STOCHASTIC MODELLING AND ANALYSIS OF SOME GENERALIZED QUEUEING NETWORKS AND THEIR APPLICATIONS

Thesis submitted to the

University of Calicut

for the award of the Degree of

DOCTOR OF PHILOSOPHY

IN

STATISTICS

under the faculty of Science

by

ANJALE RAMESH

under the guidance of

Prof. (Dr.) M. MANOHARAN

DEPARTMENT OF STATISTICS

UNIVERSITY OF CALICUT

KERALA - 673 635

INDIA

February 2025

DEPARTMENT OF STATISTICS

UNIVERSITY OF CALICUT

Dr. M. MANOHARAN

Senior Professor (Retd.)

CALICUT UNIVERSITY P.O.

MALAPPURAM (District)

KERALA, INDIA - 673 635

Mob: 91-9447424043

Email: manumavila@gmail.com

CERTIFICATE

I hereby certify that the work presented in the thesis entitled "STOCHASTIC MODELLING AND ANALYSIS OF SOME GENERALIZED QUEUEING NETWORKS AND THEIR APPLICATIONS" is a bona fide record of research work carried out by Mrs. Anjale Ramesh, Research Scholar, Department of Statistics, University of Calicut, under my supervision and guidance for the award of the Degree of Doctor of Philosophy in Statistics from the University of Calicut. I further certify that this work has not been included in any other thesis submitted previously for the award of any degree. Additionally, I certify that the content of this thesis have been checked using an anti-plagiarism database, and no unacceptable similarity was found through the software check.

University of Calicut

Date: 08-07-2025

Dr. M. Manoharan

Research Guide

Dr. M. MANOHARAN
Senior Professor, Dept. of Statistics
University of Calicut
Calicut University P.O., Malappuram (Dt.)
Kerala. PIN-673635

DECLARATION

I hereby declare that the work presented in this thesis, entitled "STOCHASTIC MODELLING AND ANALYSIS OF SOME GENERALIZED QUEUEING NETWORKS AND THEIR APPLICATIONS" is based on the original work done by me under the guidance of Dr. M. Manoharan and has not been included in any other thesis submitted previously for the award of any degree. The contents of the thesis are undergone a plagiarism check using iThenticate software at C.H.M.K. Library , University of Calicut, and the similarity index found within the permissible limit. I also declare that the thesis is free from AI generated contents.

Anjale Ramesh

University of Calicut

Date: 08-07-2025

08/07/2025

Dr. M. MANOHARAN
Senior Professor, Dept. of Statistics
University of Calicut
Calicut University P.O., Malappuram (Dt.)
Kerala. PIN-673635

ACKNOWLEDGEMENT

I am deeply grateful to the many individuals whose guidance and support have been crucial in the completion of this thesis.

I would like to begin by expressing my heartfelt gratitude to my guide, Dr. M. Manoharan, Retd. Senior Professor, Department of Statistics, University of Calicut, for his generous assistance, insightful feedback, invaluable supervision, and constant mental support throughout the development of this thesis. Without his encouragement and guidance, this work would not have been possible.

I extend my gratitude to Dr. Krishnarani S. D., Head of the Department, Dr. K. Jayakumar, Dr. C. Chandran, Dr. Dileep Kumar and other faculty members of the department, for sharing their knowledge and providing essential facilities throughout my research.

I also thank the librarian and the non-teaching staff of the Department of Statistics, University of Calicut for their help and cooperation. I am grateful to my friends for bringing joy and support to my journey and to my colleagues for their warmth and cooperation throughout this period.

I am forever grateful to my parents, Ramesh and Mini and my sister, Amrutha, for their unwavering support, encouragement, and prayers. Their constant faith in me has been my source of strength.

Most importantly, I wish to extend my deepest love and gratitude to my husband, Dr. Ajay Dev. Without his motivation, emotional support, and encouragement, this research would not have been possible.

Above all, I thank God Almighty for being with me every step of the way, guiding and blessing me.

ABSTRACT

Queueing theory is a field dedicated to the modelling and analysis of queues or waiting lines. It provides mathematical tools to optimise various processes in service systems to enhance overall system performance. Most of the real-world service systems operate under time-varying conditions, such as fluctuating arrival and service processes. Analysing the transient behaviour of such time-varying queueing systems is significantly more challenging than steady-state analysis. This research focuses on the transient analysis of time-varying queues, which have practical applications in real-life service systems. The study investigates the transient distributional law that links the virtual workload to customer waiting times in a non-stationary general single-server queueing system. Additionally, a simulation study is conducted to validate the transient measure alongside other performance measures. The research also introduces a general algorithmic framework to derive transient performance measures in a k-station Markovian tandem network, supported by numerical studies that analyse the transient behaviour of these performance measures. Furthermore, a comparative study is presented on a Markovian non-stationary two-station tandem network with finite queue capacity, examining different blocking mechanisms. The study provides explicit expressions for transient performance measures under both BAS and BBS blocking mechanisms. Another significant contribution of this research is the exploration of time-varying approximations for performance measures in a feed-forward open queueing network comprising single-server queues with time-varying arrival rates. An algorithm is developed to compute time-varying approximations for performance measures in feed-forward open queueing networks of $G_t/G/1$ queues. The thesis concludes by emphasizing the critical role of time-varying queues in real-life service systems and offers actionable recommendations for future research directions.

Keywords: Time-varying queues, Tandem queueing networks, Transient performance measures, Blocking, General queueing networks.

സംഗ്രഹം

കൃയിങ് തിയറി, കൃകൾ അഥവാ കാത്തിരിപ്പ വരികളുടെ മാതൃകാ രൂപീകരണത്തിനും വിശകലനത്തിനുമായി സമർപ്പിച്ചിരിക്കുന്ന ഒരു പഠന ശാഖയാണ്. സേവന സംവിധാനങ്ങളിലെ വിവിധ പ്രക്രിയകളെ മെച്ചപ്പെടുത്താനും അതിലൂടെ അത്തരം കേന്ദ്രങ്ങളുടെ പ്രവർത്തനക്ഷമത വർധിപ്പിക്കാനും സഹായകമായ ഗണിതശാസ്ത്ര സാങ്കേതികവിദ്യകൾ ഈ ശാഖ പ്രദാനം ചെയ്യുന്നു. സാധാരണയായി, യാഥാർത്യജീവിതത്തിലെ ബഹുഭൂരിപക്ഷം സേവന വ്യവസ്ഥിതികളും സമയാനുസൃതമായി പ്രവർത്തിക്കുന്നവയാണ്. അത്തരം സേവനകേന്ദ്രങ്ങളിലെ ക്ഷണികമായ സ്വഭാവം വിശകലനം ചെയ്യുന്നത് സ്ഥിരാവസ്ഥ വിശകലനം ചെയ്യുന്നതിനേക്കാൾ വെല്ലുവിളികൾ നിറഞ്ഞതാണ്. ഈ ഗവേഷണം പ്രധാനമായും അത്തരം സമയാനുസൃതമായി പ്രവർത്തിക്കുന്ന കാത്തിരിപ്പ വരികളുടെ ക്ഷണികമായ വിശകലനത്തിൽ ശ്രദ്ധ കേന്ദ്രീകരിക്കുന്നു, ഈ കണ്ടെത്തലുകൾ യാഥാർഥ്യ ജീവിതത്തിലെ സേവന കേന്ദ്രങ്ങളിൽ പ്രായോഗികതലത്തിൽ ഉപയോഗിക്കാവുന്നതാണ്. ഈ പഠനം സമയാനുസൃതമായി പ്രവർത്തിക്കുന്ന ഏക സെർവർ കൃയിങ് കേന്ദ്രങ്ങളിലെ ജോലിഭാരവും ഉപഭോക്താവിന്റെ കാത്തിരിപ്പുസമയവും തമ്മിലുള്ള ബന്ധം പരിശോധിക്കുന്നു. അനുകരണ പഠനത്തിലൂടെ ഈ ബന്ധവും കൂടാതെ മറ്റ് അളവുകളെയും സാധൂകരിക്കുന്നു. ഇതിനോടൊപ്പം കെ നോഡുകളുള്ള മാർകോവിയൻ ശ്രേണി കൃകളുടെ ക്ഷണികമായ പ്രവർത്തന അളവുകൾ കണ്ടെത്തുന്നതിനായി ഒരു പൊതുവായ അൽഗോരിതമിക് ഘടന അവതരിപ്പിക്കുന്നു. ഇത് കൂടാതെ വ്യത്യസ്ത തടയൽ രീതികൾ പഠിക്കുന്നതിനായി, കൃ ശേഷിയിൽ പരിമിതിയുള്ള ശ്രേണി കൃകളുടെ ഒരു താരതമ്യ പഠനവും അവതരിപ്പിക്കുന്നു. ബിഎഎസ്, ബിബിഎസ് എന്നീ തടയൽ രീതികളിലെ ക്ഷണികമായ പ്രവർത്തന അളവുകളുടെ വ്യക്തമായ സൂത്രവാക്യങ്ങൾ ഈ ഗവേഷണത്തിലൂടെ കണ്ടെത്തുന്നു. സമയാനുസൃതമായ പ്രവേശനനിരക്കോടു കൂടിയ ഏക സെർവർ കൃയിങ് ശൃംഖലയിലെ പ്രവർത്തന അളവുകളുടെ സമയാനുസൃതമായ ഏകദേശ മൂല്യം കണ്ടെത്തുന്നതിനുള്ള അൽഗോരിതമാണ് ഈ ഗവേഷണത്തിന്റെ മറ്റൊരു സംഭാവന. ഈ കൃകൾ പൊതുസ്വഭാവം ഉള്ളവയും കൃ ശൃംഖല തിരിച്ചുവരവുകൾക്ക് നിയന്ത്രണം ഉള്ളവയുമാണ്. ഈ പ്രബന്ധത്തിന്റെ അവസാനത്തിൽ യാഥാർഥ്യ ജീവിതത്തിലെ സേവന കേന്ദ്രങ്ങളിൽ സമയാനുസൃതമായ കാത്തിരിപ്പ വരികളുടെ നിർണ്ണായക സ്ഥാനം എടുത്തുകാട്ടുന്നതിനൊപ്പം ഭാവി ഗവേഷണത്തിനായി പ്രവർത്തനക്ഷമമായ ശുപാർശകൾ നൽകുകയും ചെയ്യുന്നു.

സൂചനാ പദങ്ങൾ: സമയാനുസൃതമായ കാത്തിരിപ്പ വരികൾ, ശ്രേണി വരികളുടെ ശൃംഖലകൾ, ക്ഷണികമായ പ്രവർത്തന അളവുകൾ, തടയൽ, പൊതുസ്വഭാവമുള്ള വരികളുടെ ശൃംഖലകൾ.

Contents

List of Figures

List of Tables

CHAPTER 1

INTRODUCTION

1.1 Introduction

Queueing theory explores the dynamics of queues or waiting lines, focusing on scenarios where the demand for service exceeds the capacity to provide the service immediately. This field studies systems where entities (such as customers, calls, or patients) requiring service arrive at a facility, wait in line if service is not instantly available, and depart after being served. Queueing theory has widespread applications in diverse fields, including telecommunications, traffic engineering, and the operations of manufacturing and service industries to optimise performance and efficiency. The progress and evolution of queueing theory continue to advance, both in its methodologies and applications. Innovative analytical approaches are being developed to enhance its theoretical foundations, while new areas of application are constantly emerging. This study aims to contribute to this vast and dynamic field by enhancing the existing body of knowledge.

When the queue, a fundamental component of a queueing system, becomes excessively long, impatient entities may abandon the system. This behaviour negatively

impacts the overall efficiency of the queueing system. A lengthy queue is, therefore, a clear indicator of congestion in the system and the main motivation of studies in queueing theory is to control this congestion. This congestion in the system is primarily measured using two key metrics, that are the number of entities in the queue and the time an entity spends waiting in the queue before receiving service. In certain cases, the mathematical formulation of these metrics can become complex. This presents a significant challenge in queueing theory, requiring sophisticated analytical and computational approaches to address effectively.

In many cases, a single queue and its corresponding service mechanism may not be sufficient to meet the diverse requirements of the entities. Consequently, interconnected queues or a queueing network may be necessary to address varying needs. In these models, entities receive service from queues sequentially based on their requirements. Queueing networks are more practically applicable in many real-world scenarios because they effectively model complex systems where entities require multiple stages of service. A typical example of a queueing network in a healthcare system involves multiple interacting service points, such as the registration or admission desk, outpatient departments for consultations with specialists, laboratories and diagnostic centres for tests like blood work, X-rays or MRIs, and pharmacies for dispensing prescribed medications. These interconnected queues form a network through which patients navigate to fulfil their needs. The overall performance of a queueing network depends on the efficiency of all its service centres. The primary goal of queueing theory studies is to effectively model such service systems using queueing networks and develop mathematical formulations to minimise waiting times. In the present research, some queueing networks where conventional theories are not applicable are considered. Therefore, advanced methodologies such as approximation methods and simulation techniques are applied. These approaches

complement traditional methodologies and enhance the understanding of complex queueing networks.

1.1.1 Background of the Study

Queueing systems with constant parameters, such as fixed arrival rates, service rates, and number of servers, dominate the literature due to their mathematical tractability. These stationary models assume time-invariant behaviour, simplifying analysis and solution methods. However, this assumption often fails to represent the dynamic nature of real-world service systems, where arrival and service rates vary over time. For instance, call arrivals in customer service centres, traffic at toll plazas, and patient flows in hospitals exhibit significant fluctuations in arrival and service rates. The non-stationary queueing models provide a more realistic framework but introduce additional analytical challenges. In such models, transient analysis is more effective than steady-state analysis because it captures the dynamic behaviour of the system. Transient analysis provides detailed insights into the state of a system at specific times which enables the identification of peak congestion periods. The general purpose of this thesis is to explore on transient analysis on non-stationary queueing systems. Bertsimas and Mourtzinou (1997) and Fralix and Riaño (2010) developed the transient version of Little's Law, which is the foundational result in queueing theory, explaining powerful relationship between average waiting time and queue length, that is $E[L] = \lambda\ E[W]$ where λ is the arrival rate, $E[L]$ is the average queue length, and $E[W]$ is the average waiting time. Another important relationship, which connects the mean workload to the mean waiting time, is presented by Brumelle (1971) is, $E[Z] = \rho E[W] + \rho E[V^2]/2\mu$ where $E[Z]$ is the average workload, $E[W]$ is the average waiting time, ρ is the traffic intensity, V is the service time, and μ is the mean service time. The main idea of this thesis is initiated from these

works. This thesis attempts to derive a transient version of Brumelle's formula for a single-server queueing system and for tandem queueing networks.

Relaxing the Markovian assumptions of arrival and service processes in time-varying queueing models further complicates the analysis, as traditional memoryless property can no longer be applied. The resulting mathematical complexity in analysing general non-stationary queueing systems can be addressed using approximation methods. Whitt and You (2022) developed an approximation algorithm, the robust queueing network analyser (RQNA), to evaluate the steady-state performance in a network of non-Markovian stationary queues. The present research introduces a time-varying approximation algorithm for a network of non-Markovian non-stationary queues, building upon the results in Whitt and You (2019).

In many real-world service systems, waiting areas have a physical limit on the number of entities they can accommodate. This finite buffer causes excessive waiting times, which can degrade overall system performance. To model such systems, a queueing system with finite buffers is required. Analysing these systems is challenging because the time a customer spends in the system does not follow the Markovian property. This complexity increases when the system parameters vary over time. This study aims to advance the understanding of such systems by addressing these challenges, using an outpatient system with finite capacity queues as a reference. Choosing an appropriate blocking mechanism is crucial to maintaining the system's efficiency. This study seeks to provide insights that are both rigorous and applicable to real-world scenarios.

1.1.2 Research Objectives

The objectives of this research are as follows:

- To advance the understanding and analysis of queueing networks, particularly

in non-stationary contexts.

- To develop a theoretical framework to support the analysis of non-Markovian queueing systems, addressing the complexities associated with time-varying parameters.
- To study on various aspects of queueing systems, including those with time-varying parameters and restricted queue capacities.
- To address analytical challenges in queueing theory using programming languages and computational tools.
- To explore inferential methods for estimating the time-varying parameters of queueing models.
- To examine the practical applicability of non-stationary queueing systems across various real-world scenarios.

1.1.3 Significance of the Study

In queueing theory, the average workload represents the total amount of unfinished work in a system and it plays a crucial role in analysing and understanding the performance and behaviour of queueing systems. In the case of a non-stationary queueing system, the virtual workload is critical, as it captures the dynamic nature of the system. The virtual workload represents the total remaining service time in the system, including the work for all currently waiting and in-service customers, at the instant a hypothetical new customer arrives. Virtual workload provides insights into system congestion and helps evaluate how a queue evolves over time. In this study, an explicit integral formula for a time-varying non-Markovian system is derived. In time-varying systems, virtual workload aids in understanding transient

behaviour, making it crucial for systems with fluctuating arrival or service rates. A stochastically useful and widely applicable queueing model is the tandem queueing network. This study also extends the theoretical explanation of transient measures, such as queue length and virtual workload, to tandem queueing networks.

Most real-world service systems have restricted queue capacity. Therefore, implementing finite buffers in a time-varying tandem queueing network not only increases its practical applicability but also raises its computational complexity. Blocking mechanisms define how a system handles incoming entities when there are queue capacity restrictions. This study compares different blocking mechanisms within the context of an outpatient clinic.

When considering a general open Jackson queueing network with time-varying arrival rates, the analysis becomes extremely difficult because it does not possess the product-form property. In such cases, approximation methods are more appropriate. The time-varying approximation algorithm presented in this study provides time-dependent approximations for performance measures such as queue length, waiting time, and sojourn time.

In non-stationary queueing systems, time-varying arrival rates are not constant but are instead a function of time. This time-dependent arrival rate function, known as the intensity function, plays a critical role in the modelling and analysis of such systems. However, in real-world service systems, estimating this function can be challenging due to the fluctuations in arrival patterns over time. This study addresses the challenge by exploring an existing estimation method, through a simulation and a real-case study, with the help of advanced computational tools to improve accuracy and efficiency.

Queueing theory has a wide range of applications, and this study considers two different practical areas: healthcare systems and telecommunications systems. The

present study examines a two-node tandem queueing network with finite buffers, using the outpatient clinic case to compare different blocking mechanisms. A simulation of call centre data is also conducted, along with detailed analysis. Additionally, real bus arrival time data is utilised to study the inferential aspects of the intensity function.

1.2 Queueing Theory- Basic Concepts

The basic characteristics of a queueing system are essential for modelling and predicting queue behaviour, enabling a better understanding of how queues operate and are analysed. A queueing system is typically defined by its arrival process, service process, queue discipline, number of servers, and queue capacity. These characteristics influence how entities move through the system, the time they spend waiting, and the efficiency with which they are served. The key characteristics of a queueing system are as follows:

- **Arrival Process (Input Pattern):** This describes how entities arrive to the system. The arrival process determines the rate of arrival of entities to the service facility and it is represented by the distribution of the inter-arrival times (the time between successive arrivals). The most commonly used notations for arrival patterns include:
  - **M** - Poisson arrivals or Exponentially distributed inter-arrival times. M indicates the Markovian property.
  - **D** - Deterministic arrivals. Arrivals are scheduled or occur at exactly known times. This arrival pattern is characterised by a constant time interval between successive arrivals.

- **G** - General or arbitrary inter-arrival distribution, allowing greater flexibility in modelling various real-world scenarios.
- $\mathbf{E_k}$ - Erlang Arrivals, that is, arrivals are less variable than those described by the Poisson process. Here inter-arrival distribution is Erlang(a special case of the gamma distribution) of order k.
- $\mathbf{H_k}$ - Hyper-Exponential arrival processes where the inter-arrival times are described by a mixture of exponential distributions.
- **MAP** - Markovian arrival process. This pattern allows for arrivals to be dependent on previous events. This is useful for more complex systems where arrivals are not completely independent.

◆ **Service Process (Service Mechanism):** This refers to the manner in which how entities are served once they have reached the front of the queue. It comprises the service rate (the rate at which the server can process entities) and the distribution of service times. Some of the most commonly used service time distributions are as follows:

- **Exponential(M)** - It is often used due to its memory-less property, that is, the probability of service completion in the future is independent of the amount of service time they spent in the past.
- **Deterministic(D)** - This is the scenario where service times of the entities are fixed and known in advance. Unlike random service time distributions, deterministic service times do not vary and are the same for all entities.
- **General(G)** - The process allows service time to follow any distribution so it can be applied to a wide range of real-world scenarios where service times vary in a non-specific manner.

- **Number of Servers:** This refers to the number of serving units available in a queueing system to provide service for entities in the waiting line. This can be either one or more than one.
  - **Single Server** - Only one server is available for serving customers.
  - **Multiple Servers** - A number of servers are available, which can operate in parallel to serve multiple customers at the same time. It may be possible to have a single queue for all the servers, or to have a separate queue for each server.
- **Queue Capacity:** This refers to the maximum number of entities the queue can permit. The capacity of some systems may be infinite, which means there is no limit to the number of entities that can wait in the queue, while others may have a finite capacity, which limits the number of entities that can wait in queue.
- **Queue or Service Discipline:** This is the criteria by which entities are selected for service from the queue. The most common queue disciplines are:
  - **First-come-first-served (FCFS)** - It is a fundamental service discipline where entities are served in the order of their arrival.
  - **Last-come-first-served (LCFS)** - In this discipline, entities are served in the reverse order of their arrivals. In other words, the last-arriving customer in the queue will be served first.
  - **Priority** - Customers are served based on priority criteria, without considering the order of their arrival, that is, the high-priority customers are treated ahead of the lower-priority customers.
  - **Random Selection** - Customers are served at random, irrespective of the order of their arrivals. Thus, there is an equal chance of selection for each

customer in the queue.

Kendall's Notation

A queueing system can be described using a simple shorthand method called Kendall's notation, introduced by Kendall (1953) in 1953. The standard format of Kendall's notation is $A/B/C/X/Y$, where A describes the arrival process or inter-arrival distribution, B the service time distribution or service pattern, C represents the number of servers available in the system, X indicates capacity of the system and Y specifies the queue discipline. For example, $M/G/1/\infty/FCFS$ represents a queueing system with Poisson arrivals, general service time distribution, a single server, infinite queue capacity and FCFS queue discipline. A common simplification of Kendall's notation exclude the queue capacity and service discipline leading to the notation $A/B/C$. This model assumes infinite queue capacity and a FCFS service discipline by default. This notation has been extensively used to represent a wide range of queue types.

1.3 Performance Measures

Performance measures are considered as quantitative descriptions of the performance of a queueing system. These indicators are essential for evaluating and understanding the behaviour of the system. Input specifications that describe a queue include the information on the arrival process, the service process at servers, the number of servers, capacity of the queue and queue discipline. These parameters need to be provided to model a system and that lead to the derivation of performance measures which helps to examine the performance of the system. Waiting time plays a crucial role in the theory of queueing systems. Waiting time in a queueing system are of two types, the time an entity spends in the waiting line or queue and the total time

spends in the system(including service time), which is also called *sojourn time* or *response time*. The waiting time depends on factors such as the number of entities in the system, the number of servers available, the service discipline, and the service rate. Similarly, there are two counting measures as well, the number of entities waiting in the queue and the total number of entities in the system. The latter measure is significant when there are more servers and only one queue in the system. Another important metric is the workload, which represents the total amount of unfinished work in the system from the server's perspective. It accounts for the cumulative service time needed to complete all tasks for entities currently in service and those waiting in the queue. An idle time in a queueing system is the time during which the server is not busy. The percentage of time through which any particular server is idle or the entire system is idle is referred to as the idle-service measures. These quantitative descriptions of a queueing system differ depending on whether the system is in a steady-state or a transient state.

1.3.1 Steady-State Behaviour

A queueing system is considered to be in a steady-state or equilibrium state when the probability of the system being in a particular state is not time-dependent. Steady-state performance measures are used to examine the long-term behaviour of a queueing system when it reaches a stable condition after a sufficiently long period of operation. The equilibrium state is achieved when the arrival rate of entities to the system is equal to the departure rate of entities over time. Steady-state measures are particularly useful for analysing average or typical queueing performance under constant circumstances. For example, L and W are the long-term average number of entities in the system and long-term average time an entity spends in the system, respectively. Similarly, average number of entities waiting in the queue for service

and the amount of time they wait in the queue are described by the measures L_Q and W_Q, respectively. There is a fundamental relationship between these measures, developed by Little (1961) called Little's Law. Little related the steady-state average number of entities to the average waiting time in the system as follows, $L = \lambda W$. For the measures in the queue, this becomes, $L_Q = \lambda W_Q$, with mean arrival rate to the system λ.

1.3.2 Transient Behaviour

In most of the real life service systems, the study of transient behaviour is more meaningful since it deals with the operating behaviour of the system for a finite amount of time. In other words, transient performance measures focus on the behaviour of a queueing system over a specific period before it reaches stationary state with changes in the input conditions of the system. These measures are essential for understanding how a system responds to time fluctuations and the analysis can be challenging due to the time-dependent nature of the performance measures. Most queueing models are treated under stationary conditions, assuming constant arrival and service rates, with the system operating in a stabilized state. However, in real-world service systems, these parameters often vary over time, making the system time-dependent and requiring performance measures that capture its transient behaviour. Some key transient performance measures include time-dependent queue length, which represents the number of entities in the system at a specific time; waiting time, which is the amount of time an entity spends in the queue after arriving at a specific time; and virtual workload, which refers to the total remaining service time in the system at a given moment, including both waiting and in-service entities. These measures are essential for understanding and analysing the dynamic behaviour of time-varying queueing systems.

Even though the stationary or equilibrium distribution of a queueing system exists and can be obtained, there is a possibility that several stochastic processes to have same stationary distribution. Whitt (1983) identified the risk of using the stationary distribution alone and discussed the necessity to examine the transient performance measures.

1.4 Queueing Networks

Queueing networks are systems in which queues are interconnected, allowing entities to move between different service stations according to their needs. Most of the real-world systems such as computer networks, manufacturing systems, transportation, and telecommunication systems can be modelled and analysed by using queueing networks. In addition to the key characteristics of a queueing system discussed in Section 1.1, nodes and routing are also essential components for queueing networks. Each node in a queueing network represents a service station. Each station may have one or more servers to provide service to arriving entities. Routing is the direction of entities through the network. This may be probabilistic or deterministic. After being served at a node, if the entities are routed to other nodes according to a set of probabilities, then it is called probabilistic or random routing. Mathematical representation of the probabilities is called routing matrix, typically denoted by $\mathbf{P}$. Deterministic routing uses predetermined paths for entities through the network. A queueing network can be visualised as a system where each node and its associated queue, is interconnected by directed lines, representing the flow of entities between the nodes. The growing interest in the network of queues in recent years provides the main impetus and further motivation for studying the performance measures of typically complex stochastic models.

A network of queues can be described as a collection of k nodes, where each node $i,\ i = 1, 2, ..., k$ represents a service facility with c_i servers. The routing matrix $\mathbf{P}$ is a square matrix where each element p_{ij} represents the probability for moving an entity from node i to node j after completing the service at node i. The order of the matrix depends on the number of nodes in the network, that is, for a network with k nodes the routing matrix will be of order k. A routing matrix is always a sub-stochastic matrix, each row of the matrix sums up to 1, reflecting the total probability of an entity leaving a node to either another node or exiting the network.

1.4.1 Classification of Queueing Networks

Queueing networks can be fundamentally classified based on their network topology into three main categories: open, closed and mixed queueing networks. Henceforth, for the ease of understanding, the major component of a network, entities are assumed to be customers.

❖ **Open Queueing Network**

In open queueing networks, customers enter the network from external sources, receive service from one or more nodes, and eventually depart from the network after their service requirements are fulfilled. Networks of this type are typical in many real-world applications where customers are continuously entering and exiting. Jackson Queueing Network is a specific type of open queueing network model that is widely used to analyse Markovian systems.

❖ **Closed Queueing Network**

Closed queueing networks consist of a fixed number of customers that circulate between the nodes, there being no additional external arrivals or departures from the system. Once a customer completes the service at a node, they are routed

to another node and continue to circulate through the network. This model is especially useful in environments where the total number of jobs or items is conserved, such as computer networks or manufacturing systems. The closed Markovian network is also called the Gordon-Newell network. If in a closed network, nodes are connected in series and the customers rejoin the first node after completing service at the last node, such networks are called cyclic networks.

❖ **Mixed Queueing Network**

Mixed networks combine the features of both open and closed queueing networks. In these networks, some fixed customers may circulate within the network indefinitely (as in closed networks), while others enter and exit the network and traverse it (as in open networks). This type is less common but can be seen in more intricate systems where different types of transaction occur simultaneously.

1.4.2 Network Operations

Queueing network operations involve various mechanisms to manage customer flow and service within the network. Splitting, superposition and feedback are the basic network operations, see Whitt (1983). These are integral for addressing the dependency on customer flow within the network and can be applied to perform analytical and approximation methods based on parametric decomposition.

