \frac{\lambda}{\mu}\text{)}$$

$$B_0 = 1;\ B_k = 1 - \frac{\sum_{i=1}^{k}\lambda_i}{s\mu}. \qquad for \quad k = 1, 2, ...N$$

We use the same reasoning:

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

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

$$L_{q1} = \frac{\lambda_1}{A \cdot B_0 \cdot B_1}$$

$$= \frac{\lambda_1}{\left[s!(\frac{s\mu - \lambda}{(\lambda/\mu)^s}) \cdot \frac{(\lambda/\mu)^0}{1!} + s\mu \right] \cdot 1 \cdot \left[1 - \frac{\lambda_1}{s\mu} \right]}$$

$$= \frac{L_q \cdot \frac{\lambda_1}{\mu}(1 - \frac{\lambda}{s\mu})}{(1 - \frac{\lambda_1}{\lambda}\frac{\lambda}{s\mu})}$$

$$= \frac{L_q \cdot Fract1 \cdot (1 - \frac{\lambda}{s\mu})}{(1 - Fract1 \cdot \frac{\lambda}{s\mu})}.$$

Similarly, we can have derive L_{q2} , L_{q3}, L_{q4} etc. By Little's rule, we can have $W_{q1}, W_{q2,} W_{q3,}$ *and* W_{q4} *etc.*

For the $G/G/s$ priority queue, we use the similar approximate method analogous to the $M/M/s$ priority queue. We conjecture that the mean waiting time for each priority class has the same relations of those of the $M/M/s$ priority queue. In other words, we assume for the $G/G/s$ queue, the above formulas also hold for each priority class.

The above formulas are used in our spreadsheet to calculate average flow times in non-preemptive priority queues. For each priority class, we have the following relation to calculate and standard deviation of waiting time.

$$c_q = \sqrt{1 + \frac{4(1 - P(T_q > 0)(\alpha + 2)}{3P(T_q > 0)(\alpha + 1)}}$$

We conjecture the interpolation models hold as well. So, we can estimate the standard deviation of waiting time for each priority class.

6.2 Queuing networks

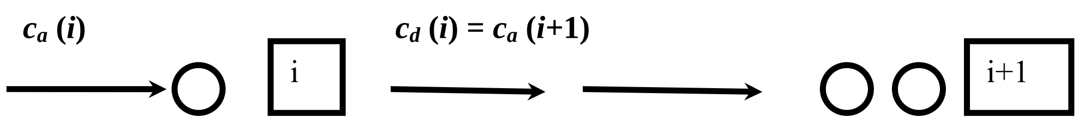
For queuing networks, our model is an open network of single queues in series. Each customer arrives according to an arrival process and is served once at each queue, with the order of the queues being the same for all customers. Each queue has unlimited waiting space, the FIFO discipline, and i.i.d service times that are independent of the other random quantities in the model. The problem is to determine, for a given fixed external arrival process, the standard deviation of flow time in the system per customer. More generally, the object is to determine whether variability, utilization and server numbers actually matter.

The approach of queuing networks approximation is parametric-decomposition: the queues in the network are treated as independent $G/G/s$ models, each partially specified by the basic 5 parameters at that queue. The goal is to use the two arrival parameters at each queue to capture the main effects of the dependence among the queues and the actual properties of the arrival process at each queue.

The basic assumptions for queuing networks:

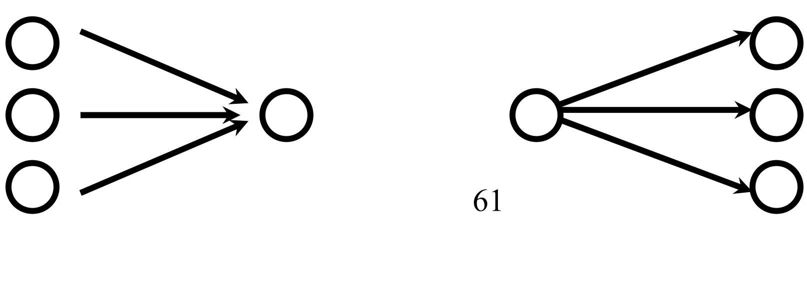
(1) Arrival process of a queue is approximated as the output of its previous queue

$$c_a(i+1) = c_d(i)$$



(2) Each queue in the network is treated as independent $G/G/s$ models, so the standard deviation of total flow time is calculated by formula $\sigma_t = \sqrt{\sigma_1^2 + \sigma_2^2 + ... + \sigma_n^2}$.

(3) In complex queues when the arrival at a queue is the output of several other queues or a random selection of departures from one or more queues (see the graphs below), we assume Poison arrivals.



Since each queue is regarded as a $G/G/s$ queue specified by the basic parameters, the approximation here can be applied directly. More ever, because inter-arrival times at a queue in a network of queues are rarely independent (unless the arrival process is nearly Poisson) and because extra information about the arrival process at each queue is usually unavailable, the partially characterized $G/G/s$ is appropriate.

Interarrival at each queue is not typically independent, but the two parameter characterization is an approximation by a renewal process (having independent interarrival times). The idea is not to ignore the dependence among successive interarrival times, but to try to capture its essential properties with the variability parameters.

Specifically, our models estimate all queuing network situations by using entering departure c_d. So, we can estimate all kinds of $G/G/s$ network queues. When $G/G/s$ models appear as sub-models, simple closed form analytic formulas are useful. For multi-class jobs, we use the law of total variance to calculate the pooled average and pooled variance of flow time.

For the $M/M/s$ queuing series, the departure time distribution from $M/M/s$ queue is identical to the inter-arrival time distribution, namely, exponential. Hence, all stations are $M/M/s$ models.

For the $G/G/s$ queues, our model can estimate all situations by using the proceeding departure c_d as entering c_a (Hopp and Spearman 2002).It doesn't require any type of iterative algorithm to solve and is therefore easily implement-able in a spreadsheet program. This makes it possible to couple the single-station approximation with the multiple linking equations to create a spreadsheet tool for analyzing the performance of a series of queues.

The next step is to characterize the departures from a workstation. We can use measures analogous to those used to describe arrivals, namely, the mean time between departure t_d, the departure rate $r_d = 1/t_d$, and the departure c_d. In a serial queue, where all the output from queue i becomes input to queue $i+1$, the departure rate from i must equal the arrival rate to $i+1$, so

$t_a(i+1) = t_d(i)$ in a serial line where departures from i becomes arrival to $i+1$, the departure c of workstation i is the same as the inter-arrival c of queue $i+1$, $c_a(i+1) = c_d(i)$

The one remaining issue to resolve concerning flow variability is how to characterize the variability of departures from a station in terms of information about the variability of arrivals and process times. Variability from a departure is the result of both variability in arrivals to the station and variability in the process times. The relative contribution of these two factors depends on the utilization of the workstation.

Notice ρ increases with both the arrival rate and the mean effective process time. An obvious upper limit on the utilization is one (that is 100 percent), which implies that the effective process times must satisfy $\rho = \lambda/(\mu s) \leq 1$.If ρ is close to one, then the station is almost always busy (heavy traffic). Therefore, under these conditions, the inter-departure from the queue will be essentially identical to the service times. Thus, we could expect the departure coefficient of variation to be the same as that of the service time (that is $c_d = c_s$).

At the other extreme, when ρ is close to zero, the station is very lightly loaded. Virtually every time a job is finished, the queue has to wait a long time for another arrival. Because process time is a small fraction of the time between departures, interdependent times will be almost identical to inter-arrival times. Thus, under these conditions, we could expect the arrival and departure c to be the same (that is $c_d = c_a$) . A good, simple method for interpreting between these two extremes is to use the square of the utilization as follows:

$$c_d{}^2 = \rho^2 c_s{}^2 + (1-\rho^2) c_a{}^2 .$$

If the server is always busy, so that, ρ =1, then $c_d{}^2 = c_s{}^2$. Similarly, if the machine is almost always idle, so that, ρ =0 then $c_d{}^2 = c_a{}^2$. For intermediate utilization levels, $0 < \rho < 1$, the departure c_d is a combination of the inter-arrival c_a and the service time c_s .

When there is more than one server at a queue (that is s>1), the following is a reasonable way to estimate c_d (Hopp and Spearman 2002).

$$c_d^{\ 2} = 1 + (1-\rho^2)(c_a^{\ 2} - 1) + \frac{\rho^2}{\sqrt{s}}(c_s^{\ 2} - 1).$$

Note, this reduces to the above equation when s=1. This formula is used in our spreadsheet. The above results for flow time variability are building blocks for characterizing the effects of variability in the overall queuing networks.

With our approximation procedure, we know each distribution is partially characterized by its first two moments, or equivalently, by its mean and squared coefficient of variation. The closed-form formulas give an approximate squared coefficient for the arrival process to each queue and an approximate expected steady-state waiting time. The expressed steady-state waiting time for queues in series actually depends on the distributions beyond their first two moments, but experience indicates that fairly good approximation can often be obtained given this partial information.

Given the 5 basic parameters, assuming ρ <1, we have a proper steady state queue. Our results show that the models yield a satisfactory approximation (in the order of 10 percent relative error), providing that the variability parameters c_a and c_s are either equal or less than 1. The violation of any of these conditions should be a clear warning. If one needs more accuracy, additional information about the distributions is needed.

Chapter 7
Simulation

7.1. Test the accuracy of the approximation by simulation

We have discussed the analytical approach, i.e., when it is possible to describe a queuing situation analytically and obtain also by analytical methods and useful expressions, such as $M / M / s$, the average and standard deviation of waiting time and the average length of the queue, from which many useful measures are available.

Due to the characteristics of the input or service mechanism and the nature of the queuing discipline, or combinations of the above, for $G / G / s$ queue, it is impossible to model analytically. The alternative methods are to simulate the system. The experiment must be repeated sufficiently often to obtain large samples and a variety of answers, which are then taken together in some manners to obtain a value for what is desired. This is a very useful method in practice whenever complicated problems require immediate answers.

While simulation may offer a "way out" for many analytical intractable models, it is not in itself a panacea. There are considerable numbers of pitfalls one may encounter in using simulation. Great care is required to obtain correct simulation with enough samples and to properly combine the results to obtain an answer.

Since simulation is comparable to analysis by experimentation, one has all the usual problems associated with running experiments in order to make inferences concerning the real world, and must be concerned with such things as run length, number of replications, and statistical significance.

To achieve meaningful results, a great deal of care and thought must go into planning and running the simulators, especially in the areas run-length determination and the interpretation of the output. Another drawback to simulation analysis occurs if one is interested in optimal design of queuing systems. How close one gets to optimality in a simulation study often depends on how clever the analysis is in considering the alternatives to be investigated.

Because of this, simulation has often been referred to as art. Nevertheless, simulation can be an extremely important tool and is often the only procedure that can be used in analyzing many of the complex queuing systems encountered in practice. The success or failure of a simulation study often lies in how it is used and how the output is interpreted.

The purpose of conducting simulation in our research is to test the accuracy of the approximations. Since no closed-form analytical results are available for $G/G/s$ models, to evaluate the accuracy of our approximations, we conduct simulation experiments using the Extend simulation program. The testing of our approximations has been based on extensive simulation experiments. These simulation experiments are indispensable parts of our research on the $G/G/s$ queue.

To verify the quality of the approximations, it is necessary to resort to analyses by simulation. It should be emphasized , however, that if analytical models are achievable, they should be used and that simulation should be relied upon only in cases where analytical models are either not achievable and approximations not acceptable or they are so complex that solution is prohibitive.

A simulation model can be considered as consisting of three basic phases: data generation, bookkeeping, and output analysis. Data generation involves the production of representative inter-arrival times and service times where needed throughout the queuing system. Generally, this involves producing representative observations from pre-specified probability distributions, and it is this aspect to which the term Monte Carlo has been applied. Thus, a Monte Carlo simulation is one in which it is necessary to generate at least one stream of random observations from some specified probability distribution (either a theoretical or empirical distribution).Most queuing simulations are of the Monte Carlo type. For the $G/G/s$ queue, we conduct a Monte Carlo simulation due to its inherent random nature.

7.2 Approach for using simulation

Simulation is indeed a process. We basically follow the following steps suggested by the Extend program in our simulation process.

1. Identify the problem, set objectives of the model, and plan the project.

The purpose is to understand important cause/effect relationships so that this outcome can be improved, also evaluate proposed changes or decision alternative, or to forecast system behavior under different input conditions.

2. Define the system

Identify factors such as system components, descriptive variables, and interactions (logic) that constitute the system. Agree on a level of details and boundaries of the system. A good way to develop this understanding is through the use of a Cause-Effective analysis, where the effect is the critical process metric.

3. Create conceptual models by identifying factors and determining functional relationships.

Identify factors to model, with a mind toward simplification.

4. Plan the experiments. Tie the model to the system and select what is to be varied.

Tie the model to the system by defining output messages of the system, which should be generated by the model for purposes of comparison and decision-making. This sets up some pre-work necessary for establishing the credibility of a model so that predictions under conditions untried in the current system can be corroborated.

A critical metric system measures performance relative to a critical system requirement. It is in terms of this metric that the gap between current performance and target should be stated to justify the simulation effort. A critical metric is used to validate the model and to judge the effect of system changes. (Our research uses this measurement).

5. Prepare the input data.

Identify and collect data to model the descriptive input variables. Create a statistical model to characterize the variables, two situations we meet are:

(1) No Existing data, where we must use our best conjecture and some knowledge of the typical distributions that associated with different kinds of processes. (For priority and two workstations)

(2) Existing data, where a fit a distribution to the available data. Using knowledge of the distributions associated with different kinds of process to help us.

6. Formulate the simulation model.

We have to determine the appropriate kind of simulation model to build. Some choices (discussed in more details later on in this chapter) are discrete event, continuous, or mixed.

7. Verify and validate the model.

Confirm that the model works the way we intend it to. Then confirm that the model is representative of the actual system. Validation is done by comparing symptom and critical metric output data from the model with the same output from the actual system. This is most effective if system and model outputs are generated under a wide range of input conditions. The model may have to be refined at this point if the difference between model and system results does not meet the criteria set in step 4.

8. Design the experiments to run.

Determine the final experimental design. Issues beyond the factors/levels to run are warm-up period, length of run, number of runs of each alternative, etc.

