UNIVERSITI PUTRA MALAYSIA

A SOLUTION OF SINGLE SERVER AND MULTISERVER GENERAL QUEUEING SYSTEMS USING A LINEAR ALGEBRAIC APPROACH

MOHAMMAD SALEH MUSTAFA

A SOLUTION OF SINGLE SERVER AND MULTISERVER GENERAL QUEUEING SYSTEMS USING A LINEAR ALGEBRAIC APPROACH

BY

MOHAMMAD SALEH MUSTAFA

Thesis Submitted in Fulfilment of the Requirements for the Degree of Master of Science in the Faculty of Science and Environmental Studies, Universiti Pertanian Malaysia

December 1995

UNIVERSITI PUTRA MALAYSIA

A SOLUTION OF SINGLE SERVER AND MULTISERVER GENERAL QUEUEING SYSTEMS USING A LINEAR ALGEBRAIC APPROACH

MOHAMMAD SALEH MUSTAFA

قال تعالى و قضى ربك الآ تعبدوا إلا اياه و بالوالدين احسانا

اهداء الى نبع الحنان و رمز التضحية و العطاء

الى والدي و والدتي الأحبة

Dedicated to
My Parents

ACKNOWLEDGMENTS

First, I would like to thank the chairman of my Supervisory Committee, Assoc. Prof. Dr. Abu Talib Othman. His excellent supervision, continuous encouragement, guidance and numerous discussions were instrumental for the completion of the thesis.

My special thanks also go to the members of the Committee, Dr. Abdul Azim Abd. Ghani and Dr. Ramlan Mahmod, for their comments, help and encouragement. I am grateful to Dr. Yazid Mohd. Saman for translating the abstract of this thesis into Bahasa Malaysia. I also wish to extend my thanks Dr. Ali Mamat, head of the Computer Science Department, and all the lecturers and staff in the department.

I am thankful to the very kind man, Mr. Mohammad Noor Azman, the system administrator, for all the help he offered and is still offering.

To my parents, brothers and sisters, my heartfelt thank for all that no one can give but you.

My gratitude to a very special person without whom, I would not have been able to continue my higher education in this country. My thanks to Belal Mustafa Abu Ata (Abu Al Bara') and his family for all the help, support and encouragement.

I am grateful to the personnel in the Graduate School Office in UPM for their kindness and cooperation.

The financial support provided by the Public Services Department of Malaysia (JPA), is highly appreciated and gratefully acknowledged.

TABLE OF CONTENTS

LIST OF FIGURES

Abstract of thesis submitted to the Senate of Universiti Pertanian Malaysia in fulfilment of the requirements for the degree of Master of Science.

A SOLUTION OF SINGLE SERVER AND MULTISERVER GENERAL QUEUEING SYSTEMS USING A LINEAR ALGEBRAIC APPROACH

by

MOHAMMAD SALEH MUSTAFA

December, 1995

Chairman : Assoc. Prof. Dr. Abu Talib Othman

Faculty : Science and Environmental Studies

Performance evaluation of computer systems and communication networks is very important due to the essential need of finding such systems which can meet all the objectives, for which these resources are built, at an acceptable cost. Many methods have been applied to describe the performance of these resources and to find possible ways of improving performance. Each of these methods has its own characteristics and procedure in dealing with the performance concept. These procedure describe these systems such that a previously specified technique can be used to evaluate useful performance measurements which will help decision making.

The queueing network model is the most widely used computer systems and communication networks description method. Traditional techniques used in solving the above method use unrelated mathematical techniques. The linear algebraic approach has been chosen as the mathematical approach in this thesis. This approach transforms the queueing theory problems from those of integral equations into those of algebraic equations over a finite dimensional vector space. The solutions obtained using this approach can now make use of the high-speed parallel processors since the components of these solutions are in matrix-vector format. In this research, the linear algebraic approach is used to derive explicit solutions for single server and multiserver general queueing systems. The research approach uses the generalized exponential to represent the G-type distribution because of the tractable facilities that it offers e.g., it facilitates the derivation of product form solutions. The performance measurements derived are the queue length probabilities, mean queue length, throughput, mean time a customer spends in the system and departure and arrival probabilities. The derived solutions can be comprehended, give a clear insight about the behaviour of the systems and are easy to evaluate for any given system parameters. Numerical examples showing the effects of changing the parameters of the solutions obtained on the performance of the systems are studied.

Abstrak tesis yang dikemukakan kepada Senat Universiti Pertanian Malaysia bagi memenuhi keperluan untuk Ijazah Master Sains.