❑ **Splitting (Thinning) -** Splitting occurs when a queue of customers after receiving service from a particular node is divided into multiple sub-queues, which then proceed independently through the network.

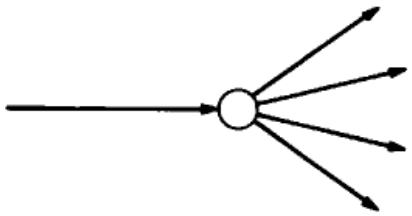


Figure 1.1: Splitting

- **Superposition (Merging) -** Superposition in queueing networks refers to the merging of customers from different nodes into a single queue for receiving service from a particular node.

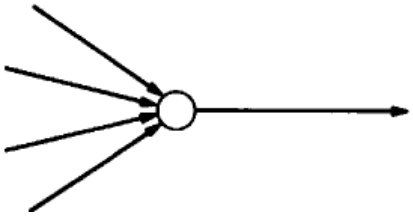


Figure 1.2: Superposition

- **Departure -** Departure is the flow of customers from a node after service. This can be arrivals to other nodes. For a series or tandem queueing network, a departure process is again an arrival process to the next node.

Figure 1.3: Departure

- **Feedback -** Customer feedback refers to a mechanism whereby customers return to a node and join the queue that they have previously visited. A feedback can be of two types, namely immediate feedback and delayed feedback (after visiting some other nodes).

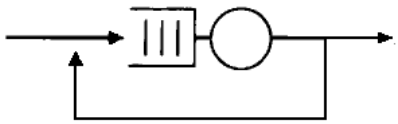


Figure 1.4: Feedback

- **Blocking -** Customer blocking occurs when a customer cannot enter a queue or proceed to the next node in the network due to capacity constraints or other limitations. There are different types of blocking mechanisms, including blocking-after-service(BAS) and blocking-before-service(BBS). Figure 1.5 demonstrates BBS blocking mechanism which is most common in communication systems.

Figure 1.5: Blocking

1.4.3 Classification of Customers

In queueing networks, customers are often categorised into different classes based on their service requirements, priorities, and routing behaviours. This classification improves the accuracy of the model, particularly in complex service systems that offer multiple types of services, such as those in telecommunications, healthcare, and computer systems. The network may have two classes of customers, with one class receiving priority over the other. For example, in healthcare systems, emergency services are prioritized over non-emergency patients. Customers can also be classified according to the type of service or service rate they require. Another common classification is based on the route they follow, that is, different classes may follow

distinct paths within the network. A mixed queueing network can also be defined in situations with multiple job classes within the system, incorporating both open and closed systems. Rosti et al. (1997) examined closed-queue product form networks with two classes of customers and two single-server fixed-rate stations. Kelly (1976) studied the equilibrium behaviour of queueing networks with different types of customers.

1.4.4 Prominent Types of Queueing Networks

Queueing networks are most commonly classified as open, closed, or mixed depending on whether customers are allowed to enter and leave the system, whether a fixed number of customers circulate within it, or a combination of both. Another important criterion for classifying queueing networks is based on the Markovian property. Queueing networks can be categorized as Markovian or non-Markovian, depending on the distributional assumptions for arrival and service processes. Cox (1955) was the pioneer in the analysis of non-Markovian process. Filipowicz and Filipowicz and Kwiecień (2008) reviewed most popular queueing networks which are used for the performance analysis of various systems including computer, communications, transportation networks and manufacturing. In the 1950s, Jackson introduced fundamental models in queueing theory, including the Jackson Open Queueing Network (OQN) and tandem queues. In 1967, Gordon (1967) introduced the concept of closed queueing networks. In 1968, Mandelbaum and Avi-Itzhak (1968) introduced a new class of queueing systems involving the splitting and merging of queues, later termed Fork-Join systems. By 1975, multi-class queueing networks gained attention, particularly Baskett et al. (1975) developed BCMP networks. This section explores these key types of queueing networks that are widely used to model real-world service systems.

Tandem Network(Series Queues)

A series queue or tandem queue is a type of open queue network in which customers enter the system from outside at node 1, flow always in one direction, and leave after receiving services from required nodes. Customers visit each station sequentially, one after another, in a predefined manner. The servers at each node can be either single or multiple.

Consider a simple series of Markovian queues as an example. There is no restriction on waiting space, and customers are not allowed to revisit previously visited nodes (feed-forward). Customers arrive at the first node according to a Poisson process, and the service time at each node follows an exponential distribution. Bruke's theorem(Burke (1956)) states that, *for a queueing system in which the input process is Poisson, the service process is exponential, and there is no capacity restriction on the queues, the output distribution is the same as the input distribution.* Therefore, in a tandem queueing system with Markovian queues, the departure time distribution is equally as important as the arrival time distribution, since the departure time from one node is identical to the arrival time at the next node. Burke's theorem enables the decomposition of queues, facilitating the analysis of each queue separately when multi-server queues are connected in series with feedback restrictions. Some examples of tandem queues include a registration process where a registrant visits multiple desks sequentially to complete the process, a physical examination where patients undergo a series of tests, and an assembly line in manufacturing where a product moves through various stages, such as machining, inspection, and packaging Shortle et al. (2018).

Open Jackson Network

Jackson (1957) introduced the mathematical framework for open Jackson networks

to analyse systems of interconnected queues. This was the pioneering study on networks of queues under the Markovian assumption. These networks are widely applied in various service systems with multiple service stations. In an open Jackson network, customers arrive from outside the system according to a Poisson process, receive service at one or more stations in the order of arrival (FIFO), and eventually leave the system. The service times at each station are mutually independent, exponentially distributed, and independent of the arrival processes. The flow of customers through the network follows Markovian routing. That is, after completing service at station i, customers move to station j with probability p_{ij}, independent of the system's state. Each queue in the network can be modelled as an independent $M/M/c/FIFO/\infty$, queue, where c is the number of servers.

A remarkable feature of the Jackson OQN is that the steady-state joint probability distribution of the number of customers at each station has a product-form solution, as described by Jackson's theorem. It states that, the joint distribution of states across all queues can be expressed as the product of the marginal distribution of individual queues. Product-form networks are those with simple closed form expressions for the steady-state distribution, which makes it possible to evaluate average performance measures. Another implication of Jackson's theorem is that once the flow rates are determined, the queues can be considered independently. Melamed (1979) extended Bruke's theorem in open Jackson networks and demonstrated that the departures from internal nodes and the sum of departures follow a Poisson distribution. A detailed study of the Jackson open queueing network is presented in the next chapter.

Closed Jackson Network(Gordon-Newell Network)

A Gordon-Newell network is a closed network of Markovian queues where customers do not enter or leave the system. In this type of closed network, a fixed number of customers circulate through the network, following probabilistic routing. Customers are served based on an exponential service time distribution and follow a First-Come-First-Served (FCFS) service discipline. As a closed system, the external arrival rate is zero. The probability of a customer moving from node i to node j, denoted as p_{ij}, is independent of the state of the nodes. The sum of the transition probabilities from a given node to itself and all other nodes is equal to unity, making the routing matrix $\mathbf{P}$ stochastic in nature.

In a network with k nodes, the system states are represented by $n_1, n_2, n_3.....n_k$, where the total number of customers is fixed, $n_1 + n_2 + n_3 + ... + n_k = M$. Similar to the open Jackson network, the closed Gordon-Newell network also exhibits the product-form property. This means that the joint probability distribution of the states at each node can be expressed as the product of the state distributions at each node. The Gordon-Newell network is widely used to model closed systems with fixed populations, such as manufacturing lines, computer systems, communication networks for performance analysis, and resource optimization.

BCMP Network

A BCMP network is a general closed queueing network with multiple classes, introduced by Baskett et al. (1975). In this network, a fixed number of customers (or jobs), categorized into different classes, circulate through the nodes. The transition probability $p_{ir,js}$ denotes the probability that a customer of class r at node i moves to node j and changes to class s. The service rate for a customer i in class r is denoted by μ_{ir}. Depending on service discipline, service time distribution, and

customer classes, nodes can be divided into four types.

- Type 1: Service centres with multiple servers and exponentially distributed service times, all classes of customers serve identically and according to FCFS service discipline.

- Type 2: Service centres with single server, different classes of customers have a distinct service time distribution and follow processor sharing(PS) service discipline. The service time distribution is general and has rational Laplace transforms.

- Type 3: Service centres with a sufficient number of servers such that customers do not need to wait before service(infinite servers), and different classes of customers have a distinct service time distribution. The service time distribution is general and has rational Laplace transforms.

- Type 4: Service centres with single-server different classes of customers have a distinct service time distribution and follow LCFS service discipline. The service time distribution is general and has rational Laplace transforms.

BCMP networks are widely applicable for modelling service systems with multiple classes of customers interacting with various service stations, such as computer systems, transportation networks, and manufacturing units (Balsamo et al. (2015), Iglesias et al. (2019), Mizuno et al. (2024)). For example, the chemotherapy unit in an oncology hospital can be modelled using a BCMP network, as it is typically administered in several treatment sessions, with rest periods between sessions. The modelling aspects of a chemotherapy unit are discussed in Filipowicz and Kwiecień (2008).

Fork-Join Network

A Fork-Join Queueing Network is a specialized type of queueing system that is used to model parallel processing systems. In such networks, customers or jobs are initially decomposed into multiple independent queues and processed separately at different service stations. After completion of service at all service stations, customers or jobs join together to form a single queue. The process of splitting jobs or customers is called the forking process, and reassembly of customers after the service is complete is called the joining process. Joining occurs only when all forked tasks have completed their service at the parallel nodes. Thus, the joining time equals the maximum of the arrival times from the parallel service stations. Fork-join networks are queueing systems without feedback. The key elements of these networks are queues, forks, and joins. Parallel nodes typically consist of single-server systems with unlimited waiting space, operating under a first-come, first-served (FCFS) service discipline. Konstantopoulos and Walrand (1989) discussed the stationarity and stability of fork-join networks, while Varki and Dowdy (1996) studied the properties of balanced fork-join queues.

Based on whether synchronization is required after service completion at parallel nodes, fork-join systems can be divided into two types: fork-join systems with synchronised queues and fork-join systems without synchronized queues. In a fork-join system with synchronised queues, each parallel node has an associated queue that serves as a waiting space for customers who have completed their service at the node. This ensures that the parallel nodes remain available for subsequent customers. Conversely, if a synchronisation queue is absent, customers who complete service at the parallel nodes are forced to wait at the same node, this will block the next customers from being served (see, Bose (2013)). Fork-join queues are widely used to model systems that involve parallel processing, such as cloud computing,

telecommunication networks, and manufacturing systems.

1.5 Organisation of the Chapters

Queueing networks are commonly used to model large-scale service systems. In most real-world scenarios, service systems exhibit time-varying arrival and / or service rates, which typically fluctuate throughout the day. Time-varying queueing systems offer a more realistic framework for analysing systems where parameters are not constant but change over time. This study focuses primarily on time-varying queues and their performance measures. The thesis is organised into eight chapters.

Chapter 1 serves as the introductory chapter and provides a foundational framework for the study. It presents an overview of the basic concepts and ideas of queueing systems and queueing networks.

Chapter 2 addresses analytical issues related to complex queueing networks. It provides a detailed discussion on non-product form queueing networks and the analytical techniques used to study them. In addition, this chapter also includes a literature survey and explores various applications of queueing networks.

Chapter 3 introduces time-varying queueing systems and examines the inferential aspects of the intensity function in a non-homogeneous Poisson process. The chapter includes both simulation study and real data analysis to address and validate the estimation problem.

Chapter 4 explores transient performance measures in time-varying single-server queues. It includes the derivation of an explicit integral formula for virtual workloads in such queues, accompanied by a simulation study to demonstrate its application.

Chapter 5 presents the formulation of transient performance measures in tandem queueing networks, such as the number of customers in the system and virtual

workload, along with numerical illustrations.

Chapter 6 examines the impact of queue capacity constraints, which can lead to blockage of entities in service systems. It investigates the comparative time-varying performance of the blocking-after-service (BAS) and blocking-before-service (BBS) mechanisms by modelling a hospital emergency department using a tandem queueing network with finite capacity queues.

Chapter 7 focuses on an open single-server general queueing network with time-varying arrival rates. It presents a general framework for an algorithm to obtain time-varying approximations of performance measures.

The final chapter provides concluding remarks and offers a brief overview of potential directions for future research. In the end, a comprehensive list of references used in the study is included.

CHAPTER 2

SOME ANALYTICAL ISSUES ON QUEUEING NETWORKS

2.1 Introduction

Queueing networks were developed to model and predict the behaviour of systems with interconnected queues that provide service for randomly arising demands. The earliest problems studied were related to telephone congestion in 1909. Since then, researchers have tackled many practical challenges by building strong foundations for queueing networks, focusing on assumptions and techniques for analysis. However, queueing networks, especially those with complex configurations often present significant analytical challenges. Key issues include non-Markovian assumptions, interdependence between queues, finite queue capacity or blocking, multi-class systems, feedback or routing dependencies, and non-stationary systems.

Traditional queueing models assume that inter-arrival and service times follow exponential distributions, based on Markovian assumptions. When these assumptions are relaxed, that is when inter-arrival or service times follow general or deterministic

distributions, the analytical complexity increases significantly. Queueing networks have extensive practical applications in telecommunications and manufacturing systems. A major challenge in such queueing systems is the presence of limited buffer capacities, which can lead to the blocking of entities. Handling such systems is analytically challenging.

As the number of service stations in a queueing network increases, exact analysis becomes cumbersome. Decomposing such large-scale networks into smaller subnetworks can simplify the analysis. While traditional models assume common service discipline such as FCFS, more complex systems may employ various service disciplines such as processor sharing, priority queues, or Last-Come-First-Serve with preemption. These diverse service disciplines require specialised methods for performance analysis. Feedback of entities and state-dependent routing are common in healthcare scenarios, further complicating the analysis. Another significant source of analytical complexity arises from non-stationary or time-varying characteristics of systems. Analysing these systems often requires time-dependent models or approximation methods. To effectively model and analyse the performance of such networks, a combination of approximation techniques and simulation approaches may be necessary. This chapter discusses these analytical challenges in detail and describes the analytical and approximation techniques, as well as simulation methods, that can be applied to study complex queueing networks. Additionally, it explores practical applications of queueing networks.

2.2 Literature Review

The foremost queueing problems were based on telephone traffic congestion, which was investigated by the Danish mathematician Erlang (1909). Erlang published

'The theory of Probabilities and Telephone Conversations' in 1909, which laid the bedrock for Poisson distribution in queueing theory. Erlang introduced some of the most eminent concepts and techniques which are fundamental in the study of queueing systems. The Erlang loss models and flow balance equations (later called the Chapman-Kolmogorov equations) are such contributions. The investigation of the theory of telephone communication continued after Erlang. The practical problem of congestion was the main motivation behind the works in the twenties and thirties done by Erlang and others. Molina (1927) and Fry (1928) provided extensions to Erlang's contributions. Over the next two decades, several statisticians became attracted to these problems and developed general models that can be applied to more complex systems. In the early 1930s, Pollaczek did some significant works on single-server queue with Poisson input and general service-time distributions and established a well-known formula, see Pollaczek (1930). Two years later Khintchine modified the formula in probabilistic terms, and then it is known as Pollaczek–Khintchine formula. Other researchers who have made significant contributions include Palm (1937); Palm (1938); Palm (1947), Crommelin (1932); Crommelin (1934) and Feller (1949).

Most of the basic queueing theory models have been extensively studied during the 1950s and 1960s. Some notable works are done by Kendall (1951); Kendall (1953), Syski et al. (1960), Saaty (1961), and Bhat (1968). In equilibrium state, since the flow balance equations are simple, it is relatively easy to derive a limiting distribution for the queue size. But to study the time-dependent behaviour of a system, there is a need for advanced mathematical techniques. Bailey (1954) addressed this problem, utilised generating functions and Laplace transforms to derive the time-dependent differential equations. Study of more general queueing systems like M/G/1 and G/M/c systems has been initiated by introducing imbedded

Markov chains. Lindley (1952) derived integral equations for waiting time distributions in general single-server queue, GI/G/1. Cox (1955) proposed analysis of non-Markovian process by adding a supplementary variable.

Some pioneering works in queueing networks were carried out by Jackson (1957, 1963), who studied open queueing networks under the Markovian assumption. This network model later became known as the open Jackson network. In subsequent work, Jackson generalized his results and performance measures for open queueing networks under specific assumptions. Kingman (1969) provided significant theoretical insights into systems modelled using Markovian frameworks, aiding in the analysis of complex queueing networks. Halfin and Whitt (1981) investigated systems with many servers under heavy traffic conditions, which are common in large-scale systems such as hospitals and cloud computing. Ward Whitt made substantial contributions to queueing networks, focusing on the performance analysis of stochastic systems and their applications in telecommunications, healthcare, and service systems. The Queueing Network Analyzer (QNA), developed by Whitt (1983), is a computational tool that analyses queueing network performance using decomposition techniques. Disney and Konig (1985) reviewed random processes in queueing networks, including queue length processes, sojourn times, and flow processes, presenting results for continuous-time and embedded processes. Sigman (1990) proved that, to achieve stability in open queueing networks, it is sufficient to assume that service time distributions have finite first moments. In the literature on non-stationary queueing models, Pointwise Stationary Approximation (PSA) has been widely used as a method for analysing time-varying systems. Green and Kolesar (1991) and Whitt (1991) addressed the approach of PSA to approximate non-stationary performance measures by using a stationary model with a constant arrival rate.

Major developments in queueing networks during the first decade of the 21st

century include extensions to non-Markovian models, heavy traffic approximations, and fluid models. Whitt provided a comprehensive guide to applying stochastic process limits in queueing theory, focusing on scaling and approximations in Whitt (2002). Bramson and Dai (2001) proved heavy traffic limit theorems for various families of multi-class queueing networks, including single-server systems and re-entrant queues operating under different service disciplines. Bolch et al. (2006) provided both theoretical foundations and practical methods for the performance analysis of systems modelled using queueing networks and Markov chains. Meyn (2008) addressed the stability and control of queueing networks, particularly using fluid models, and offered an in-depth treatment of Lyapunov stability theory and optimization. Boucherie and Van Dijk (2010) focused on exact analytical results for queueing networks. They also discussed fluid limits for analysing system stability and diffusion approximations for multi-server systems. Queueing network modelling of patient flow in hospitals can significantly enhance performance. Armony et al. (2015) conducted an exploratory data analysis (EDA) on a large hospital dataset, revealing critical features that are not readily explained by existing models. By the end of the second decade, time-varying queues gained considerable attention in the queueing literature, see (Whitt (2015); Whitt (2018); Whitt and You (2019); Pender (2017); Schwarz et al. (2016)). Recent trends in queueing theory literature include the implementation of machine learning techniques for the performance analysis of queueing systems. Kyritsis and Deriaz (2019) presented a machine learning approach for predicting waiting times in queueing scenarios, using bank queues as an example. To build an accurate simulation model, Pan et al. (2021) proposed a novel two-level routing component for queueing network models and utilized machine learning tools to calibrate the routing components. A detailed review of an approach combining traditional methods of queueing theory with various machine learning algorithms

is presented in Vishnevsky and Gorbunova (2022). They concluded that the application of machine learning methods is highly effective and promising for further research. Efrosinin et al. (2024) explored supervised machine learning, particularly artificial neural networks, to estimate queue lengths when only a small number of tagged customers is observable. They addressed this problem in the context of vehicle queues at traffic lights.

2.3 Jackson Queueing Network

A Jackson Queueing Network is a foundational model in queueing theory, widely used to analyse complex service systems comprising interconnected service centres. An open Jackson network represents a specific type of network where customers can enter the system from outside, move through various interconnected service nodes, and eventually leave the system. The following assumptions define the conditions under which a queueing network qualifies as a Jackson queueing network. Consider an open queueing network with k nodes, where the i^{th} node has c_i servers.

- Customers arrive from outside at any node according to Poisson process. Let the arrival rate be γ_i, for $i = 1, 2, ..., k$.
- Service times at each node follows exponential distribution with rate μ_i and mean $1/(\mu_i)$, for $i = 1, 2, ..., k$.
- Customers after receiving service at node i, proceed to the node j with probability p_{ij}. Customers depart from the network through node i with probability, $1 - \sum_{j=1}^{k} p_{ij}$.
- Let λ_i be the total average rate of customer arrivals to node i. This can be written as the sum of external Poisson arrival rate to the node i and the arrival rate of

customers from other nodes to node i, $\sum_{j=1}^{k} p_{ji}\lambda_j$, that is

$$\lambda_i = \gamma_i + \sum_{j=1}^{k} p_{ji}\lambda_j, \quad i = 1, 2, ..., k. \tag{2.3.1}$$

These equations are known as traffic rate equations, flow balance equations or conservation equations. The existence of the preceding equations is necessary for the existence of steady-state distribution in a Jackson queueing network, see Medhi (2003).

- The network contains only one class of customers.

Jackson's Theorem provides a fundamental result for obtaining the steady-state distribution in open Jackson networks. For the state $(n_1, n_2, ..., n_k)$, where n_i represents the number of customers at node ii (including both those in the queue and those being served), the theorem states that the joint probability distribution in the equilibrium state is given by:

$$p(n_1, n_2, ..., n_k) = p_1(n_1)p_2(n_2)....p_k(n_k) \tag{2.3.2}$$

where $p_i(n_i)$ denotes the marginal probability that there are n_i customers at node i. For an $M/M/c_i$ queue with c_i servers at node ii, the marginal probability $p_i(n)$ is given by:

$$p_i(n) = \begin{cases} p_i(0)\frac{(\lambda_i/\mu_i)^n}{n!} & \text{for } n = 0, 1, 2, ..., c_i \\ p_i(0)\frac{(\lambda_i/\mu_i)^n}{[c_i!\ c_i^{n-c_i}]} & \text{for } n = c_i + 1, ... \end{cases} \tag{2.3.3}$$

Here assume that the network is stable with traffic intensity,

$$\rho_i = \lambda_i/\mu_i < 1, \quad i = 1, 2, ..., k$$

Once the average flow rates are determined, the queues in a Jackson network can be analysed independently, even in the presence of feedback within the network. The states of the individual queues behave as if they are independent, which means the joint probability of the system's states can be expressed as the product of the state probabilities of the individual queues. Additionally, although the flows entering individual queues may not be strictly Poisson due to feedback paths, they behave as if they are Poisson. These are the most significant implications of Jackson's theorem, see Bose (2013).

The steady-state balance equations in (2.3.2) are generally referred to as the product-form solution. Since the product form property allows independent examination of the queues, they are feasible to exact analytical methods and are easy to analyse. Due to this useful theory, Jackson network has been extensively studied and applied to several service systems such as vehicle routing, cloud computing system, healthcare related systems, etc.

2.3.1 Product-Form Networks

Product-Form Queueing Networks are a class of queueing models in which the steady-state joint probability distribution of the network state can be expressed as the product of the marginal probabilities of individual nodes. A queueing network is said to be of product form if its joint probability distribution is represented as,

$$p(n_1, n_2, ..., n_k) = \prod_{j=1}^{k} p_j(n_j)$$

For a queueing network to exhibit product-form property, it must satisfy the following conditions (see Sahner et al. (1996)).

- Markovian routing: The routing of customers (or jobs) between service centres must be Markovian.
- Queueing disciplines: The network can employ specific queueing disciplines, such as FCFS, Last-Come-First-Served Preemptive-Resume (LCFSPR) and processor sharing.
- Service time distributions: Service time distributions must be differentiable. For FCFS centres, the service time distribution must specifically be exponential with the same mean for all customer types.
- Mixed networks: The network can be a mix of open and closed networks. The external arrivals to open networks must be Poisson.

Jackson (1963) introduced the concept of the product-form property in open Jackson queueing networks to derive simple closed-form solutions for the steady-state probability distribution. Consequently, open product-form queueing networks are often referred to as Jackson networks. The extension of product-form solutions to closed queueing networks was first explored by Gordon (1967). They proposed a set of assumptions on the model's characteristics and provided closed-form expressions for the steady-state distribution and average performance measures. Baskett et al. (1975) contributed a significant result in queueing theory with the BCMP theorem, which generalises product-form solutions to BCMP queueing networks. These networks are characterised by open or closed models, various service disciplines, general service time distributions, and multiple customer classes. In the case of closed networks, the steady-state probability distribution in product-form solutions is expressed with a normalising constant.

Product-form networks exhibit several useful properties. The most important of these is the arrival theorem, which states that a customer arriving at a queue in

an open network observes the stationary state distribution of the network. In contrast, a customer arriving at a queue in a closed network sees the stationary state distribution with one less customer than the overall network state. Another notable property is insensitivity, which states that the steady-state distribution and average performance measures depend on the service time requirements only through their mean. Similarly, the performance measures depend on customer routing only through the average visit ratio to each service centre. Another important property is that the aggregation method produces exact results. The aggregation theorem, first introduced by Chandy et al. (1975), allows for the substitution of a single-server subnetwork. The aggregated network maintains the same performance characteristics, ensuring that the overall behaviour remains consistent (see Balsamo (2000)).

Product-form queueing network models are a very useful class of models for assessing system performance. In contrast, Non-Product-Form Queueing Networks are those where the steady-state joint probability distribution of the number of customers across nodes cannot be expressed as the product of the marginal probabilities at individual nodes. These networks typically involve complexities or interactions that violate the assumptions necessary for product-form solutions. Examples of non-product-form networks include those with general distributional assumptions, finite buffers, priority service disciplines, or state-dependent routing probabilities.

2.3.2 Generalized Jackson Queueing Network

When the Markovian assumptions on inter-arrival and service times are relaxed, Jackson networks extend to Generalized Jackson Queueing Networks (GJQNs). Therefore this can be refer to extensions or generalisation of the classic Jackson queueing network. In a standard Jackson network, each queue is typically modelled as an M/M/c queue, where arrivals follow a Poisson process, service times are exponen-

tially distributed. In this generalized version, service times or inter-arrival times are not necessarily follow exponential distribution. GJQNs can be seen as an open network generalisation of the GI/GI/c queue, where both inter-arrival and service times follow general distributions. This can be Phase-type, hyper-exponential or deterministic. However, the routing policy remains Markovian with a sub-stochastic routing matrix. So the arrival times(mean of renewal process), service times of customers and routing mechanism together are enough to fully describe the evolution of the queueing network. In GJQNs, multiple classes of customers can be handled. Each class may have its own queueing discipline, service rate, and priority.

GJQNs are useful for modelling more realistic service systems as they incorporate a wider range of system characteristics and complexities. Their flexibility in handling general arrival and service times, multi-class customer flows, and feedback loops makes them highly applicable in real-world scenarios such as telecommunication networks, manufacturing systems, healthcare systems, transportation systems, and more. Unlike Jackson networks, the steady-state distribution of a GJQN typically does not have an explicit analytical form, as they do not satisfy the product-form property. Therefore, approximation methods and simulation techniques are often employed for effective analysis. Some works in this area include contributions by Chen and Yao (2001), Wein (1989), and Blanchet and Chen (2019).

2.4 Blocking and Methods for Analysis

Most real-world service systems operate with limited queue capacities. Blocking occurs when a customer or job cannot proceed to the next queue due to these capacity constraints. In other words, if there is insufficient space in the target queue, the customer will encounter blocking and cannot proceed further. This phenomenon is

crucial for understanding real-world systems with finite resources, such as manufacturing lines, computer networks, or call centres. A blocking mechanism defines the rules that determine when a node becomes blocked, what happens during the blocking period, and the conditions under which the node is unblocked. Various blocking mechanisms, categorized based on their behaviour, are discussed in the literature and given below.

Blocking-After-Service(BAS): Suppose a customer, after completing service at node i, attempts to enter the next node j, but the queue capacity at the destination node is full. In this case, the customer, after receiving service at node i, is forced to wait before joining the queue at the destination node. In this blocking mechanism, the customer becomes unblocked only when a departure occurs from the destination node or when the number of customers at the destination drops below its maximum capacity. This mechanism is also referred to as manufacturing blocking or production blocking in the literature.

Blocking-Before-Service(BBS): Suppose a customer declares their destination node, j, before receiving service at the source node i. If the customer finds that the destination node j is full, the service at the source node i does not start, and the customer becomes blocked at the source node until space becomes available in the queue of the destination node. The service at node i becomes unblocked as soon as a departure occurs at the destination node j. The BBS blocking mechanism can be further divided into two subcategories based on server space utilization during the blocking period: BBS-SO (Server Occupied) and BBS-SNO (Server Not Occupied). In BBS-SO, the service space is used to hold the blocked customer, while in BBS-SNO, the server space cannot be used to hold the blocked customer. The BBS blocking mechanism has been widely used to model telecommunication and computer systems. It is also referred to as service blocking or immediate blocking

in the literature.