9. Run experiments, analyze data, and interpret results.

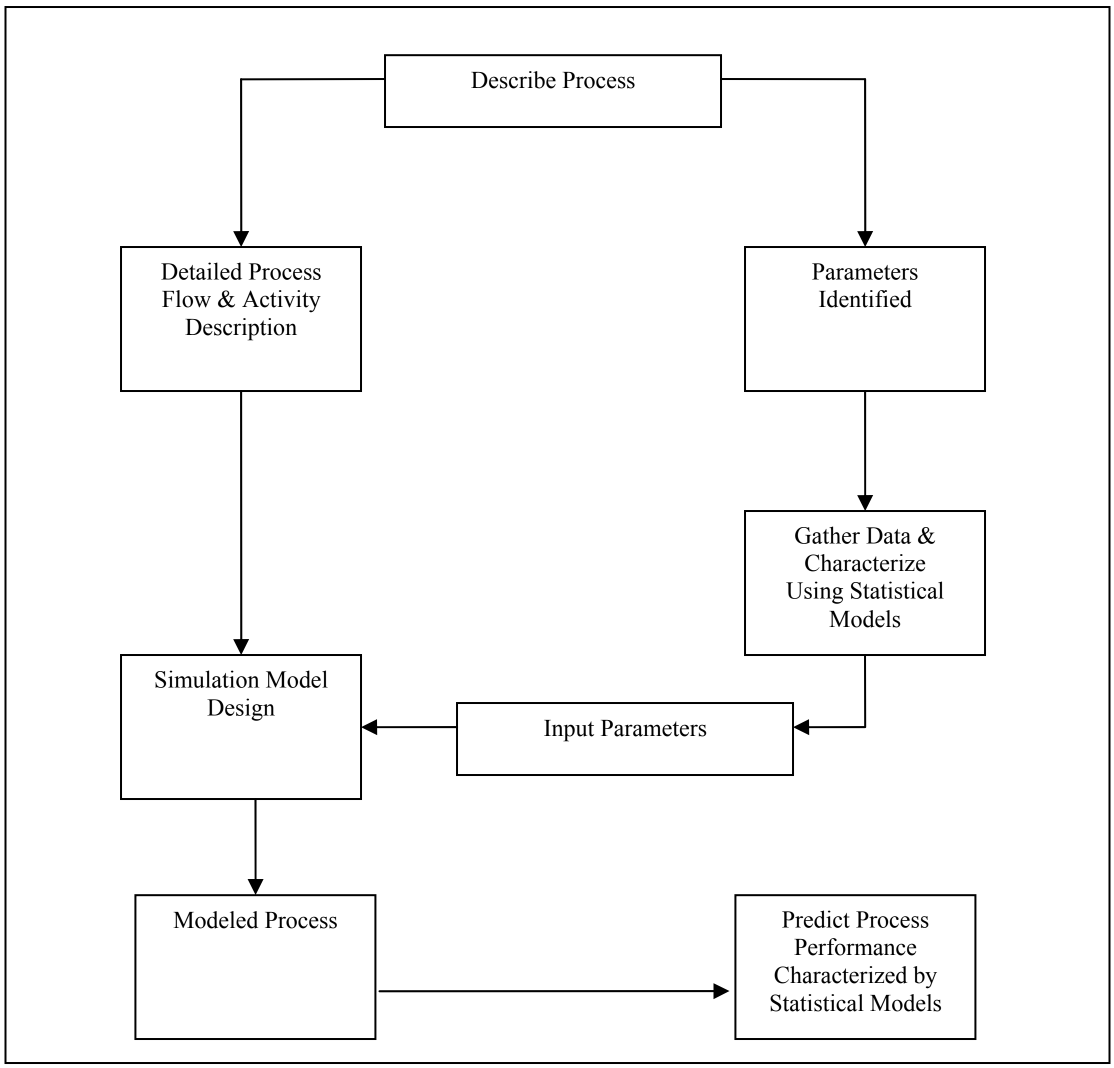
Run the experiments, and draw inferences from the data generated.

10. Implement the decisions. Make use of the findings.

11. Document and maintain the model.

An executive view of the approach is shown below (Extend software manual).

Process Modeling Method



7.3 Analysis of the results

We compare the approximations with the simulation values of the standard deviations of waiting time (See appendices). These numerical comparisons show that our approximation performs remarkably well.

In this simulation research, we concentrate on a single queue with and without priority, and the special case of only two queues. We use Manzana case study to discuss queuing networks.

To estimate the mean and standard deviation of steady-state waiting times, we conduct 4 experiments using the Extend simulation program. In each case, we performed independent replications using 54000 minutes of simulation time and estimated 95 % confidence intervals.

The four experiments are:

(1) single queue without priority
(2) single queue with 4 priority classes
(3) two tandem queues to test Central Limit Theory(covariance)
(4) Manzana case study to test queuing networks

For each experiment, we first use Excel spreadsheet model to formulate our approximation results. Then we use Extend software to simulate corresponding spreadsheets so that we can compare the two results.

We characterize the queuing models by the parameters c_a, c_s, ρ *and* s . Here c_a is the coefficient of variation of an inter-arrival time; c_s is the coefficient of variation of the service time; ρ is the utilization and s is the number of servers. We specify the distributions to go with the first two moments. We considered various parameters for all combinations of the utilization ρ =0.8 and ρ =0.9.

For each queue, we consider 4 values of c_a and c_s : 0, 0.5, 1, and 1.5. Thus, with two utilizations ρ =0.8 and ρ =0.9, we have 32 cases.

$$(4 \text{ values of } c_a) \times (4 \text{ values of } c_s) \times 2 = 32$$

The number of servers could be 1, 2 or 3. So we have 32 × 3= 96 scenarios. For the first 3 experiments, we totally have 96 × 3=288 scenarios. The last experiment is the combination of first 3 experiments to test queuing networks by using the Manzana case.

When coefficient of variation c=0, we use a deterministic distribution; c=0.5 and 1.5 Gamma distribution; and c=1 exponential distribution. For a deterministic distribution, we can calculate constant inter-arrival rate and constant service rate. For an exponential distribution, we calculate interarrival rate λ and process rate μ. For the Gamma distribution, we first calculate scale and shape parameters. We then key in the parameters in the Extend simulation blocks to obtain different queuing models.

Weibull, Erlang, lognormal or Pareto were used as the $G/G/s$ queue in simulation literature (Whitt 2004). In our research, Gamma distribution is used as general distribution. When shape parameter k is positive integer, Gamma is called Erlang. When k =1, it is exponential. When $k \to \infty$, it is deterministic.

PDF of Gamma distribution:

$$f(x;k;\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k\Gamma(k)} \qquad \text{for x>0.}$$

Where k (>0) is the shape parameter; $\theta(>0)$ is the scale parameter; $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$

CDF of Gamma distribution:

$$F(x;k;\theta) = \int_0^x f(x;k;\theta) = \frac{r(k, x/\theta)}{\Gamma(k)}.$$

We calculate shape and scale parameters and input the simulation blocks. (See appendices).

Extend is a widely used simulation software.With Extend, we create a block diagram of a process where each block describes one part of the process. In Extend, we lay out our process in a twodimensional drawing environment. Extend provides the equivalent of a moving picture. We use Extend's iterative technique to create models of real-world processes that are too complex to be easily represented in a spreadsheet.

In Extend, the Generator block from the Generators submenu of the Discrete Event library is used to provide items at exponential inter-arrival times (and many other inter-arrival times as well). The Queue FIFO block holds the items, releases them first-in, first-out, and can have a maximum queue length specified in the dialog. The Activity Delay (from the Activities submenu of the Discrete Event library), Machine, and Station blocks (from the Activities submenu of the Manufacturing library) represent servers: you specify an exponential service time by connecting an Input Random Number block (Inputs/Outputs submenu of the Generic library) to the *D* (delay) connector on those blocks.

The Discrete Event and Manufacturing libraries allow us to select the type of queue (FIFO, LIFO, priority, or queuing by matching attribute names and/or values) required for our models.

For a single queue without priority, we use 7 Extend blocks: generator, timer, queue (FIFO), mean & variance, input (random number), activity (multiple).

For the single queue with priority, we consider 4 priority classes with workload fraction 0.25 for each class. We use the queue (attribute) block to measure priority classes. All other blocks are the same as without priority queue blocks. We collect data of different priority classes and use Excel to calculate the mean and standard deviation of each class.

To test Central Limit Theory (covariance), we consider two tandem queues. Two queues have same service rate and number of servers, as well as same c. Just as before, we use deterministic, exponential and Gamma distributions when $c=0$, $c=1$ and $c=0.5$ and 1.5 respectively.

Finally, we demonstrate the use of these results by applying the approximations to an analysis of the Harvard Business Case *Manzana Insurance* and compare the results of the analysis to those obtained via a Monte Carlo simulation.
We use data from Manzana Insurance to develop spreadsheet models. We develop three models: (1) the current system with and without priority queue. (2) Combined UT with and without

priority queue. (3) Moving one policy writer to distribution. Then, we conduct the corresponding simulations using Extend, exploiting same data from Manzana so as to compare the results of spreadsheets with those from simulations. Last, we estimate the quality of the approximations by comparing total flow times, standard deviations and "worst cases" of different scenarios. Extensive simulations show that our approximation methods are simple yet fairly good in their performance.

For multi-classes of jobs (in Manzana Case**)**, we use the law of total variance to calculate pooled average and pooled variance.

Coefficient of variation calculation for multi-products (see appendices)

Conditional Variance (Law of Total Covariance)

$$Var(x) = E\left[Var(x|y)\right] + Var\left[E(x|y)\right]$$

Conditional Expectation

$$E(x) = E\left[E(x|y)\right] = E\left[E(x_i|y_i)\right] = \sum_i p_i \mu_i$$

Proof:

$$\begin{aligned} Var(x) &= E(x^2) - \left[E(x)\right]^2 \\ &= E\left[E(x^2|y)\right] - E\left[E(x|y)\right]^2 \\ &= E\left[Var(x|y)\right] + E\left[E(x|y)^2\right] - E\left[E(x|y)\right]^2 \qquad \left(\because E\left[Var(x|y)\right] = E\left[E(x^2|y)\right] - E\left[E(x|y)^2\right]\right) \\ &= E\left[Var(x|y)\right] + Var\left[E(x|y)\right] \end{aligned}$$

Therefore

$$\begin{aligned} Var(x) &= E\left[Var(x_i|y_i)\right] + Var\left[E(x_i|y_i)\right] \\ &= \sum_i p_i \sigma_i^{\ 2} + E\left(\mu^2\right) - \left[E(\mu_i)\right]^2 \\ &= \sum_i p_i \sigma_i^{\ 2} + \sum_i P_i \mu_i^{\ 2} - \left(\sum_i p_i \mu_i\right)^2 \end{aligned}$$

The above formula is used in the *Manzana* case study to calculate total variance of different types of arrivals so that we can calculate total coefficient of variation of service time.

Numerical comparison

We present a representative set of tables comparing the approximations with exact (simulation) values. Before discussing these tables in detail, we comment how we evaluate the quality of the approximations.

There are two standard ways to measure the quality of queuing approximations: absolute difference and relative percentage error (Whitt 1993). We contend that neither procedure alone is usually suitable over the entire range of values. We can obtain satisfactory results if either the absolute difference is below a critical threshold or the relative percentage error is below another critical threshold. Thus, a final adjusted measure of error (AME) might be:

$$Error = \min\left\{A\left|eaxct - approx.\right|,\ \ 100(\left|exact - approx.\right|)/exact.\right\}.$$

A is a constant chosen in each instance to reflect the relative importance of absolute difference and the relative percentage of error.

In our comparisons, we choose A=1 for simplicity. Although we don't display the calculations of any specific measures of errors, our discussion explains the goals. Either the relative percentage error or the absolute difference should be small.

Here we have 4 simulation results corresponding to 4 different experiments.
Table 1 contains the simulation for queue without priority. Table 2 contains the simulation results with priority. Table 3 contains results for testing CLT. Table 4 contains queuing networks with the Manzana case study.

These tables display expected mean and standard deviation of flow time in specific queuing systems. The difference and relative error analysis are displayed in a separate spreadsheet.

We compare the approximations for the standard deviation of waiting time with simulation values generated by Extend simulators. The cases considered are $G/G/s$ queue with $\rho = 0.8$ and $\rho = 0.9$ respectively. For these cases, in which both $c_a \le 1$ and $c_s \le 1$, the approximations appear to be remarkably accurate. (The calculation is exact for $M/M/s$ queue and imbedded Markov queues ($E_k/M/1$ and $M/E_\alpha/1$).

Consistent with remarks by Hopp and Spearman (2002) and Whitt (2002), but deserving more emphasis, we conclude that the key factor is variability. The results indicate that if the coefficient of variation (either interarrival or service time) is 1.5, our approximations are not precise. However, when the coefficient of variation is small, we can see simulation results match with spreadsheet results well regardless of utilization and server number.

In general, the accuracy improves as coefficient of variation decreases. The weak part of approximation scheme seems to be priority queues with lowest priority class when the utilization is high. Overall, the approximations seem to be sufficiently accurate for practical operations purposes.

Chapter 8
Summary

8.1 Contributions to knowledge

In this research, we have developed mathematically tractable expressions for the standard deviation of waiting time for $G/M/s$ and $M/G/1$ queues. We provide an approximation for the standard deviation of flow time in system for a general multi-server queue with infinite waiting capacity ($G/G/s$). The approximation requires only the mean and standard deviation or the coefficient of variation of the inter-arrival and service time distributions, and the number of servers. We also extend the approximations to the $G/G/s$ priority queues and queuing networks. The quality of the approximations is not the same for all cases, but in comparisons to Monte Carlo simulations has proven to give good approximations (within $\pm$ 10%) for cases in which the coefficients of variation for the inter-arrival and service times are between 0 and 1. A significant feature of the approximation methods is that it is mathematically intractable and can be implemented in a spreadsheet format. The following are the outlines of the contributions:

1. We derive the standard deviation of waiting time in system for $M/M/1$ and $M/M/s$ queues, as well as imbedded Markov chain queues ($G/M/1, G/M/s$). We found that for all these queue models, the following relation holds regardless of distribution $\sigma_q = \sqrt{\frac{2-P(T_q>0)}{P(T_q>0)}} \cdot W_q$. σ_q is just a function of $P(T_q>0)$ and W_q, i.e. Standard deviation of waiting time is just a function of probability of waiting and average waiting time.

2. We present a general expression for the coefficient of variation of waiting time (c_q), which is applicable to the $G/M/s$ and $M/G/1$ queues. We conjecture that this expression provides a good approximation for $G/G/s$ queues and have validated this conjecture via computer simulations. For $G/M/s$ and $M/G/1$ queues:

$$c_q = \sqrt{1+\frac{4E[s^3]}{3\lambda}\frac{(1-P(T_q>0))}{(E[s^2])^2}} \tag{8.1}$$

Where $P(T_q > 0)$ is the probability of waiting, $E[s^2]$ and $E[s^3]$ are the second and third moments of the service time distribution.

3. We examine the sensitivity of the formula (1) to errors in estimating $E[s^3]$, given that the other parameters $P(T_q > 0)$ and $E[s^2]$ are known. We find that the formula is relatively insensitive to the errors in estimating $E[s^3]$.Suppose a small change in $E[s^3]$, expressed as a proportion P, is $\Delta E[s^3] = P \cdot E[s^3]$, the resulting change in c_q is at most $P/2$, $\frac{\Delta c_q}{c_q} \le 0.5P$.

Similarly, we examine the sensitivity of the formula (8.1) to errors in estimating $P(T_q = 0)$. Suppose a small change in $P(T_q = 0)$, expressed as a proportion P,

is $\Delta P(T_q = 0) = P \cdot P(T_q = 0)$, the resulting change in c_q is at most $P/2$, that is $\frac{\Delta c_q}{c_q} \le 0.5P$.

4. For $M/M/s$, we derived $P(T_q > 0) = (s\mu - \lambda)W_q$.

For $G/G/s$, we develop point based interpolation model to estimate the probability of waiting in $G/G/s$ queue: $P(T_q > 0) = \dfrac{r^k c_s(1-c_a) + (\lambda/\mu')(1-c_s)c_a}{1 - c_s c_a}$

5. We develop a queuing system performance predictor based upon the above results. The prediction generalizes the approximations proposed in our research. For these models, we only need the basic 5 parameter $s, \lambda, \mu, c_a, and\ c_s$ to measure the performances of all kinds of steady-state unlimited capacity queues. We believe that our two moment approximation will be beneficial to those practitioners who like simple and quick answers to their queuing systems.

8.2 Limitations and future directions

(1) For priority queues, we have testes 4 priority classes and our approximation methods indicate that the performance for the lowest class in the $G/G/s$ queue is not accurate and satisfactory. We need to test more classes, such two and three classes to see if we can obtain the same conclusion.

(2) For coefficient of variations of interarrival time or service time greater than 1, the approximations are less reliable. Its performance tends to deteriorate as the c_s and c_a get further away from 1, especially in the case of light traffic. Currently, we know of no general models for the standard deviation of waiting time with the coefficients of variation outside this range $c_a, c_s \leq 1$ .Also no computer package is commonly available that would enable us to compute exact performances numerically. For these cases, they have not yet been studied sufficiently and such descriptions evidently depend more critically on the missing information (the discussions beyond the first two moments). More sophisticated numerical procedures are needed for those cases.