PENYELESAIAN SISTEM GILIRAN A-M SATU PELAYAN DAN MULTIPELAYAN MENGGUNAKAN PENDEKATAN ALJABAR LINEAR

Oleh

MOHAMMAD SALEH MUSTAFA

Disember, 1995

Pengerusi: Prof Madya. Dr. Abu Talib bin Othman

Fakulti: Sains dan Pengajian Alam Sekitar

Penilaian prestasi untuk sistem komputer dan rangkaian komunikasi adalah sangat mustahak. Ini disebabkan oleh keperluan utama untuk mencari satu sistem dengan segala penggunaan sumber yang mahal dilaksanakan. Tambahan pula, semua objektif untuk pembinaan sumber ini dapat dicapai dengan kos yang munasabah. Terdapat banyak kaedah untuk menerangkan prestasi sumber ini dan mencari jalan untuk memperbaikinya. Setiap kaedah mempunyai ciri-ciri dan prosedurnya dalam mengendali konsep prestasi ini. Prosedur ini menerangkan sistem berdasarkan teknik yang terdahulu untuk menilai pengukur prestasi yang akan membantu dalam menetapkan keputusan.

Selalunya, teknik tradisi dalam penyelesaian model rangkaian giliran (seperti dalam kaedah penerangan dalam sistem komputer dan rangkaian komunikasi) juga terlibat dalam teknik Matematik yang tiada kaitannya. Pendekatan aljabar linear dipilih sebagai pendekatan Matematik dalam tesis ini. Ia menterjemah masalah dalam teori giliran dari persamaan integral ke dalam persamaan algebra untuk ruang vektor berdimensi terhad. Hasil dari penggunaan pendekatan ini boleh digunakan dalam pemproses selari yang laju kerana komponen hasil berformat matrik-vektor. Di dalam penyelidikan ini pendekatan aljabar linear digunakan untuk mencari penyelesaian tersurat untuk sistem giliran satu pelayan dan multi pelayan. Penyelidikan ini menggunakan kaedah eksponen yang am untuk mewakili taburan jenis-G kerana kemudahan yang ada berbentuk hasil darab. Pengukuran prestasi yang diperolehi terdiri dari probabiliti panjang giliran, purata panjang giliran truput, purata masa pelanggan menunggu dalam sistem dan probabiliti ketibaan dan keluar. Hasil yang dicapai boleh difahami, memberikan pandangan yang jelas berkenaan sistem dan mudah untuk dinilai dengan memberikan sebarang parameter. Contoh numerikal yang menunjukkan kesan perubahan parameter hasil yang diperolehi dari prestasi sistem juga dikaji.

CHAPTER I

INTRODUCTION

Computer communication network can be viewed as a facility that makes possible communication between computers and other devices. According to Hammond and O'Reilly (1986), it came into being when intelligent devices such as the devices of a data communication system including control and graphic terminals, rotation and applications servers, multiplexing and switching devices, modems, videos, facsimile, scanners and printers were moved from the main frame to remote locations.

Increased use of sophisticated computer and communication systems which are expensive resources, makes it imperative that they are used efficiently. To achieve this aim, performance evaluation of such systems must be considered at all stages of the system life cycle (design, development, configuration and tuning).

Computer and communication systems performance methods are divided into two main areas: performance measures and performance modeling methods (Ferrari, 1978). The measurement method is applied by using a direct measure to the performance of the system using hardware dedicated and/or software dedicated tools. This method is intractable since the system needs to be operational. As a result, the performance of the system during design and development stages can not be analyzed by

using this method. Moreover, this method is very complex because it involves considerable human and machine cost. Hence, modeling techniques are preferred.

According to Buchholz (1992), performance analysis of complex computer systems is often accomplished by means of models which can be mapped on Markov processes. Evaluation and prediction of the stochastic behaviour of work flow in the system are undertaken using these models. As a result, useful system performance measurements can be obtained to aid decision making. In specifying these models, description techniques such as queueing networks (QN) (Lavenberg, 1983), generalized Petri nets (PN) (Ajmone-Marsan *et al.*, 1984) or networks of stochastic automata (Plateau, 1985) are used.

A computer system can be viewed as a network of queues. It is generally characterized as a set of software and hardware resources (limited number) and a set of multiple jobs or tasks using these resources, in which a queueing situation will occur. From this point of view, a computer system can be thought of as a queueing system in which computer resources represent the service centres and the jobs represent the customers. The results developed from this queueing system such as queue lengths, queue delay, etc., are then used to predict the performance of the real system.