Repetitive-Service-Blocking(RS): Suppose a customer completes service at the source node i and finds that the queue at the destination node, j is full. In this case, the customer at node i starts receiving a new and independent service based on the service discipline. The RS blocking mechanism can be further classified into two categories based on how a blocked customer chooses the new destination node: RS-RD (Random Destination) and RS-FD (Fixed Destination). In RS-RD, after completing service at node i, the customer randomly selects a new destination. In RS-FD, the blocked customer declares their destination before service starts at node i. The RS blocking mechanism is primarily used to model telecommunication systems. Since the customer is rejected by the destination node due to its maximum capacity being reached, this mechanism is also referred to as rejection blocking in the literature.

Exact analysis of queueing networks with finite capacities and blocking can be achieved by modelling the system as a stochastic Markov process, specifically as a continuous-time Markov chain. This analysis enables the evaluation of various average performance measures, such as the joint queue length distribution at arbitrary times or arrival times, as well as passage time and cycle time distributions. In certain special cases, queueing networks with blocking admit a product-form solution, provided specific constraints on network parameters and blocking types are satisfied. Detailed discussions on product-form solutions for networks with blocking, as well as equivalence properties among different blocking models, can be found in Balsamo and Personè (1994), Perros (1994) and Simonetta Balsamo (2001). The Convolution Algorithm and Mean Value Analysis (MVA), widely used for analysing product-form closed networks, can also be effectively applied to closed networks with blocking.

Since general queueing networks with blocking do not exhibit product-form solu-

tions, they are typically analysed using approximate analytical methods or simulation techniques. Various approximation methods have been proposed for both open and closed queueing networks with finite-capacity queues to derive average performance measures and queue length distributions(see Perros (1994)). For closed queueing networks with finite capacities, notable approximation methods are Throughput Approximation (TA) (Onvural and Perros (1989)), Network Decomposition (ND) (Frein and Dallery (1989)), Approximate MVA (AMVA) (Akyildiz (1988)), and Maximum Entropy Algorithm (ME) (Kouvatsos and Awan (1998)). Approximation techniques to handle open queueing networks of finite capacity queues with various blocking mechanisms include Tandem Exponential Network Decomposition (TED) (Dallery and Frein (1993)), Tandem Phase-Type Network Decomposition (TPD) (Perros and Altiok (1986)), Acyclic Network Decomposition (AND) (Lee et al. (1998)) and Maximum Entropy Algorithm for Open networks (ME-O) (Kouvatsos and Xenios (1989)). For a comprehensive discussion of analytical methods for queueing networks with finite-capacity queues, see Simonetta Balsamo (2001), Balsamo (2011), and Bose (2013).

2.5 The Concept of Feedback Flows

In queueing networks, feedback occurs when entities that have completed service at a node are routed back to the same node for additional processing. This mechanism is particularly useful for modelling real-world scenarios where repeated service is required. For example, in telecommunications, data packets may need to be retransmitted due to errors; in manufacturing, defective products are returned to the assembly line for rework; and in hospitals, patients revisit healthcare professionals for follow-up consultations. Immediate feedback and delayed feedback are two

types of feedback mechanisms commonly used to model such scenarios. Immediate feedback occurs when an entity, after completing service at a node, is routed back to the same node immediately (without any delay) for additional processing. This mechanism is often used to model systems with real-time error detection, such as in manufacturing or telecommunications. Delayed feedback, on the other hand, occurs when there is a time lag between the completion of service and the entity's return to the system for further processing. This mechanism is applicable in scenarios where external factors influence actions, such as follow-up consultations in healthcare systems. For a feedback queue, the sojourn time (or response time) refers to the total time an entity spends in the system, from its arrival until it finally departs.

The concept of feedback was initially introduced by Finch (1959). According to Finch, there are two types of customer feedback in a cyclic queueing system: terminal feedback and single-server feedback, distinguished by whether or not the customer leaves the system after service. Takacs (1963); Takács (1977) presented analyses for both types of feedback systems: immediate and delayed. They employed the generating function method to analyse two-dimensional Markov models of a single-server queueing system with an unlimited queue and infinite orbit volume. D'Avignon and Disney (1977), Dudin et al. (2005), and Melikov et al. (2016b) focused on queueing models with immediate feedback. Delayed feedback was considered by Foley and Disney (1983), Pekoz and Joglekar (2002), and Van Do (2010). Meanwhile, studies by Melikov et al. (2016a) and Melikov and Aliyeva (2019) analysed queueing models with both types of feedback. In these studies, hierarchical phase integration algorithms were developed to compute the stationary distribution of the corresponding multidimensional Markov chains. Nazarov et al. (2021) used asymptotic analysis to study queueing systems with both types of feedback.

2.6 Analytical and Approximation Techniques in Queueing Networks

There are a wide range of mathematical techniques to analyse queueing systems with non-Markovian distributional assumptions. In such systems, since exact analytical solutions are intractable, advanced approximation and analytical methods have been employed. There are some approximation methods to be discussed in this section, including decomposition approximation, heavy-traffic approximation, sequential bottleneck approximation and Robust queueing approximation.

The decomposition technique is one of the most commonly used approximation methods for analysing non-product-form queueing networks. It involves studying individual queues separately and then combining the results to compute the overall network performance. In queueing networks, this method reduces the complexity of the network by breaking it down into subsystems and analysing each subsystem independently. The technique is motivated by the product-form property of Markovian queueing networks. Decomposition techniques are particularly useful for networks with general arrival and service processes, networks with blocking, and large-scale networks. Networks with non-exponential distributions or state-dependent parameters do not adhere to the Markovian assumptions of exponential inter-arrival and service times, making exact analysis more challenging. The decomposition method simplifies this complexity by analysing each node in isolation while approximating the interactions between them. Unlike infinite-capacity queues, where customers can always enter the system, finite-capacity queues with blocked or rerouted customers introduce complexities that impact the flow of customers between nodes. Decomposition helps approximate such networks by identifying the blocking mechanisms involved and incorporating blocking probabilities and effective arrival rates. For

large-scale networks with many interconnected nodes, exact analysis becomes impractical. In decomposition approximation, breaking down the network into smaller, more manageable sub-networks simplifies the evaluation of performance measures.

Kuehn (1979) introduced the decomposition approximation method for analysing general open queueing networks. This approach involves decomposing the overall network into subsystems, such as single-server general queues (GI/G/1), and analysing the subsystems individually. In this method, all related processes are considered only with respect to their first two moments. While decomposition approximation performs well, it can face challenges due to the interdependence of customer arrivals at different nodes. To address this issue, Whitt (1983) developed the Queueing Network Analyser (QNA) software, which approximately characterise the arrival processes by first two moments(mean and squared coefficient of variation) and then analyse the individual nodes separately. QNA can be considered a decomposition method or an extended-product-form solution, as it attempts to capture the dependencies among nodes. In addition to Whitt's work, several other researchers have contributed to decomposition approximation techniques for open general queueing networks. Pujolle and Ai (1986), Whitt (1994), and Marie (1979) have presented decomposition methods for both single-server and multi-server queues. Frein and Dallery (1989) focused on decomposition methods applicable to tandem queueing networks with exponential service times and blocking-after-service. Other contributors, such as Takahashi et al. (1980), Altiok and Perros (1987), and Perros and Snyder (1989), have developed decomposition approximation methods for analyzing open exponential queueing networks with capacity restrictions.

Heavy traffic limit approximation is another approach that can be applied to analyse entire queueing networks. Iglehart and Whitt (1970a,b) and Harrison (1973) established heavy traffic limits for feedback-restricted open queueing networks, while

Reiman (1984) extended these results to general open queueing networks. A relatively tractable method for approximating open queueing networks is the use of the Reflected Brownian Motion (RBM) process. Harrison and Nguyen (1990) developed a software package called QNET for performance analysis of queueing networks, motivated by heavy traffic theory. In QNET, the original network is approximated by a multidimensional RBM model, and its steady-state mean queue length is determined by calculating the stationary distribution of the multidimensional RBM. Harrison and Nguyen (1990) also introduced a simpler approach called ΠNET, which uses a product-form or decomposition approximation for the stationary distribution of the Brownian system model. Harrison and Reiman (1981), Harrison and Williams (1987) and Dai and Harrison (1992) studied the theoretical and numerical aspects of steady-state distribution of multidimensional RBM.

To overcome the computational complexity of the QNET algorithm in analysing large networks, Dai et al. (1994) developed a hybrid method called Sequential Bottleneck Decomposition (SBD), which combines decomposition approximation with heavy-traffic theory. This method generally outperforms other approximation techniques, including QNA and QNET. In SBD, the network is first partitioned into several ordered bottleneck sub-networks such that their traffic intensities are approximately equal. These sub-networks are then sequentially analysed using heavy-traffic theory. The analysis of a sub-network with k stations involves formulating a corresponding k-dimensional RBM and determining its steady-state distribution. Reiman (1990) proposed an alternative approach called Individual Bottleneck Decomposition (IBD) for general queueing networks. In IBD, the partitioned sub-networks consist of a single station. The main motivation behind this approach is that a single-station sub-network can be approximated by a one-dimensional RBM, which has a known stationary (exponential) distribution.

An alternative approach for analysing the performance of generalized queueing networks with multiple server queues, based on robust optimization, is the Robust Queueing (RQ) approximation, proposed by Bandi et al. (2015). The robust optimization approach provides closed-form solutions for system time in multi-server queues, even when dealing with potentially heavy-tailed arrival and service processes. The basic idea behind RQ is to model uncertainty in arrivals and services using a suitable uncertainty set and to analyse the worst-case performance. Even though the RQ is simple and useful, it remains challenging to identify suitable uncertainty sets and connect them with the original queueing system. Whitt and You (2018b) addressed this issue by proposing a new non-parametric formulation for approximating the continuous-time workload process in a single-server queue based on indices of dispersion. They exploited the powerful relation between the arrival index of dispersion(IDC) and the normalized workload in a single-server queue discussed by Fendick and Whitt (1989). This approach involves decomposing the network into individual general single-server queues, where arrival and service processes are partially characterized by their rates and IDCs. The analysis also incorporates three essential network operations superposition, splitting, and departure to evaluate the overall network performance.

2.7 Numerical Methods and Simulation

Numerical studies provide essential insights into performance of the queueing system, particularly when exact analytic solutions are complex or difficult to derive. In queueing theory, Markov chains and Laplace transforms are useful tools to obtain analytical solutions. Markov chains capture the dependencies within the processes involved in the queueing system, while Laplace transforms provide a convenient ap-

proach for solving differential equations. When considering time-dependent queueing systems which are modelled using first-order difference-differential equations, transient states may be difficult to analyse. In these cases, numerical methods can be more practical and efficient solution. Numerical methods are advantageous because they allow easy estimation of errors due to truncation of higher-order terms. Ashour and Jha (1973) examined transient solutions to non-homogeneous Markov processes, where parameters change over time. Two well-known numerical methods, the Runge-Kutta and Hamming techniques, are applied for solving these processes.

A major disadvantage of numerical techniques is that all system parameters are required to be specified numerically before the answers can be obtained. This will make the calculations to be redone when the values of the parameters are changed. Accordingly, a powerful technique to analyse complex queueing systems is to use computer simulation. But when a numerical solution is possible, it is typically much more cost-effective than using simulation.(Shortle et al. (2018))

Simulation is a powerful tool for studying complex queueing systems. It is an experiment which mimics the characteristics of a system to study the system empirically. In simulation, we produce the synthetic data related to the system rather than taking the real-world events to study the dynamic behaviour of the system. Since the basic idea behind simulation is conceptually simple to understand, it is accessible to a wide range of users. Conventional numerical methods solve specific systems of equations structured to each expectation type, while simulation provides a flexible approach to compute expectations of various system functionals without changing the basic approach, see Glynn and Iglehart (1988). Discrete event simulation is a method used to simulate systems where events occur at distinct points in time.

2.7.1 Discrete Event Simulation

Discrete event simulation(DES) is the most common form of simulation which quantitatively represents a queueing system, simulates the dynamics of the system by discrete events such as arrivals, services and departures, that involves the performance of the system. DES is particularly useful for the systems with unpredictable events and when the dependency in the interactions makes the analysis complicated. The main components of DES are entities(customers, calls, patients,vehicle etc), events(arrival, completion of service, departures) which make changes in the system's state, simulation time, servers and queues. Setting up and executing a DES involves several systematic steps, which are initialisation, simulation loop, data collection and termination.

Consider an example of implementing DES on a bank service desk with a single server. Primarily, set the initial condition such as empty queue and an idle service desk. Process the events by adjusting the simulation time and system states. Generate a new event(service start of the new customer) based on the current event(service end of the previous customer). Throughout the simulation, collect data on queue length and waiting times. Finally end the simulation based on the simulation time or any other criteria set on the number of customers. Obtained data can be used to asses the necessity of adding another server in the service desk. 'SimPy' is a library commonly used to implement DES in python.

2.8 Non-Stationary Queueing Systems

In most real-world service systems, customer demand is non-stationary or time-varying. There are several time-varying characteristics or system parameters, such as arrival rate, service rate, number of servers, and routing probabilities, that change

dynamically over time. Therefore, it is difficult to imagine real service systems with stationary characteristics. However, most of the theoretical queueing models assume stationary system parameters, since the mathematics of time-varying stochastic processes is highly complex. Time-varying queueing systems have been extensively used to model various service systems, including communication and transportation systems. As a result, the evaluation of queue behaviour in such systems is significant for practical problems. Some notable pioneering works include those by Koopman (1972), Kolesar et al. (1975), and Newell (1968).

Green et al. (1991) examined some effects of non-stationarity on the performance of multi-server queueing systems. Most common scenario is when the arrival process exhibits strong non-stationarity which needs to deal with an explicit analysis. There are several approaches to be employed. One approach is to focus on the period of peak arrivals and perform a separate (stationary) analysis, using an average arrival rate from the peak period. Another method is to divide the entire time period into segments, estimate average arrival rates for each segment, and apply a series of independent (stationary) analyses. A third approach is to explicitly capture the non-stationarity of the process using a simulation model. The choice among these approaches depends on understanding how non-stationarity impacts the accuracy of stationary approximations.

In general, the analysis of time-varying queueing systems is computationally complex, and it is difficult to produce closed-form expressions. Steady-state analysis is often insufficient for such systems, as it fails to capture their dynamic nature. Transient analysis, approximation techniques, and simulation methods are frequently employed to address this challenge. The remainder of this thesis mainly focuses on the study of time-varying queues.

2.9 Some Applications of Queueing Networks

Queueing networks has a wide range of well-known, but non-trivial, applications across various industries and fields. Since queueing theory is originally introduced to optimise telephone traffic operations in early 1900s, it is considered as the first area of application. Numerous other applications discovered since then.

Real-world telecommunication networks are characterised by large-scale dimensions and their ability to provide a wide variety of services with heterogeneous traffic. These systems can be effectively modelled mathematically using queueing networks. Queueing models also play a crucial role in predicting and managing network congestion by analysing the arrival rates of data packets and the service times at routers. Some notable studies on the applications of queueing theory in telecommunication systems include Palm (1988), Giambene (2014), Lakatos et al. (2013), and Afolalu et al. (2021). When it comes to the modelling of call centres, queueing theory helps to solve traffic congestion problems, minimise the count of dropped calls and minimising waiting times. By analysing customer arrival patterns, service times, and resource availability, queueing models help to predict call blocking probabilities and improve overall service efficiency. Key contributions in this area include the works of Koole and Mandelbaum (2002), Mandelbaum and Zeltyn (2007), and Kim and Whitt (2013b).

Hospitals and outpatient clinics utilise queueing operations to manage patient flow and improve system performance. Effective queue management enhances patient satisfaction, reduces waiting times, and optimises the utilization of facilities. Queueing networks are also employed to model emergency departments, including ambulances and emergency medical teams, to minimise waiting times and optimise service levels. Several studies, including those by Yom-Tov and Mandelbaum (2014),

Shi et al. (2016), and Dai and Shi (2017), have explored the applications of queueing networks in healthcare systems.

Another interesting application of queueing theory is in transportation systems. Queueing networks can be used to analyse and manage the flow of aircraft by predicting and minimising delays caused by congestion. They are also employed to optimise operations such as baggage handling, check-in counters, security lines, and gate assignments by reducing waiting times. Some notable works include, Galliher and Wheeler (1958) and Ebert et al. (2019). Queueing models are further utilised in public transportation systems, including buses, trains, and metro systems, to manage flow of passengers at stations, as explored by Newell (1965), Kurzhanskiy and Varaiya (2010), and Ran and Boyce (2012). Queueing networks are also applied in vehicle management on roads to model and optimise traffic flow, reduce congestion, and improve overall efficiency.

In addition, there is a wide range of applications, most of them are well documented in the literature. Some of the most important of these are computer networks (Robertazzi (2000)), banking and financial services (Olu (2019), Xiao and Zhang (2010)), toll booths (Wang (2017)), cloud computing (Jafarnejad Ghomi et al. (2019)) etc.

2.10 Summary of the Chapter

In this chapter, various analytical issues related to queueing networks have been considered. A review of the literature on queueing networks is presented, along with an extensive discussion of the concept of Jackson queueing networks and their generalisations. The importance of product form queueing networks, which is a prominent class of queueing networks with significant properties, is also discussed.

The two main operations in queueing networks, blocking and feedback, along with related literature, have described in detail.

Exact solutions for complex queueing networks are often intractable. This highlighting the necessity of approximation methods and simulation techniques. Accordingly, existing analytical approximation methods and simulation techniques have reviewed, followed by a discussion of some practical applications of queueing networks. Another significant challenge in queueing network analysis is the non-stationarity of queueing systems, which complicates the analytical process. The remainder of the thesis focuses on queueing systems with non-stationary characteristics.

CHAPTER 3

ASPECTS OF TIME-VARYING QUEUES

3.1 Introduction

Stationary queueing models assume that the key parameters governing the system, such as arrival rates, service rates, and system capacity, remain constant over time. These models are widely used due to their mathematical tractability and the simplicity they offer in analysing queueing systems. The majority of queueing systems discussed in the literature focus on stationary or constant model parameters, with notable contributions from Odoni and Roth (1983), Brandt et al. (1990), Baccelli and Brémaud (2012), and others.

Stationary models are effective in scenarios where system conditions do not fluctuate significantly, providing valuable insights into steady-state performance measures such as average waiting time, queue length, and system utilization. However, most large-scale service systems in real life are subject to time-varying conditions, including fluctuating arrival rates, service rates, and other factors that influence system performance. Stationary queueing models do not adequately represent real-world service systems that change over time. This limitation makes stationary queueing

models less suitable for applications involving highly dynamic environments where parameters change over time. To effectively represent such systems, non-stationary queueing models are necessary, as they can accommodate the time-varying nature inherent in real-world scenarios.

Emergency departments (EDs) in hospitals are classic examples of non-stationary queueing systems due to fluctuating patient arrival rates throughout the day. For instance, EDs often see higher patient inflows during evenings and weekends, when primary care facilities are closed. Notable studies on non-stationary queueing models in EDs include Flottemesch et al. (2007), McCarthy et al. (2008), and Defraeye and Nieuwenhuyse (2011). Non-stationary queueing models play a critical role in optimising ED operations, reducing patient waiting times, and ensuring timely care for critical cases. In outpatient clinics, the number of patient arrivals fluctuates throughout the day, with greater flow observed in the morning and evening compared to mid-day or late evening.

To effectively analyse such time-varying service systems, it is essential to capture the variations in arrival and service patterns over time. The rest of this thesis is dedicated to the study of time-varying queues, emphasising their importance and applicability in modelling real-life service systems.

3.2 Related Literature

Earlier works in stochastic systems with time-varying queues have done by Rothkopf and Oren (1979), Massey (1985), Newell (1982). Time-varying queues have been widely used to model complex service systems such as communication (Neely et al. (2003), Leung et al. (1994), Shakkottai et al. (2004)), (Newell (1965), Kurzhanskiy and Varaiya (2010), Ran and Boyce (2012)), healthcare (Yom-Tov and Mandelbaum

(2014), Shi et al. (2016), Dai and Shi (2017)) etc. Comprehensive surveys of the literature on time-varying queues have been provided by Defraeye and Nieuwenhuyse (2011), Defraeye and Van Nieuwenhuyse (2016), Schwarz et al. (2016) and Whitt (2018).

Various approaches have been proposed in the literature to approximate non-stationary queues by using related stationary queueing models. One of the simplest and most straightforward methods is the Simple Stationary Approximation (SSA) introduced by Green et al. (1991). This approach involves ignoring the time-varying nature of the arrival process and using the average arrival rate over the entire time period as an input parameter for a single stationary model. Another widely used method is the Pointwise Stationary Approximation (PSA), proposed by Green and Kolesar (1991) and Whitt (1991). In PSA, steady-state performance measures are calculated for each moment in time by using the instantaneous arrival rate at that moment within a stationary model. This method is particularly suitable for queues with a finite number of servers. A notable refinement of PSA is the Modified Offered Load (MOL) approximation, originally developed by Jagerman (1975) and later advanced by Massey and Whitt (1994). The MOL method operates similarly to PSA but replaces the actual arrival rate with the MOL arrival rate, which provides a more accurate representation of the offered load in non-stationary systems. MOL is significantly more precise than PSA for systems with longer service times, while it reduces to PSA when service times are shorter. The primary advantage of using stationary model approaches is their simplicity and computational efficiency, as these models are generally quick and straightforward to implement. However, because they rely on approximations, they do not always perform well. In certain scenarios, stationary models fail to capture the changes in the system over time, making it necessary to use methods that explicitly account for the time-varying nature of the

system for better accuracy.

Steady-state analysis of time-varying general queueing systems in the literature mainly focuses on analytical approximation methods and simulation methods, since exact mathematical analysis of such systems is intractable. Transient analysis of general queueing models is considered challenging and becomes much more complex when we account for the arrival and/or service rates changing over time. Little's law, established by John D. C. Little (1961) became the fundamental theorem in queueing theory due to its theoretical and practical importance. Bertsimas and Mourtzinou (1997) derived the transient version of Little's Law. They developed the concept of time-varying Little's law to establish a distributional relationship between the number of customers present in a queueing system at time t and the waiting time of all customers that arrive in the system until the time t. Fralix and Riaño (2010) extended the time-varying generalisation of Little's Law by using Palm measures and discussed higher order moment expressions. Kim and Whitt (2013a) utilised time-varying Little's Law to estimate the average waiting time in an infinite server queueing system and reduce the bias in the estimator.

Some research has focused on non-homogeneous Poisson process since it is advantageous to model arrival processes when the arrival rate varies over time. Li et al. (2019) proposed a method for estimating the model parameters of computer systems with job arrivals following a NHPP by using CPU utilization data. Kim et al. (2015) applied statistical tests to arrival data from an endocrinology clinic, where arrivals are consistent with an NHPP within each day. Queues with periodic arrival rate functions have been studied extensively in the literature due to their applicability in real-world systems where the arrival of jobs, customers, or data follows a predictable cycle, such as daily, weekly, or seasonal patterns. Early works of the periodic $M_t/GI/1$ model was done by Hasofer (1964), Harrison and Lemoine (1977), Lemoine

(1981). Recently Whitt (2014) established conventional heavy-traffic limits for the number of customers in a $G_t/GI/1$ queue with a periodic arrival process. Xiong et al. (2015) developed algorithms for the perfect sampling of time-varying queues with periodic Poisson arrivals under the first come first served (FCFS) discipline. In their assumptions, service duration followed periodic and time-dependent exponential or homogeneous distributions. Whitt and You (2019) developed a time-varying robust queueing algorithm for the continuous-time workload in a single-server queue with a periodic arrival-rate function.

Over the last few decades, tandem queueing networks have gained attention due to their practical applications in communication networks, manufacturing, toll-booths, supply chains etc. Recent research on applications of tandem queueing networks includes the works of Gerum and Baykal-Gürsoy (2022), Gangadhar and Kadambi (2022), Sinu Lal et al. (2020) and others. The transient behaviour of a single server two-station tandem network is investigated in Prabhu (1967) and transient analysis of k-node tandem queueing model with load dependent service rates is discussed in Rama Murthy M. et al. (2018). Zychlinski et al. (2018a) analysed tandem networks with general time-varying arrival rate and blocking, using time-varying fluid models.

3.3 Key Notations

Time-varying queues refer to queueing systems in which one or more parameters vary over time. In contrast to stationary queueing models, where parameters remain constant, time-varying queues account for fluctuations in key components such as arrival rates, service rates, and the number of servers. This section covers various notations and concepts associated with time-varying queueing systems.

1. According to Kendall's notation discussed in the first chapter, standard format to denote a time-varying queueing system is adding the component of time with the inter-arrival and service distributions. For example, $M_t/M_t/s/\infty$, where arrivals are according to non-stationary Markov process and service process is non-stationary and exponentially distributed.

2. The time-varying arrival process is denoted by $\{A_t, t \geq 0\}$. The associated arrival rate function, $\lambda(t)$, represents the instantaneous rate, i.e., the expected number of arrivals per unit time at time t. The most commonly studied arrival process is non-homogeneous Poisson arrivals with intensity function $\lambda(t)$. In the following section, this will be discussed in detail.

   The cumulative arrival rate is denoted as,

$$\Lambda(t) = \int_0^t \lambda(u)du, \tag{3.3.1}$$

   this represents the expected total number of arrivals to the system during the time interval $[0, t]$.

3. The time-varying service process is denoted by, $\{S_t, t \geq 0\}$ with service rate function $\mu(t)$. The cumulative service rate is given by,

$$M(t) = \int_0^t \mu(u)du, \tag{3.3.2}$$

   where $M(t)$ is the maximum total amount of service that can be provided in the time interval $[0, t]$.

4. Instantaneous traffic intensity $\rho(t)$ can be written as,

$$\rho(t) = \lambda(t)/\mu(t) \qquad t \geq 0, \tag{3.3.3}$$

indicating the load on the system at time t and time-varying traffic intensity can be expressed as,

$$\rho^*(t) = \sup_{0 \leq u \leq t} \{\Lambda(u)/M(u)\}, \tag{3.3.4}$$

see Massey (1985). The heavy traffic behaviour can be described as, if $\rho^*(t) > 1$, overloaded; if $\rho^*(t) < 1$, under-loaded; and if $\rho^*(t) = 1$, balanced Whitt (2018).

3.4 Non-Homogeneous Poisson arrivals

When modelling a real-world service system, it is crucial to capture its realistic features. A key aspect of them is the time variability of parameters, especially the arrival rate since it is the initial component of every system. A most important counting process for modelling the arrivals into a system is NHPP, as it relaxes the stationary increments assumption in a Poisson process. It allows the arrival rate to vary over time instead of being constant. The NHPP is a generalisation of Poisson process, which is also called non-stationary Poisson process. For a classical Poisson process, the arrival rate, λ is assumed to be constant, does not vary over time. An NHPP occurs when the arrival rate is allowed to depend on time, i.e., λ is a function of t.

A counting process, $\{N(t), t > 0\}$ is said to be an NHPP with intensity function $\lambda(t)$, $t > 0$, if

(a) $N(0) = 0$.

(b) $\{N(t), t \geq 0\}$ has independent increments, i.e., for $t_0 < t_1 < t_2 < .....$, the random variables $N(t_1) - N(t_0), N(t_2) - N(t_1), ...$ are independent. It does not have stationary increments, unlike the homogeneous Poisson process, because the distribution of the number of events in an interval depends on the location of the interval, not just its length.

(c) $P\{N(t+h) - N(t) \geq 2\} = o(h)$.

(d) $P\{N(t+h) - N(t) = 1\} = \lambda(t)\ h + o(h)$.(Medhi (2003), Ross (2014)).

The mean value function, $m(t)$ of the NHPP is defined by,

$$m(t) = \int_0^t \lambda(u)du$$

which is same as the cumulative rate of events during the interval $[0, t]$.

From a modelling point of view, the major weakness of the Poisson process is its assumption that events are just as likely to occur in all intervals of equal size. A generalisation, which relaxes this assumption, leads to the non-homogeneous or non-stationary process, see Ross (2022).

When using mathematical and statistical models to analyse data, it is common to deal with unknown parameters that need to be estimated. For a homogeneous Poisson process, the unknown parameter is a real-valued number. Several methods exist for estimating this parameter, including the method of moments, the method of least squares, and the method of maximum likelihood (see Dudewicz and Mishra (1988)). In the case of a non-homogeneous Poisson process, the process is parametrised by its intensity function, $\lambda(t)$, which adds complexity to the modelling process. Several studies have explored different methods for estimating the intensity function in NHPP. These include maximum likelihood estimation (Zhao and Xie (1996), Drazek (2013)), Bayesian inference (Jensen II (2022)), and non-parametric approaches (Bigot et al. (2013), Slimacek and Lindqvist (2016)). The following section focuses on the maximum likelihood estimation of the intensity function, $\lambda(t)$, in a non-homogeneous Poisson process.