(3) For simulation testing we have considered different combinations of four values of c_a and c_s respectively: 0, 0.5, 1, 1.5 with two utilizations $\rho = 0.8$ and $\rho = 0.9$. The other combination values of c_a and c_s , such as 0.25, 0.75 needs to be tested to make sure our approximations can be used in a wide range of applications. In the literature, we have seen Seelan and TIJM (1984) and Whitt (1989) used Erlang and H (hyperexponential) distribution to represent general distribution. In our simulation experiments, we have used gamma distribution to represent general distribution. We can test other distributions, such as hyperexponential, Weibull and normal distributions.

(4) The research results can be extended to estimate the performances of batch, balking, optimal design and other queuing system applications.

(5) The approximation models presented in this research could be used in scheduling, inventory, insurance management, reliability and maintenance, and many other operations and supply chain systems.

Bibliography

Bibliography

Abate J., G. L. Choudhury, and W.Whitt. Waiting-time tail probabilities in queues with long-tail service-time distributions. Queuing Systems 16, 1994.

Baccelli, F., P. Bremaud. 1994. Elements of Queuing Theory. Springer-Verlag, New York.

Bertsimas, D. An exact FCFS waiting time analysis for a general class of G/G/s queuing systems. Queuing Systems 3. 1988.

Bertsimas ,Dimitris, An Analytic Approach to a General Class of G/G/s Queueing Systems , Operations Research Vol. 38, No. 1 (Jan., 1990)

Bitran, G. R., S. Dasu. A review of queuing network models of manufacturing systems. Queuing Systems 12. 1992.

Boxma, O. J. Cohen, J.W. and Huffels, "Approximations of the mean waiting time in an M/G/s queuing system" Operations Research. 27, 1979.

Buzacott, J. A. 1996. Commonalities in reengineered business processes: models and issues. Management Science. 42

___, J. G. Shanthikumar. 1992. Design of manufacturing systems using queuing models. Queuing Systems 12.

___. 1993. Stochastic Models of Manufacturing Systems. Prentice-Hall, Englewood Clifs, NJ.

Coffman, Jr., E.G., L. Flatto, W. Whitt. 1996. Stochastic limit laws for schedule make spans. Stochastic Models 12.

Cohen, J.W. The Single Server Queue, 2nd edition. North-Holland, 1969

Cooper, R.B. Introduction to Queuing Theory. 2nd edition. Macmillan, 1972

Duffield, N. G., W. Whitt 1997. Control and recovery from rare congestion events in a large multi-server system. Queuing Systems 26

Feller, W. 1971. An Introduction to Probability Theory and its Applications. Vol. II, 2nd ed. Wiley, New York.

Griffiths , J. D. ,The Coefficient of Variation of Queue Size for Heavy Traffic
The Journal of the Operational Research Society Vol. 47, No. 8 (Aug., 1996), pp. 1071-1076

Gross, D., C. M. Harris. 1985. Fundamentals of Queuing Theory, 2nd ed. Wiley, New York.

Gross, D., C. M. Harris. 2002. Fundamentals of Queuing Theory, 3nd ed. Wiley, New York.

Hall, R. W. 1991. Queuing Methods Science for Services and Manufacturing. Prentice Halal, Englewood Cliffs, NJ.

Hillier and Lieberman, Operations Research (1986)

Hopp and Spearman. Factory Physics. 2002.

Hui, M. K., D. K. Tse. 1996. What to tell customers in waits of different lengths: an integrative model of service evaluation. J. Marketing 60.

Kendall, D. G. “Some Problems in the Theory of Queue” Journal of Royal Statistics
Society .Ser.B13 No2.1951

Kleinrock, L. Queuing Systems, Volume I & II: Theory. John Wiley and Sons, 1975 & 1976

Katz, K. L., B. M. Larson, R. C. Larson. 1991. Prescription for the waiting-in-line blues: entertain, enlighten and engage. Sloan Management Rev. 32.

Kimura , Toshikazu ,A Two-Moment Approximation for the Mean Waiting Time in the GI/G/s Queue ,Management Science Vol. 32, No. 6 (Jun., 1986)

Kingman, J.F.C. (1970) Inequalities in the Theory of Queues”, Journal of the Royal Statistical Society. Series B 32.

Nasreddine Tabet-Aouel; Demetres D. Kouvatsos ,On an Approximation to the Mean Response Times of Priority Classes in a Stable G/G/c/PR Queue The Journal of the Operational Research Society Vol. 43, No. 3 (Mar., 1992)

Reiman M.I. and Simon B. An interpolation approximation for queuing system with Poisson Input Bell Laboratories. 1984

Rothkopf, M. H., P. Rech. 1987. Perspectives on queues: combining queues is not always beneficial. Oper. Res. 35.

Sakasegawa, H. 1977. An approximation formula L[subq] = alphabeta[supp]/(1- rho). Ann. Inst. Statist. Math. 29

Saaty, T.L. Elements of Queuing Theory. McGraw-Hill, 1961

Seelen, L.P. and Tijms, H.C. Approximations for the conditional waiting times. Operations research letters. October 1984

Shanthikumar, J. G.**;** Buzacott, J. A., On the approximations to the single server queue. International Journal of Production Research, Nov/Dec80, Vol. 18 Issue 6

Shimshak ,Daniel G.; Georghios P. Sphicas ,Waiting Time in a Two Station Series Queueing System: The Effect of Dependent Interarrival Times ,
The Journal of the Operational Research Society Vol. 33, No. 8 (Aug., 1982)

Shore , Haim, Simple Approximations for the GI/G/c Queue-I: The Steady-State Probabilities The Journal of the Operational Research Society Vol. 39, No. 3 (1988)

Stanford, D. A., B. Pagurek, and C. M. Woodside. 1983. Optimal prediction of times and queue lengths in the GI/M/1 queue. Operations. Res. 31

Streams David A. Stanford, Waiting and Inter-departure Times in Priority Queues with Poisson- and General-Arrival, Operations Research Vol. 45, No. 5 (Sep., 1997)

Suresh ,S.; W. Whitt ,Arranging Queues in Series: A Simulation Experiment , Management Science *Vol. 36, No. 9 (Sep., 1990)*

Toshikazu Kimura, Approximating the Mean Waiting Time in the GI/G/s Queue The Journal of the Operational Research Society .Vol. 42, No. 11.1986

Taylor, S. 1994. Waiting for service: the relationship between delays and evaluations of service. J. Marketing 58.

Whitt Ward ,Planning Queueing Simulations, Management Science Vol. 35, No. 11, Focussed Issue on Variance Reduction Methods in Simulation (Nov., 1989)

Whitt, Ward, A Diffusion Approximation for the G/GI/n/m Queue. Operations Research, Nov/Dec2004, Vol. 52 Issue 6

Whitt, Ward Predicting Queueing Delays. Management Science Vol. 45, No. 6 (Jun., 1999)

Whitt, W. 1983. Comparison conjecture about the M/G/s queue. Operations research letters

----1985. The best order for queues in series. Management Science. 31 475-487.

----1992. Understanding the efficiency of multi-server service systems. Management Science. 38 708-723.

----1993. Approximations for the GI/G/m queue. Production and Operations Management 2 114-161.

Whitt, Ward. An approximation approach for the mean workload in a GI/G/1 queue. Operations Research, Nov/Dec89, Vol. 37 1991)

Whitt W. 1984a. Departures from a queue with many busy servers. Math of Operations. Research.

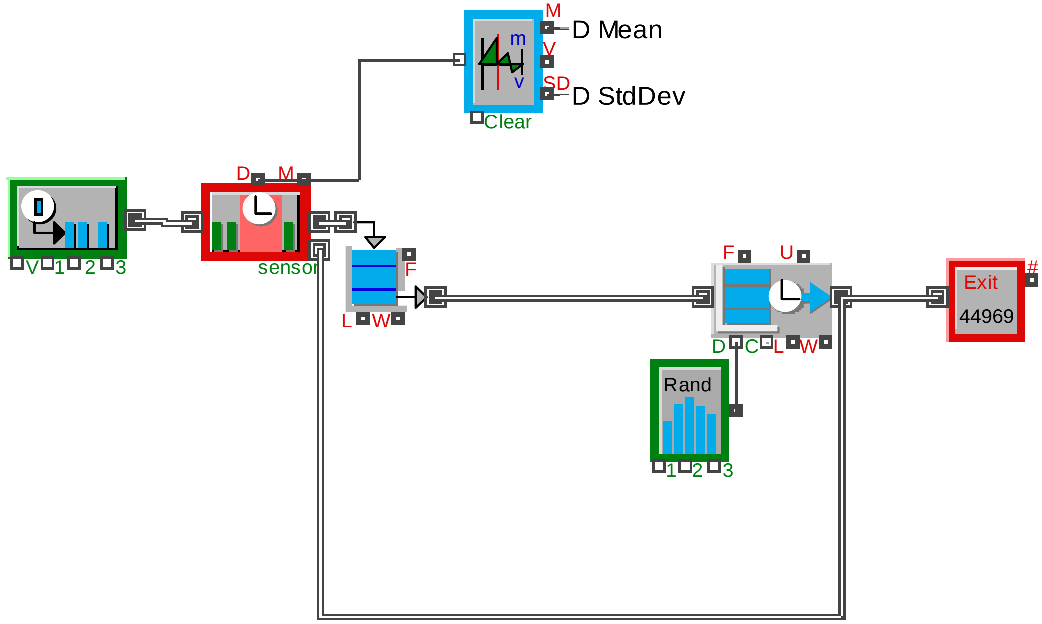
----. 1984b. Approximations for departure processes and queue in series. Naval. Res. Logistics. Quart.

----. 1999. Improving service by informing customers about anticipated delays. Management Science. 45.

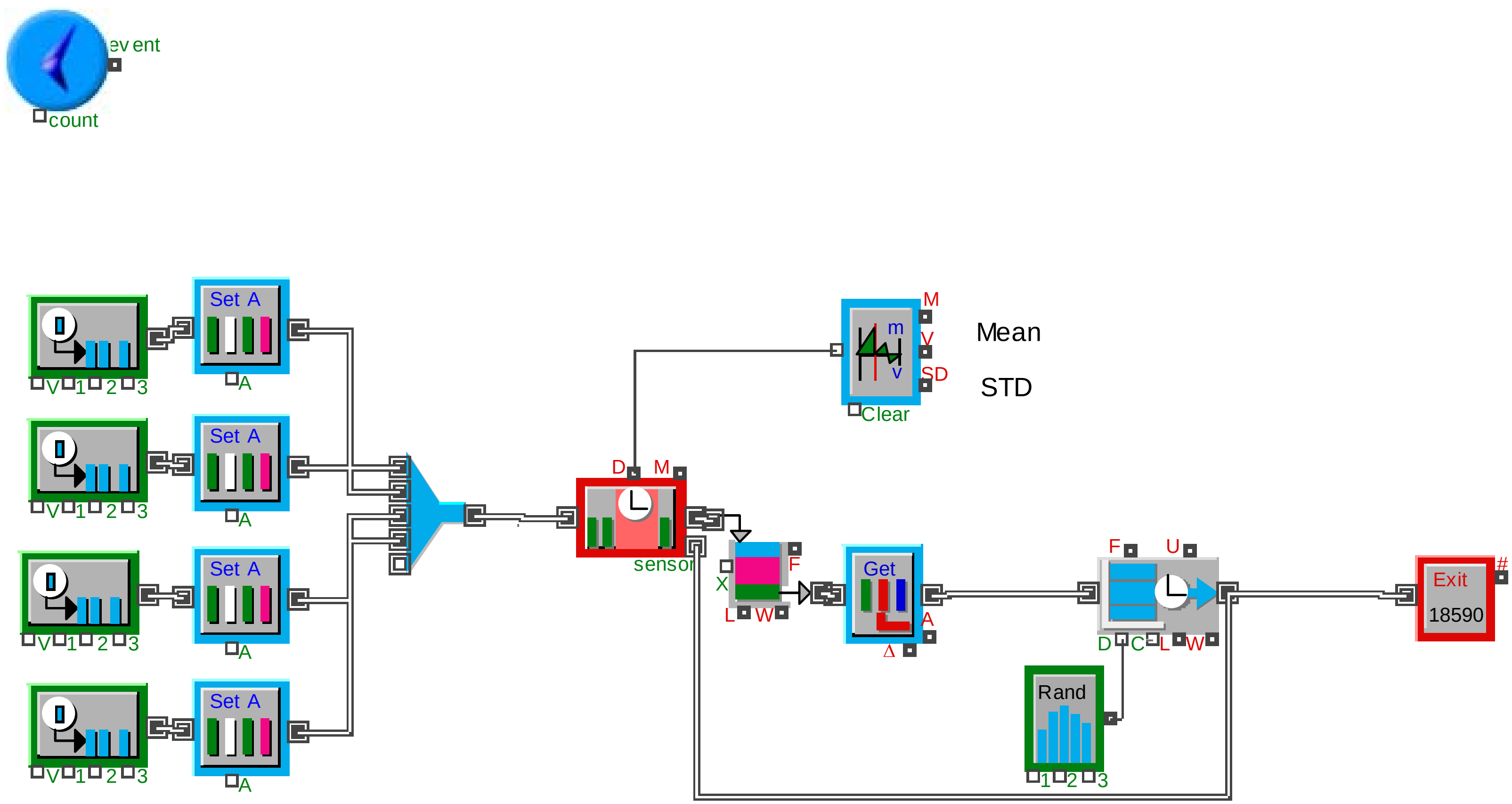
William G. Marshal; Carl M. Harris, A Modified Erlang Approach to Approximating GI/G/1 Queues , Journal of Applied Probability. Vol. 13, No. 1 (Mar., 1976)