According to Lazowska *et al.* (1984), the importance of queueing network models as modeling tools lies in their ability to achieve a favourable balance between accuracy and efficiency in most applications.

For this reason, queucing network models are used in different application areas such as telecommunication systems, transportation systems, banking systems, production systems, computer systems, health care systems and in many service systems where a queueing situation occurs. Even simple queueing network models are able to yield fairly accurate predictions of the behaviour of complex systems. This has helped in stimulating queueing network research.

The application of probability theory to these problems can be traced to 1917 when a Danish engineer (Erlang, 1917) analyzed the behaviour of simple queues to assist him in designing a telephone exchange (King, 1990). Jackson (1957) was responsible for a key paper entitled "Network of Waiting Lines". A review of related literature covering a sixty-year period until 1967 can be found in Bhat (1969). The use of queueing network models to evaluate the performance of computer systems and computer communication networks began in the seventies (Kleinrock, 1975; Allen, 1978). Extensive research was carried out on this field in the seventies and eighties. Disney and König (1985) reviewed more than 300 related articles while Prabhu (1987) compiled a bibliography of books and survey papers on queueing systems which included the surveys by Lavenberg (1983) and Kleinrock (1988).

As a final comment, performance evaluation cannot be studied in isolation as it cannot be separated from queueing systems and networks. It is generally agreed that a research in performance evaluation research is realy research on queueing theory. Moreover, this theory has had an

important effect on communication systems and networks. Indeed, queueing theory was first applied in communication systems, where it is still in use today (Pekergin, 1990).

Background of the Problem

A queueing network can be modeled using three main techniques. These are simulation, approximation, and analytical techniques. Continuous and discrete-event queueing model simulations (Mitrani, 1982; Sauer and MacNair, 1983) are effective techniques for evaluating system performance. However, simulation techniques are costly and have two main constraints. The first is that the level of details in the model and computation is time consuming while the second is the difficulty of creating the simulation programme and verifying its accuracy.

Approximate methods are often used for solving queueing networks in cases where the exact solution is not known or computationally intractable. These techniques are applied only under certain considerations which cannot be satisfied for a wide range of systems. For instance, diffusion methods, which are used to approximate the discrete flow of customers by continuous flow, are only accurate under heavy traffic conditions. On the other hand, the maximum entropy method needs certain known results which have to be obtained independently by some other technique (Harrison and Patel, 1993).

descriptions of the system under consideration. These techniques are preferred to than the other techniques because they (i) provide quantitative predictions and insights into the structure and behaviour of the real system, (ii) provide cost effective methods, (iii) are applicable in all stages of the system life-cycle, and (iv) a considerable amount of cost savings can be obtained as a result of early recovery and elimination of distressed designs from performance viewpoint in these stages.

Some of the computer systems and communication networks aspects can be modeled as product form queueing networks (Kelly, 1979; Walrand, 1988; Van Dijk, 1993) which is an important class of stochastic analytic modeling technique. These tractable models can be solved efficiently with a variety of computational algorithms such as mean value analysis (MVA) (Reiser and Lavenberg, 1980), recursion by chain (Conway and Georaganas, 1986), and convolution (Buzen, 1973).

However, not all of computer and communication network aspects can be modeled as product form networks. In addition, to that, there is another general and important class of stochastic models which can be modeled by the presence of imbedded Markov chains that are two dimensional generalizations of elementary M/G/1 and G/M/1 queues. These are the matrix-geometric techniques. These techniques enable the construction of computationally efficient algorithms which are applicable for a large number of general queueing networks and lead to closed-form solutions (Neuts, 1981; Neuts, 1989; Nelson, 1991).

Throughout this thesis, the linear algebraic approach is used. In this approach queueing theory problems are transformed from those of integral equations into those of algebraic equations over a finite dimensional vector space (Lipsky, 1985a). This is done by representing each nonexponential service time (or interarrival time) distribution by a collection of exponential servers, with the constraint that the collection can only be accessed by one customer at a time (Carrol *et al.*, 1982). In using the traditional techniques to solve the queueing system, problems arise from the fact that these techniques usually involve using unrelated mathematical techniques in which integrals play a very important rule. Solutions to problems in simple systems can be understood using such techniques. However, when problems arise in more complex systems, such techniques become very complicated, difficult to analyze and in many cases, it can be quite difficult to evaluate the solution for a given system parameters.