3.4.1 ML Estimation of Intensity Function

For a real world dynamic system, arrivals occur according to a time-dependent arrival rate function. When we collect data from such systems, it would be difficult to find an arrival pattern or appropriate arrival rate function. In such cases, we need to estimate the arrival rate function according to the observed data. Markovian non-stationary queueing models assume that arrivals are governed by an NHPP. In other words, an NHPP with intensity function $\lambda(t)$ is a Poisson process which is obtained by allowing the arrival rate at time t to be a function of t. Here we consider Maximum likelihood estimation of intensity function $\lambda(t)$ discussed by Drazek (2013) to obtain the arrival rate function to the system. The same method has been applied to model NHPP in many disciplines, e.g, credit card fraud detection (Izotova and Valiullin (2021)), monitoring the quality of service in hospital emergency departments (Das et al. (2019)) and in the modelling of seasonal rainfall events (Ngailo et al. (2016)).

Drazek (2013) developed an estimation method for the intensity function by using the principle of maximum likelihood estimation(MLE). Suppose we have a sample of size n, $t_1, t_2, t_3, ..., t_n$. Let $\Lambda = \int_0^T \lambda(t)dt$ be the cumulative of arrivals over $[0, T]$, for the intensity function $\lambda : [0, T] \longrightarrow R_{\geq 0}$. Since we look at the Poisson process, probability of observing n points is $\frac{e^{-\Lambda}\Lambda^n}{n!}$. The probability density of a single observation is $\frac{\lambda(t_i)}{\Lambda}$. Since the observations are independent, the joint probability distribution of the sample is the product of the probability density of single observations, so the probability density of n observations becomes $\prod_{i=1}^{n} \frac{\lambda(t_i)}{\Lambda}$. Then the likelihood of the sample is the product of these two terms,

$$L(\lambda; t_1, t_2, t_3, ..., t_n) = \frac{e^{-\Lambda}\Lambda^n}{n!} \prod_{i=1}^{n} \frac{\lambda(t_i)}{\Lambda}. \tag{3.4.1}$$

If the sample is ordered, then the likelihood of the ordered sample can be obtained by multiplying (3.4.1) with $n!$. i.e.,

$$L(\lambda) = \frac{e^{-\Lambda}\Lambda^n}{n!} \prod_{i=1}^{n} \frac{\lambda(t_i)}{\Lambda} n! = e^{-\Lambda} \prod_{i=1}^{n} \lambda(t_i) \tag{3.4.2}$$

Log-likelihood function corresponding to (3.4.1) is

$$\log(L(\lambda)) = -\int_0^T \lambda(t)\ dt + \sum_{i=1}^{n} \log \lambda(t_i) \tag{3.4.3}$$

As the principle of MLE, our aim is to find a function $\lambda(t)$ which maximises the log-likelihood, $\log(L(\lambda))$.

3.4.2 Simulation Study

Considering simulated data from a call centre in which calls arrive according to NHPP, indicates the time-dependent arrival rate of calls. The call centre provides 24/7 service. In order to conduct the study, the data of a single day is generated, which processed 1433 call arrivals over this period. Over the course of the day, the arrival rate is not stationary, but it appears to be high during the middle part of the day. Assume that the arrival rate function is unknown, and the goal is to estimate this function using the available data. Since the arrivals are time-dependent, the problem is to find the intensity function corresponding to the non-homogeneous Poisson arrivals. In this study, Python 3.10.9 is the programming language of choice.

The data contains 1433 data values over a single day period. The first step is to visually present the data in a meaningful way. For that, partition the data into time periods of equal length and count the number of events within each interval. Here the length of each partition is one hour. Assume that the number of arrivals in each partition follows homogeneous Poisson process. For each partition, estimate

the parameter of Poisson distribution, together with the upper and lower limits of confidence intervals. The value of the estimates depends on the length of the time period of the partition.

For each partition, parameters are estimated and they are illustrated with its 95% confidence intervals in Figure 3.1. Based on the figure, it appears to be a positive half cycle of a periodic function. Now using the principle of Maximum likelihood to fit a best intensity function $\lambda(t)$ for the data. Among the four classes of functions discussed by Drazek (2013), consider using the exponential Fourier series function since it will ensure that the intensity function to be non-negative.

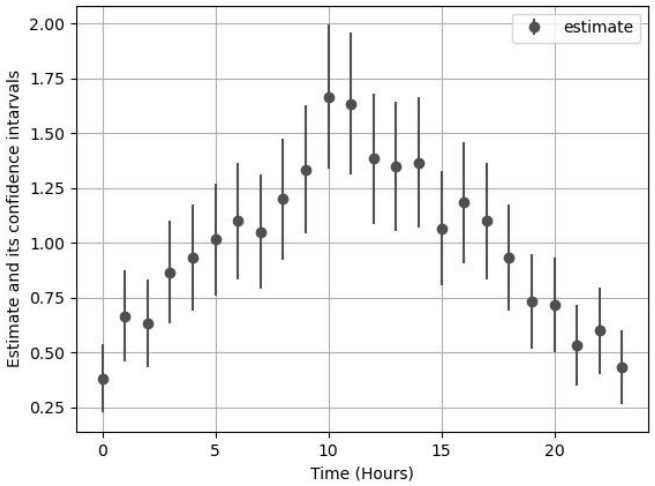


Figure 3.1: Estimates and their 95% confidence intervals with one hour grouping.

Let the exponential Fourier series function be,

$$\lambda(t) = \exp\left(a_0 + \sum_{j=1}^{m} b_j \sin\frac{2\pi t f_j}{T} + \sum_{j=1}^{m} c_j \cos\frac{2\pi t f_j}{T}\right) \tag{3.4.4}$$

Now the exponential Fourier model is to be fitted to the entire dataset and superimpose the fitted intensity curve with the estimates in Figure 3.1. If the fitted curve closely corresponds with the data, it will be considered a good fit. The parameter m

represents the number of frequency components in the Fourier series expansion. For $m = 0$, the fitted intensity function will be a constant, *i.e.,* homogeneous Poisson process with $\lambda(t) = exp(a_0)$. This is a horizontal line that passes through the centre of the estimates as an overall average. For $m = 1$, the exponential Fourier model becomes,

$$\lambda(t) = \exp\left(a_0 + b_1 \sin\frac{2\pi t f_1}{T} + c_1 \cos\frac{2\pi t f_1}{T}\right).$$

Figure 3.2 displays the fitted intensity function and the actual intensity function superimposed on the data, showing how closely the fitted model aligns with the data.

The fitted parameters of the exponential Fourier model for $m = 0$ and $m = 1$ are included in Table 3.1.

m	a_0	b_1	c_1
0	-0.005		
1	-0.065	0.053	-0.493

Table 3.1: Fitted parameters for the exponential Fourier model.

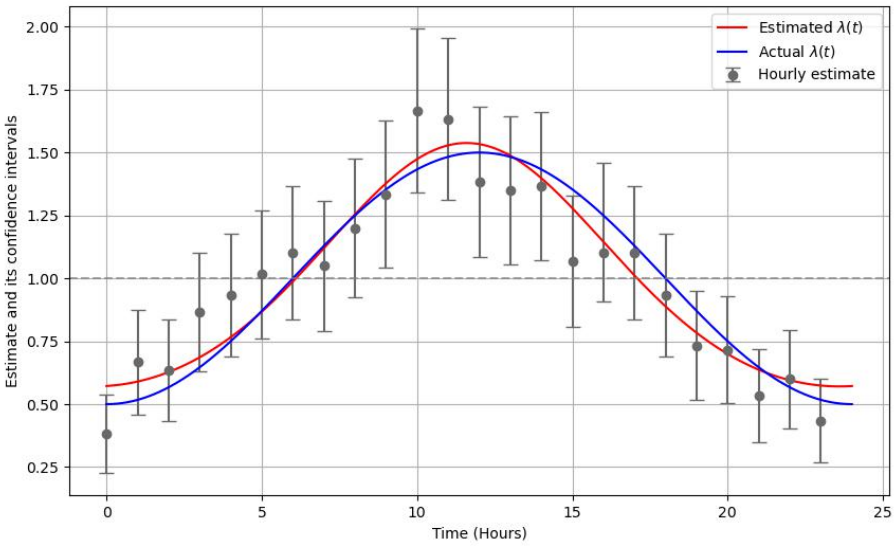


Figure 3.2: The fitted model superimposed on estimates and their 95% confidence intervals.

3.4.3 Real Case Study

This section examines a real dataset of bus arrival times in Kandy, the largest city in central Sri Lanka (Ratneswaran and Thayasivam (2023a,b)). As part of the Strategic Cities Development Project (SCDP) by Ministry of Urban Development of Sri Lanka, GPS devices were installed on buses to collect raw GPS data aimed at addressing traffic issues within the Kandy city area. Data was collected over a five-month period, from October 1, 2021, to February 28, 2022, for the bus routes between Kandy and Digana. In this study, the data of the month January 2022 is specifically considered for the analysis. There are mainly two directions: Direction 1 is from Kandy(Terminal 1) to Digana(Terminal 2), while Direction 2 is the reverse route, from Digana(Terminal 2) to Kandy(Terminal 1). The main focus of this study is the "end time" in the dataset, which records the arrival times of buses at the successive bus terminals.

There are 1203 bus arrivals in the terminal 2(direction 1) and 1209 bus arrivals in terminal 1(direction 2). In January 2022, bus arrival times at Terminal 1 range from 5:00 AM to 7:00 PM, while at Terminal 2, they range from 6:00 AM to 10:00 PM. However, these time ranges vary across different days. First, partition the dataset into equal one-hour intervals and count the bus arrivals within each interval. As in the previous simulation study, assume that the number of bus arrivals in each interval follows a homogeneous Poisson process. Next step is to estimate the arrival rate and its confidence intervals in each partition. The arrival rates are estimated together with its 95% confidence intervals for both directions and illustrated in Figure 3.3. There are a number of peak periods during the month, as well as inter-peak periods (intermediate peak periods). There is no linear trend evident in the monthly estimates of arrival rates, and it is also difficult to identify any other pattern present in Figure 3.3. For simplicity, consider each day of the month separately when partitioning the data into one-hour intervals and counting the bus arrivals. Among the 31 days in January 2022, focus only on the days that show stronger evidence of periodicity, in two groups, Direction 1 (Kandy to Digana) and Direction 2 (Digana to Kandy). For Direction 1, the estimated arrival rates appear more periodic on 10/01/2022, 12/01/2022, 18/01/2022, and 25/01/2022. In Direction 2, the days showing a periodic pattern in the estimated arrival rates are 05/01/2022, 15/01/2022, 25/01/2022, and 26/01/2022. Since the intensity function is assumed to be non-negative, the exponential Fourier series function is the best fit to the data. The parameter m in equation 3.4.4 represents the number of harmonics in the Fourier series expansion. As shown in Figure 3.3, a simple periodic pattern is observed, characterized by a single cycle of peaks and troughs repeating over the period. Therefore, the parameter m is set to be 1.

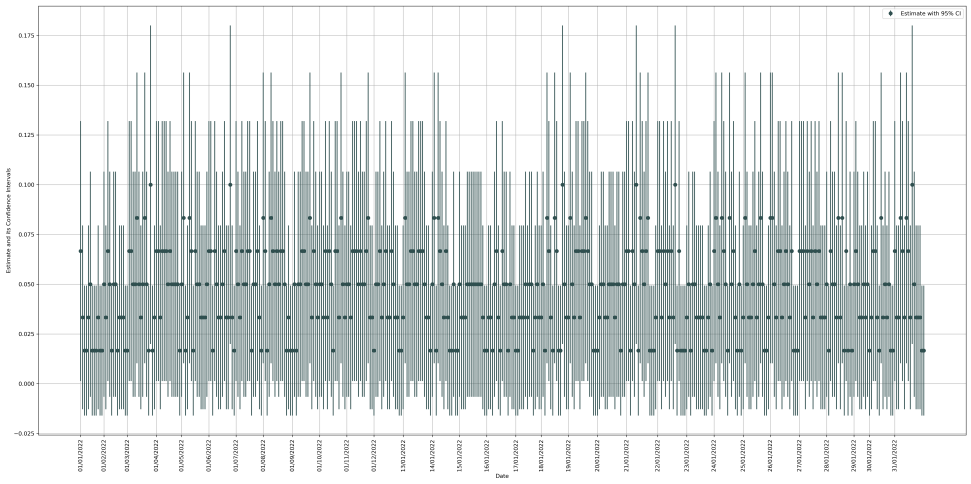
Let the expression for exponential Fourier series function, $\lambda(x)$ with periodicity,

m=1 be,

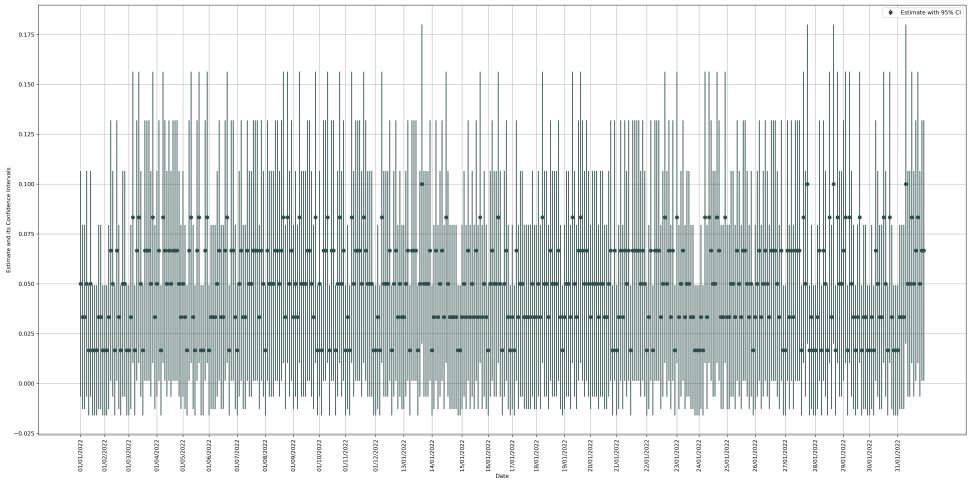
$$\lambda(t) = \exp\left(a_0 + b_1 \sin\frac{2\pi t f_1}{T} + c_1 \cos\frac{2\pi t f_1}{T}\right).$$

To fit the model, the parameters, a_0, b_1 and c_1 in the expression are estimated using MLE. The exponential Fourier model is fitted to the full dataset of a single day, and the fitted intensity curve is superimposed on the hourly estimates of arrival rates and their corresponding confidence intervals. Figures 3.4 and 3.5 illustrate how the fitted exponential Fourier function is superimposed on the data of two groups(direction 1 and direction 2). These figures help to assess how well the model fits the observed data. It is considered a good fit when the fitted curve closely follows the data points and falls within the confidence intervals.

To strengthen the validation, we additionally applied the Pearson's Chi-Square Goodness-of-Fit Test. The observed and expected arrival counts in the partitioned one-hour intervals were compared. The test resulted in a p-value greater than 0.05, which confirms that the fitted model adequately captures the arrival process and aligns well with the observed data.

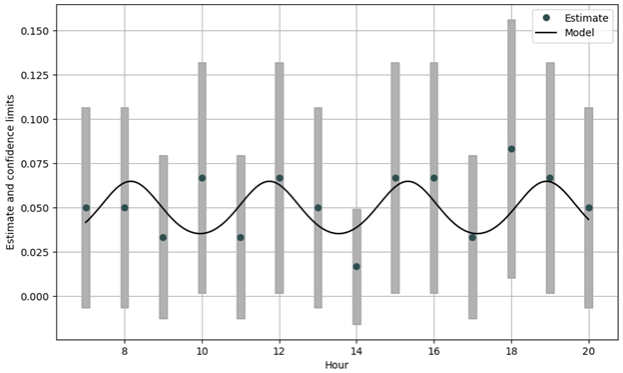


(a) Estimated rate of bus arrivals with 95% confidence intervals at Terminal 2 (Direction 1).

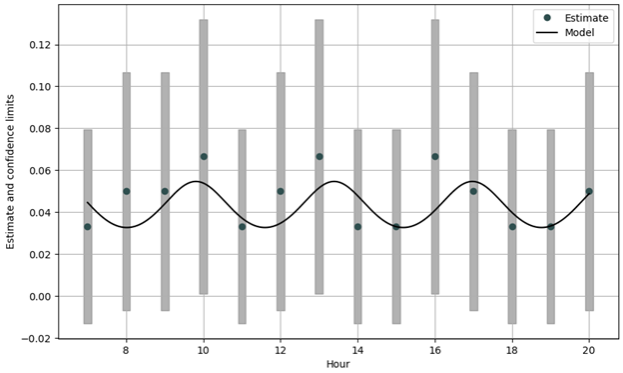


(b) Estimated rate of bus arrivals with 95% confidence intervals at Terminal 2 (Direction 1).

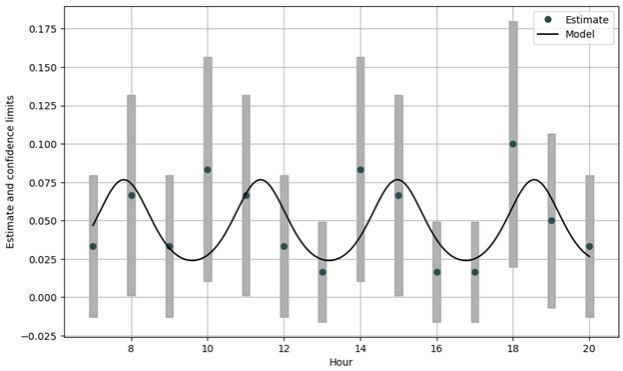
Figure 3.3: Hourly estimates with 95% confidence intervals for January 2022.



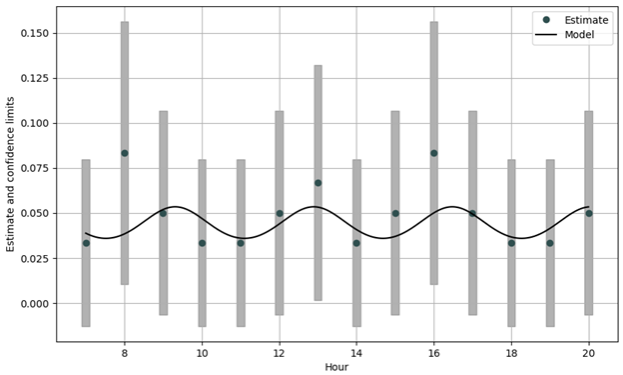
(a) Fitted exponential Fourier series for bus arrivals on 10/01/2022, with coefficients $a_0 = -3.042$, $b_1 = 0.3$ and $c_1 = -0.054$.



(b) Fitted exponential Fourier series for bus arrivals on 12/01/2022, with coefficients $a_0 = -3.162$, $b_1 = -0.256$ and $c_1 = -0.02$.

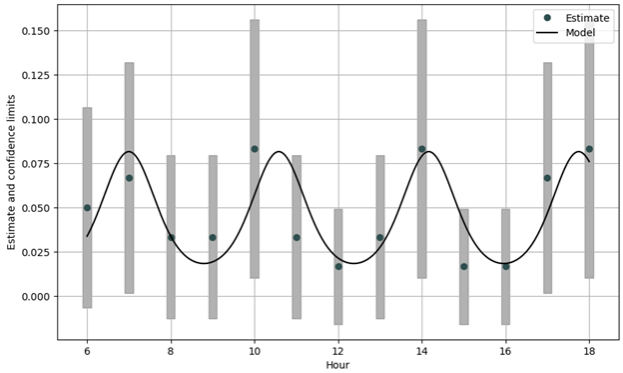


(c) Fitted exponential Fourier series for bus arrivals on 18/01/2022, with coefficients $a_0 = -3.151$, $b_1 = 0.524$ and $c_1 = 0.252$.

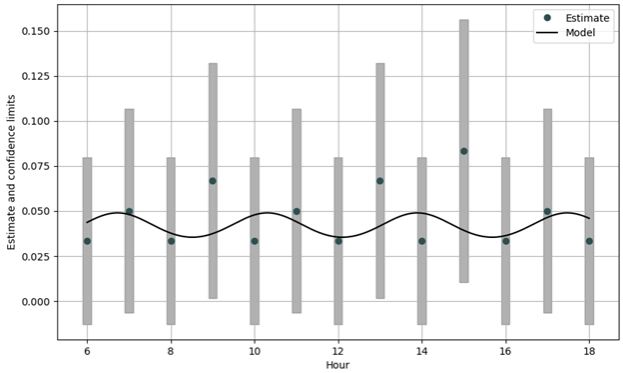


(d) Fitted exponential Fourier series for bus arrivals on 25/01/2022, with coefficients $a_0 = -3.129$, $b_1 = -0.115$ and $c_1 = -0.162$.

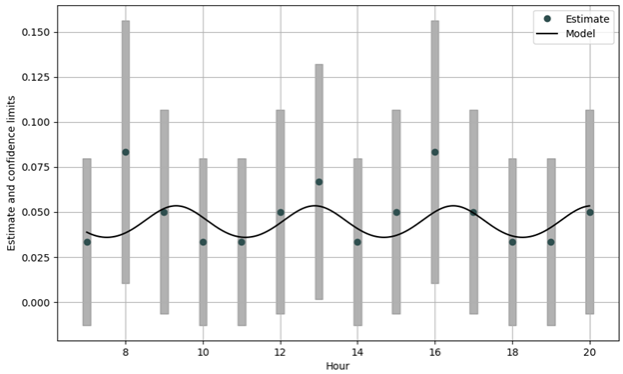
Figure 3.4: Estimates and confidence intervals for bus arrivals at Terminal 2 (Direction 1), with fitted intensity functions, illustrating arrival patterns.



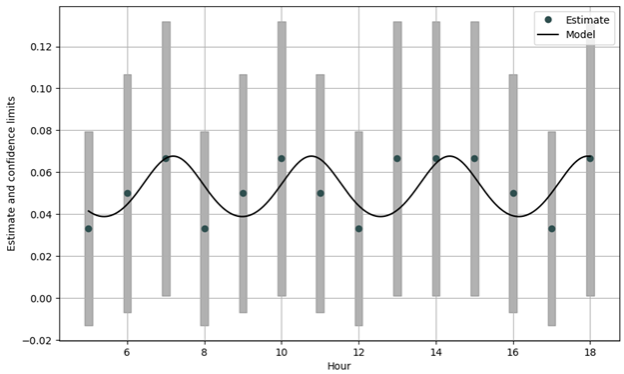
(a) Fitted exponential Fourier series for bus arrivals on 05/01/2022, with coefficients $a_0 = -3.254$, $b_1 = -0.216$ and $c_1 = 0.716$.



(b) Fitted exponential Fourier series for bus arrivals on 15/01/2022, with coefficients $a_0 = -3.180$, $b_1 = -0.113$ and $c_1 = 0.116$.



(c) Fitted exponential Fourier series for bus arrivals on 25/01/2022, with coefficients $a_0 = -3.138$, $b_1 = 0.269$ and $c_1 = 0.057$.



(d) Fitted exponential Fourier series for bus arrivals on 26/01/2022, with coefficients $a_0 = -2.969$, $b_1 = 0.0130$ and $c_1 = 0.276$.

Figure 3.5: Estimates and confidence intervals for bus arrivals at Terminal 1 (Direction 2), with fitted intensity functions, illustrating arrival patterns.

3.5 Challenges in Analysing Time-Varying Queues.

Although time-varying queueing models provide a more nuanced understanding of system performance under non-stationary conditions, they are not usually mathematically tractable. The computational methods and approximation techniques used in time-varying queueing problems have long been considered challenging. The analysis of such systems requires advanced mathematical techniques such as fluid approximations, numerical algorithms, and simulations in order to capture the dynamic behaviour accurately. Transient behaviour of a time-varying queueing model can differ significantly from the transient behaviour of a stationary model. Continuous-time Markov chains (CTMCs) can be used to study the transient behaviour in time-varying queues. This is achieved by allowing the transition rate matrix to depend on time and using the Kolmogorov forward and backward differential equations. Many algorithms have been developed to compute the transient performance of a stationary model for general initial conditions. These algorithms can be applied to calculate the time-varying performance of a time-varying model if we approximate the time-varying arrival-rate function by a piece-wise constant arrival rate function. Then we can recursively compute the time-varying performance on each interval by letting the initial distribution on each interval be specified by the terminal distribution on the previous interval. Another approach is using fluid models with time-varying parameters. The implementation of fluid models has been successful in modelling a variety of non-stationary service systems. When the deterministic variability in the arrival rate and departure rate becomes more significant than the stochastic variability involved in the rates, it may be reasonable to ignore the stochastic component of the model altogether. Additionally, if the number of customers or jobs in the system fluctuates over a wide range, the discrete nature of individual entities

might also be ignored. This leads to continuous deterministic fluid models being used as alternatives or approximations for discrete stochastic time-varying queueing models (see, Whitt (2018)). Addressing the challenges posed by the complexity and dynamism of time-varying queueing systems requires advanced statistical methods such as CTMC and fluid models. Additionally, simulation techniques such as discrete event simulation and adaptive algorithms play a crucial role in efficiently managing and analysing these dynamic systems.

3.6 Summary of the Chapter

This chapter provided a comprehensive discussion on time-varying queueing systems. A literature survey on time-varying queueing systems is included, covering existing analytical methods for modelling such systems. The chapter explored the concept of the non-homogeneous Poisson process and its inferential aspects, with a particular focus on the maximum likelihood method for estimating the intensity function.

A customer care call centre dataset is simulated, where arrivals followed a periodic function. The intensity function is estimated from the data and compared with the actual intensity function. The estimated intensity curve was closely aligned with the actual intensity function, suggesting a good fit. In addition, a real-world case study of bus arrival times at two terminals in Sri Lanka is conducted by estimating the intensity function. The fitted curve closely followed the hourly estimates of the data points. The chapter concluded by addressing some challenges associated with time-varying queueing systems.

CHAPTER 4

TRANSIENT PERFORMANCE MEASURES FOR TIME-VARYING SINGLE-SERVER QUEUES

4.1 Introduction

Traditional queueing models assume constant parameters, whereas time-varying queues take into account fluctuations in arrivals or service rates, making them more suitable for modelling real-world scenarios. The arrival rate of the entities varies depending on factors such as rush hours, seasonal demand, or service spikes. Similarly, service rates change based on server efficiency, resource utility, or staffing constraints. Since the system is not stationary, transient analysis is more important to study how queue lengths, waiting times, and other performance measures evolve over time.

The literature on time-varying single-server queues is extensive and it can be categorised into three main areas: (i) structural results, as illustrated by Harrison and Lemoine (1977), Heyman and Whitt (1984) and Rolski (1989); (ii) numerical algorithms, discussed by Choudhury et al. (1997), Kolesar et al. (1975) and Koopman

(1972); and (iii) asymptotic methods and approximations, explored by Keller (1982), Massey (1985), Mandelbaum and Massey (1995) and Whitt and You (2019).

Simulation is a powerful tool for analysing transient behaviour of non-stationary complex queueing systems. In simulation, a method for representing the time-dependent behaviour of queueing systems is to numerically integrate the time differential difference equations related to the system state probabilities. Taaffe and Gordon (1982) presented surrogate distribution approximation approach to reduce the number of differential difference equations numerically integrated to represent such systems. Taaffe and Ong (1984) explained the same approach for approximating the time-dependent distribution of queue size for queueing models with non-stationary phase type distributions. Using Volterra integral equations, Zhang and Coyle (1991) studied the transient solution of a single server queue with non-stationary arrival and service rates. Knessl and Yang (2002) considered a single-server queueing system with time dependent arrival rate and service rate, obtained an explicit analytic expression for the probability distribution function. Whitt (2015) considered a class of general $G_t/G_t/1$ single-server queues, including Markovian queues, with unlimited waiting space, time-varying arrival rate and the service rate at each time is subject to control. They introduced rate-matching control, which makes the service rate proportional to the arrival rate with a fixed traffic intensity, thereby discussed the stabilization of the queue length distribution and virtual waiting time. Ma and Whitt (2019) modified the concept of rate-matching control in order to minimize the maximum expected waiting time.

Bertsimas and Mourtzinou (1997) developed the transient version of Little's Law (Little (1961)), which is one of the fundamental principles of queueing theory. A time-varying version of Little's law was developed to establish a relationship between the number of customers present in a queue at a particular time and the waiting time

for all customers who arrive in the queue until that moment. Similar to Little's Law, Brumelle (1971) developed a useful formula, which relates average work in the system, service time and waiting time of a customer in a system with stationary inputs. In this chapter, a transient distributional law that relates the virtual workload in the system to the waiting times of customers is derived for a non-stationary single-server queueing system. This formula is shown to subsume Brumelle's formula when relax the time-varying assumptions and go on with the stationary regime. In addition, a simulation study is conducted in order to validate the transient distributional law.

4.2 Model Description

Considering a single-server non-Markovian queueing model, $G_t/G_t/1$ with time-varying arrivals and service processing. A_t is the arrival process with arrival rate $\lambda(t)$. V_t is the service requirement of a customer who arrive at time t with service rate $\mu(t)$. There is an unlimited waiting space in the queue and rendering service in the order of their arrival (FCFS service discipline). $\rho(t) = \frac{\lambda(t)}{\mu(t)}$ is the instantaneous traffic intensity at time t.