Appendices

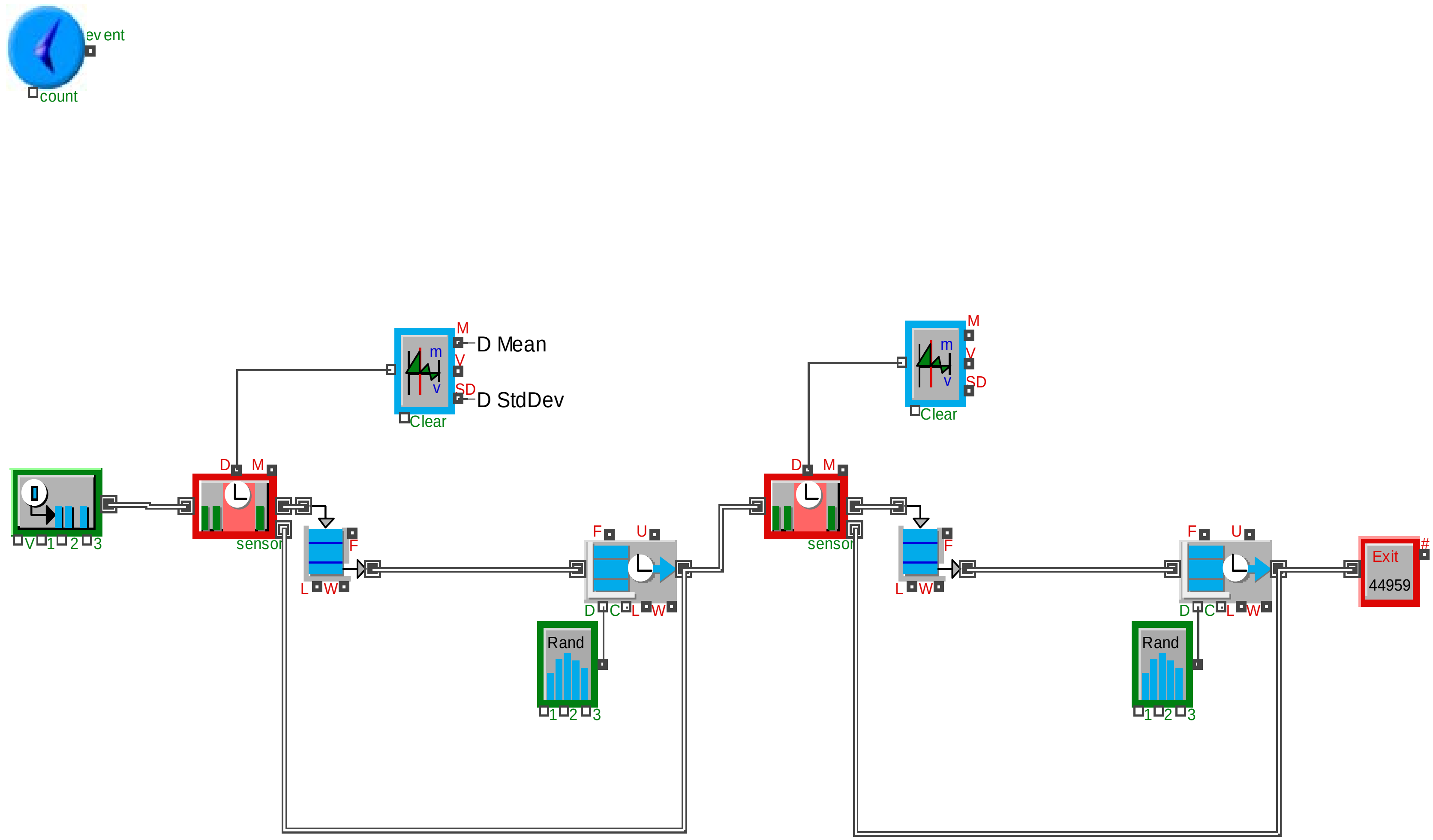
Appendix 1 Simulation experiment for a single queue without priority



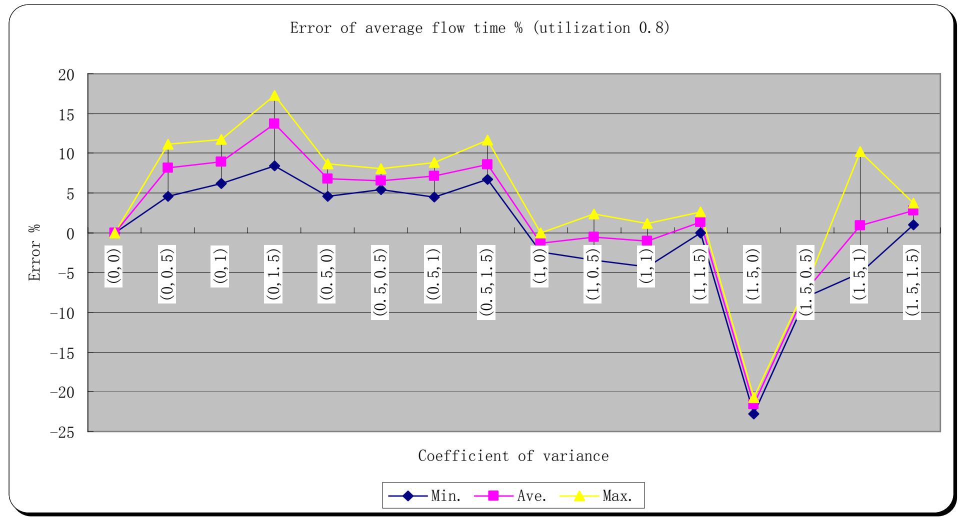
Appendix 2 Simulation experiment for a single queue with 4 priority classes

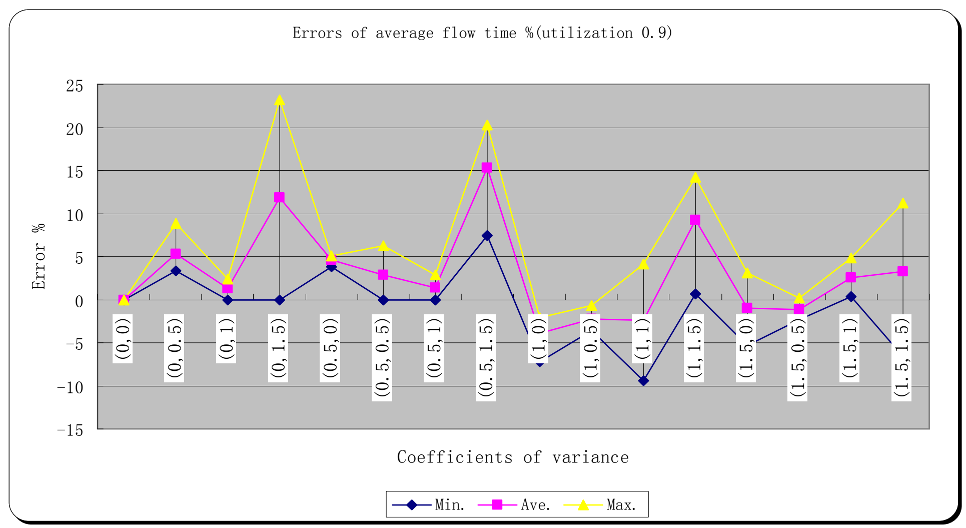


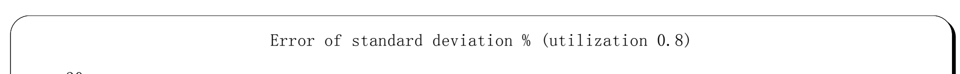
Appendix 3 Simulation experiment for a two queues in a series



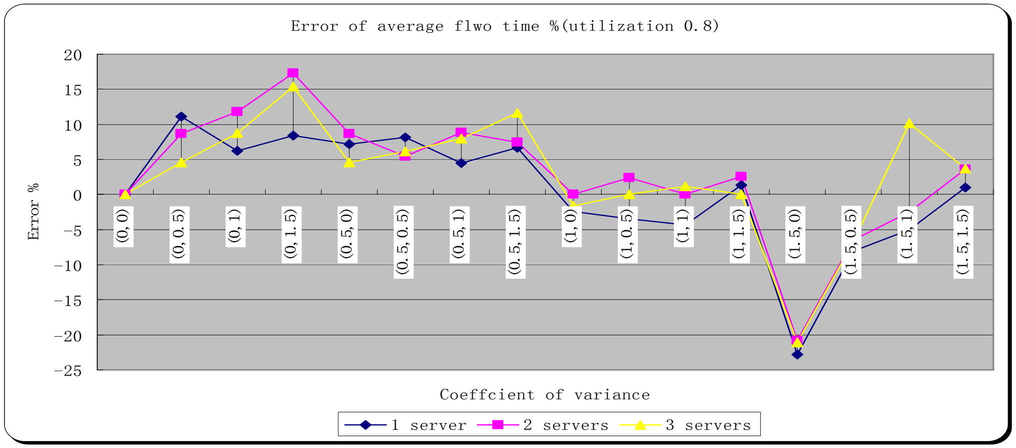
Note: all errors are relative error=(approx.-simu.)/simu.

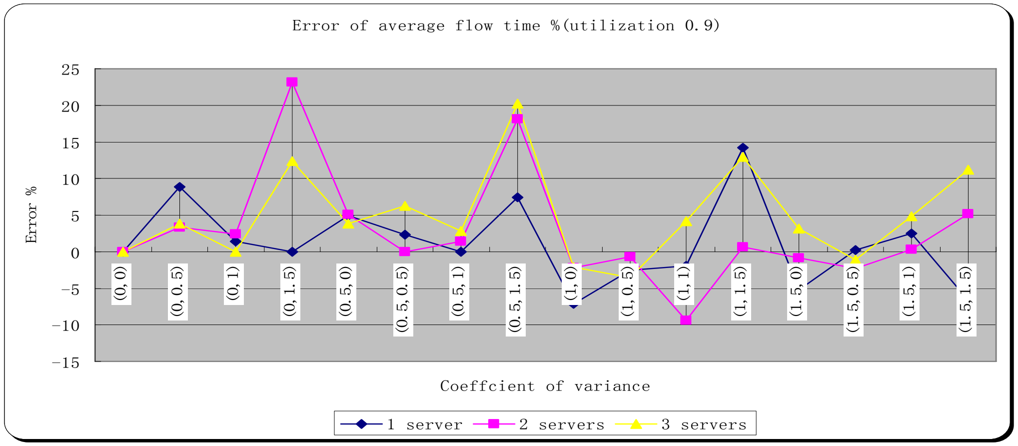


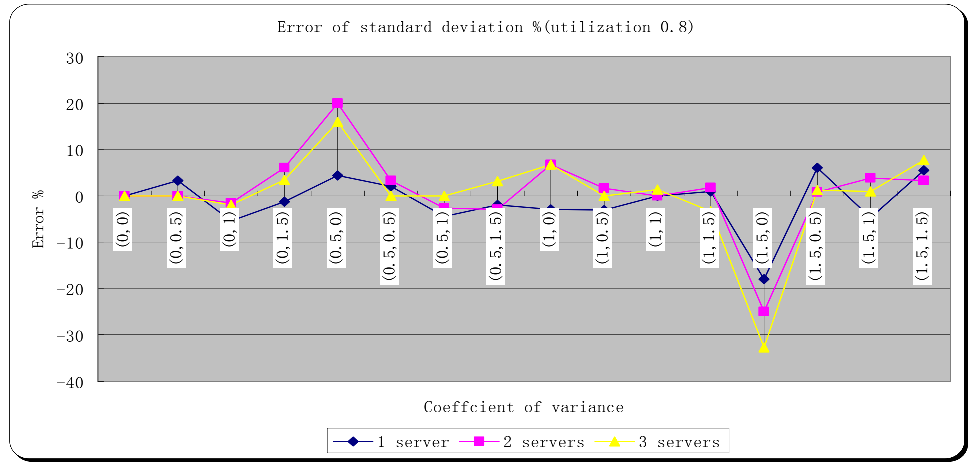




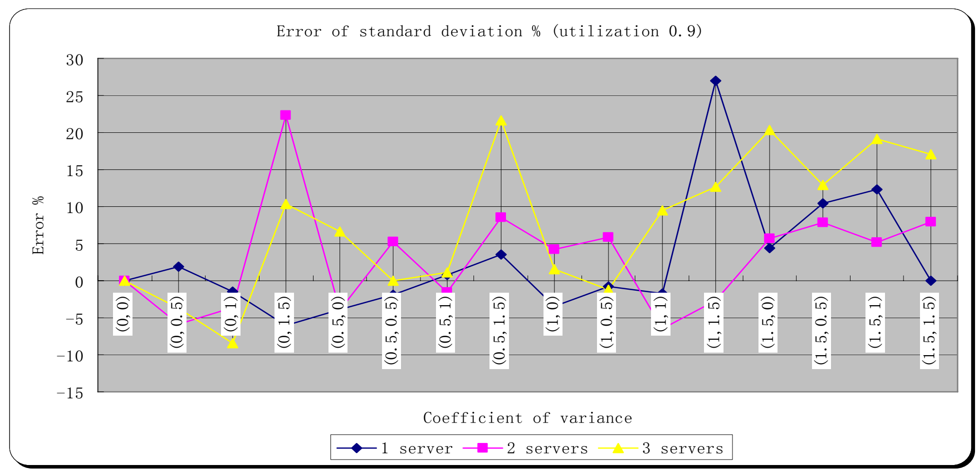
Different number of servers







Error of standard deviation %(utilization 0.8)
30
20
10
0
-10
-20
-30
-40
Error %
(0,0)
(0,0.5)
(0,1)
(0,1.5)
(0.5,0)
(0.5,0.5)
(0.5,1)
(0.5,1.5)
(1,0)
(1,0.5)
(1,1)
(1,1.5)
(1.5,0)
(1.5,0.5)
(1.5,1)
(1.5,1.5)
Coeffcient of variance
1 server
2 servers
3 servers



Error of standard deviation % (utilization 0.9)
30
25
20
15
10
5
0
-5
-10
-15
Error %
(0,0)
(0,0.5)
(0,1)
(0,1.5)
(0.5,0)
(0.5,0.5)
(0.5,1)
(0.5,1.5)
(1,0)
(1,0.5)
(1,1)
(1,1.5)
(1.5,0)
(1.5,0.5)
(1.5,1)
(1.5,1.5)
Coefficient of variance
1 server
2 servers
3 servers

Test Covariance of Two que(including 1,2,and 3 servers)

Note:all errors are relative error=(approx-simu.)/simu.

Case1	CVa=0	CVs=0						arrival mean	service mean
utilization=0.8					Total				0.2
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		
	1	mean	0.2	0.2	0.4	0.4	0.4		
	lamda=4	std	0	0	0	0	0	0.25	
	mu=5								
	2	mean	0.2	0.2	0.4	0.4	0.4		
	lamda=8	std	0	0	0	0	0	0.125	
	mu=5								
	3	mean	0.2	0.2	0.4	0.4	0.4		
	lamda=12	std	0	0	0	0	0	0.08333	
	mu=5								

Case2	CVa=0	CVs=0.5						arrival mean	service
utilization=0.8									scale=0.05
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		shape=4
	1	mean	0.3	0.36	0.66	0.58	0.58		
	lamda=4	std	0.16	0.22	0.27	0.23	0.22	0.25	
	mu=5								
	2	mean	0.25	0.3	0.55	0.49	0.48		
	lamda=8	std	0.12	0.17	0.21	0.17	0.17	0.125	
	mu=5								
	3	mean	0.23	0.27	0.50	0.45	0.45		
	lamda=12	std	0.11	0.14	0.18	0.15	0.15	0.08333	
	mu=5								

Case3	CVa=0	CVs=1						arrival mean	service
utilization=0.8									mean=0.2
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		
	1	mean	0.6	0.86	1.46	1.28	1.29		
	lamda=4	std	0.53	0.83	0.99	0.87	0.87	0.25	
	mu=5								
	2	mean	0.38	0.5	0.88	0.79	0.76		
	lamda=8	std	0.31	0.44	0.54	0.47	0.5	0.125	
	mu=5								
	3	mean	0.31	0.38	0.69	0.62	0.61		
	lamda=12	std	0.25	0.33	0.41	0.36	0.36	0.08333	
	mu=5								

Case4	CVa=0	CVs=1.5						arrival mean	service
utilization=0.8									scale=0.45
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		shape=0.44
	1	mean	1.1	1.68	2.78	2.44	2.43		
	lamda=4	std	1.14	1.83	2.16	1.75	1.63	0.25	
	mu=5								
	2	mean	0.61	0.83	1.44	1.18	1.27		
	lamda=8	std	0.62	0.88	1.08	0.87	0.92	0.125	
	mu=5								
	3	mean	0.45	0.57	1.02	0.9	0.9		
	lamda=12	std	0.46	0.61	0.76	0.67	0.67	0.08333	
	mu=5								

Case5	CVa=0.5	CVs=0						arrival	service mean
utilization=0.8									0.2
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		
	1	mean	0.3	0.24	0.54	0.48	0.47	scale=0.0625	
	lamda=4	std	0.12	0.04	0.13	0.11	0.1	shape=4	
	mu=5								
	2	mean	0.25	0.25	0.5	0.43	0.43	scale=0.03125	
	lamda=8	std	0.06	0.07	0.09	0.05	0.06	shape=4	
	mu=5								
	3	mean	0.23	0.24	0.47	0.42	0.42	scale=0.0208	
	lamda=12	std	0.04	0.06	0.07	0.03	0.03	shape=4	
	mu=5								