The linear algebraic approach is used to derive the solution of single server general queueing systems from the type M/G/1 and G/M/1 and the solution of multiserver general queueing networks using the generalized exponential (GE) pdf (see Appendix A) to represent the G-type interarrival or service time distribution with known first two moments. The generalized exponential (GE) distribution which is defined as (El-Affendi and Kouvatsos, 1983; Kouvatsos, 1986a; 1988),

$$f(t) = \left(\frac{C^2 - 1}{C^2 + 1}\right) u_0(t) + \frac{4\mu}{(C^2 + 1)^2} \exp\left(\frac{-2\mu t}{C^2 + 1}\right), \qquad t \geq 0 \qquad [1.1]$$

where μ is the mean service rate, C is the coefficient of variation and $u_0(t)$ is the unit impulse function. This has been chosen as the best choice to represent the G-type distribution. According to Kouvatsos (1988), the interest in the GE distribution as a universal model to approximate G-type distributions (when only the first two moments are known) has been motivated by some of its robust and versatile properties which make it particularly useful for the analysis of multiserver queues and general queuing networks. Moreover, it facilitates the derivation of product form approximations for isolated queues and networks (Kouvatsos, 1986a; 1986b).

The solutions derived for the systems mentioned with the (GE) distribution representing the G-type using the linear algebraic techniques can be understood, give a clear insight about the system and are easily evaluated for any given system parameters.

Objectives of the Study

This study has been undertaken work in this thesis had been done to achieve four main objectives. The first is to prove the power and ability of the linear algebraic approach in the queueing theory by two means: (i) comparing different techniques used to represent the theory to show the potential and limitations of each, and (ii) representing the linear algebraic approach and its mechanism to derive the solutions. The second objective is to derive the solution of single server general queueing

networks the M/G/1 and G/M/1 with the generalized exponential 'GE' representing the G-type pdf. Deriving an explicit solution for the multiserver general queueing systems from the type G/M/C with the (GE) representing the G-type pdf is the third objective of this thesis. The final objective is to provide a clear and proven method for on-going research on this topic by suggesting a way to explore an explicit solution for the M/G/C queueing system with the 'GE' representing the G-type pdf. The approach is discussed briefly in the section on further research on this queueing system.

The Approach

The research methodology starts with formulation of the algebraic components describing the general queueing system using the hyperexponential (H_2) pdf to represent the G-type. Then the solution of the system using these algebraic components is constructed. There are infinitely many two-stage models representing the G-type with the same mean value and squared coefficient of variation. The previous point shows the need to control the manner in which such representations are constructed. This necessitates the changing of parameters (El-Affendi and Kouvatsos, 1983), which leaves us with the mean, squared coefficient of variation and a new parameter k. Subsequantly, the two-stage convergence to the (GE) pdf is followed through by taking the limit when the parameter k approaches infinity. The resultant solution is simplified and rearranged to achieve the explicit solution. As a result, the

system can be studied under various conditions to give an instant view of the behaviour of the real system easily, without encountering evaluation difficulties.

Organization of the Thesis

The thesis consists of five chapters, organized in a manner that gives a clear indication of the focus of the research, paving the way for further research. The outline of the chapters follows:

Chapter I: This chapter presents an overview of the importance of queueing theory in the performance evaluation of computer systems and communication networks. The potential and limitations of the various techniques used to evaluate a computer system or a communication network are discussed. This chapter alsohighlights the power of the linear algebraic approach as the approach used throughout this study. This chapter takes a brief look at the methodology used and the objectives of the thesis and concludes with an outline of the thesis.

Chapter II: The basis of the linear algebraic approach of the queueing theory is discussed in this chapter by a review of the approach stages until the current stage. Literature which relate to the analysis and modeling of the single server general queueing networks and multiserver general queueing network are reviewed and documented. Finally, the linear algebraic approach is discussed briefly showing the construction of the algebraic components of a given system and the mechanism to derive the solution from these components.

Chapter III: The solution of single server general queueing networks M/G/1 and G/M/1 with the (GE) pdf describing the G-type is derived in this chapter using the linear algebraic approach. The performance measurements of the systems such as queue length probabilities, throughput, etc. are given. The systems are compared under different conditions which may reflect the behaviour of real life systems.

Chapter IV: This chapter discusses the queue length probabilities departure and arrival probabilities of multiserver general queueing systems of the type GE/M/C i.e., for load dependent multiserver general queueing systems when the (GE) pdf represents the G-type. The system is analyzed using this approach under different conditions to give an insight into real systems performance.

Chapter V: This chapter concludes the research with a summary of the main ideas. Further work is encouraged on the exploration of the solution of M/GE/C queueing system using the linear algebraic approach.