4.3 Performance Measures

Transient performance measures in queueing systems provide insights into how a queueing system behaves over time before it reaches a steady-state. In this section, explicit integral formula related to some important transient performance measures are discussed.

4.3.1 Number of Customers

In a queueing system, the queue length, or the number of customers waiting in the line before receiving service, is critically important. As the arrival rate fluctuates over time, the number of customers in the queue can vary dynamically. Let $L(t)$ be the number of customers in the system at time t, $t \geq 0$ and let $W(s)$ be the waiting time of the customer who arrive at time s, $0 \leq s \leq t$. Then the expected number of time-varying number of customers in the system or transient generalisation of Little's Law in Bertsimas and Mourtzinou (1997) is,

$$E(L(t)) = \int_0^t F_s^c(t-s)\lambda(s)ds, \tag{4.3.1}$$

where $F_t(x) = P\{W(t) \leq x|A_t\},\ x \geq 0$ is the cumulative distribution function of the waiting time for a customer who arrive at time t and $F_t^c(x) = P\{W(t) > x|A_t\},\ x \geq 0$. A_t is the number of arrivals at time t.

4.3.2 Virtual Workload

The virtual workload in time-varying queues is the amount of work remaining in the system, i.e., the amount of time required to clear all customers currently waiting in the queue for service. The virtual workload captures the dynamic nature of the system's load in a time-varying queue, where arrival and service rates change over time. In queueing theory, service requirement and service time are related but distinct concepts. Service requirement V is the total amount of service that a customer requires. While service time S refers to the actual time a customer spends receiving service at a server. The virtual workload process is denoted by, $Z = \{Z(t), t \geq 0\}$, the amount of work remaining in the system at time t. For a single server queue, it is same as the amount of time a customer, who arrived at

time t has to wait until service. i.e., the virtual waiting time.

Theorem 4.3.1. *Let $Z(t)$ be the virtual workload in the system at time t and V_s be the service requirement of the customer who arrive at time s. Then expected virtual workload at time t is,*

$$E(Z(t)) = \int_0^t F_s^c(t-s)\rho(s)ds + \int_0^t \frac{E(V_s^2)}{2}\lambda(s)ds. \tag{4.3.2}$$

Proof. Consider the interval $[0,t]$. Starting with a reverse-time construction. Let $T_{-k}(t)$ be the k^{th} arrival before time t. i.e., $T_{-(k+1)}(t) < T_{-k}(t) \leq t$, $\forall\ n \geq k \geq 1$. Let $W_{-k}(t) = W(T_{-k}(t))$ be the waiting time of a customer who arrived at time $T_{-k}(t)$. Then the workload in the system at time t can be expressed as,

$$Z(t) = \sum_{k=1}^{\infty}(I_{\{W(T_{-k}(t)\geq t-T_{-k}(t)\}}V_{T_{-k}(t)}) + \frac{V^2_{T_{-k}(t)}}{2}, \tag{4.3.3}$$

where $I_{\{W(T_{-k}(t)\geq t-T_{-k}(t)\})} = \begin{cases} 1 & \text{if } W(T_{-k}(t) \geq t - T_{-k}(t)) \\ 0 & \text{otherwise} \end{cases}$. The first term in (4.3.3) gives the service time of a customer who arrived at $T_{-k}(t)$ and waiting for service at time t. $\frac{V^2_{T_{-k}(t)}}{2}$ is the remaining service time of the customer in the server, seen by a customer who arrived at $T_{-k}(t)$. If the server is idle then this term becomes zero.

(4.3.3) can be written as,

$$\begin{aligned} Z(t) &= \int_0^t (I_{\{W(s)>t-s\}}V_s + \frac{V_s^2}{2})\ dA_s \\ &= \int_0^t I_{\{W(s)>t-s)\}}V_s\ dA_s + \int_0^t \frac{V_s^2}{2}\ dA_s \end{aligned} \tag{4.3.4}$$

By using Campbell-Mecke formula in Fralix and Riaño (2010) for taking expectations

of stochastic integrals,

$$E(Z(t)) = \int_0^t P\{W(s) > t-s)\}E(V_s)\lambda(s)ds + \int_0^t \frac{E(V_s^2)}{2}\lambda(s)ds \tag{4.3.5}$$

For $F_s^c(x) = P\{W(s) > x|A_s\},\ x \geq 0$ and $\rho(s) = \frac{\lambda(s)}{\mu(s)}$

$$E(Z(t)) = \int_0^t F_s^c(t-s)\rho(s)ds + \int_0^t \frac{E(V_s^2)}{2}\lambda(s)ds.$$

By the definition of squared coefficient of variation of service time, this can be written as

$$E(Z(t)) = \int_0^t F_s^c(t-s)\rho(s)ds + \int_0^t \frac{c_s^2+1}{2}\ \frac{\lambda(s)}{\mu(s)^2}ds$$

$$= \int_0^t F_s^c(t-s)\rho(s)ds + \int_0^t \frac{c_s^2+1}{2\ \mu(s)}\ \rho(s)ds.$$

□

Remark 4.3.1. *If the time-varying conditions are relaxed, it can be seen that the well-known Brumelle's formula in Brumelle (1971) immediately follows from this result. i.e., assuming* $\tilde{Z} = \{\tilde{Z}(t); t \geq 0\}$ *to be stationary, then*

$$E(\tilde{Z}(0)) = \int_{-\infty}^0 P_s(\tilde{W}(s) > -s)\rho\ ds + \int_{-\infty}^0 \frac{c_s^2+1}{2\mu}\rho\ ds$$

$$= \rho\int_{-\infty}^0 P_0(\tilde{W}(0) > -s)\ ds + \int_{-\infty}^0 \frac{c_s^2+1}{2\mu}\rho\ ds$$

$$E(\tilde{Z}(0)) = \rho E(\tilde{W}(0)) + \rho\frac{c_s^2+1}{2\mu}.$$

4.3.3 Other Analytical Results

There are explicit formulations designed for general single-server time-varying queues. This section explores some of the analytical results provided in Whitt (2015).

An alternative way to represent the queue length process is

$$L(t) = A(t) - D(t),$$

where $A(t)$ denotes the total number of arrivals during the interval $[0, t]$ and $D(t)$ represents the total number of departures in the same interval. The departure process, $D(t)$, can be explicitly expressed as,

$$D(t) = N\Big(\int_0^t (I_{\{L(s)>0\}} \mu(s)\ ds \Big).$$

Here, the indicator function $I_{\{L(s)>0\}}$ equals 1 when $L(s) > 0$ and 0 otherwise, and $\mu(s)$ represents the service rate at time s.

Let $\{V_k; k \geq 1\}$ denote the sequence of service requirements, and $\{S_k; k \geq 1\}$ represent the sequence of actual service times for the customer k, respectively. Assuming the system starts empty, let A_k, B_k and D_k denote the times at which the k^{th} customer arrives, begins service, and departs, respectively. Then,

$$D_0 = 0, \qquad B_1 = A_1, \qquad B_k = max\{D_{k-1}, A_k\}, \qquad D_k = B_k + S_k.$$

Therefore, V_k is defined as,

$$V_k = \int_{B_k}^{B_k+S_k} \mu(s) ds.$$

The service time of customer k can then be expressed as, $S_k = V_k/\mu(B_k)$, where

$\mu(B_k)$ is the service rate at the time when the k^{th} customer begins service.

4.4 Simulation Study

In this section, simulated customer care call centre data is considered. An analysis is conducted using the data, in which the arrival rate of calls is characterised by an NHPP. The call centre provides 24/7 service. This study is based on 1433 call arrivals during a single day, processed over a period of 24 hours. The arrival rate function follows a periodic function, $1 + 05\ \sin(\pi t/1440 - \pi/2)$ and constant service rate as $\mu = 2$.

Figure 4.1 shows the number of calls waiting in the system including the one in service as a new call arrives. Y-axis denotes the the number of customers and X-axis denotes time in minutes. Here it can be seen that there is a peak in the middle of the day and almost same for starting and ending of the day. An arriving call finds maximum number of customers in the system at 10:33 AM.

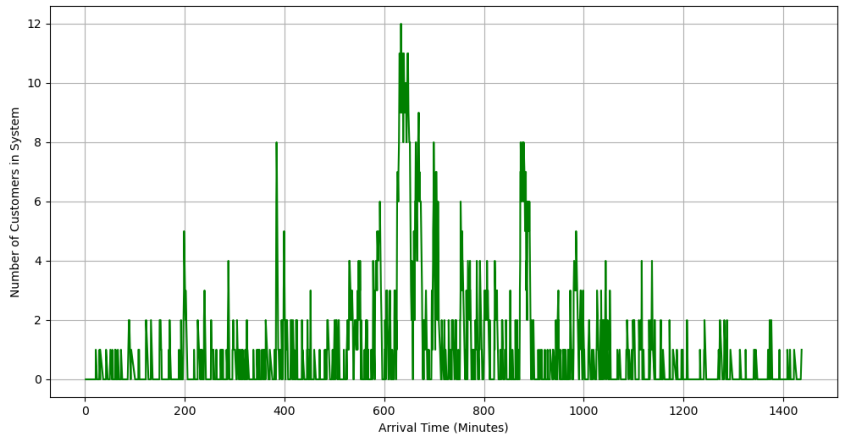


Figure 4.1: Number of customers in the system at the time of arrivals.

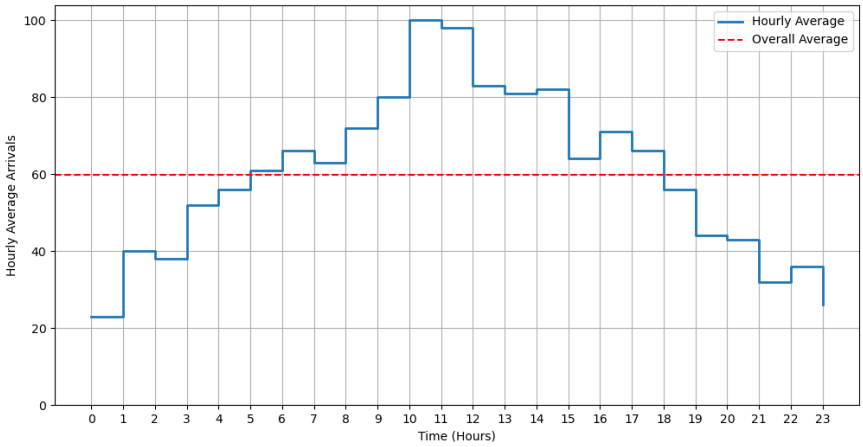
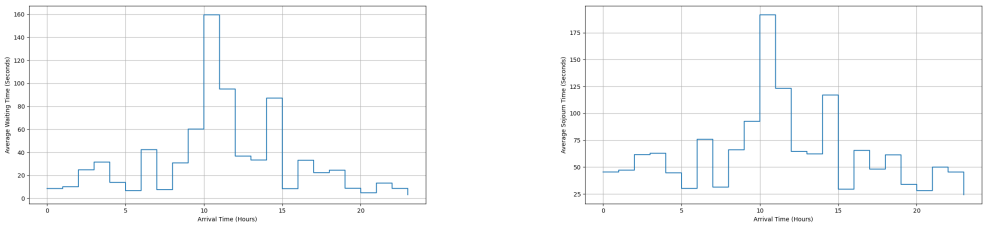


Figure 4.2: Hourly average arrivals throughout the day.

Figure 4.2 shows the overall average and hourly averages of arrival rates over the day. It exhibits the non-stationary nature of arrivals. Overall average of arrivals is 60 customers. The hourly averages of waiting times and sojourn times are also plotted in Figure 4.3. It shows that the hourly averages of the waiting time and sojourn time is high in the interval [8,15].



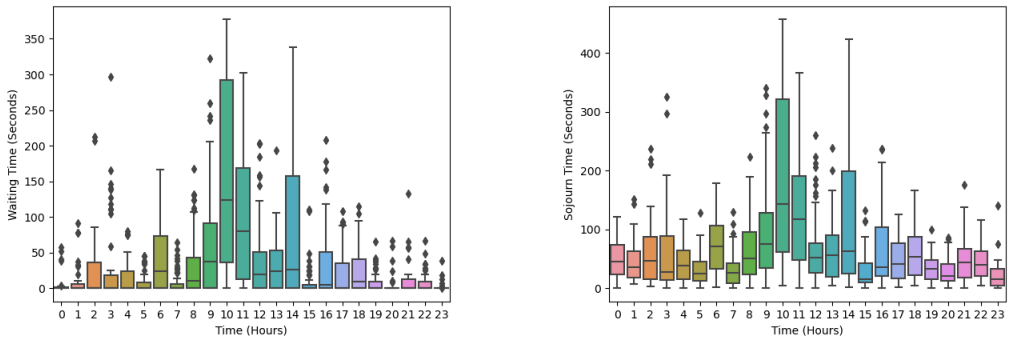
(a) Average waiting time by hour.

(b) Average sojourn time by hour.

Figure 4.3: Step function of average waiting time and sojourn time.

The box-plots in Figure 4.4 illustrate the variation in service and waiting times throughout the day. There is noticeable variation in median values across different hours, suggesting fluctuations in waiting times and sojourn times throughout the day. In both plots, high median value observed between 9th hour and 15th hour of

the day. During off-peak hours, the waiting time plot contains more outliers than the sojourn time plot. It indicates extreme cases where waiting times were exceptionally high. Figure 4.5 shows the scatter-plot of total time spent by the arrivals.



(a) waiting time variation throughout the day. (b) sojourn time variation throughout the day.

Figure 4.4: Box-plots of waiting time and sojourn time over the day.

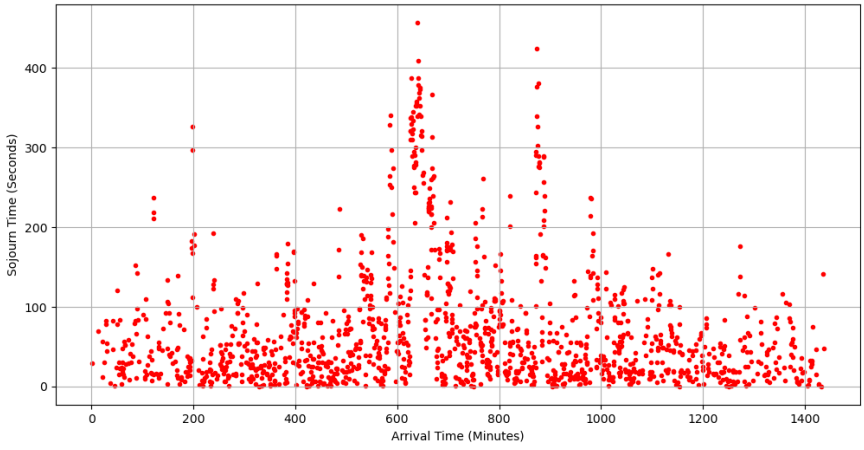


Figure 4.5: Sojourn time of customers throughout the day.

From all the above figures, it is clear that there is a busy period of 8 hour interval [8,15]. The distribution of waiting times in [8,15] follows an exponential distribution with parameter 69.30. Now it can be applied on equation 4.3.2 and virtual workload, $Z(t)$ is obtained.

Virtual workload during the time period is illustrated in Figure 4.6. i.e., an arriving customer during the interval finds that the server is busy and the time required the server to be idle is demonstrated here. It is same as the amount of time a customer who enter the system at time t, has to spent in the system.

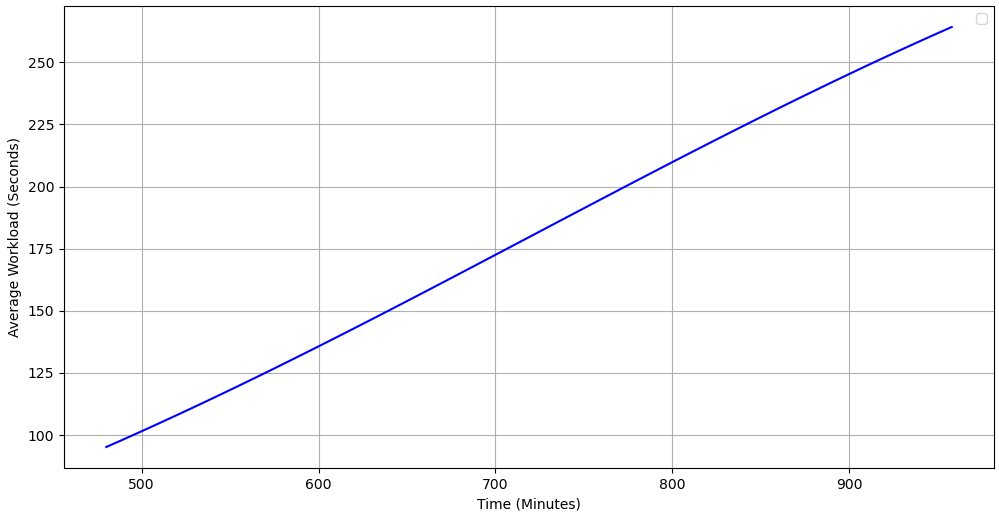


Figure 4.6: Virtual workload over the interval [8,15]

4.5 Summary of the Chapter

In this chapter, some significant transient performance measures for a time-varying single-server queueing system are discussed. These measures refer to the evaluation of queue behaviour over time, particularly when the characteristics of the system are not static. First discussed the transient generalisation of Little's Law, which can be used to evaluate time-dependent queue length. Another most important concept is virtual workload. The virtual workload at a specific time t is the total service time required to serve all the customers waiting in the system at time t. In this chapter an explicit integral formula to obtain virtual workload is presented. This

can be viewed as a time-varying generalisation of well-known Brumelle's formula. This measure is particularly useful in managing congestion and improving overall system performance. A detailed analysis on a simulated data of a 24/7 working call centre is presented.

CHAPTER 5

TRANSIENT PERFORMANCE MEASURES FOR TIME-VARYING TANDEM QUEUEING NETWORKS

5.1 Introduction

Tandem queueing networks refer to a type of open queueing network where entities or tasks move through a series of service stations or nodes arranged linearly. The output of one node serves as the input to the next, creating a sequential flow. These networks are particularly useful for modelling systems that involve multiple stages of service, such as manufacturing, healthcare systems, or logistics operations. By analysing tandem queues, it becomes possible to balance workloads across service stations, thereby reducing delays and improving efficiency. For example, airports have sequential processes that passengers must go through, such as security checks, immigration control, and boarding. Passengers navigate through each station and proceed in an order, reflecting a tandem queueing network. Similarly, in hospitals or outpatient clinics, patients move through the nodes such as registration, consul-

tation, diagnostic tests, and pharmacy, from one to the next based on their needs.

Jackson (1954) was the pioneer to propose queues in series as a queueing system for the overhaul of aircraft engines, in which successive operations include stripping, inspecting, repairing, assembling, and testing component parts. These models are widely used in various real-world systems such as manufacturing, computer networking, and service operations to analyse and optimize the flow of tasks, see, Veatch and Wein (1994), Sun et al. (2022), Higasa et al. (2023). Some recent research include, Sinu Lal et al. (2020), Gerum and Baykal-Gürsoy (2022), Gangadhar and Kadambi (2022) etc. To alleviate the complexity in modelling traffic flow within time-varying tandem queues, the idea of network decomposition in Queueing Network Analyser(QNA) developed by Whitt (1983) can be used. The decomposition method allows each node to be treated individually and analysed independently to reduce computational effort. Time-dependent performance analysis is crucial in understanding the dynamic behaviour of queueing systems, especially in non-stationary environments. Gerhardt and Nelson (2009) provided an algorithm for approximating a tandem queueing network where the external arrival process is a non-stationary MAP_t. Nasr and Taaffe (2013) considered tandem queues with time-dependent Markovian (M_t) service distributions, multiple servers, and time-dependent phase-type (Ph_t) arrival distributions and presented an algorithm to numerically solve the first two moments of the time-dependent departure count process. Pender (2015) proposed a method for approximating the time-varying dynamics of a two dimensional tandem queue with coupled processors in which the external arrival process is assumed to be a non-homogeneous Poisson process. Badrinath and Balakrishnan (2017) proposed an optimal control methodology to control a non-stationary tandem queue in order to mitigate surface congestion at large airports. Rama Murthy M. et al. (2018) conducted a transient analysis of K-node series and parallel queueing

model with load dependent service rates. Zychlinski et al. (2018a) analysed tandem networks with general time-varying arrival rate and blocking, using time-varying fluid models. Rao and Aparajitha (2019) considered a tandem queueing model with non-homogeneous Poisson service process and analysed the system performance by deriving the performance measure such as, queue size, average waiting time of a customer in the queue and in the system, the throughput of transmitters, and the variance of the number of customers in the system.

Throughout this chapter, the waiting space before each service node is assumed to be sufficiently large to accommodate any number of waiting entities. Transient performance measures such as the number of customers in the system and virtual workload are formulated for a two-station non-stationary Markovian tandem queueing network, and this formulation is extended to a k-station tandem queueing network. Then a general framework of algorithm to obtain the transient performance measures is also presented. Finally, numerical studies are conducted on three-station and five-station tandem network models, analysing the transient behaviour of their performance measures.

5.2 Tandem Network of Two Stations

In this section, considering a model of queueing network with two stations in tandem and unlimited waiting capacity at each queue, as illustrated in Figure 5.1. Arrival and service processes in each queue in the system are time-dependant and we restrict the distributional assumptions associated with the system to Markovian($M_t/M_t/1$) framework.

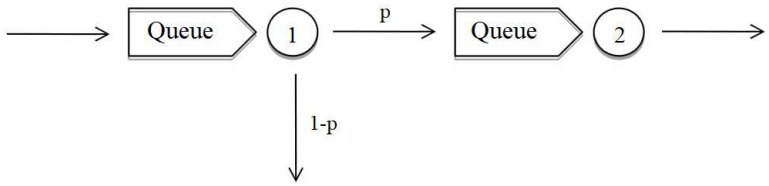


Figure 5.1: A two station tandem queueing network model

5.2.1 Model Description

The model is characterised by the following parameters;

1. $\{A_{t,i}, t \geq 0\}$ is the arrival process with $E(A_{t,i}) = \lambda_i(t)$; where $\lambda_i(t)$ is the arrival rate of station i at time t, for $i = 1, 2$.

2. $V_i(t)$ is the service requirement of a customer who arrive at time t with service rate $\mu_i(t)$(mean=$1/\mu_i(t)$), $i = 1, 2$.

3. Transition from station 1 to station 2 occurs with probability p, $0 \leq p \leq 1$ and customer departs from station 1 with probability $1 - p$.

4. Number of servers, $N_i = 1$, $i = 1, 2$.

5. $W_i(t)$ is the waiting time of a customer who arrive at time t, $i = 1, 2$.

6. $Z_i(t)$ is the virtual workload in the i^{th} station at time t.

7. The counting process $L_i = \{L_i(t), t \geq 0\}$, the number of customers present in station i waiting for service, $i = 1, 2$.

8. $D_{t,i}$ denotes the departure of customers from station 1 to station 2, therefore $D_{t,1}$ is merely the pattern of arrivals to the station 2($A_{t,2}$). $D_{t,2}$ denote the departures from station 2 and $D_{t,3}$ denote the departures from station 1 to out of the network.

5.2.2 Performance Measures

$$L_1(t) = \int_0^t (I_{\{W_1(s)>t-s)\}}\ dA_{s,1}.$$

$$E(L_1(t)) = \int_0^t (P\{W_1(s) > t-s)\})\ \lambda_1(s)\ ds$$

$$= \int_0^t F_1^c(t-s)\ \lambda_1(s)\ ds. \tag{5.2.1}$$

$$L_2(t) = \int_0^t (I_{\{W_2(s)>t-s)\}}\ dD_{s,1}.$$

$$E(L_2(t)) = \int_0^t (P\{W_2(s) > t-s)\})\ \lambda_2(s)\ ds$$

$$= \int_0^t F_2^c(t-s)\ p\ \mu_1(s)\ ds. \tag{5.2.2}$$

where $I_{\{W_i(s)>t-s)\}}$ represents the number of customers entered at station i at time s and waiting for service at time t, $0 \le s \le t$ and $F_i^c(t-s) = 1 - F_i(t-s) = P(W_i(s) > t-s))$. In equation (5.2.2) the arrival rate at station 2, $\lambda_2(t)$ is replaced by $p\ \mu_1(s)$.

$$Z_1(t) = \int_0^t I_{\{W_1(s)>t-s)\}}\ V_{s,1}\ dA_{s,1} + \int_0^t \frac{V_{s,1}^2}{2}\ dA_{s,1}$$

$$E(Z_1(t)) = \int_0^t F_1^c(t-s)\ \frac{\lambda_1(s)}{\mu_1(s)}\ ds + \int_0^t \frac{c_1^2+1}{2}\ \frac{\lambda_1(s)}{\mu_1(s)^2} ds. \tag{5.2.3}$$

Similarly, for the second station,

$$Z_2(t) = \int_0^t I_{\{W_2(s)>t-s)\}}\ V_{s,2}\ dD_{s,1} + \int_0^t \frac{V_{s,2}^2}{2}\ dD_{s,1}$$

$$E(Z_2(t)) = \int_0^t F_2^c(t-s)\ \frac{\lambda_2(s)}{\mu_2(s)}\ ds + \int_0^t \frac{c_2^2+1}{2}\ \frac{\lambda_2(s)}{\mu_2(s)^2} ds$$

$$= p\Big(\int_0^t F_2^c(t-s)\ \frac{\mu_1(s)}{\mu_2(s)}\ ds + \int_0^t \frac{c_2^2+1}{2\mu_2(s)}\ \frac{\mu_1(s)}{\mu_2(s)}\ ds \Big). \tag{5.2.4}$$

where c_i^2, $i = 1, 2$ is the coefficient of variation of service process in station i.

5.3 Tandem Network of k Stations

In this section, extending the two station non-stationary Markovian tandem model to a k station tandem network model, as illustrated in Figure 5.2. The transition probability $p_{i,i+1}$ denotes the probability that a customer transfer from station i to station $i+1$ after service, $i = 1, 2, 3, ..., k-1$. Let $\{A_{t,i}, t \geq 0\}$ and $\{D_{t,i}, t \geq 0\}$ $i = 1, 2, 3, ..., 2k-1$ be the arrival and departure processes respectively. Here $D_{t,i}$ for $i = 1, 3, 5, ..., 2k-3$ are departure processes as well as arrival processes and $D_{t,i}$ for $i = 2, 4, 6...., 2k-2$ and $2k-1$ are departures from the network.

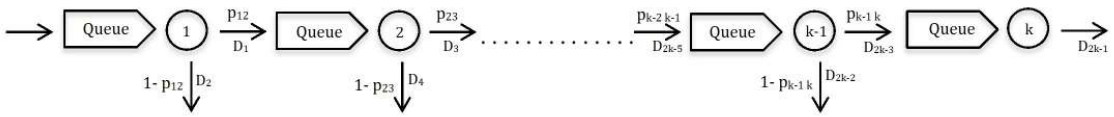


Figure 5.2: A k station tandem queueing network model

$$E(L_1(t)) = \int_0^t F_1^c(t-s)\ \lambda(s)\ ds.$$

$$E(L_i(t)) = \int_0^t F_i^c(t-s)\ p_{i-1,i}\ \mu_{i-1}(s)\ ds.$$

$$E(L_k(t)) = \int_0^t F_k^c(t-s)\ p_{k-1,k}\ \mu_{k-1}(s)\ ds. \tag{5.3.1}$$

$$E(Z_1(t)) = \int_0^t F_1^c(t-s)\ \frac{\lambda_1(s)}{\mu_1(s)}\ ds + \int_0^t \frac{c_1^2+1}{2}\ \frac{\lambda_1(s)}{\mu_1(s)^2} ds.$$

$$E(Z_i(t)) = p_{i-1,i}\Big(\int_0^t F_i^c(t-s)\ \frac{\mu_{i-1}(s)}{\mu_i(s)}\ ds + \int_0^t \frac{c_i^2+1}{2\mu_i(s)}\ \frac{\mu_{i-1}(s)}{\mu_i(s)}\ ds\Big).$$

$$E(Z_k(t)) = p_{k-1,k}\Big(\int_0^t F_k^c(t-s)\ \frac{\mu_{k-1}(s)}{\mu_k(s)}\ ds + \int_0^t \frac{c_k^2+1}{2\mu_k(s)}\ \frac{\mu_{i-1}(s)}{\mu_i(s)}\ ds\Big). \quad (5.3.2)$$

where $c_i^2\ i = 1, 2, ..., k$ is the coefficient of variation of service process in station i

5.4 Algorithm

A general framework of algorithm to obtain transient performance measures such as number of customers and average virtual workload for k-station non-stationary tandem network model is presented in this section. Initially discussing some prerequisites for developing the algorithm.