Case6	CVa=0.5	CVs=0.5						arrival	service
utilization=0.8									scale=0.05
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		shape=4
	1	mean	0.4	0.4	0.8	0.72	0.75	scale=0.0625	
	lamda=4	std	0.26	0.26	0.37	0.34	0.37	shape=4	
	mu=5								
	2	mean	0.29	0.32	0.61	0.55	0.55	scale=0.03125	
	lamda=8	std	0.16	0.18	0.24	0.21	0.22	shape=4	
	mu=5								
	3	mean	0.26	0.28	0.54	0.5	0.5	scale=0.0208	
	lamda=12	std	0.13	0.15	0.2	0.18	0.16	shape=4	
	mu=5								

Case7	CVa=0.5	CVs=1						arrival	service
utilization=0.8									mean=0.2
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		
	1	mean	0.7	0.89	1.59	1.47	1.51	scale=0.0625	
	lamda=4	std	0.64	0.87	1.08	0.97	1.04	shape=4	
	mu=5								
	2	mean	0.43	0.51	0.94	0.84	0.82	scale=0.03125	
	lamda=8	std	0.36	0.46	0.59	0.52	0.5	shape=4	
	mu=5								
	3	mean	0.34	0.39	0.73	0.67	0.68	scale=0.0208	
	lamda=12	std	0.28	0.34	0.44	0.39	0.42	shape=4	
	mu=5								

Case8	CVa=0.5	CVs=1.5						arrival	service
utilization=0.8									scale=0.45
	server No.		queue1	queue2	spreadsheet	simulat	simu.2		shape=0.44
	1	mean	1.2	1.71	2.91	2.7	2.82	scale=0.0625	
	lamda=4	std	1.26	1.88	2.26	2	2.2	shape=4	
	mu=5								
	2	mean	0.65	0.84	1.49	1.25	1.26	scale=0.03125	
	lamda=8	std	0.67	0.9	1.12	0.91	0.9	shape=4	
	mu=5								
	3	mean	0.48	0.58	1.06	0.93	0.96	scale=0.0208	
	lamda=12	std	0.49	0.62	0.79	0.68	0.69	shape=4	
	mu=5								

Case9	CVa=1	CVs=0						arrival mean	service mean
utilization=0.8									0.2
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		
	1	mean	0.6	0.34	0.94	0.8	0.8		
	lamda=4	std	0.49	0.18	0.52	0.49	0.47	0.25	
	mu=5								
	2	mean	0.38	0.3	0.68	0.58	0.58		
	lamda=8	std	0.24	0.13	0.27	0.24	0.23	0.125	
	mu=5								
	3	mean	0.31	0.27	0.58	0.51	0.51		
	lamda=12	std	0.16	0.1	0.19	0.15	0.15	0.08333	
	mu=5								

Case10	CVa=1	CVs=0.5						arrival mean	service
utilization=0.8									scale=0.05
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		shape=4
	1	mean	0.7	0.51	1.21	1.16	1.17		
	lamda=4	std	0.62	0.39	0.73	0.72	0.7	0.25	
	mu=5								
	2	mean	0.43	0.36	0.79	0.77	0.75		
	lamda=8	std	0.32	0.24	0.4	0.4	0.38	0.125	
	mu=5								
	3	mean	0.34	0.31	0.65	0.64	0.62		
	lamda=12	std	0.22	0.18	0.28	0.29	0.28	0.08333	
	mu=5								

Case11	CVa=1	CVs=1						arrival mean	service
utilization=0.8									mean=0.2
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		
	1	mean	1	1	2	2.2	2		
	lamda=4	std	1	1	1.41	1.55	1.31	0.25	
	mu=5								
	2	mean	0.56	0.56	1.12	1.1	1.07		
	lamda=8	std	0.52	0.52	0.74	0.72	0.69	0.125	
	mu=5								
	3	mean	0.42	0.42	0.84	0.8	0.81		
	lamda=12	std	0.37	0.37	0.52	0.5	0.51	0.08333	
	mu=5								

Case12	CVa=1	CVs=1.5						arrival mean	service
utilization=0.8									scale=0.45
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		shape=0.44
	1	mean	1.5	1.82	3.32	3.09	3.03		
	lamda=4	std	1.62	2.01	2.58	2.2	2.22	0.25	
	mu=5								
	2	mean	0.79	0.89	1.68	1.57	1.54		
	lamda=8	std	0.84	0.96	1.28	1.18	1.12	0.125	
	mu=5								
	3	mean	0.56	0.61	1.17	1.08	1.14		
	lamda=12	std	0.59	0.66	0.89	0.82	0.86	0.08333	
	mu=5								

Case13	CVa=1.5	CVs=0						arrival	service mean
utilization=0.8									0.2
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		
	1	mean	1.1	0.52	1.62	1.38	1.4	scale=0.563	
	lamda=4	std	1.1	0.4	1.17	1.04	1.08	shape=0.444	
	mu=5								
	2	mean	0.61	0.38	0.99	0.83	0.9	scale=0.282	
	lamda=8	std	0.51	0.24	0.56	0.49	0.58	shape=0.444	
	mu=5								
	3	mean	0.45	0.32	0.77	0.67	0.69	scale=0.188	
	lamda=12	std	0.35	0.17	0.39	0.32	0.34	shape=0.444	
	mu=5								

Case14	CVa=1.5	CVs=0.5						arrival	service
utilization=0.8									scale=0.05
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		shape=4
	1	mean	1.2	0.69	1.89	2.03	1.83	scale=0.563	
	lamda=4	std	1.23	0.61	1.37	1.56	1.38	shape=0.444	
	mu=5								
	2	mean	0.65	0.45	1.1	1.06	1.05	scale=0.282	
	lamda=8	std	0.61	0.34	0.69	0.68	0.64	shape=0.444	
	mu=5								
	3	mean	0.48	0.36	0.84	0.84	0.82	scale=0.188	
	lamda=12	std	0.41	0.24	0.48	0.47	0.45	shape=0.444	
	mu=5								

Case15	CVa=1.5	CVs=1						arrival	service
utilization=0.8									mean=0.2
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		
	1	mean	1.5	1.18	2.68	2.95	2.75	scale=0.563	
	lamda=4	std	1.6	1.22	2.01	2.52	2	shape=0.444	
	mu=5								
	2	mean	0.79	0.64	1.43	1.38	1.41	scale=0.282	
	lamda=8	std	0.81	0.62	1.02	0.98	0.94	shape=0.444	
	mu=5								
	3	mean	0.56	0.47	1.03	1.06	1.12	scale=0.188	
	lamda=12	std	0.55	0.43	0.7	0.7	0.77	shape=0.444	
	mu=5								

Case16	CVa=1.5	CVs=1.5						arrival	service
utilization=0.8									scale=0.45
	server No.		queue1	queue2	spreadsheet	simu.1	simu.2		shape=0.44
	1	mean	2	2	4	4.6	4.5	scale=0.563	
	lamda=4	std	2.22	2.22	3.14	3.56	3.79	shape=0.444	
	mu=5								
	2	mean	1.01	0.97	1.98	2.08	1.85	scale=0.282	
	lamda=8	std	1.12	1.07	1.55	1.68	1.39	shape=0.444	
	mu=5								
	3	mean	0.7	0.66	1.36	1.36	1.34	scale=0.188	
	lamda=12	std	0.77	0.72	1.05	1.01	1.03	shape=0.444	
	mu=5								

Single queue without priority

Note:all errors are relative error=(approx.-simu.)/simu.							
Case1	CVa=0	CVs=0				arrival mean	service mean
utilization=0.8							0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	0.2	0.2	0.2		
	lamda=4	std	0	0	0	0.25	
	mu=5						
	2	mean	0.2	0.2	0.2		
	lamda=8	std	0	0	0	0.125	
	mu=5						
	3	mean	0.2	0.2	0.2		
	lamda=12	std	0	0	0	0.08333	
	mu=5						
Case2	CVa=0	CVs=0.5				arrival mean	service
utilization=0.8							scale=0.05
	server No.		spreadsheet	simulation1	simulation2		shape=4
	1	mean	0.3	0.27	0.27		
	lamda=4	std	0.16	0.16	0.15	0.25	
	mu=5						
	2	mean	0.25	0.23	0.23		
	lamda=8	std	0.12	0.12	0.12	0.125	
	mu=5						
	3	mean	0.23	0.22	0.22		
	lamda=12	std	0.11	0.11	0.11	0.08333	
	mu=5						
Case3	CVa=0	CVs=1				arrival mean	service
utilization=0.8							mean=0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	0.6	0.56	0.57		
	lamda=4	std	0.53	0.56	0.56	0.25	
	mu=5						
	2	mean	0.38	0.33	0.35		
	lamda=8	std	0.31	0.3	0.33	0.125	
	mu=5						
	3	mean	0.31	0.29	0.28		
	lamda=12	std	0.25	0.26	0.25	0.08333	
	mu=5						
Case4	CVa=0	CVs=1.5				arrival mean	service
utilization=0.8							scale=0.45
	server No.		spreadsheet	simulation1	simulation2		shape=0.44
	1	mean	1.1	0.99	1.04		
	lamda=4	std	1.14	1.11	1.2	0.25	
	mu=5						
	2	mean	0.61	0.52	0.52		
	lamda=8	std	0.62	0.59	0.58	0.125	
	mu=5						
	3	mean	0.45	0.4	0.38		
	lamda=12	std	0.46	0.45	0.44	0.08333	
	mu=5						

Case5	CVa=0.5	CVs=0				arrival	service mean
utilization=0.8							0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	0.3	0.28	0.28	scale=0.0625	
	lamda=4	std	0.12	0.11	0.12	shape=4	
	mu=5						
	2	mean	0.25	0.23	0.23	scale=0.03125	
	lamda=8	std	0.06	0.05	0.05	shape=4	
	mu=5						
	3	mean	0.23	0.22	0.22	scale=0.0208	
	lamda=12	std	0.04	0.035	0.034	shape=4	
	mu=5						
Case6	CVa=0.5	CVs=0.5				arrival	service
utilization=0.8							scale=0.05
	server No.		spreadsheet	simulation1	simulation2		shape=4
	1	mean	0.4	0.37	0.37	scale=0.0625	
	lamda=4	std	0.26	0.26	0.25	shape=4	
	mu=5						
	2	mean	0.29	0.28	0.27	scale=0.03125	
	lamda=8	std	0.16	0.16	0.15	shape=4	
	mu=5						
	3	mean	0.26	0.24	0.25	scale=0.0208	
	lamda=12	std	0.13	0.13	0.13	shape=4	
	mu=5						
Case7	CVa=0.5	CVs=1				arrival	service
utilization=0.8							mean=0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	0.7	0.68	0.66	scale=0.0625	
	lamda=4	std	0.64	0.68	0.66	shape=4	
	mu=5						
	2	mean	0.43	0.41	0.38	scale=0.03125	
	lamda=8	std	0.36	0.38	0.36	shape=4	
	mu=5						
	3	mean	0.34	0.32	0.31	scale=0.0208	
	lamda=12	std	0.28	0.29	0.27	shape=4	
	mu=5						
Case8	CVa=0.5	CVs=1.5				arrival	service
utilization=0.8							scale=0.45
	server No.		spreadsheet	simulation1	simulation2		shape=0.44
	1	mean	1.2	1.17	1.08	scale=0.0625	
	lamda=4	std	1.26	1.3	1.27	shape=4	
	mu=5						
	2	mean	0.65	0.61	0.6	scale=0.03125	
	lamda=8	std	0.67	0.69	0.69	shape=4	
	mu=5						
	3	mean	0.48	0.42	0.44	scale=0.0208	
	lamda=12	std	0.49	0.46	0.49	shape=4	
	mu=5						

Case9	CVa=1	CVs=0				arrival mean	service mean
utilization=0.8							0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	0.6	0.62	0.61		
	lamda=4	std	0.49	0.5	0.51	0.25	
	mu=5						
	2	mean	0.38	0.38	0.38		
	lamda=8	std	0.24	0.22	0.23	0.125	
	mu=5						
	3	mean	0.31	0.32	0.31		
	lamda=12	std	0.16	0.15	0.15	0.08333	
	mu=5						
Case10	CVa=1	CVs=0.5				arrival mean	service
utilization=0.8							scale=0.05
	server No.		spreadsheet	simulation1	simulation2		shape=4
	1	mean	0.7	0.72	0.73		
	lamda=4	std	0.62	0.63	0.65	0.25	
	mu=5						
	2	mean	0.43	0.42	0.42		
	lamda=8	std	0.32	0.32	0.31	0.125	
	mu=5						
	3	mean	0.34	0.34	0.34		
	lamda=12	std	0.22	0.22	0.22	0.08333	
	mu=5						
Case11	CVa=1	CVs=1				arrival mean	service
utilization=0.8							mean=0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	1	1.08	1.01		
	lamda=4	std	1	1.03	0.97	0.25	
	mu=5						
	2	mean	0.56	0.54	0.58		
	lamda=8	std	0.52	0.5	0.54	0.125	
	mu=5						
	3	mean	0.42	0.41	0.42		
	lamda=12	std	0.37	0.35	0.38	0.08333	
	mu=5						
Case12	CVa=1	CVs=1.5				arrival mean	service
utilization=0.8							scale=0.45
	server No.		spreadsheet	simulation1	simulation2		shape=0.44
	1	mean	1.5	1.45	1.51		
	lamda=4	std	1.62	1.7	1.51	0.25	
	mu=5						
	2	mean	0.79	0.79	0.75		
	lamda=8	std	0.84	0.84	0.81	0.125	
	mu=5						
	3	mean	0.56	0.54	0.58		
	lamda=12	std	0.59	0.58	0.64	0.08333	

Case13	CVa=1.5	CVs=0				arrival	service mean
utilization=0.8							0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	1.1	1.4	1.45	scale=0.563	
	lamda=4	std	1.1	1.31	1.37	shape=0.444	
	mu=5						
	2	mean	0.61	0.78	0.76	scale=0.282	
	lamda=8	std	0.54	0.72	0.72	shape=0.444	
	mu=5						
	3	mean	0.45	0.59	0.55	scale=0.188	
	lamda=12	std	0.35	0.54	0.5	shape=0.444	
	mu=5						
Case14	CVa=1.5	CVs=0.5				arrival	service
utilization=0.8							scale=0.05
	server No.		spreadsheet	simulation1	simulation2		shape=4
	1	mean	1.2	1.28	1.33	scale=0.563	
	lamda=4	std	1.23	1.13	1.19	shape=0.444	
	mu=5						
	2	mean	0.65	0.69	0.7	scale=0.282	
	lamda=8	std	0.61	0.6	0.61	shape=0.444	
	mu=5						
	3	mean	0.48	0.52	0.51	scale=0.188	
	lamda=12	std	0.41	0.41	0.4	shape=0.444	
	mu=5						
Case15	CVa=1.5	CVs=1				arrival	service
utilization=0.8							mean=0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	1.5	1.62	1.54	scale=0.563	
	lamda=4	std	1.6	1.66	1.7	shape=0.444	
	mu=5						
	2	mean	0.79	0.81	0.81	scale=0.282	
	lamda=8	std	0.81	0.76	0.8	shape=0.444	
	mu=5						
	3	mean	0.65	0.6	0.58	scale=0.188	
	lamda=12	std	0.55	0.55	0.54	shape=0.444	
	mu=5						
Case16	CVa=1.5	CVs=1.5				arrival	service
utilization=0.8							scale=0.45
	server No.		spreadsheet	simulation1	simulation2		shape=0.44
	1	mean	2	1.93	2.03	scale=0.563	
	lamda=4	std	2.22	2.04	2.17	shape=0.444	
	mu=5						
	2	mean	1.01	0.96	0.99	scale=0.282	
	lamda=8	std	1.12	1.08	1.09	shape=0.444	
	mu=5						
	3	mean	0.7	0.7	0.65	scale=0.188	
	lamda=12	std	0.77	0.76	0.67	shape=0.444	

Case17	CVa=0	CVs=0				arrival mean	service mean
utilization=0.9							0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	0.2	0.2	0.2		
	lamda=4.5	std	0	0	0	0.22	
	mu=5						
	2	mean	0.2	0.2	0.2		
	lamda=9	std	0	0	0	0.11	
	mu=5						
	3	mean	0.2	0.2	0.2		
	lamda=13.5	std	0	0	0	0.07	
	mu=5						
Case18	Cva=0	CVs=0.5				arrival mean	service
utilization=0.9							scale=0.05
	server No.		spreadsheet	simulation1	simulation2		shape=4
	1	mean	0.43	0.4	0.39		
	lamda=4.5	std	0.27	0.27	0.26	0.22	
	mu=5						
	2	mean	0.31	0.3	0.3		
	lamda=9	std	0.16	0.17	0.17	0.11	
	mu=5						
	3	mean	0.27	0.26	0.26		
	lamda=13.5	std	0.13	0.13	0.14	0.07	
	mu=5						
Case19	CVa=0	CVs=1				arrival mean	service
utilization=0.9							mean=0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	1.1	1.08	1.09		
	lamda=4.5	std	1.01	1.02	1.03	0.22	
	mu=5						
	2	mean	0.63	0.63	0.6		
	lamda=9	std	0.53	0.57	0.53	0.11	
	mu=5						
	3	mean	0.47	0.45	0.49		
	lamda=13.5	std	0.38	0.39	0.44	0.07	
	mu=5						
Case20	CVa=0	CVs=1.5				arrival mean	service
utilization=0.9							scale=0.45
	server No.		spreadsheet	simulation1	simulation2		shape=0.44
	1	mean	2.23	2.13	2.33		
	lamda=4.5	std	2.26	2.29	2.52	0.22	
	mu=5						
	2	mean	1.17	0.94	0.96		
	lamda=9	std	1.15	0.93	0.95	0.11	
	mu=5						
	3	mean	0.82	0.72	0.74		
	lamda=13.5	std	0.8	0.72	0.73	0.07	
	mu=5						

Case21	CVa=0.5	CVs=0				arrival	service mean
utilization=0.9							0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	0.43	0.4	0.42	scale=0.0555	
	lamda=4.5	std	0.25	0.24	0.28	shape=4	
	mu=5						
	2	mean	0.31	0.3	0.29	scale=0.02775	
	lamda=9	std	0.12	0.13	0.12	shape=4	
	mu=5						
	3	mean	0.27	0.26	0.26	scale=0.01853	
	lamda=13.5	std	0.08	0.07	0.08	shape=4	
	mu=5						
Case22	CVa=0.5	CVs=0.5				arrival	service
utilization=0.9							scale=0.05
	server No.		spreadsheet	simulation1	simulation2		shape=4
	1	mean	0.65	0.65	0.62	scale=0.0555	
	lamda=4.5	std	0.51	0.54	0.5	shape=4	
	mu=5						
	2	mean	0.41	0.4	0.42	scale=0.02775	
	lamda=9	std	0.27	0.28	0.29	shape=4	
	mu=5						
	3	mean	0.34	0.33	0.31	scale=0.01853	
	lamda=13.5	std	0.19	0.2	0.18	shape=4	
	mu=5						
Case23	CVa=0.5	CVs=1				arrival	service
utilization=0.9							mean=0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	1.33	1.36	1.3	scale=0.0555	
	lamda=4.5	std	1.26	1.28	1.22	shape=4	
	mu=5						
	2	mean	0.74	0.74	0.72	scale=0.02775	
	lamda=9	std	0.65	0.68	0.64	shape=4	
	mu=5						
	3	mean	0.54	0.53	0.52	scale=0.01853	
	lamda=13.5	std	0.46	0.45	0.46	shape=4	
	mu=5						
Case24	CVa=0.5	CVs=1.5				arrival	service
utilization=0.9							scale=0.45
	server No.		spreadsheet	simulation1	simulation2		shape=0.44
	1	mean	2.45	2.34	2.22	scale=0.0555	
	lamda=4.5	std	2.51	2.6	2.25	shape=4	
	mu=5						
	2	mean	1.27	1.05	1.1	scale=0.02775	
	lamda=9	std	1.27	1.12	1.22	shape=4	
	mu=5						
	3	mean	0.89	0.7	0.78	scale=0.01853	
	lamda=13.5	std	0.87	0.68	0.75	shape=4	
	mu=5						

Case25	CVa=1	CVs=0				arrival mean	service mean
utilization=0.9							0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	1.1	1.14	1.23		
	lamda=4.5	std	0.99	0.95	1.1	0.2222	
	mu=5						
	2	mean	0.63	0.66	0.63		
	lamda=9	std	0.49	0.49	0.45	0.1111	
	mu=5						
	3	mean	0.47	0.48	0.48		
	lamda=13.5	std	0.33	0.31	0.34	0.0741	
	mu=5						
Case26	CVa=1	CVs=0.5				arrival mean	service
utilization=0.9							scale=0.05
	server No.		spreadsheet	simulation1	simulation2		shape=4
	1	mean	1.33	1.35	1.38		
	lamda=4.5	std	1.25	1.25	1.27	0.2222	
	mu=5						
	2	mean	0.74	0.77	0.72		
	lamda=9	std	0.63	0.6	0.59	0.1111	
	mu=5						
	3	mean	0.54	0.57	0.55		
	lamda=13.5	std	0.42	0.43	0.42	0.0741	
	mu=5						
Case27	CVa=1	CVs=1				arrival mean	service
utilization=0.9							mean=0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	2	1.95	2.13		
	lamda=4.5	std	2	1.78	2.29	0.2222	
	mu=5						
	2	mean	1.06	1.16	1.18		
	lamda=9	std	1.01	1.14	1.02	0.1111	
	mu=5						
	3	mean	0.75	0.73	0.71		
	lamda=13.5	std	0.69	0.63	0.63	0.0741	
	mu=5						
Case28	CVa=1	CVs=1.5				arrival mean	service
utilization=0.9							scale=0.45
	server No.		spreadsheet	simulation1	simulation2		shape=0.44
	1	mean	3.13	2.65	2.83		
	lamda=4.5	std	3.25	2.38	2.74	0.2222	
	mu=5						
	2	mean	1.59	1.62	1.54		
	lamda=9	std	1.64	1.79	1.58	0.1111	
	mu=5						
	3	mean	1.09	0.99	0.94		
	lamda=13.5	std	1.11	1.05	0.92	0.0741	
	mu=5						

Case29	CVa=1.5	CVs=0				arrival	service mean
utilization=0.9							0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	2.23	2.21	2.5	scale=0.5	
	lamda=4.5	std	2.24	1.87	2.42	shape=0.444	
	mu=5						
	2	mean	1.17	1.25	1.11	scale=0.25	
	lamda=9	std	1.11	1.18	0.92	shape=0.444	
	mu=5						
	3	mean	0.82	0.8	0.79	scale=0.167	
	lamda=13.5	std	0.74	0.63	0.6	shape=0.444	
	mu=5						
Case30	CVa=1.5	CVs=0.5				arrival	service
utilization=0.9							scale=0.05
	server No.		spreadsheet	simulation1	simulation2		shape=4
	1	mean	2.45	2.32	2.57	scale=0.5	
	lamda=4.5	std	2.49	2.2	2.31	shape=0.444	
	mu=5						
	2	mean	1.27	1.32	1.28	scale=0.25	
	lamda=9	std	1.24	1.18	1.12	shape=0.444	
	mu=5						
	3	mean	0.89	0.89	0.91	scale=0.167	
	lamda=13.5	std	0.83	0.71	0.76	shape=0.444	
	mu=5						
Case31	CVa=1.5	CVs=1				arrival	service
utilization=0.9							mean=0.2
	server No.		spreadsheet	simulation1	simulation2		
	1	mean	3.13	3.38	2.73	scale=0.5	
	lamda=4.5	std	3.24	3.24	2.53	shape=0.444	
	mu=5						
	2	mean	1.59	1.56	1.61	scale=0.25	
	lamda=9	std	1.62	1.53	1.55	shape=0.444	
	mu=5						
	3	mean	1.09	0.96	1.12	scale=0.167	
	lamda=13.5	std	1.09	0.88	0.95	shape=0.444	
	mu=5						
Case32	CVa=1.5	CVs=1.5				arrival	service
utilization=0.9							scale=0.45
	server No.		spreadsheet	simulation1	simulation2		shape=0.44
	1	mean	4.25	4.62	4.48	scale=0.5	
	lamda=4.5	std	4.49	4.18	4.8	shape=0.444	
	mu=5						
	2	mean	2.13	2.08	1.97	scale=0.25	
	lamda=9	std	2.25	2.03	2.14	shape=0.444	
	mu=5						
	3	mean	1.44	1.31	1.28	scale=0.167	
	lamda=13.5	std	1.51	1.28	1.3	shape=0.444	
	mu=5						

Single queue with priority (4 classes)

Note:all errors are relative error=(approx.-simu.)/simu.

Case1	CVa=0	CVs=0	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.2	0.2	0.2	0.2
std	0	0	0	0
Simulation	P1	P2	P3	P4
mean	0.211872	0.421872	0.6118716	0.59957
std	0.063454	0.063454	0.0634539	0.430448
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.2	0.2	0.2	0.2
std	0	0	0	0
Simulation	P1	P2	P3	P4
mean	0.2	0.2	0.4	0.405506
std	0	0	0	0.044525
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.2	0.2	0.2	0.2
std	0	0	0	0
Simulation	P1	P2	P3	P4
mean	0.2	0.2	0.2667	0.4
std	0	0	6.569E-09	0.020206
Case2	CVa=0	CVs=0.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.225	0.241667	0.2833333	0.45
std	0.104583	0.112268	0.1428869	0.322102
Simulation	P1	P2	P3	P4
mean	0.219033	0.409336	0.6442162	0.846505
std	0.10866	0.152104	0.242687	0.309154
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.211307	0.218845	0.2376902	0.313071
std	0.101121	0.103084	0.1118269	0.180409
Simulation	P1	P2	P3	P4
mean	0.198656	0.236541	0.3592864	0.480883
std	0.090296	0.119992	0.1166144	0.164845
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.206927	0.211545	0.2230895	0.269269
std	0.10048	0.101329	0.1052153	0.140115
Simulation	P1	P2	P3	P4

Case3	CVa=0	CVs=1	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.3	0.366667	0.5333333	1.2
std	0.234521	0.285774	0.4546061	1.240967
Simulation	P1	P2	P3	P4
mean	0.286009	0.488804	0.7596804	1.164371
std	0.249155	0.328093	0.4564877	0.810955
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.245228	0.27538	0.3507609	0.652283
std	0.208824	0.223654	0.2829913	0.633053
Simulation	P1	P2	P3	P4
mean	0.241539	0.295359	0.4763156	0.867042
std	0.256341	0.233028	0.3785754	0.948618
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.227707	0.246179	0.292358	0.477074
std	0.203817	0.210431	0.2390066	0.440587
Simulation	P1	P2	P3	P4
mean	0.20004	0.265297	0.3249968	0.435377
std	0.185152	0.221045	0.2544316	0.314877

Case4	CVa=0	CVs=1.5	1 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.425	0.575	0.95	2.45
std	0.407354	0.548578	0.9663074	2.771958
Simulation	P1	P2	P3	P4
mean	0.417973	0.749632	1.3659989	3.772663
std	0.475782	0.810152	1.5372437	4.10759
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.301764	0.369606	0.539212	1.217636
std	0.329034	0.375142	0.5412259	1.384316
Simulation	P1	P2	P3	P4
mean	0.246474	0.348504	0.5398057	0.91892
std	0.263432	0.377178	0.5522902	0.931734
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.262342	0.303903	0.4078055	0.823417
std	0.312733	0.334175	0.4203463	0.932855
Simulation	P1	P2	P3	P4
mean	0.241959	0.285798	0.3610528	0.5361
std	0.316848	0.318982	0.4030489	0.549791

Case5	CVa=0.5	CVs=0	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.225	0.241667	0.2833333	0.45
std	0.030619	0.051031	0.1020621	0.306186
Simulation	P1	P2	P3	P4
mean	0.239293	0.293085	0.3016703	0.481443
std	0.035945	0.083793	0.1695819	0.355393
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.211307	0.218845	0.2376902	0.313071
std	0.015016	0.025026	0.0500525	0.150158
Simulation	P1	P2	P3	P4
mean	0.231336	0.238988	0.2453326	0.299448
std	0.021602	0.030083	0.067163	0.148197
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.206927	0.211545	0.2230895	0.269269
std	0.009814	0.016357	0.0327148	0.098144
Simulation	P1	P2	P3	P4
mean	0.217197	0.227464	0.2284022	0.250035
std	0.027098	0.041592	0.0465466	0.101548
Case6	CVa=0.5	CVs=0.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.25	0.283333	0.3666667	0.7
std	0.11726	0.142887	0.227303	0.620484
Simulation	P1	P2	P3	P4
mean	0.289248	0.320128	0.3901203	0.730306
std	0.136082	0.178239	0.2553792	0.819445
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.222614	0.23769	0.2753804	0.426141
std	0.104412	0.111827	0.1414956	0.316527
Simulation	P1	P2	P3	P4
mean	0.224465	0.248468	0.2997347	0.420767
std	0.120823	0.134206	0.1675979	0.387812
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.213854	0.22309	0.246179	0.338537
std	0.101908	0.105215	0.1195033	0.220294
Simulation	P1	P2	P3	P4
mean	0.215772	0.221832	0.2435648	0.279026
std	0.106739	0.106885	0.1164263	0.185597

Case7	CVa=0.5	CVs=1	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.325	0.408333	0.6166667	1.45
std	0.251868	0.324198	0.5481028	1.54394
Simulation	P1	P2	P3	P4
mean	0.310922	0.395537	0.5729503	0.959754
std	0.257831	0.312836	0.6087902	1.25052
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.256535	0.294226	0.3884511	0.765353
std	0.213628	0.235919	0.3203612	0.77697
Simulation	P1	P2	P3	P4
mean	0.247308	0.294943	0.3732238	0.559284
std	0.196765	0.243761	0.3813483	0.608672
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.234634	0.257724	0.3154475	0.546343
std	0.205932	0.216077	0.2583727	0.529913
Simulation	P1	P2	P3	P4
mean	0.233475	0.248313	0.3114381	0.411018
std	0.206281	0.198719	0.2492543	0.431667

Case8	CVa=0.5	CVs=1.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.45	0.616667	1.0333333	2.7
std	0.428661	0.59196	1.0637982	3.076524
Simulation	P1	P2	P3	P4
mean	0.463191	0.576829	1.1315395	1.096745
std	0.487139	0.718707	1.9962061	1.361113
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.313071	0.388451	0.5769022	1.330707
std	0.335481	0.390681	0.5835455	1.53125
Simulation	P1	P2	P3	P4
mean	0.32084	0.406615	0.5772726	1.170426
std	0.332583	0.425013	0.6907666	1.605913
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.269269	0.315448	0.430895	0.892685
std	0.315646	0.341696	0.443876	1.026271
Simulation	P1	P2	P3	P4
mean	0.291163	0.350127	0.4111786	0.884789
std	0.335609	0.439516	0.4367234	1.29503