- The principle of rate-matching control, as discussed in Whitt (2015), is applied to determine the service rate function. In rate-matching control, the service rate is chosen to be proportional to the arrival rate, for a constant traffic intensity, ρ. ie., for a given traffic intensity ρ_i, the service rate becomes,

$$\mu_i(t) \equiv \lambda_i(t)/\rho_i, \quad i = 1, 2, 3, ..., k, \quad t \geq 0. \quad (5.4.1)$$

- For an $M_t/M_t/1$ model, distribution of the waiting time $W(u)$, i.e., the probability that the waiting time of a customer who arrive at u, is larger than x is,

$$P(W(u) > x) = \rho e^{-(1-\rho)\Lambda_t(u)/\rho}, \quad (5.4.2)$$

where $\Lambda_t(u) = \Lambda(t+u) - \Lambda(u)$, $\Lambda(.)$ is the cumulative arrival rate function defined as,

$$\Lambda(u) = \int_0^u \lambda(r)dr, \quad r \geq 0,$$

and $\Lambda_t(u)$ need to be strictly increasing and continuous, see Whitt (2015). Therefore, the expected waiting time, $E(W(t))$ in the $M_t/M_t/1$ model is

$$E(W(t)) = \int_0^\infty P(W(u) > x) = \rho \int_0^\infty e^{-(1-\rho_i)\Lambda_t(u)/\rho}. \tag{5.4.3}$$

- Probability that, the waiting time of a customer who arrived at time s to be $t-s$ is,

$$P(W(s) > t-s) = \rho e^{-(1-\rho)\Lambda_t(s)/\rho}$$

, where $\Lambda_t(s) = \Lambda(t) - \Lambda(s)$. Here for station i, $\Lambda_{t,i}(s) = \Lambda_i(t) - \Lambda_i(s)$ and

$$P(W_i(s) > t-s) = \rho_i e^{-((1-\rho_i)\Lambda_{t,i}(s))/\rho_i} \quad i = 1, 2, 3, ..., k.$$

- The term squared coefficient of variation of service time(c^2) is involved in (5.3.2). Since the model considered here is non-stationary and Markovian, c^2 can be assumed to be 1.

Algorithm 1

Require: External arrival rate $\lambda_1(t)$, transition probabilities $p_{i,j}$, $i = 1, 2, 3, ..., k-1$, $j = 2, 3, ..., k$ and traffic intensity, $\rho = \{\rho_1, \rho_2, ..., \rho_k\}$

Ensure: Transient performance measures, $E(L_i(t))$ and $E(Z_i(t))$ for $i = 1, 2, 3, ..., k$.

1: Compute $\lambda_i(u) = p_{i-1,i}\frac{\lambda_{i-1}(u)}{\rho_{i-1}} \quad i = 1, 2, 3, ..., k$
2: Compute $\mu_i(u) = \frac{\lambda_i(u)}{\rho_i} \quad i = 1, 2, 3, ..., k$
3: Compute $\Lambda_i(t, s) = \int_0^t \lambda_i(u)\, du - \int_0^s \lambda_i(u) du. \quad i = 1, 2, 3, ..., k$
4: Compute $P(W_i(s) > t-s) = \rho_i e^{-((1-\rho_i)\Lambda_{t,i}(s))/\rho_i} \quad i = 1, 2, 3, ..., k$
5: Apply the formulas in (5.3.1) and (5.3.2) to obtain expected number of customers and virtual workload at each station.

5.5 Numerical Study

This section discusses two examples of tandem networks. The algorithm developed in the previous section is used to compute performance measures and analyse the

transient behaviour of the models. Specifically, the effects of traffic intensity on the transient performance measures are analysed.

5.5.1 A Three-Station Example

A three-station tandem network of non-stationary Markovian $M_t/M_t/1$ queues with the following characteristics is considered.

- Let the external arrival rate to station 1, $\lambda_1(t)$ be the identity function, t, $t \geq 0$.

$$\lambda_i(t) = p_{i-1,i}\mu_{i-1}(t), \quad i = 2, 3, \tag{5.5.1}$$

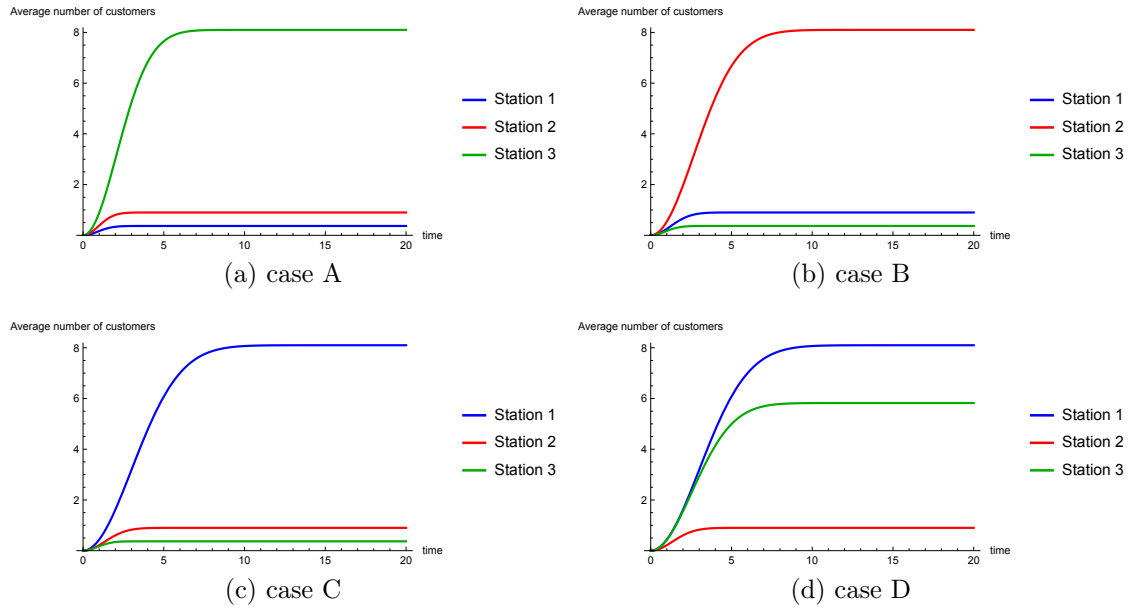
  where $\mu_i(t)$, $i = 1, 2, 3$ is the service rate function obtained from (5.4.1) and $p_{i-1,i}$ is the transition probability from station i to $i+1$, $i = 1, 2$. In this study, $p_{i-1,i} = p = 0.75$ is taken, ie., equal probability for all transitions.

- In this study, four cases, A, B, C, and D are considered by taking arbitrary values for traffic intensities, as shown in Table 5.1. Stations with traffic intensities close to 1 are considered bottleneck stations. Case D is more challenging because of two bottleneck stations, whereas all the other cases have only one bottleneck station.

Case	ρ_1	ρ_2	ρ_3
A	0.45	0.60	0.90
B	0.60	0.90	0.45
C	0.90	0.60	0.45
D	0.90	0.60	0.87

Table 5.1: Four cases of traffic intensities for three stations.

Figure 5.3 presents the time-varying number of customers in each station in the three station tandem network. As can be seen from the figure, stations with high

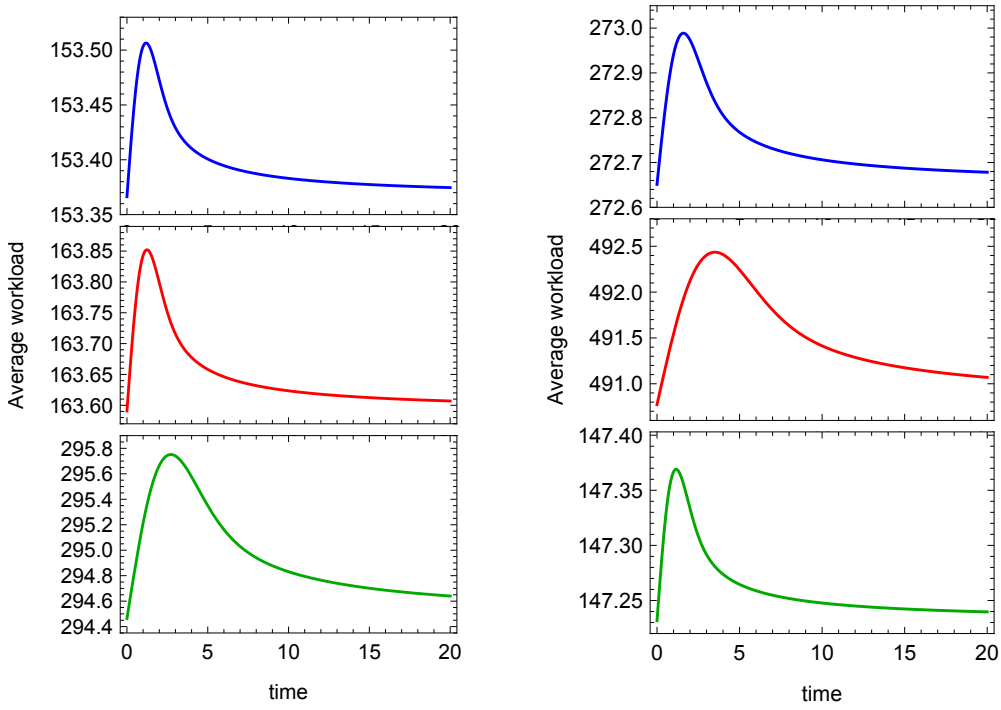


(a) case A (b) case B

(c) case C (d) case D

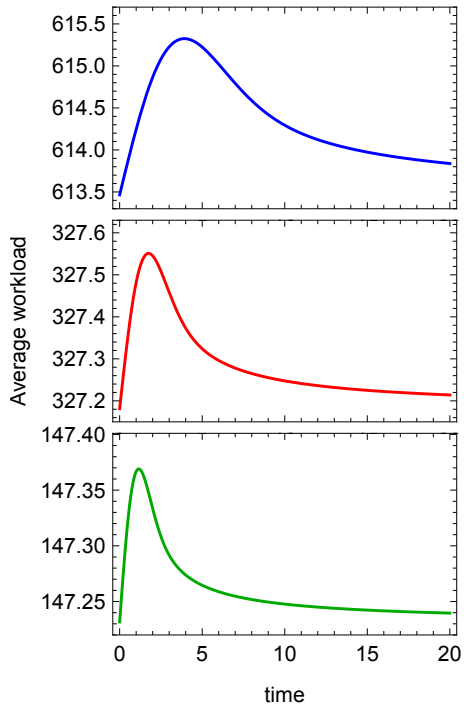
Figure 5.3: Average number of customers at time t for four different cases considered in this study

traffic intensity have a large number of customers in the queue. Figure 5.4 shows the average virtual workload in each station. The workload is heavy for bottleneck stations when compared to other stations. If the bottleneck station is located in the first position, the workload becomes heavier than if it is located in the second or third position.

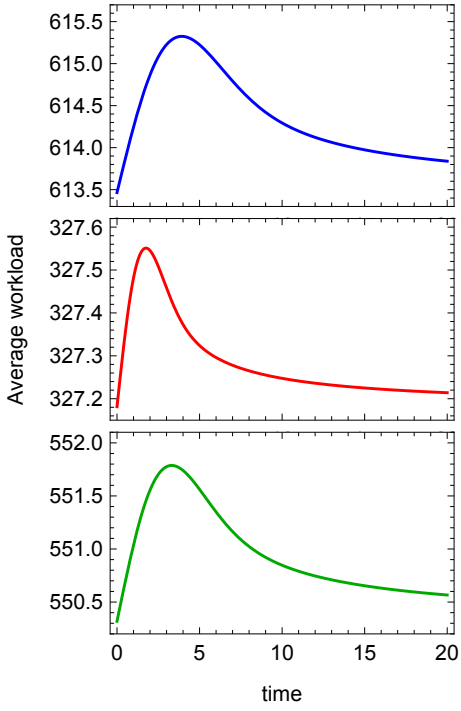


(a) case A

(b) case B



(c) case C



(d) case D

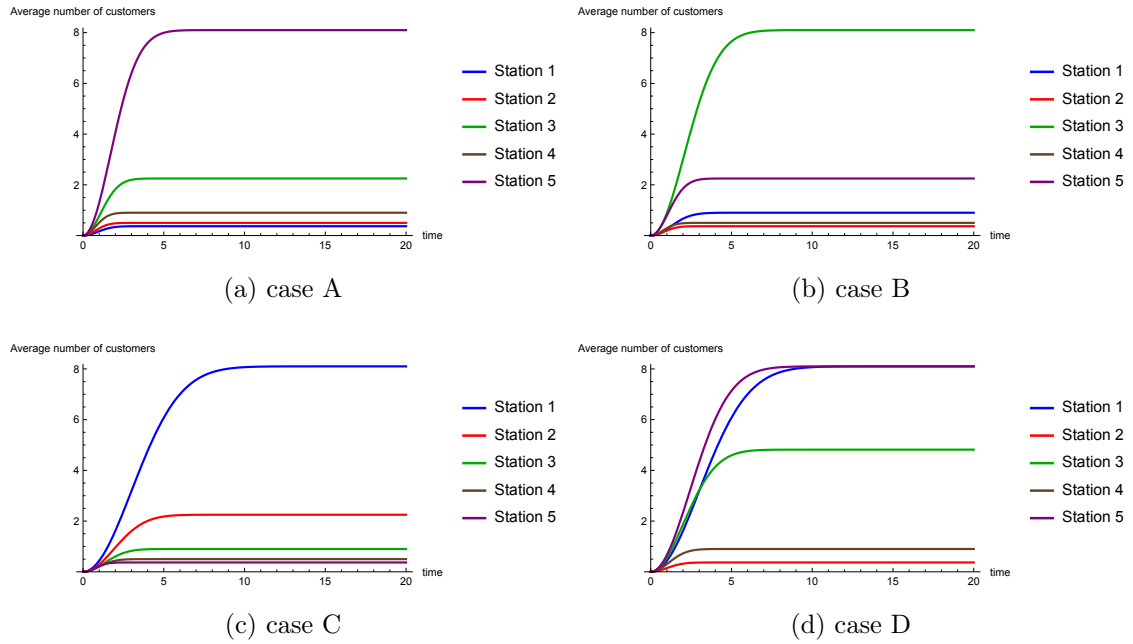
Figure 5.4: Average workload at time t for different cases. Here blue, red and green figures represent station 1, station 2 and station 3 respectively.

5.5.2 A Five-Station Example

Another example of a five-station tandem network model with similar external arrival rate and transition probability is considered. The following table, Table 5.2 summarises four different cases of traffic intensity for each station. As in the previous example, Figure 5.5 illustrates the time-varying number of customers and Figure 5.6 presents the average virtual workload in the five-station tandem queueing network.

case	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5
A	0.45	0.50	0.75	0.60	0.90
B	0.60	0.45	0.90	0.50	0.75
C	0.90	0.75	0.60	0.50	0.45
D	0.90	0.45	0.85	0.60	0.90

Table 5.2: Four cases of traffic intensities for five stations.



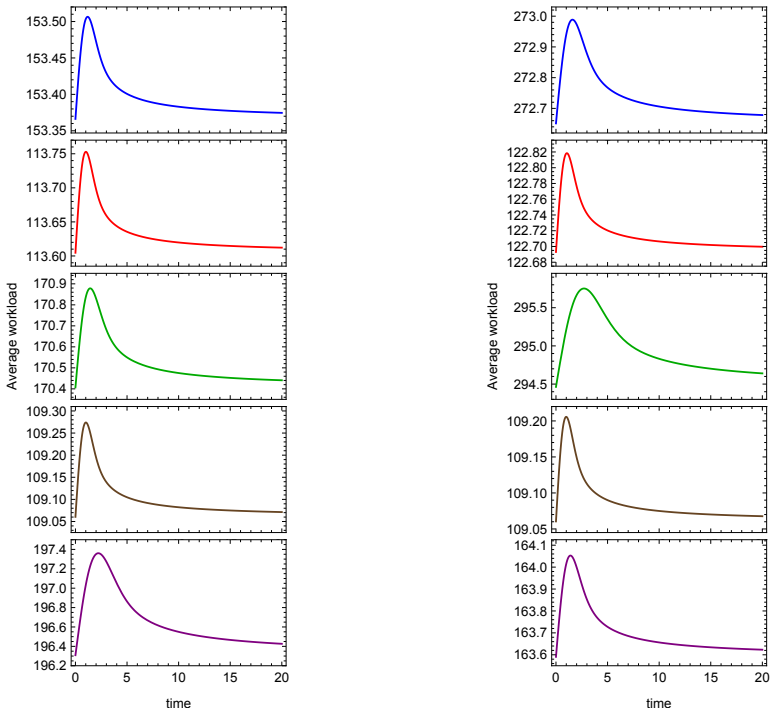
(a) case A

(b) case B

(c) case C

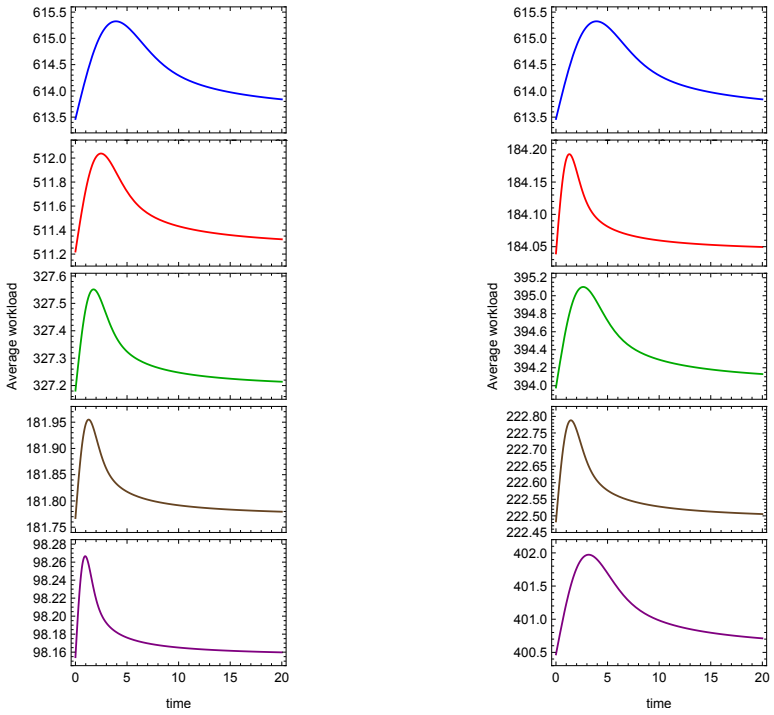
(d) case D

Figure 5.5: Average number of customers at time t for four different cases considered in this study



(a) case A

(b) case B



(c) case C

(d) case D

Figure 5.6: Average workload at time t for different cases. Here blue, red, green, brown and purple figures represent stations 1, 2, 3, 4 and 5 respectively.

The results from the numerical study provide insight into the relation between location of bottleneck stations and transient performance measures. It is evident in this example that if the bottleneck station is located first, the workload in both the first station and the next stations will be heavy. As a result, customers will have to spend more time in the system. Interestingly, when the bottleneck station is at the end of the series, there is no effect on the average workload in the previous stations and hence, the customers will be able to obtain the service without experiencing any delay. Whereas, the average workload was found to be consistently high when there are multiple bottleneck stations within the system. The results are illustrated by implementing algorithm 1 in Wolfram Mathematica 12.3.

5.6 Summary of the Chapter

In this chapter, transient distributional laws that characterise the performance of a time-varying tandem queueing network are discussed. To derive the virtual workload in the system, we initially considered a general single server queueing system with unlimited waiting capacity. Subsequently, extended the model from single server queue to tandem queueing network of k servers and formulated transient performance measures such as, number of customers and average virtual workload in stations. Further introduced an algorithm that provides a general framework for obtaining transient performance measures in a k station tandem network, and implemented the algorithm through numerical study. The results exhibited the relationship between performance measures and traffic intensities.

CHAPTER 6

TIME-VARYING TANDEM QUEUEING NETWORK WITH BLOCKING

6.1 Introduction

In queueing theory, capacity restriction on waiting line is a crucial aspect to study and it is common for real-world service systems to have queues of finite capacity. In such systems, the flow of customers from the source node will be blocked if the waiting room at the destination node is full. There are mainly two blocking mechanisms, i.e., blocking-after-service(BAS) and blocking-before-service(BBS) that describe different scenarios when there are restrictions on waiting rooms. BAS occurs when an entity after service from a node finds the waiting room of the next station is full(saturated). So, they are being blocked before entering the waiting room of the next node. They must have to wait until the space becomes available. In BBS system, before starting service at current node, entities are blocked if there is no available space in the waiting room of next station. Once the space is available at the destination node, the blocked entity resumes service at the source node.

Tandem queueing networks with finite capacity are useful for modelling healthcare, communication, and manufacturing systems de Bruin et al. (2007); Seo et al. (2008); Seo and Lee (2011); Meerkov and Yan (2016). A BAS mechanism is also known as manufacturing blocking. In manufacturing and production lines, items move through the workstations which can only process a limited number of items at a time. If the next workstation is full, it is not possible to move the items that have been processed from the current workstation, leading to a blockage. A steady-state analysis under the BAS mechanism with two single-server queues connected in tandem was conducted in Avi-Itzhak and Yadin (1965). In Avi-Itzhak (1965), the same model is extended to a k-station tandem network with general arrival times, deterministic service times and finite waiting room between stations. Transient behaviour of a two station tandem network with no restriction on first station and no queue allowed for second station was investigated in Prabhu (1967). Zychlinski et al. (2018a) developed time-varying fluid models for tandem network with a general time-varying arrival rate, a finite waiting room before first station and no intermediate waiting room. The BBS mechanism, which is also known as communication blocking, is commonly used in telecommunication networks Suri and Diehl (1984); Frein and Dallery (1989). A detailed description of the different types of BBS mechanisms is presented in Simonetta Balsamo (2001). In communication networks, data packets move through a series of nodes/router, where each node processes the packets and forward it to the next node. If the buffer of the next node is full, the current node cannot forward the packet causing the packet to be blocked before the current node. Healthcare systems can also use the BBS mechanism in short medical procedures, such as cataract surgery, laparoscopic surgery and cardiac catheterization. These procedures can only begin when the room is available in the recovery area. Avi-Itzhak and Levy (1995) introduced a k-stage blocking scheme as

a generalisation of the results presented in Avi-Itzhak and Yadin (1965); Avi-Itzhak (1965). In Avi-Itzhak and Halfin (1993), they analysed the steady-state performance measures of a k-station single-server network with no intermediate queue and an unlimited buffer prior to the first station under both BAS and BBS mechanisms. Fluid limits for tandem model of time-varying multi-server queues with finite buffers before the first station and between stations under BBS mechanism is considered in Zychlinski et al. (2018b). To facilitate comparison, they also developed steady-state closed-form expressions for system performance measures under the BAS and BBS mechanisms.

There has been extensive research conducted on tandem networks with finite capacity queues. However there is limited research on time-varying tandem queues with blocking. In this study, we provide an analytical comparison between BBS and BAS in time-varying tandem queues, with special reference to a case of healthcare system. We develop a stochastic model for a two station finite capacity tandem network under different blocking mechanisms. In the second section, explicit expressions for transient performance measures such as number of patients and average virtual workload at time t under BAS and BBS mechanisms are discussed. We also conducted a numerical study with the blocking mechanisms under different traffic intensity and queue capacity.

6.2 Model Description

We consider a health care system, such as a hospital emergency department with a triage system where patients arrive according to a non-homogeneous Poisson process. The model considered here is a tandem network with two stations. Patients are first assessed by a triage nurse to determine their level of need for medical assis-

tance. Subsequently the patients are sent to the consultation room, where medical professionals provide the necessary treatment. There is an unlimited waiting room for triage and finite waiting room for treatment. Suppose that only a limited number of patients can wait in the treatment area due to space and resource constraints. In the above situation, there are two ways to manage heavy traffic of patients. In BAS, if the waiting space of treatment room is full (saturated), patients cannot join the queue for treatment after the triage process, causing a blockage. In BBS, even though the triage nurse is present, the treatment waiting room is full (saturated), preventing patients from being triaged and causing them to be blocked. In this section, we establish transient performance measures for the two-station tandem queueing network model, under the two blocking mechanisms.

6.2.1 Blocking After Service(BAS)

Initially, we model a healthcare system with a two-station tandem queueing network with finite queue capacity in second station. In the first-come-first-served (FCFS) model illustrated in the Figure 6.1, patients arrive according to non-homogeneous Poisson process and the time-dependent service time following an exponential distribution, i.e., triage node is $M_t/M_t/1/\infty$ and treatment node is $M_t/M_t/1/K$. Following are the parameters that characterise the model.

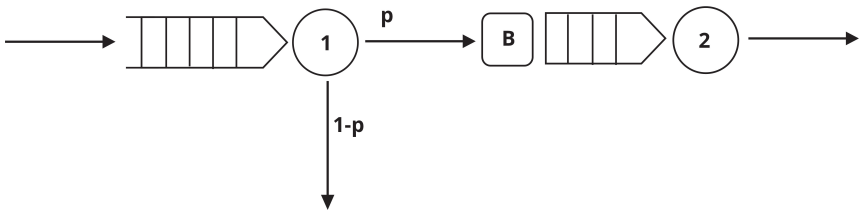


Figure 6.1: *A BAS two station tandem queueing network model*

1. $\{A_{t,1}, t \geq 0\}$ is the external arrival process to the triage node with arrival rate $\lambda_1(t)$.

2. V_t is the service requirement of the patient, i.e., the total amount of service (in terms of time) that a patient requires. $V_{t,1}$ is the service requirement of a patient arriving at the triage at time t. Similarly, $V_{t,2}$ is the service requirement of a patient arriving at the treatment area at time t. $\mu_i(t),\ i = 1, 2$ is the service processing rate at time t.

3. After triage process, patients move to the treatment area with probability p and leave the system with probability $1 - p$ and at a rate of $(1 - p)\mu_1(t)$.

4. $A_{t,2}$ denote the arrival of patients to the treatment area after triage process. Therefore, the arrival rate is $\lambda_2(t) = p\mu_1(t)$.

   The instantaneous traffic intensity at the triage node, $\rho_1(t)$ is defined as,

   $$\rho_1(t) = \lambda_1(t)/\mu_1(t).$$

   Thus, the two arrival rate functions are related as $\lambda_2(t) \ = p\ \lambda_1(t)/\rho_1(t)$.

5. In this study, the instantaneous traffic intensity, $\rho(t) = \lambda(t)/\mu(t)$, which measures the utilisation of a service system is chosen to be invariant of time. This adaptation is made to choose service rate function properly in order to adjust with the arrival rate and traffic intensity.

   We use the principle of rate-matching control, as discussed in Whitt (2015), to determine the service rate function. In rate-matching control, the service rate is set to be proportional to the arrival rate for a fixed traffic intensity ρ. Thus, for a constant traffic intensity ρ_i, the time-dependent service rate function, $\mu_i(t)$ can be written as,

   $$\mu_i(t) \equiv \lambda_i(t)/\rho_i, \quad i = 1, 2 \quad t \geq 0. \tag{6.2.1}$$

6. There is an infinite waiting room at the triage node and a finite waiting room at the treatment node with a maximum capacity of K. If the finite buffer at treatment node is full, patients will be blocked. When the capacity of the waiting room of treatment node is $K-1$, blocked patients will join the queue for consultation.

The following are the formulations of some transient performance measures associated with the model under consideration.

1. $\{W_1(t), t \geq 0\}$ is the waiting time of a patient who arrives at triage node at time t. An explicit expression for the probability distribution of waiting time $W(t)$ for $M_t/M_t/1/\infty$ is derived by Whitt (2015). If a patient who arrives at time s is still waiting for service in the queue at time t, then we can express the probability that the waiting time of the patient who arrives at time s, is larger than $t-s$, for $0 \leq s \leq t$ as,

$$P(W_1(s) > t-s) = \rho_1 e^{-((1-\rho_1)\Lambda_{t,1}(s))/\rho_1}, \tag{6.2.2}$$

where $\Lambda_{t,1}(s) = \Lambda_1(t) - \Lambda_1(s)$, $\Lambda(.)$ is the cumulative arrival rate function defined as,

$$\Lambda_1(u) = \int_0^u \lambda_1(r)\, dr, \quad r \geq 0 \tag{6.2.3}$$

and $\Lambda_{t,1}(u)$ need to be strictly increasing and continuous, see Whitt (2015).

2. $\{W_2(t), t \geq 0\}$ is the waiting time of a patient who joins the queue of treatment area at time t. Since the queue capacity is finite, i.e., $M_t/M_t/1/K$, here we derive a closed form expression for probability distribution of waiting time.