Case9	CVa=1	CVs=0	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.3	0.366667	0.5333333	1.2
std	0.122474	0.204124	0.4082483	1.224745
Simulation	P1	P2	P3	P4
mean	0.299088	0.370729	0.496491	0.818307
std	0.096435	0.202558	0.3627338	0.772839
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.245228	0.27538	0.3507609	0.652283
std	0.060063	0.100105	0.20021	0.60063
Simulation	P1	P2	P3	P4
mean	0.249426	0.265631	0.3396764	0.541722
std	0.05786	0.089649	0.1721694	0.44539
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.227707	0.246179	0.292358	0.477074
std	0.039258	0.06543	0.1308592	0.392578
Simulation	P1	P2	P3	P4
mean	0.237636	0.247519	0.2911296	0.434771
std	0.04396	0.060049	0.1246644	0.38905
Case10	CVa=1	CVs=0.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.325	0.408333	0.6166667	1.45
std	0.182859	0.274051	0.520016	1.534194
Simulation	P1	P2	P3	P4
mean	0.339033	0.383827	0.5866874	1.530921
std	0.189318	0.253608	0.5339369	1.519861
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.256535	0.294226	0.3884511	0.765353
std	0.125047	0.160181	0.269502	0.757418
Simulation	P1	P2	P3	P4
mean	0.254833	0.284686	0.376308	0.90907
std	0.118098	0.158247	0.2770266	0.936797
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.234634	0.257724	0.3154475	0.546343
std	0.111392	0.129186	0.1917198	0.500808
Simulation	P1	P2	P3	P4
mean	0.252154	0.252563	0.2990125	0.467583
std	0.113152	0.133986	0.1817725	0.427106

Case11	CVa=1	CVs=1	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.4	0.533333	0.8666667	2.2
std	0.316228	0.454606	0.8406347	2.457641
Simulation	P1	P2	P3	P4
mean	0.405728	0.5425	0.873978	1.928142
std	0.308001	0.456907	0.8581763	1.956361
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.290457	0.350761	0.5015218	1.104565
std	0.233303	0.282991	0.4475893	1.217795
Simulation	P1	P2	P3	P4
mean	0.287123	0.346092	0.4775061	0.976939
std	0.231844	0.284629	0.46555	1.151725
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.255415	0.292358	0.384716	0.754148
std	0.21486	0.239007	0.3293882	0.810228
Simulation	P1	P2	P3	P4
mean	0.254416	0.308859	0.3753421	0.829999
std	0.199321	0.241677	0.321248	0.827653
Case12	CVa=1	CVs=1.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.525	0.741667	1.2833333	3.45
std	0.498435	0.728083	1.3603002	3.99171
Simulation	P1	P2	P3	P4
mean	0.483105	0.555762	0.8998337	3.250077
std	0.44566	0.562208	1.1395663	3.78665
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.346992	0.444986	0.6899729	1.669919
std	0.357917	0.442546	0.7165108	1.974966
Simulation	P1	P2	P3	P4
mean	0.353543	0.495611	0.7798234	2.242156
std	0.346427	0.536793	0.8528775	2.208839
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.290049	0.350082	0.5001636	1.100491
std	0.326004	0.367721	0.5204553	1.310673
Simulation	P1	P2	P3	P4
mean	0.281916	0.295953	0.5934931	1.216568
std	0.330241	0.351605	0.5356382	1.523726

Case13	CVa=1.5	CVs=0	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.425	0.575	0.95	2.45
std	0.275568	0.459279	0.9185587	2.755676
Simulation	P1	P2	P3	P4
mean	0.392597	0.372905	0.5831058	2.161167
std	0.222596	0.359526	0.8689836	3.071262
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.301764	0.369606	0.539212	1.217636
std	0.135142	0.225236	0.4504725	1.351418
Simulation	P1	P2	P3	P4
mean	0.381581	0.38141	0.6493988	1.919903
std	0.165856	0.186664	0.5790002	2.708002
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.262342	0.303903	0.4078055	0.823417
std	0.08833	0.147217	0.2944333	0.8833
Simulation	P1	P2	P3	P4
mean	0.242339	0.26292	0.3135201	0.917638
std	0.051781	0.067798	0.1265887	0.858194
Case14	CVa=1.5	CVs=0.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.45	0.616667	1.0333333	2.7
std	0.322102	0.520016	1.025508	3.063495
Simulation	P1	P2	P3	P4
mean	0.367929	0.40494	0.6406149	1.172886
std	0.204763	0.201673	0.646746	1.16096
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.313071	0.388451	0.5769022	1.330707
std	0.180409	0.269502	0.5104168	1.504901
Simulation	P1	P2	P3	P4
mean	0.322817	0.315642	0.3829598	0.633224
std	0.197428	0.181283	0.2957532	0.712546
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.269269	0.315448	0.430895	0.892685
std	0.140115	0.19172	0.3420905	0.986526
Simulation	P1	P2	P3	P4
mean	0.25401	0.262426	0.3180955	0.587175
std	0.119447	0.129798	0.1966103	0.60165

Case15	CVa=1.5	CVs=1	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.525	0.741667	1.2833333	3.45
std	0.445463	0.692895	1.3417961	3.985442
Simulation	P1	P2	P3	P4
mean	0.434747	0.507944	0.6797057	2.268456
std	0.388574	0.40074	0.5787079	2.802311
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.346992	0.444986	0.6899729	1.669919
std	0.279473	0.381899	0.6807259	1.962267
Simulation	P1	P2	P3	P4
mean	0.34778	0.394706	0.5731842	1.154074
std	0.24687	0.268367	0.5273045	1.341609
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.290049	0.350082	0.5001636	1.100491
std	0.237231	0.291922	0.4699721	1.291458
Simulation	P1	P2	P3	P4
mean	0.25255	0.339212	0.3858216	0.997509
std	0.194693	0.269565	0.3765349	1.156249
Case16	CVa=1.5	CVs=1.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.65	0.95	1.7	4.7
std	0.627495	0.966307	1.861451	5.519511
Simulation	P1	P2	P3	P4
mean	0.553842	0.59179	1.1448214	2.809364
std	0.410002	0.556641	1.3484565	3.720443
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.403527	0.539212	0.878424	2.235272
std	0.403798	0.541226	0.9495799	2.719433
Simulation	P1	P2	P3	P4
mean	0.406669	0.326128	0.5855596	1.066846
std	0.385279	0.314208	0.5913283	1.238229
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.324683	0.407806	0.6156111	1.446833
std	0.34815	0.420346	0.6608811	1.791891
Simulation	P1	P2	P3	P4
mean	0.26502	0.245567	0.4562427	0.729457
std	0.278797	0.29022	0.5871684	1.077554

Case17	CVa=0	CVs=0	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.2	0.2	0.2	0.2
std	0	0	0	0
Simulation	P1	P2	P3	P4
mean	0.25433	0.45433	0.6543298	0.718706
std	0.062265	0.062	0.0685697	0.458565
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.2	0.2	0.2	0.2
std	0	0	0	0
Simulation	P1	P2	P3	P4
mean	0.220837	0.273466	0.4209183	0.428662
std	0.041301	0.071089	0.0593475	0.279991
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.2	0.2	0.2	0.2
std	0	0	0	0
Simulation	P1	P2	P3	P4
mean	0.201862	0.215721	0.2200975	0.245627
std	0.009623	0.035972	0.0460026	0.082579
Case18	CVa=0	CVs=0.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.229032	0.252786	0.3258741	0.892308
std	0.105025	0.115782	0.1713629	0.77188
Simulation	P1	P2	P3	P4
mean	0.270622	0.467019	0.677974	1.081169
std	0.133439	0.168429	0.2013339	0.900428
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.213845	0.225172	0.2600259	0.530143
std	0.101267	0.104128	0.1216217	0.393641
Simulation	P1	P2	P3	P4
mean	0.223477	0.234519	0.4051202	0.64364
std	0.107882	0.114054	0.137443	0.564057
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.208869	0.216125	0.2384511	0.411481
std	0.100559	0.101835	0.11003	0.271522
Simulation	P1	P2	P3	P4
mean	0.210789	0.214162	0.2350976	0.229409
std	0.10701	0.114281	0.1120375	0.111011

Case19	CVa=0	CVs=1	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.316129	0.411144	0.7034965	2.969231
std	0.237661	0.30739	0.5914761	3.068026
Simulation	P1	P2	P3	P4
mean	0.308818	0.521276	0.7846929	1.301564
std	0.265183	0.347632	0.5144313	1.305793
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.255379	0.300689	0.4401037	1.520571
std	0.209949	0.231264	0.341569	1.535985
Simulation	P1	P2	P3	P4
mean	0.245847	0.321209	0.5240144	1.242406
std	0.20494	0.240305	0.4901503	1.338926
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.235474	0.264499	0.3538043	1.045924
std	0.204433	0.214307	0.2714875	1.029364
Simulation	P1	P2	P3	P4
mean	0.258398	0.258496	0.3729887	0.613405
std	0.223054	0.210031	0.2529857	0.550072
Case20	CVa=0	CVs=1.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.46129	0.675073	1.3328671	6.430769
std	0.416466	0.604855	1.2878607	6.894904
Simulation	P1	P2	P3	P4
mean	0.413688	0.681846	1.2211994	2.605109
std	0.410953	0.612936	1.2951927	2.738728
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.324602	0.42655	0.7402334	3.171284
std	0.332638	0.397816	0.6914759	3.439651
Simulation	P1	P2	P3	P4
mean	0.328345	0.428728	0.7540351	2.384726
std	0.395839	0.385142	0.6189385	1.734248
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.279817	0.345122	0.5460597	2.103328
std	0.314765	0.346421	0.5105231	2.291653
Simulation	P1	P2	P3	P4
mean	0.263779	0.395711	0.607777	2.315865
std	0.306918	0.410444	0.6589621	2.279692

Case21	CVa=0.5	CVs=0	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.229032	0.252786	0.3258741	0.892308
std	0.032096	0.058357	0.1391591	0.765375
Simulation	P1	P2	P3	P4
mean	0.285024	0.338645	0.3840257	0.757988
std	0.067628	0.129396	0.185811	0.557798
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.213845	0.225172	0.2600259	0.530143
std	0.015966	0.029029	0.0692231	0.380727
Simulation	P1	P2	P3	P4
mean	0.247041	0.262827	0.2947916	0.45997
std	0.045249	0.063431	0.1041193	0.373317
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.208869	0.216125	0.2384511	0.411481
std	0.010586	0.019247	0.0458976	0.252437
Simulation	P1	P2	P3	P4
mean	0.230223	0.232824	0.2489209	0.353592
std	0.039142	0.041644	0.0567059	0.248268
Case22	CVa=0.5	CVs=0.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.258065	0.305572	0.4517483	1.584615
std	0.118831	0.153695	0.2957381	1.534013
Simulation	P1	P2	P3	P4
mean	0.303319	0.337835	0.4514875	0.900149
std	0.130515	0.172811	0.3549941	1.160268
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.227689	0.250344	0.3200519	0.860285
std	0.104975	0.115632	0.1707845	0.767992
Simulation	P1	P2	P3	P4
mean	0.227443	0.265152	0.3302868	0.67707
std	0.104211	0.130697	0.1768526	0.526552
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.217737	0.232249	0.2769021	0.622962
std	0.102217	0.107153	0.1357437	0.514682
Simulation	P1	P2	P3	P4
mean	0.219739	0.244827	0.2841976	0.468018
std	0.100186	0.115414	0.1850602	0.498596

Case23	CVa=0.5	CVs=1	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.345161	0.46393	0.8293706	3.661538
std	0.256426	0.353749	0.7239691	3.832097
Simulation	P1	P2	P3	P4
mean	0.347263	0.418196	0.7337323	1.228244
std	0.247432	0.363178	0.5446971	1.272978
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.269223	0.325861	0.5001297	1.850713
std	0.215343	0.247118	0.3997449	1.914113
Simulation	P1	P2	P3	P4
mean	0.274028	0.324598	0.5439006	1.289341
std	0.211772	0.264654	0.3693441	1.175023
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.244343	0.280623	0.3922554	1.257405
std	0.206886	0.221949	0.3044089	1.277931
Simulation	P1	P2	P3	P4
mean	0.235355	0.333746	0.3971582	1.200525
std	0.186616	0.241444	0.2741108	1.213498
Case24	CVa=0.5	CVs=1.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.490323	0.727859	1.4587413	7.123077
std	0.439338	0.656166	1.4235607	7.659627
Simulation	P1	P2	P3	P4
mean	0.467051	0.665967	1.1827305	3.202776
std	0.427144	0.597776	1.4459406	3.015529
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.338447	0.451722	0.8002594	3.501427
std	0.33984	0.417455	0.7544427	3.819072
Simulation	P1	P2	P3	P4
mean	0.335625	0.45148	0.7134679	2.252362
std	0.335377	0.46993	0.7966012	2.328145
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.288686	0.361246	0.5845107	2.314809
std	0.31813	0.356435	0.5483238	2.542132
Simulation	P1	P2	P3	P4
mean	0.282286	0.346959	0.5300306	0.99033
std	0.323125	0.3626	0.4772232	1.591662

Case25	CVa=1	CVs=0	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.316129	0.411144	0.7034965	2.969231
std	0.128385	0.233428	0.5566363	3.0615
Simulation	P1	P2	P3	P4
mean	0.323103	0.416838	0.7379082	1.894836
std	0.111545	0.232913	0.6135549	1.892755
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.255379	0.300689	0.4401037	1.520571
std	0.063864	0.116116	0.2768924	1.522908
Simulation	P1	P2	P3	P4
mean	0.257254	0.293951	0.4343899	1.156408
std	0.049812	0.090999	0.2956724	1.316063
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.235474	0.264499	0.3538043	1.045924
std	0.042344	0.07699	0.1835904	1.009747
Simulation	P1	P2	P3	P4
mean	0.240641	0.265142	0.3589679	0.944023
std	0.043011	0.076505	0.1910533	0.902559
Case26	CVa=1	CVs=0.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.345161	0.46393	0.8293706	3.661538
std	0.189088	0.308445	0.7029447	3.828181
Simulation	P1	P2	P3	P4
mean	0.343659	0.47049	0.8514605	3.070389
std	0.179172	0.308333	0.7761197	2.965379
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.269223	0.325861	0.5001297	1.850713
std	0.127956	0.176259	0.3602721	1.90626
Simulation	P1	P2	P3	P4
mean	0.273648	0.334368	0.4593579	1.284029
std	0.128141	0.195642	0.3524453	2.619181
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.244343	0.280623	0.3922554	1.257405
std	0.113144	0.138786	0.2503293	1.266139
Simulation	P1	P2	P3	P4
mean	0.249226	0.287267	0.3870632	0.736719
std	0.11119	0.131296	0.2711628	0.687984

Case27	CVa=1	CVs=1	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.432258	0.622287	1.206993	5.738462
std	0.325471	0.507892	1.131095	6.126265
Simulation	P1	P2	P3	P4
mean	0.446893	0.582404	1.1982967	3.318919
std	0.33954	0.532146	1.1434043	3.835289
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.310758	0.401377	0.6802075	2.841141
std	0.237307	0.306483	0.5887934	3.052376
Simulation	P1	P2	P3	P4
mean	0.313908	0.382824	0.7393206	4.476637
std	0.240098	0.306314	0.6276351	6.370704
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.270948	0.328997	0.5076086	1.891847
std	0.217191	0.252408	0.4181169	2.029374
Simulation	P1	P2	P3	P4
mean	0.267456	0.331865	0.5195698	1.290027
std	0.20923	0.307042	0.5120135	1.476577
Case28	CVa=1	CVs=1.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.577419	0.886217	1.8363636	9.2
std	0.513907	0.815804	1.8337741	9.954396
Simulation	P1	P2	P3	P4
mean	0.532272	0.641735	1.5844902	6.909173
std	0.449391	0.612869	1.4835444	6.121674
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.379981	0.527238	0.9803372	4.491855
std	0.364802	0.482093	0.9485888	4.958536
Simulation	P1	P2	P3	P4
mean	0.391368	0.57758	1.1049117	6.275922
std	0.360506	0.52413	0.9457178	4.491439
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.315291	0.40962	0.699864	2.949252
std	0.330059	0.390651	0.6678426	3.295363
Simulation	P1	P2	P3	P4
mean	0.249904	0.333698	0.5106505	0.803872
std	0.268931	0.304034	0.4953043	0.909161

Case29	CVa=1.5	CVs=0	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.46129	0.675073	1.3328671	6.430769
std	0.288867	0.525213	1.2524317	6.888375
Simulation	P1	P2	P3	P4
mean	0.39482	0.569819	1.2704824	3.664624
std	0.20248	0.432212	1.1104888	3.591971
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.324602	0.42655	0.7402334	3.171284
std	0.143694	0.261261	0.6230079	3.426544
Simulation	P1	P2	P3	P4
mean	0.291399	0.340556	0.5112078	1.25736
std	0.079775	0.16025	0.380075	1.154735
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.279817	0.345122	0.5460597	2.103328
std	0.095275	0.173226	0.4130785	2.271931
Simulation	P1	P2	P3	P4
mean	0.267787	0.366794	0.5522461	1.84875
std	0.068488	0.218546	0.5744588	1.728643
Case30	CVa=1.5	CVs=0.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.490323	0.727859	1.4587413	7.123077
std	0.336181	0.592076	1.3951792	7.654403
Simulation	P1	P2	P3	P4
mean	0.449145	0.841609	1.7935929	6.039941
std	0.236845	1.06751	2.4722492	3.600938
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.338447	0.451722	0.8002594	3.501427
std	0.188391	0.307032	0.6994168	3.808584
Simulation	P1	P2	P3	P4
mean	0.300851	0.39846	0.7322282	2.308465
std	0.145233	0.249365	0.6998263	2.157829
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.288686	0.361246	0.5845107	2.314809
std	0.145624	0.216901	0.4697436	2.526348
Simulation	P1	P2	P3	P4
mean	0.297572	0.384809	0.618187	8.606267
std	0.130211	0.221251	0.495376	5.728924

Case31	CVa=1.5	CVs=1	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.577419	0.886217	1.8363636	9.2
std	0.462709	0.784562	1.8200899	9.951884
Simulation	P1	P2	P3	P4
mean	0.453781	0.883394	1.4540904	4.006175
std	0.30105	0.791046	1.7635438	3.404116
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.379981	0.527238	0.9803372	4.491855
std	0.288236	0.427099	0.9218572	4.953491
Simulation	P1	P2	P3	P4
mean	0.355551	0.475798	0.8198105	3.146254
std	0.232597	0.360433	0.7233071	2.764739
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.315291	0.40962	0.699864	2.949252
std	0.242773	0.320325	0.6292962	3.287768
Simulation	P1	P2	P3	P4
mean	0.245809	0.328977	0.4063662	1.19801
std	0.190089	0.261264	0.3480926	1.228484
Case32	CVa=1.5	CVs=1.5	1 server	
Spreadsheet	P1	P2	P3	P4
mean	0.722581	1.150147	2.4657343	12.66154
std	0.650982	1.092427	2.5227646	13.78002
Simulation	P1	P2	P3	P4
mean	0.608159	1.016241	1.9284807	5.069782
std	0.488982	0.921477	2.4370266	6.174875
			2 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.449204	0.653099	1.2804669	6.142568
std	0.415441	0.60252	1.2816223	6.859651
Simulation	P1	P2	P3	P4
mean	0.476303	0.675032	0.9417284	3.169708
std	0.444397	0.641957	0.9290555	2.730521
			3 servers	
Spreadsheet	P1	P2	P3	P4
mean	0.359634	0.490244	0.8921193	4.006656
std	0.3554	0.45829	0.8789398	4.553756
Simulation	P1	P2	P3	P4
mean	0.337356	0.571744	1.4653689	4.661931
std	0.303772	0.527435	1.3723056	4.512371

Approximate Formula for Steady-State, Infinite Capacity Queues

Basic Inputs:

Number of Servers, S =	1
Arrival Rate, λ =	4.00
Service Rate Capacity of each server, μ =	5.00
Coefficient of Variation of Inter-arrival time, COV(a) =	**0.5**
Coefficient of Variation of Service time, COV(s) =	**0.5**

Inputs

The Waiting Line:

Average Number Waiting in Queue (Lq) =	**0.80**
Average Waiting Time (Wq) =	**0.20**
Probability of customer waiting P(t) =	**0.30**
Probability of customer waiting no time P(0) =	**0.70**
Standard deviation of waiting time σ =	**0.48**

Service:

Average Utilization of Servers (rho) =	**0.80**
Average Number of Customers Receiving Service =	**0.80**

The Total System (waiting line plus customers being served):

Average Number in the System (L) =	**1.60**
Average Time in System (W) =	**0.40**
Standard deviation of time in system σs =	**0.49**
"Worst Case" Flow time = Average time in system + 3 σs =	**1.86**

An Option: Multiple Classes of Customers

		priority 1	priority 2	priority 3	priority 4
	Class	1	2	3	4
	Workload fraction	**0.25**	**0.25**	**0.25**	**0.25**
Average number in Queue	*Lq (k)*	**0.05**	**0.08**	**0.17**	**0.50**
Average Time in Queue	*Wq (k)*	**0.05**	**0.08**	**0.17**	**0.50**
Average Number in System	L(k)	**0.25**	**0.28**	**0.37**	**0.70**
Average Time in System	W(k)	**0.25**	**0.28**	**0.37**	**0.70**
Standard Deviation of Time in System		**0.16**	**0.22**	**0.41**	**1.20**
Average + 3*Standard Deviations		**0.72**	**0.95**	**1.60**	**4.29**

FLOW
Arrivals/Day
2.9
15
3.8
17.3
39

DISTRIBUTION		
Number of Clerks=	4.00	
Available Minutes=	1800.00	
Avg. ServiceTime	Load	
68.5	198.65	12.43%
50	750	46.92%
43.5	165.3	10.34%
28	484.4	30.31%
	1598.35	1
Utilization	0.89	

		Number of Servers	4.00
CV of Serv.Time	0.54386081	Arrival Rate (jobs/day)	39.00
Avg. Serv. time	40.9833333	Service rate (jobs/day)	10.98

FLOW
Arrivals/Day
1.4
6.3
1.6
5.3
14.6

UNDERWRITING 1		
Team	1	
Available Minutes=	450	
ServiceTime=	Load	
43.60	61.04	14.01%
38.00	239.40	54.94%
22.60	36.16	8.30%
18.70	99.11	22.75%
	435.71	1
Utilization	0.97	

CV of Serv.Time	0.829093
Avg. Serv. time	29.84315
Arrival rate	14.6
Service rate	15.07884

FLOW
Arrivals/Day
0.8
4.3
1.1
7
13.2

UNDERWRITING 2		
Team	1	
Available Minutes=	450	
ServiceTime=	Load	
43.60	34.88	9.85%
38.00	163.40	46.15%
22.60	24.86	7.02%
18.70	130.90	36.97%
	354.04	1
Utilization	0.79	

CV of Serv.Time	0.891027
Avg. Serv. time	26.82121
Arrival rate	13.2
Service rate	16.77777

FLOW
Arrivals/Day
0.7
4.4
1.1
5
11.2

UNDERWRITING 3		
Team	1	
Available Minutes=	450	
ServiceTime=	Load	
43.60	30.52	9.66%
38.00	167.20	52.90%
22.60	24.86	7.87%
18.70	93.50	29.58%
	316.08	1
Utilization	0.70	

CV of Serv.Time	0.856447
Avg. Serv. time	28.22143
Arrival rate	11.2
Service rate	15.94533

FLOW
Arrivals/Day
2.9
15
3.8
17.3
39

RATING		
Number of Clerks=	8	
Available Minutes=	3600	
ServiceTime=	Load	
75.5	218.95	7.98%
64.7	970.5	35.36%
65.5	248.9	9.07%
75.5	1306.15	47.59%
	2744.5	1
Utilization	0.76	

		Number of Servers	8.00
CV of Serv.Time	0.199483	Arrival Rate (jobs/day)	39.00
Avg. Serv. time	70.37179	Service rate (jobs/day)	6.39

FLOW
Arrivals/Day
2.9
15
3.8
17.3
39

FLOW
Arrivals/Day
2.9
2.3
3.8
17.3
26.3

12.7

POLICY WRITING		
Number of Clerks=	5	
Available Minutes	2250	
ServiceTime	Load	
71	205.9	14.29%
71	163.3	11.33%
54	205.2	14.24%
50.1	866.73	60.14%
	1441.13	1
Utilization	0.64	

		Number of Servers	5.00
CV of Serv.Time	0.22906	Arrival Rate (jobs/day)	26.30
Avg. Serv. time	54.79582	Service rate (jobs/day)	8.21

		FLOW
		Arrivals/Day
From Agent	**RUNS**	2.9
	RAPS	15
	RAINS	3.8
From Computer	**RERUNS**	17.3
	Total	39

Avg. Serv. time	40.98333
CV of Arriv.Time	1
CV of Serv.Time	0.543861

DISTRIBUTION		
Number of Clerks=	4.00	
Available Minutes=	1800.00	
ServiceTime	Load	
68.5	198.65	12.43%
50	750	46.92%
43.5	165.3	10.34%
28	484.4	30.31%
	1598.35	1
Utilization	0.89	
Number of Servers	4.00	
Arrival Rate (jobs/day)	39.00	
Service rate (jobs/day)	10.98	

FLOW
Arrivals/Day
2.9
15
3.8
17.3
39

Avg. Serv. time	28.35462
CV of Arriv.Time	0.849921
CV of Serv.Time	0.858019

DISTRIBUTION	priority 1	priority 2	priority 3	priority 4
WIP	0.50	2.16	0.62	4.23
Avg. Flow Time (days)	0.10	0.12	0.15	0.36
STD. Flow Time	0.05	0.06	0.09	0.34

Combined UTs with Priority			
	Total Flow	Standard	Worst Case =
Priority Class	Time (Days)	Deviation	Average + 3 std. dev
1	0.47	0.09	0.73
2	0.49	0.10	0.79
3	0.56	0.14	0.98
4	0.88	0.40	2.08

Combined UT's		
No. of Teams	3	
Available Minutes=	1350	
ServiceTime=	Load	
43.60	126.44	11.43%
38.00	570.00	51.54%
22.60	85.88	7.77%
18.70	323.51	29.25%
	1105.83	1
Utilization	0.82	
Number of Servers	3.00	
Arrival Rate (jobs/day)	39.00	
Service rate (jobs/day)	15.87	

FLOW
Arrivals/Day
2.9
15
3.8
17.3
39

Avg. Serv. time	70.37179
CV of Arriv.Time	0.89758
CV of Serv.Time	0.199483

RATING		
Number of Clerks=	8	
Available Minutes=	3600	
ServiceTime=	Load	
75.5	218.95	7.98%
64.7	970.5	35.36%
65.5	248.9	9.07%
75.5	1306.15	47.59%
	2744.5	1
Utilization	0.76	
Number of Servers	8.00	
Arrival Rate (jobs/day)	39.00	
Service rate (jobs/day)	6.39	

FLOW
Arrivals/Day
2.9
15
3.8
17.3
39

FLOW
Arrivals/Day
2.9
2.3
3.8
17.3
26.3
12.7

Avg. Serv. time
CV of Arriv.Time
CV of Serv.Time

Combined UT's	priority 1	priority 2	priority 3	priority 4
WIP	0.33	1.75	0.35	2.31
Avg. Flow Time (days)	0.07	0.09	0.12	0.20
STD. Flow Time	0.06	0.06	0.09	0.20

RATING	priority 1	priority 2	priority 3	priority 4
WIP	0.50	2.23	0.58	3.35
Avg. Flow Time (days)	0.16	0.16	0.16	0.18
STD. Flow Time	0.03	0.03	0.04	0.06

	FLOW	
		Arrivals/Day
From Agent	RUNS	2.9
	RAPS	15
	RAINS	3.8
From Computer	RERUNS	17.3
	Total	39

DISTRIBUTION		
Number of Clerks=	5.00	
Available Minutes=	2250.00	
ServiceTime	Load	
68.5	198.65	12.43%
50	750	46.92%
43.5	165.3	10.34%
28	484.4	30.31%
	1598.35	1
Utilization	0.71	

FLOW
Arrivals/Day
2.9
15
3.8
17.3
39

Number of Servers	5.00
Arrival Rate (jobs/day)	39.00
Service rate (jobs/day)	10.98

Avg. Serv. time	40.98333
CV of Arriv.Time	1
CV of Serv.Time	0.543861

Avg. Serv. time	28.35462
CV of Arriv.Time	0.849921
CV of Serv.Time	0.858019

DISTRIBUTION	priority 1	priority 2	priority 3	priority 4
WIP	0.47	1.84	0.44	1.49
Avg. Flow Time (days)	0.10	0.10	0.11	0.13
STD. Flow Time	0.05	0.05	0.06	0.08

Moving 1 Policy Writer to Distribution			
	Total Flow	Standard	Worst Case =
Priority Class	Time (days)	Deviation	Average + 3 std. dev
1	0.46	0.09	0.73
2	0.49	0.10	0.77
3	0.53	0.12	0.89
4	0.70	0.25	1.45

Combined UT's		
No. of Teams	3	
Available Minutes=	1350	
ServiceTime=	Load	
43.60	126.44	11.43%
38.00	570.00	51.54%
22.60	85.88	7.77%
18.70	323.51	29.25%
	1105.83	1
Utilization	0.82	

FLOW
Arrivals/Day
2.9
15
3.8
17.3
39

RATING		
Number of Clerks=	8	
Available Minutes=	3600	
ServiceTime=	Load	
75.5	218.95	7.98%
64.7	970.5	35.36%
65.5	248.9	9.07%
75.5	1306.15	47.59%
	2744.5	1
Utilization	0.76	

FLOW
Arrivals/Day
2.9
15
3.8
17.3
39

FLOW
Arrivals/Day
2.9
2.3
3.8
17.3
26.3
12.7

Number of Servers	3.00
Arrival Rate (jobs/day)	39.00
Service rate (jobs/day)	15.87

Avg. Serv. time	70.37179
CV of Arriv.Time	0.89758
CV of Serv.Time	0.199483

Number of Servers	8.00
Arrival Rate (jobs/day)	39.00
Service rate (jobs/day)	6.39

Avg. Serv. time
CV of Arriv.Time
CV of Serv.Time

Combined UT's	priority 1	priority 2	priority 3	priority 4
WIP	0.33	1.75	0.35	2.31
Avg. Flow Time (days)	0.07	0.09	0.12	0.20
STD. Flow Time	0.06	0.06	0.09	0.20

RATING	priority 1	priority 2	priority 3	priority 4
WIP	0.50	2.23	0.58	3.35
Avg. Flow Time (days)	0.16	0.16	0.16	0.18
STD. Flow Time	0.03	0.03	0.04	0.06

Vita

Xiaofeng Zhao was born in China. He obtained his B.S. degree in engineering and M.A. degree in science and technology studies respectively from Xian Jiaotong University and Northwest University in China. Upon completion of an MBA degree at Indiana University of Pennsylvania, he entered the doctoral program in management science at the University of Tennessee in 2003.

Before coming to U.S. in 2000, he was an associate professor in China. He spent one year (1996-1997) as a visiting scholar in business administration at University of Illinois at Urbana-Champaign (Freeman fellow). He has published over a dozen of academic articles.

Xiaofeng is a member of Beta Gamma Sigma International Honor Society and Institute of Operations Research and Management Science (INFORMS). He will earn his Ph.D. in management science from the University of Tennessee in August 2007.