   Let P_n be the probability that there are n patients in the queue in front of

treatment area, see Shortle et al. (2018).

$$P_n = \frac{(1-\rho_2)\ \rho_2^n}{1-\rho_2^{K+1}},$$

where ρ_2 and K are the traffic intensity and queue capacity of the treatment area. Let Q_n be the probability of the arrival point, that is, the probability that there are n patients in the queue at the time of arrival, for $n < K$. This is derived in Shortle et al. (2018) using Baye's theorem.

$$Q_n = \frac{P_n}{1-P_{K+1}}. \tag{6.2.4}$$

Then the probability distribution of waiting time for the stationary $M/M/1/K$ system can be obtained by reducing the expression for multi-server system to single server, i.e.,

$$P\{W > t\} = \sum_{n=1}^{K-1} Q_n \sum_{i=0}^{n-1} \frac{(\mu t)^i\ e^{-\mu t}}{i!}$$

$$= \sum_{n=1}^{K-1} Q_n \sum_{i=0}^{n-1} \frac{(\lambda t/\rho)^i\ e^{-(\lambda t/\rho)}}{i!}. \tag{6.2.5}$$

The parameter μ is replaced by λ/ρ. "From this, the waiting time distribution for a non-stationary system can be derived using Corollary 5.1 from Whitt (2015). Specifically, under the assumptions of a non-stationary system, λt in equation 6.2.5 becomes the cumulative arrival rate function $\Lambda(t)$.

For the non-stationary system $M_t/M_t/1/K$, let $P(W_2(s) > t-s)$ denote the probability that the waiting time of a patient joining the queue of the treatment node at time s exceeds $t-s$, for $0 \le s \le t$,

$$P(W_2(s) > t - s) = \sum_{n=1}^{K-1} Q_n \sum_{i=0}^{n-1} \frac{(\frac{\Lambda_{t,2}(s)}{\rho_2})^i \ e^{\left(\frac{\Lambda_{t,2}(s)}{\rho_2}\right)}}{i!}. \tag{6.2.6}$$

The cumulative rate function $\Lambda_{t,2}(s) = \Lambda_2(t) - \Lambda_2(s)$ with

$$\Lambda_2(u) = \int_0^u \lambda_2(r)\ dr, \quad r \geq 0. \tag{6.2.7}$$

This gives the waiting time distribution for an $M_t/M_t/1/K$ system under constant traffic intensity.

3. To account for the effects of blocking, we incorporate a blocking probability into the formulation of transient performance measures, i.e., the probability that a patient arriving at time s is blocked after triage at time t. In other words, probability that the blocking time of a patient who arrive at time s is greater than t-s, $P\{B(s) > t - s\}$.

4. $L_1(t)$ represents the number of patients present at the triage node. These patients arrived during the interval $[0, t]$ and have not yet completed their service.

$$L_1(t) = \int_0^t (I_{\{W_1(s)>t-s\}})\ dA_{s,1},$$

where $I_{\{W_1(s)>t-s)\}}$ denotes the number of patients who entered the queue in front of the triage node at time s and are still waiting for service at time t, $0 \leq s \leq t$.

By using Campbell–Mecke formula in Fralix and Riaño (2010) for taking expectations of stochastic integrals, we get the average number of patients present

at the triage node at time t, i.e.,

$$E(L_1(t)) = \int_0^t (P\{W_1(s) > t - s\})\ \lambda_1(s)\ ds, \tag{6.2.8}$$

where $E(I_{\{W_1(s)>t-s\}}) = P\{W_1(s) > t - s\}$ and $E(dA_{s,1}) = \lambda_1(s)$ ds.

5. $L_2(t)$ denotes the number of patients at treatment area, including those in the blocking space. These patients completed their service at triage node and moved to treatment area during the interval $[0, t]$. They are either waiting to join the queue in front of the treatment area or already in the queue. This corresponds to the arrivals $\{A_{s,2},\ \ 0 \leq s \leq t\}$, i.e.,

$$L_2(t) = \int_0^t (I_{\{W_2(s)>t-s\}} + I_{\{B(s)>t-s\}})\ dA_{s,2},$$

where $I_{\{W_2(s)>t-s)\}}$ represents the number of patients who entered the queue of the treatment area at time s and are still waiting for service at time t, $0 \leq s \leq t$. Similarly, $I_{\{B(s)>t-s\}}$ denotes the number of patients who entered the blocking space in front of the treatment area at time s and are still waiting to join the queue of treatment area at time t.

While applying expectations on both sides, Campbell–Mecke formula together with additive property of expectation we get $E(I_{\{W_2(s)>t-s\}} + I_{\{B(s)>t-s\}}) = P\{W_2(s) > t - s\} + P\{B(s) > t - s\}$. Therefore,

$$E(L_2(t)) = \int_0^t \big(P\{W_2(s) > t - s\} + P\{B(s) > t - s\}\big)\ p\ \mu_1(s)\ ds. \tag{6.2.9}$$

This represents the average number of patients present in both the blocking space and the queue of the treatment area at time t.

6. $Z_1(t)$ represents the time required to triage all patients who arrived at first node up to time t, i.e.,

$$Z_1(t) = \int_0^t I_{\{W_1(s)>t-s\}}\ V_{s,1}\ dA_{s,1} + \int_0^t \frac{V_{s,1}^2}{2}\ dA_{s,1}.$$

While applying expectations on both sides, Campbell–Mecke formula, we get,

$$E(Z_1(t)) = \int_0^t P\{W_1(s) > t-s\}\ E(V_{s,1})\ E(A_{s,1})\ ds + \int_0^t \frac{E(V_{s,1}^2)}{2}\ E(A_{s,1})ds. \tag{6.2.10}$$

The squared coefficient of variation c^2, can be rewritten as,

$$c^2 = E(V_s^2)\ (E(V_s))^2 - 1 = \frac{E(V_s^2)}{\mu(s)^2} - 1. \tag{6.2.11}$$

Therefore the average workload at the triage node at time t can be represented as,

$$E(Z_1(t)) = \int_0^t P\{W_1(s) > t-s\}\ \frac{\lambda_1(s)}{\mu_1(s)}\ ds + \int_0^t \frac{c_1^2+1}{2}\ \frac{\lambda_1(s)}{\mu_1(s)^2}ds, \tag{6.2.12}$$

where c_1^2 is the squared coefficient of variation or relative variability in service times at triage node.

7. $Z_2(t)$ denotes the time required to complete the consultation of patients who arrived up to time t after triage, taking into account the queue and blocking space, i.e.,

$$Z_2(t) = \int_0^t \left(I_{\{W_2(s)>t-s\}} + I_{\{B(s)>t-s\}}\right)\ V_{s,2}\ dA_{s,2} + \int_0^t \frac{V_{s,2}^2}{2}\ dA_{s,2}.$$

By using campell-Mecke formula and additive property of expectation, $E(Z_2(t))$ can be written as,

$$E(Z_2(t)) = \int_0^t \big(P\{W_2(s) > t-s\} + P\{B(s) > t-s\}\big)E(V_{s,2})E(A_{s,2})ds + \int_0^t \frac{E(V_{s,2}^2)}{2}\, E(A_{s,2})ds.$$

Let c_2^2 be the squared coefficient of variation of service times in treatment node and applying equation (6.2.11),

$$E(Z_2(t)) = \int_0^t \big(P\{W_2(s) > t-s\} + P\{B(s) > t-s\}\big)\, \frac{\lambda_2(s)}{\mu_2(s)}\, ds + \int_0^t \frac{c_2^2+1}{2}\, \frac{\lambda_2(s)}{\mu_2(s)^2} ds. \tag{6.2.13}$$

This represents the average workload at the treatment node, taking into account the patients in the blocking space and the queue at time t.

6.2.2 Blocking Before Service(BBS)

Here, we examine the application of the BBS mechanism in a hospital emergency department by modelling it with a two-station tandem queueing network. This model, illustrated in Figure 6.2, is built under the non-stationary Markovian assumption, and patients are served according to the FCFS discipline.

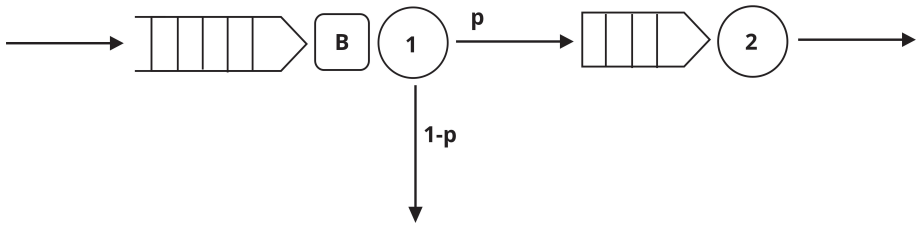


Figure 6.2: A BBS two station tandem queueing network model

Similar to the BAS tandem model, $\{A_{t,i}, t \geq 0\}, i = 1, 2$ is the arrival process

with arrival rates $\lambda_1(t)$ and $\lambda_2(t) = p\ \mu_1(t)$. p represents the probability of moving to treatment after triage, while patients leave the system with probability $1 - p$. $V_{t,i}, i = 1, 2$ is the service requirement of a patient arriving at the triage node and treatment node at time t with service processing rate at time t, $\mu_i(t),\ i = 1, 2$. An infinite waiting room is available for patients at the triage node, whereas the treatment node has a finite waiting room with a maximum capacity of K.

Before providing service from triage node, nurse checks whether the waiting room in front of treatment area is saturated or not. If it contains less than K patients, triage continues. If the waiting room has attained maximum K patients (saturated), stops service at triage node until the next waiting room can accommodate a new patient.

The following are the formulations of some transient performance measures related to the BBS tandem model considered here.

1. $\{W_1(t), t \geq 0\}$ is the waiting time of a patient who arrives at triage node at time t. Since the queue capacity is infinite, the probability distribution of waiting time is defined similar to BAS system, i.e.,

$$P(W_1(s) > t - s) = \rho_1 e^{-((1-\rho_1)\Lambda_{t,1}(s))/\rho_1},$$

where $\Lambda_{t,1}(s) = \Lambda_1(t) - \Lambda_1(s)$ and $\Lambda_1(.)$ is defined in (6.2.3).

2. $\{W_2(t), t \geq 0\}$ is the waiting time of a patient who arrives at the queue of treatment area at time t. Since the queue capacity is finite, i.e., $M_t/M_t/1/K$, the probability that the waiting time of the patient who arrives at time s, is larger than $t - s$, for $0 \leq s \leq t$ is,

$$P(W_2(s) > t - s) = \sum_{n=1}^{K-1} Q_n \sum_{1=0}^{n-1} \frac{(\frac{\Lambda_{t,2}(s)}{\rho_2})^i\ e^{\left(\frac{\Lambda_{t,2}(s)}{\rho_2}\right)}}{i!},$$

where $\Lambda_{t,2}(s) = \Lambda_2(t) - \Lambda_2(s)$.

3. $L_1(t)$ represents the number of patients present at the triage node, including those blocked before service. These patients arrived during the interval $[0, t]$, i.e., $\{A_{s,1},\ \ 0 \leq s \leq t\}$ and have not yet completed their service. They have not completed their service at the triage node, either because they are in the queue or are blocked due to capacity constraints at the treatment node. Therefore,

$$L_1(t) = \int_0^t (I_{\{W_1(s)>t-s\}} + I_{\{B(s)>t-s\}})\ dA_{s,1},$$

where $I_{\{W_1(s)>t-s)\}}$ represents the number of patients who entered the queue of the triage node at time s and are still waiting for service at time t, $0 \leq s \leq t$. Similarly, $I_{\{B(s)>t-s\}}$ denotes the number of patients who entered the blocking space in front of the triage node at time s and are still waiting for triage at time t. When taking expectations,

$$E(L_1(t)) = \int_0^t \big(P\{W_1(s) > t-s\} + P\{B(s) > t-s\}\big)\ \lambda_1(s)\ ds. \quad (6.2.14)$$

This represents the average number of patients present at the triage node at time t, including those blocked before service.

4. $L_2(t)$ represents the number of patients present at treatment node at time t. These patients moved after triage to treatment node during the interval $[0, t]$ and have not yet completed their service.

$$L_2(t) = \int_0^t (I_{\{W_2(s)>t-s\}})\ dA_{s,2},$$

where $I_{\{W_2(s)>t-s)\}}$ represents the number of patients who entered the queue

of the treatment area at time s and are still waiting for service at time t, $0 \leq s \leq t$. Then the average number of patients present in the queue of the treatment area at time t is,

$$E(L_2(t)) = \int_0^t \left(P\{W_2(s) > t - s\}\right)\ p\ \mu_1(s)\ ds. \tag{6.2.15}$$

5. $Z_1(t)$ represents the time required to triage all patients who arrived at first node up to time t, including those blocked patients, i.e.,

$$Z_1(t) = \int_0^t \left(I_{\{W_1(s)>t-s\}} + I_{\{B(s)>t-s\}}\right)\ V_{s,1}\ dA_{s,1} + \int_0^t \frac{V_{s,1}^2}{2}\ dA_{s,1}.$$

Then the average workload at the triage node, taking into account the patients in the blocking space and the queue at time t can be represented as,

$$E(Z_1(t)) = \int_0^t \left(P\{W_1(s) > t - s\} + P\{B(s) > t - s\}\right)\ \frac{\lambda_1(s)}{\mu_1(s)}\ ds + \int_0^t \frac{c_1^2 + 1}{2}\ \frac{\lambda_1(s)}{\mu_1(s)^2} ds. \tag{6.2.16}$$

6. $Z_2(t)$ denotes the time required to complete the consultation of patients who arrived up to time t after triage, i.e.,

$$Z_2(t) = \int_0^t (I_{\{W_2(s)>t-s\}})\ V_{s,2}\ dA_{s,2} + \int_0^t \frac{V_{s,2}^2}{2}\ dA_{s,2}.$$

Then the average workload at the treatment node at time t can be represented as,

$$E(Z_2(t)) = \int_0^t P\{W_2(s) > t - s\}\ \frac{\lambda_2(s)}{\mu_2(s)}\ ds + \int_0^t \frac{c_2^2 + 1}{2}\ \frac{\lambda_2(s)}{\mu_2(s)^2} ds, \tag{6.2.17}$$

where c_i^2, $i = 1, 2$ is the coefficient of variation of service process in station i.

6.3 Numerical Study

We consider a two-station tandem network with non-stationary Markovian queues, in which first station has infinite queue capacity and second station has finite queue capacity. We compare BAS system and BBS system by computing the transient performance measures. First, we outline the prerequisites for conducting the numerical study.

1. A more realistic choice for the arrival rate function would be a sinusoidal or periodic function, which is useful for modelling daily or weekly fluctuations in patient arrivals. However, for simplicity, we choose the identity function as the external arrival rate.

   Let the time-dependent arrival rate of patients to the triage node (external arrival rate), $\lambda_1(t)$ be the identity function, t, $t \geq 0$.

2. Transition rate or arrival rate of patients from triage station to treatment station, $\lambda_2(t)$ is,

$$\lambda_2(t) = p\ \mu_1(t) = p\ \lambda_1(t)/\rho_1 \tag{6.3.1}$$

   where p is the transition probability from station 1 to 2. Here we take $p = 0.75$.

3. The squared coefficient of variation of service time $(c_i^2),\ i = 1, 2$ appears in expressions of virtual workload. Since we are considering a Markovian queueing model, c_i^2 is assumed to be 1.

4. In the BAS system, blocking occurs only if a customer completes service at the triage node and attempts to move towards treatment node, but finds the

queue is full. Therefore, the probability depends on the queue capacity and traffic intensity of the treatment area.

For a BAS system, the blocking probability is defined by,

$$P(B_{BAS}) = \frac{(1-\rho_2)\rho_2^K}{1-\rho_2^{K+1}}, \tag{6.3.2}$$

where K is the queue capacity and ρ_2 is the traffic intensity of the treatment area. Shortle et al. (2018) and Ziya (2008) developed formulations for blocking probability.

Now we present an approximation for the blocking probability in the BBS system. In the BBS system, blocking occurs before the triage starts based on the availability of space in queue of treatment area. This means all patients might be blocked regardless of whether they would have moved for treatment or left the system after triage. As a result blocking probability is inflated.

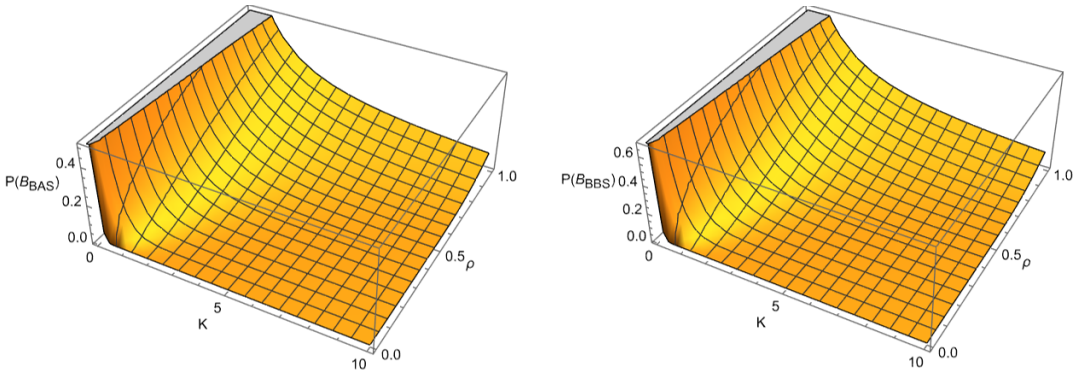
Since $P(B_{BAS})$ is applying only to transitioning customers, we can approximate $P(B_{BBS})$ by scaling $P(B_{BAS})$ with the proportion of transition, p. Here we assume that patient's decision to leave the system after triage or transition is independent of the congestion in the treatment area, so it only depends on the service at triage. $P(B_{BAS})$ is already calculated on transitioned patients, while formulating $P(B_{BBS})$, we need to undo this effect by dividing the proportion of transition, i.e.,

$$P(B_{BBS}) \approx \frac{P(B_{BAS})}{p} = \frac{(1-\rho_2)\rho_2^K}{(1-\rho_2^{K+1})\ p}. \tag{6.3.3}$$

In this study, we have chosen a constant queue capacity and a constant traffic intensity using the rate-matching control principle. Consequently, the param-

eters in the above expressions are time-invariant, making them applicable to a non-stationary model.

The Figure 6.3 illustrates the relationship between the queue capacity and traffic intensity and how their combined variation influences the blocking probabilities.



(a) BAS (b) BBS

Figure 6.3: Blocking probability for queue capacity K = 0 to 10 and traffic intensity $\rho = 0$ to 1.

5. In this study, we consider four cases, A, B, C, and D, by taking arbitrary values for traffic intensities and queue capacity of the second station, as shown in the Table 6.1. The blocking probabilities corresponding to each case are calculated and included in the table. The stations with traffic intensity close to 1 are considered bottleneck stations.

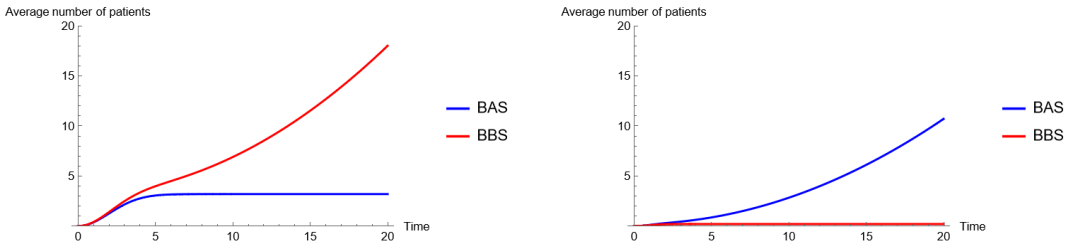
Cases	ρ_1	ρ_2	K	$P(B_{BAS})$	$P(B_{BBS})$
A	0.80	0.60	4	0.056	0.074
B	0.70	0.90	8	0.070	0.093
C	0.90	0.90	6	0.101	0.136
D	0.90	0.90	10	0.050	0.067

Table 6.1: Four cases of traffic intensities, queue capacity, and corresponding blocking probabilities.

The probability of blocking is observed to be relatively high in cases where the traffic intensity at the second station increases significantly and the queue capacity decreases. Among the cases considered, cases A and D exhibit the lowest blocking probabilities. In contrast, Cases B and C show comparatively higher probabilities of blocking due to the combination of high traffic intensity and low queue capacity.

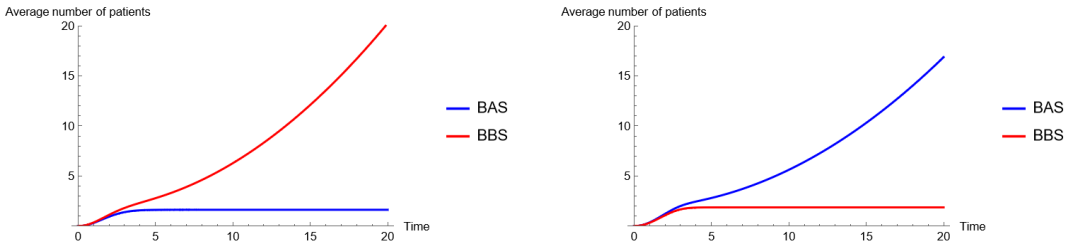
Figure 6.4 illustrates the number of patients at both nodes under the two mechanisms. Blocking after service occurs at the treatment node, while blocking before service occurs at the triage node. As a result, the number of patients increases at the corresponding nodes with time. When the traffic intensity at a node is high, it leads to a significant increase in the number of patients. A similar trend is observed in the average workload, as shown in Figure 6.5. Bottleneck nodes exhibit a relatively higher workload compared to others. In both cases, the treatment node experiences a higher workload under the BAS and BBS mechanisms, as it serves as a bottleneck.

Figure 6.6 illustrates the average number of patients in the system over time. In all cases, the total number of patients is consistently higher in the BBS system.



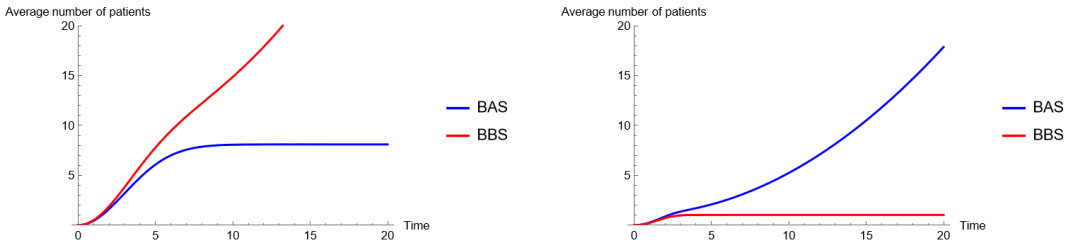
(a) case A (Triage node)

(b) case A (Treatment node)



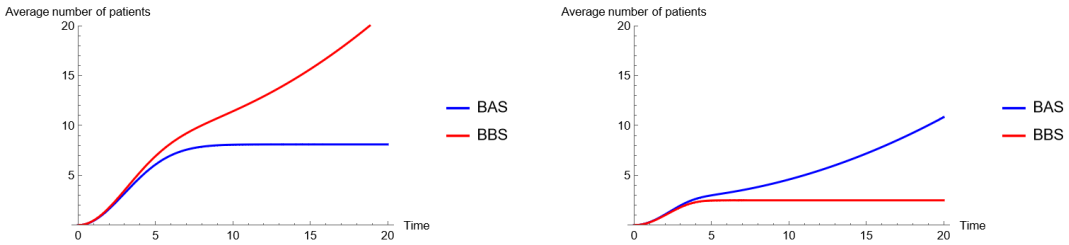
(c) case B (Triage node)

(d) case B (Treatment node)



(e) case C (Triage node)

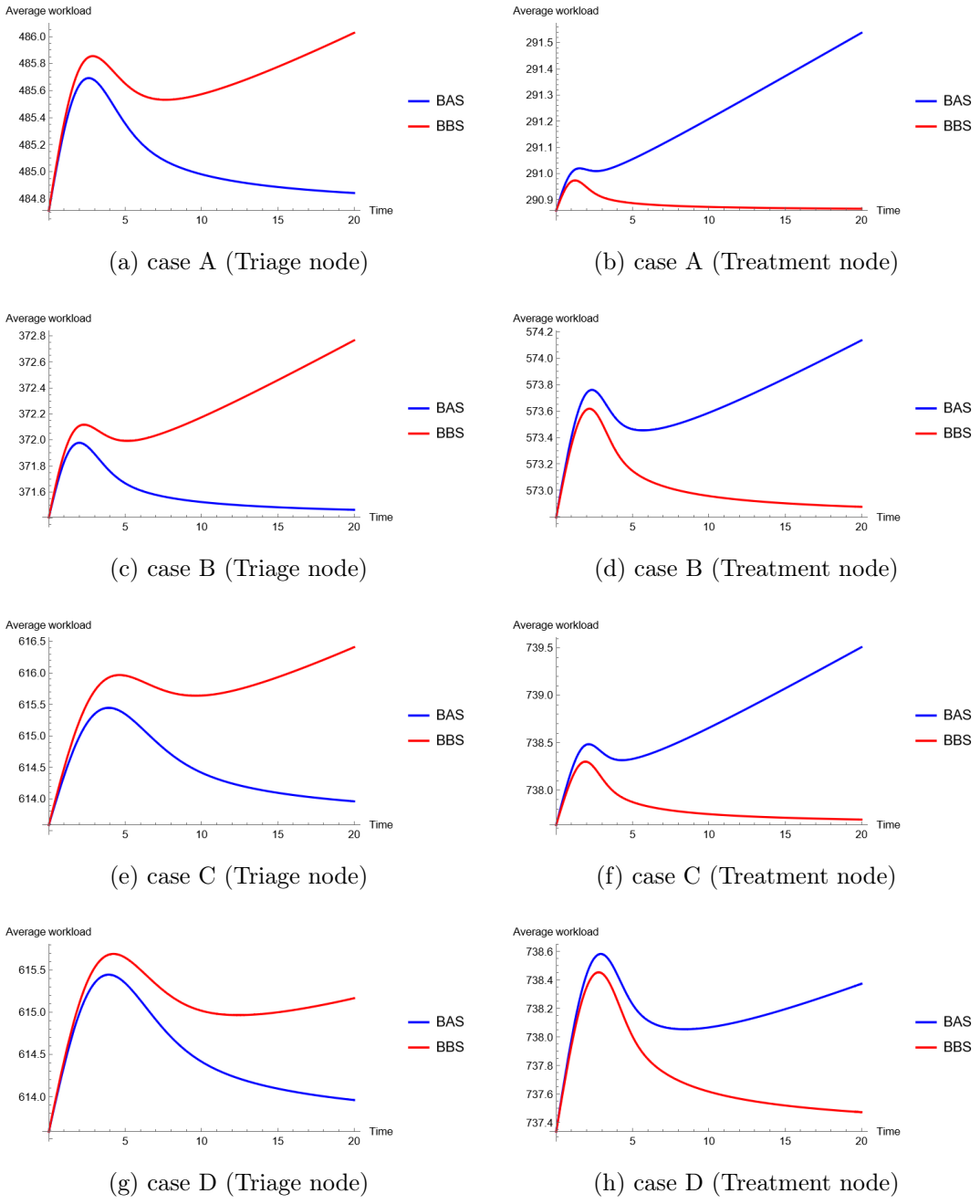
(f) case C (Treatment node)



(g) case D (Triage node)

(h) case D (Treatment node)

Figure 6.4: Average number of patients under BAS and BBS mechanisms for cases we considered in this study.



(a) case A (Triage node)

(b) case A (Treatment node)

(c) case B (Triage node)

(d) case B (Treatment node)

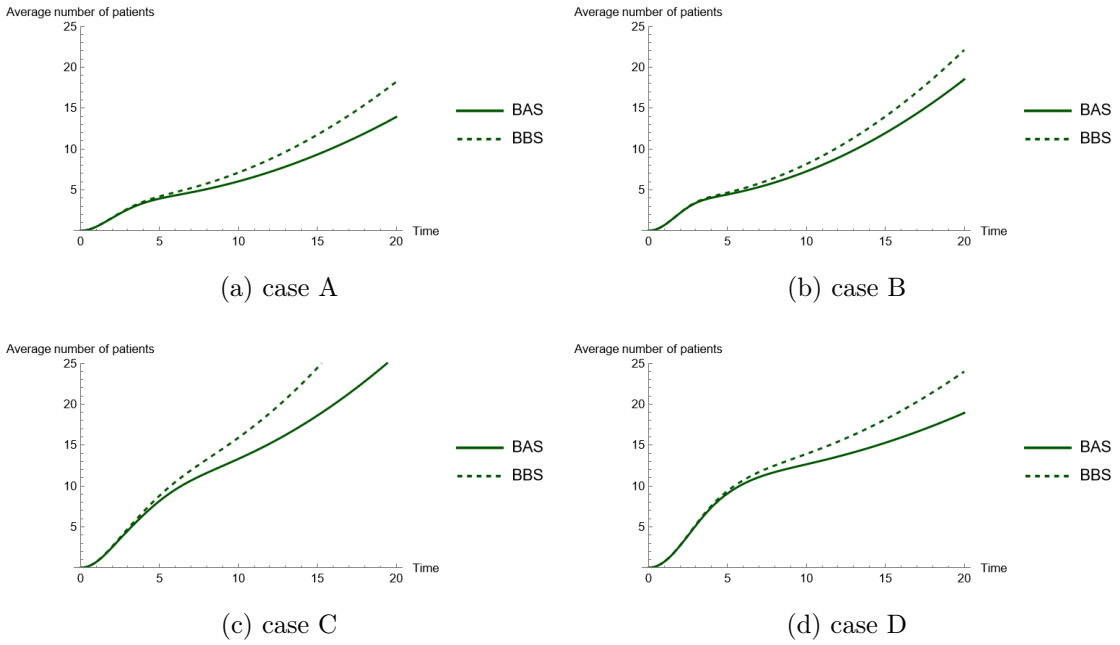
(e) case C (Triage node)

(f) case C (Treatment node)

(g) case D (Triage node)

(h) case D (Treatment node)

Figure 6.5: Average workload under BAS and BBS mechanisms for cases we considered in this study.



(a) case A (b) case B

(c) case C (d) case D

Figure 6.6: Average number of patients in the system (including both stations) for the four cases considered in this study.

6.4 Summary of the Chapter

Finite capacity queues are a realistic and common feature of many real-world service systems, such as hospitals, call centres, and manufacturing units, due to physical and budgetary constraints. However, in most of the theoretical studies of queues, infinite capacity is often assumed to simplify the analysis and results. In this study, we examined various blocking mechanisms that can be applied in queueing systems with capacity restrictions. Most commonly, BAS and BBS blocking mechanisms are used in such systems. Here, we have modelled a time-dependent hospital emergency department using a tandem network of two stations and derived transient performance measures based on these mechanisms. These measures enable an effective comparison of the BAS and BBS mechanisms through numerical study. The results highlight the impact of capacity restrictions on both mechanisms, demonstrating

how they influence the time-varying number of patients and the virtual workload in the system. In our model, patients who do not need immediate emergency treatment can leave after triage. But in the BBS system, these patients also get blocked, causing unwanted congestion in the system. So, BAS is slightly better than BBS in the above scenario. Generally, BAS is preferred when intermediate congestion can be managed, whereas BBS is preferred when smooth flow of entities through the system is more important.

CHAPTER 7

TIME-VARYING APPROXIMATION IN GENERAL QUEUEING NETWORKS

7.1 Introduction

Networks of queues have played a crucial role in modelling and analysing complex real-life systems for decades. The majority of queueing systems in the literature assume stationary or constant parameters, whereas in reality, most queueing systems exhibit time-varying properties. To analyse such dynamic service systems, capturing the time-dependent nature of arrival and service patterns is essential. Traditional queueing network models often rely on Markovian assumptions, which simplify analysis and provide closed-form solutions due to the product-form property. However, these assumptions limit the ability to model real-world scenarios where inter-arrival and service times follow general distributions. Such non-Markovian queueing networks allow more realistic modelling of service systems by incorporating renewal processes for arrivals and non-exponential service times. In these networks, customers can move according to different routing policies and service requirements.

These features complicate the computation of performance measures, making traditional analytical methods inadequate. Simulation can provide an alternative to analysing these systems, but it requires a lot of computational resources, especially for optimizing complex networks. To address these challenges, numerical techniques and analytical approximation methods have considered as powerful tools for evaluating the performance of generalized queueing networks.

There is a substantial literature available on queueing network theory. Jackson (1957), Jackson (1963), Baskett et al. (1975), Gordon (1967) and Kelly (1976) are some breakthroughs in queueing network theory. One of the basic approaches for the analysis of open queueing networks is the parametric decomposition method, see Kuehn (1979), Whitt (1983), Shortle et al. (2018). Decomposition approximations for non-Markov open queueing networks (OQN) are developed by the motivation of product-form property of Markov OQNs. Heavy traffic limit approximation is another approach to analyse performance of OQN, see Harrison (1973), Reiman (1984) and Whitt (1985). Harrison and Nguyen (1990) proposed an analytical algorithm for steady-state performance analysis, called QNET which is based on Brownian approximation and heavy traffic theory. Dai et al. (1994) developed a hybrid method, the Sequential Bottleneck Decomposition (SBD), which uses both decomposition method and heavy traffic theory. Robust queueing (RQ) is another approach to solve complexity in systems involving significant uncertainty. Bandi et al. (2015) developed RQ theory to analyse queueing performance in G/G/1 queues. Whitt and You (2018b) addressed challenges in RQ theory and introduced a new non-parametric RQ formulation that yields improved steady-state performance approximations. Whitt and You (2019) developed time-varying robust queueing algorithm for single server queues with time-varying arrival rate. They mainly focused on periodic steady-state expected workload in models having periodic arrival rate function. Whitt and You

(2022) developed a robust queueing network analyser (RQNA) algorithm to approximate the steady-state performance in network of G/G/1 queues with Markovian routing.

The research conducted on non-stationary general queueing networks is very limited due to their computational complexity. However, these queueing networks can be applied to various large-scale service systems and provide realistic modelling. The purpose of this chapter is to study how these queueing networks perform over time using a time-varying approximation algorithm. Building on the principles of the RQNA algorithm, we develop a time-varying approximation algorithm for single-server queues with time-dependent arrival rates in order to augment some new ideas to the subject matter. Primarily a $G_t/G_t/1$ time-varying (time-varying) single server queue with general arrival and service distributions is considered, where customers enter by first come, first served (FCFS) discipline and arrival and service processes are time-dependent. Then proposed a rather elegant method of time-varying approximation for this queueing system based on indices of dispersion through forward-time construction of workload process. Then extended the method to provide an algorithm to obtain time-varying approximations for performance measures in feed-forward open queueing networks of $G_t/G/1$ queues with Markovian routing. Some numerical experiments are conducted to analyse the network performances with time approximations.

7.2 Preliminaries

Whitt and You (2019) developed a time-varying robust-queueing algorithm for the continuous-time workload in a single-server queue with a time-varying arrival rate function. This chapter extends that time-varying approximation algorithm to a

general single-server queueing network with time-varying arrival rates and feedback restrictions. It begins by discussing some prerequisites from their study, which are essential for this generalisation. Whitt and You (2019) utilised a reverse-time construction approach for the workload process to formulate a time-varying representation of the steady-state workload. In contrast, this study employs a forward-time construction to approximate the workload in a single-server non-homogeneous queue, aiming to simplify subsequent network approximations.

The focus is on a general single-server queue with FCFS service discipline, unlimited waiting space and time-varying arrival and service rates, we call the $G_t/G_t/1$ queue. Let $A(t)$ be the number of arrivals in the interval $[0,t]$ with time-varying arrival rate, $\lambda(u), 0 \leq u \leq t$. Let $\mu(u)$ be the time-varying rate at which service is provided. Cumulative arrival and service rates over $[0,t]$ are denoted as $\Lambda(t)$ and $M(t)$ respectively.

Let V_k be the service requirement of the k^{th} customer. Let the net-input in the interval $[0,t]$ can be written as,

$$X(t) \equiv \sum_{k=1}^{A(t)} V_k - M(t). \tag{7.2.1}$$

Here, $Y(t) = \sum_{k=1}^{A(t)} V_k$ is the cumulative work input over $[0,t]$. Then Z_t, the workload at time t is

$$Z_t \equiv \sup_{0\leq u\leq t} X(u) \equiv \sup_{0\leq u\leq t} \sum_{k=1}^{A(u)} V_k - M(u).$$

Then the time-varying approximation for mean workload at time t is

$$Z_t^* \equiv \sup_{X\mathcal{E}\mathcal{U}_\sqcap} \sup_{0\leq u\leq t} X(u), \tag{7.2.2}$$

where $\mathcal{U}_\sqcap$ is the deterministic uncertainty set, i.e., the set of possible values that the

uncertain parameter, $X(u)$ can take.

$$\mathcal{U}_{\sqcap} \equiv \{X(u) \in \mathcal{R} : X(u) \leq E(X(u)) + b\ SD(X(u)),\ 0 \leq u \leq t,$$

with b being a specified parameter. Z_t^* can be rewritten as

$$Z_t^* \equiv \sup_{0 \leq u \leq t} E(X(u)) + b\ SD(X(u)). \tag{7.2.3}$$

Indices of Dispersion: The index of dispersion is the scaled version of variance function by the mean function. It can be considered as a continuous-time generalisation of the squared coefficient of variation since it exposes the variability over time. Index of dispersion for counts (IDC) and index of dispersion for work (IDW) are defined respectively as,

$$I_a(\Lambda(u)) \equiv \frac{Var(A(u))}{E(A(u))}. \tag{7.2.4}$$

$$I_w(\Lambda(u)) \equiv \frac{Var(Y(u))}{E(Y(u))} = \frac{Var\left(\sum_{k=1}^{A(u)} V_k\right)}{E\left(\sum_{k=1}^{A(u)} V_k\right)} \tag{7.2.5}$$

under the assumption of mean-1 service times the IDW is $\frac{Var\left(\sum_{k=1}^{A(u)} V_k\right)}{\Lambda(u)}$.

$I_a(\Lambda(u))$ describes the variability associated with arrival process $A(t)$ and $I_w(\Lambda(u))$ captures the cumulative variability of $Y(t)$ as a function of time t. The decomposition of IDW is possible by conditional variance formula. For that consider service times V_k are i.i.d and independent of arrival process $A(t)$ as a special case. Therefore equation 7.2.5 can be written as,

$$I_w(\Lambda(u)) = I_a(\Lambda(u)) + c_s^2, \tag{7.2.6}$$

where $c_s^2 = Var[V_k]/E[V_k]^2$ is the squared coefficient of variation of the service process. For $X(t)$, the net-input process in 7.2.1,

$$E(X(u)) = E\left(\sum_{k=1}^{A(u)} V_k - M(u)\right) = \Lambda(u) - M(u) \tag{7.2.7}$$

$$Var(X(u)) = Var\left(\sum_{k=1}^{A(u)} V_k\right) = \Lambda(u) I_w(\Lambda(u)) \tag{7.2.8}$$

By substituting (7.2.7) and (7.2.8) in (7.2.3), time-varying approximation of mean workload can be expressed as

$$Z_t^* \equiv \sup_{0\le u\le t} \Lambda(u) - M(u) + b\sqrt{\Lambda(u) I_w(\Lambda(u))}. \tag{7.2.9}$$

7.3 The Time-Varying Open Queueing Network

7.3.1 Model Description

For better analytic tractability, this section considers a network of k nodes where, at each node, service times are homogeneous in time. Each node represents $G_t/G/1$ queue in which arrival rates are time-dependent. Probability that a customer move from node i to node j after service is p_{ij} , where $i, j = 1, 2, \ldots, k$, $p_{i0} = 1 - \sum_{j=1}^{k} p_{ij}$ is the probability that a customer will leave from the network after service from node i and p_{0i} is the probability that a customer from outside enters the system through node i. Therefore $P = [p_{ij}]$, $1 \le i, j \le k$ is a sub-stochastic routing matrix of order k and we assume that the matrix $(I - P')$ is invertible. Here we consider a feed-forward queueing network. i.e., no customer re-enter the node which they already visited. This will be ensured by choosing the routing matrix P as an upper

triangular matrix.

Let $\Lambda_i(t)$ be the time-varying cumulative arrival rate at node i, $M_i(t)$ be the time-varying cumulative service rate at node i and $\rho_i^*(t) = \sup_{0\le u\le t}\{\Lambda_i(u)/M_i(u)\}$ be the time-varying traffic intensity at node i. Let $A_{0,i}(t), \quad t \ge 0$ be the external arrival process at node i with cumulative arrival rate $\Lambda_{0,i}(t)$ and IDC $I_a(\Lambda_{0,i}(t))$ and $A_{i,j}(t)$ be the number of arrivals from node i to node j over $[0,t]$ with cumulative rate $\Lambda_{i,j}(t)$ and IDC $I_a(\Lambda_{i,j}(t))$. For each node i, assume that service process is i.i.d and independent of arrival processes. Let $S_i(t)$ be the number of services during $[0,t]$ with cumulative rate $M_i(t)$ (mean $1/M_i(t)$) and IDC $I_s(M_i(t)) \equiv Var(S_i(t))/E(S_i(t))$. Also assume that Markovian routing is independent of arrival and service processes.

7.3.2 Time-Varying Approximation for Traffic Equations

Let $\lambda(t) \equiv (\lambda_1(t), \lambda_2(t), \ldots., \lambda_k(t))$ be the vector of total arrival rate at time t. i.e., $\lambda_i(t)$ is the total mean flow rate into node i at t and let $\lambda_0(t) \equiv (\lambda_{0,1}(t), \lambda_{0,2}(t), \ldots\ldots.., \lambda_{0,k}(t))$ be the external arrival rate vector at time. Then a time-varying approximation for traffic equations is

$$\Lambda_i(t) \approx \Lambda_{0,i}(t) + \sum_{j=1}^{k} \Lambda_{j,i}(t), \quad 1 \le i \le k, \tag{7.3.1}$$

where $\Lambda_i(t) = \int_0^t \lambda_i(u)\,du, \Lambda_{0,i}(t) = \int_0^t \lambda_{0,i}(u)du$, cumulative external arrival rate to node i over $[0,t]$ and $\Lambda_{j,i}(t) = \int_0^t \lambda_{j,i}(u)\ du = \int_0^t p_{j,i}\mu_j(u)\,du = p_{j,i}M_j(t)$, cumulative arrival rate from node j to node i over $[0,t]$. Here assume $M_j(t)$ as the cumulative service completion rate (departure rate) at node j over $[0,t]$ and $p_{j,i}$ is the transition probability from node j to i.

In matrix form (7.3.1) can be expressed as $\Lambda(t) = \Lambda_0(t) + (P')M(t)$, where

$P = [p_{ij}]$ is the routing matrix and the vectors $\Lambda(t) \equiv (\Lambda_1(t), \Lambda_2(t), \ldots, \Lambda_k(t))$, $\Lambda_0(t) \equiv (\Lambda_{0,1}(t), \Lambda_{0,2}(t), \ldots, \Lambda_{0,k}(t))$ and $M(t) \equiv (M_1(t), M_2(t), \ldots, M_k(t))$.

7.3.3 The Time-Varying Traffic Variability Equations

Basic operations in a queueing network can be classified as (i) departure operation (ii) splitting operation and (iii) superposition operation. Departure process is the flow of customers from the queues after service, splitting or decomposition is the division of flow of customers into several sub-flows and superposition or merging is combination of different customer flows into one. In this section, a system of equations is discussed, which allows the derivation of indices of dispersion and, consequently, the time-varying approximations for performance measures.

Departure

Let $I_d(M_i(t))$ be the IDC corresponding to the departure process from node i. This can be written as the convex combination of arrival IDC $I_a(\Lambda_i(t))$ and service IDC $I_s(M_i(t))$, discussed in Whitt and You (2022). *i.e.,*

$$I_d(M_i(t)) \approx w_i(t) I_a(\Lambda_i(t)) + (1 - w_i(t)) I_s(M_i(t)) \tag{7.3.2}$$

where the weight function w_i is

$$w_i(t) \equiv w^*\Big(\frac{(1-\rho_i^*(t))^2 \Lambda_i(t))}{\rho_i^*(t) c_{x,i}^2}\Big) \tag{7.3.3}$$

where $c_{x,i}^2 \equiv c_{a,i}^2 + c_{s,i}^2$, $\rho_i^*(t)$ is the time-varying traffic intensity, $\Lambda_i(t)$ is the arrival rate at node i and the canonical weight function w^* is

$$w^*(t) = \frac{1}{2t}\big((t^2 + 2t - 1)(1 - 2\Phi^c(\sqrt{t})) + 2\phi(\sqrt{t})\sqrt{t}(1+t) - t^2)\big)$$

$w^*(t)$ is obtained from correlation function of stationary reflected Brownian motion (RBM) process, see Whitt and You (2018a). Here $\phi(x)$ and $\Phi^c(x) = 1 - \Phi(x)$ are pdf and complimentary cdf of standard normal variable respectively. $w^*(t)$ is an increasing function and $0 \leq w^*(t) \leq 1$.

Splitting

Consider an indicator function,

$$\delta_{i,j}^k(t) = \begin{cases} 1 & \text{if } k^{th} \text{ customer moves to node } j \text{ from node } i \text{ during } [0,t]. \\ 0 & \text{otherwise} \end{cases}$$

with $E(\delta_{i,j}^k(t)) = p_{ij}$ and $Var(\delta_{i,j}^k(t)) = p_{ij}(1 - p_{ij})$, where $k = 1, 2, ..., D_i(t)$, $D_i(t)$ is the number of departures from node i in the interval$[0, t]$.

Therefore, the Arrival process from i to j during $[0, t]$ can be written as

$$A_{i,j}(t) = \sum_{k=1}^{D_i(t)} \delta_{i,j}^k(t)) \tag{7.3.4}$$

$$E(A_{i,j}(t)) \approx E(D_i(t))E(\delta_{i,j}^k(t))) = \frac{1}{M_i(t)}\ p_{i,j}$$

Here the expected number of departures during $[0, t]$ is approximated with cumulative service completion rate, $M_i(t)$ during $[0, t]$.

$$Var(A_{i,j}(t)) \approx E(D_i(t))Var(\delta_{i,j}^k(t))) + (E(\delta_{i,j}^k(t))))^2 Var(D_i(t))$$

$$= \frac{1}{M_i(t)}\ p_{i,j}(1 - p_{i,j}) + p_{i,j}^2 Var(D_i(t))$$

Therefore, IDC of arrival process from node i to node j is

$$I_a(\Lambda_{i,j}(t)) = \frac{Var(A_{i,j}(t))}{E(A_{i,j}(t))} = \frac{\frac{1}{M_i(t)}\ p_{i,j}(1-p_{i,j}) + p_{i,j}^2 Var(D_i(t))}{\frac{1}{M_i(t)}\ p_{i,j}}$$

$$= \frac{\frac{1}{M_i(t)} p_{i,j}(1-p_{i,j})}{\frac{1}{M_i(t)} p_{i,j}} + \frac{p_{i,j}^2 Var(D_i(t))}{p_{i,j} E(D_i(t))}$$

$$I_a(\Lambda_{i,j}(t)) = I_d(M_i(t))p_{i,j} + (1-p_{i,j}) \tag{7.3.5}$$

where $I_d(M_i(t)) = Var(D_i(t))/E(D_i(t))$

Superposition

Consider the arrival process at node i. We take into account of all possible arrivals from other nodes to node i. Here assume that all superposed customer flows are independent. Therefore

$$A_i(t) = \sum_{j=0}^{k} A_{j,i}(t) = \sum_{j=0}^{k} N(\Lambda_{j,i}(t))$$

IDC of $A_i(t)$ is

$$I_a(\Lambda_i(t)) = \frac{Var(\sum_{j=0}^{k} A_{j,i}(t))}{E(\sum_{j=0}^{k} A_{j,i}(t))}$$

$Var\big(\sum_{j=0}^{k} A_{j,i}(t)\big) = \sum_{j=0}^{k} Var(A_{j,i}(t))$ and $E\big(\sum_{j=0}^{k} A_{j,i}(t)\big) = \sum_{j=0}^{k} \Lambda_{j,i}(t)$.
Therefore,

$$I_a(\Lambda_i(t)) = \frac{\sum_{j=0}^{k} Var(A_{j,i}(t))}{\sum_{j=0}^{k} \Lambda_{j,i}(t)}$$

$$= \frac{\sum_{j=0}^{k} \Lambda_{j,i}(t) I_a(\Lambda_{j,i}(t))}{\Lambda_i(t)}$$

where $I_a(\Lambda_{j,i}(t)) = Var(A_{j,i}(t))/\Lambda_{j,i}(t)$

$$I_a(\Lambda_i(t)) = \sum_{j=0}^{k} \big(\frac{\Lambda_{j,i}(t)}{\Lambda_i(t)}\big) I_a(\Lambda_{j,i}(t)) \tag{7.3.6}$$

7.3.4 System of IDC Equations

By combining (7.3.2), (7.3.5) and (7.3.6), we get the system of IDC equations as,

$$\begin{aligned} I_a(\Lambda_i(t)) &= \sum_{j=1}^{k} (\frac{\Lambda_{j,i}(t)}{\Lambda_i(t)}) I_a(\Lambda_{j,i}(t)) + \frac{\Lambda_{0,i}(t)}{\Lambda_i(t)} I_a(\Lambda_{0,i}(t)) \\ I_a(\Lambda_{i,j}(t)) &= I_d(M_i(t))\ p_{i,j} + (1 - p_{i,j}) \\ I_d(M_i(t)) &= w_i(t) I_a(\Lambda_i(t)) + (1 - w_i(t)) I_s(M_i(t)). \end{aligned} \tag{7.3.7}$$

These equations can be solved by matrix multiplication. In the matrix form, the IDC equations can be written as

$$(\mathbf{E} - \mathbf{K}(t))\mathbf{I}(t) = \mathbf{H}(t), \tag{7.3.8}$$

where $\mathbf{E}$ is the identity matrix of order $(2k + k^2)$. $\mathbf{H}$(t) and $\mathbf{I}$(t) are $(2k + k^2) \times 1$ column vectors. $\mathbf{K}(t) \equiv (K_{m,n}(t)) \in^{(2k+k^2)^2}$ $m, n \in \{a_1, ...a_k, a_{1,1}, ...a_{k,k}, d_1, ...d_k\}$.

$$\begin{aligned} \mathbf{I}(t) \equiv \Big[& I_a(\Lambda_1(t)), \cdots, I_a(\Lambda_k(t))), I_a(\Lambda_{1,1}(t)), \cdots, I_a(\Lambda_{k,k}(t)), \\ & I_d(M_1(t)), \cdots, I_d(M_k(t)) \Big]^T \end{aligned}$$

$$\begin{aligned} \mathbf{H}(t) \equiv \Bigg[& \frac{\Lambda_{0,1}(t)}{\Lambda_1(t)} I_a(\Lambda_{0,1}(t)), \cdots, \frac{\Lambda_{0,k}(t)}{\Lambda_k(t)} I_a(\Lambda_{0,k}(t)), (1 - p_{11}), (1 - p_{12}), \cdots, \\ & (1 - p_{kk}), (1 - w_1(t)) I_s(M_1(t)), \cdots, (1 - w_k(t)) I_s(M_k(t)) \Bigg]^T \end{aligned}$$

.

$$\mathbf{K}(t) = \left[\begin{array}{cccc|cccccccc|cccc}
0 & 0 & .. & 0 & \frac{\Lambda_{11}(t)}{\Lambda_1(t)} & 0 & \cdots & 0 & \cdots & \frac{\Lambda_{k1}(t)}{\Lambda_1(t)} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 0 & \frac{\Lambda_{12}(t)}{\Lambda_2(t)} & \cdots & 0 & \cdots & 0 & \frac{\Lambda_{k2}(t)}{\Lambda_2(t)} & \cdots & 0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & \frac{\Lambda_{1k}(t)}{\Lambda_k(t)} & \cdots & 0 & 0 & \cdots & \frac{\Lambda_{kk}(t)}{\Lambda_k(t)} & 0 & 0 & \cdots & 0 \\
\hline
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & p_{11} & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & p_{12} & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & p_{1k} & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & p_{21} & \cdots & 0 \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & p_{22} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & p_{2k} & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & p_{k1} \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & p_{k2} \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & p_{kk} \\
\hline
w_1(t) & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
0 & w_2(t) & \cdots & 0 & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & w_k(t) & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0
\end{array}\right]$$

The weight function $w_i(t)$ in (7.3.3) involves the term $c_{x,i}^2$, which is the convex combination of SCV's of arrival and service processes. To solve (7.3.7), the SCVs

need to be determined first, which then allows the calculation of the weight function. For this, the approach used by Whitt and You (2022) is followed.

Let $t \to \infty$ in the system of IDC equations in (7.3.7). Consequently, the weight function $w_i(t) \to 1$. The limiting value of the IDC corresponds to the SCV, *i.e.,* $I_a(\Lambda_{j,i}(\infty)) \equiv c^2_{a,j,i}$, $I_a(\Lambda_{0,i}(\infty)) \equiv c^2_{a,0,i}$, $I_a(\Lambda_i(\infty)) \equiv c^2_{a,i}$ and $I_d(\Lambda_i(\infty)) \equiv c^2_{d,i}$. Now (7.3.7) can be written in terms of this limiting variability parameters as

$$\begin{aligned} c^2_{a,i} &= \sum_{j=1}^{k} \Big(\frac{\Lambda_{j,i}}{\Lambda_i}\Big) c^2_{a,j,i} + \Big(\frac{\Lambda_{0,i}}{\Lambda_i}\Big) c^2_{a,0,i} \\ c^2_{a,i,j} &= c^2_{d,i}\ p_{i,j} + (1 - p_{i,j}) \\ c^2_{d,i} &= c^2_{a,i} \end{aligned} \tag{7.3.9}$$

In matrix form, the limiting variability equations can be written as,

$$(\mathbf{E} - \mathbf{K}(\infty))\mathbf{c}^2 = \mathbf{H}(\infty). \tag{7.3.10}$$

Theorem 7.3.1. *The matrix* $(\boldsymbol{E} - \boldsymbol{K}(t))$ *is invertible and IDC equations in (7.3.8) have unique solution*

$$\boldsymbol{I}(t) = (\boldsymbol{E} - \boldsymbol{K}(t))^{-1}\boldsymbol{H}(t) \tag{7.3.11}$$

Also, the set of limiting variability equations in (7.3.10) has unique solution

$$\boldsymbol{c}^2 = (\boldsymbol{E} - \boldsymbol{K}(\infty))^{-1}\boldsymbol{H}(\infty) \tag{7.3.12}$$

This theorem can be easily proved using the Kronecker delta function. Theorem 1 in Whitt and You (2022) is proved in a similar manner.

Proof 7.3.1. *By substituting (7.3.2) and (7.3.5) in (7.3.6), we get*

$$I_a(\Lambda_i(t)) = \sum_{j=1}^{k} \Big(\frac{\Lambda_{j,i}(t)}{\Lambda_i(t)}\Big)\Big(\big(w_j(t)I_a(\Lambda_j(t)) + (1-w_j(t))I_s(M_j(t))\big)\ p_{j,i} + (1-p_{j,i})\Big) + \frac{\Lambda_{0,i}(t)}{\Lambda_i(t)} I_a(\Lambda_{0,i}(t)) \tag{7.3.13}$$

$$= \sum_{j=1}^{k} \Big(\frac{\Lambda_{j,i}(t)}{\Lambda_i(t)}\Big) w_j(t)\ p_{j,i}\ I_a(\Lambda_j(t)) + \sum_{j=1}^{k} (\frac{\Lambda_{j,i}(t)}{\Lambda_i(t)})(1-w_j(t))I_s(M_j(t))\ p_{j,i} + \sum_{j=1}^{k} (\frac{\Lambda_{j,i}(t)}{\Lambda_i(t)})(1-p_{j,i})) + \frac{\Lambda_{0,i}(t)}{\Lambda_i(t)} I_a(\Lambda_{0,i}(t))$$

Coefficient matrix of $I_a(\Lambda_i(t))$ becomes

$$\Big(\delta_{i,j} - \Big(\frac{\Lambda_{j,i}(t)}{\Lambda_i(t)}\Big) w_j(t)\ p_{j,i}\Big) \in \mathcal{R}^{k^2},$$

where $\delta_{i,j} = \begin{cases} 0 & for \quad i \neq j \\ 1 & for \quad i = j \end{cases}$ *is the Kronecker delta function.* $\Big(\frac{\Lambda_{j,i}(t)}{\Lambda_i(t)}\Big)$ *is the proportion of customers move from j to i in $[0,t]$. It always ranges from 0 to 1 and we have $0 \leq w_i(t) \leq 1$. Therefore $\Big(\frac{\Lambda_{j,i}(t)}{\Lambda_i(t)}\Big) w_j(t) \leq 1$ for all $t \geq 0$.*

As assumed in the model description, $(I - P')$ is invertible, coefficient matrix of $I_a(\Lambda_i(t))$ is invertible and we get unique solution for $I_a(\Lambda_i(t))$. Then using (7.3.13), unique solutions for $I_d(\Lambda_i(t))$ and $I_a(\Lambda_{i,j}(t))$ can be obtained. Similarly, $\boldsymbol{c}^2$ has unique solution.

7.4 Performance Measures

Little's law provides a fundamental relationship between average number of customers in a queueing system and waiting time of a customer in the system. This can be used to approximate and estimate various key performance measures in queueing systems. A time-varying approximation for Little's law can be written as,

$$E(Q_t) \approx \lambda(t)E(W_t), \tag{7.4.1}$$

where Q_t and W_t are time-varying approximations for queue length and waiting time respectively. $\lambda(t)$ is the time-varying arrival rate.

Brumelle (1971) derived a relationship between average workload in a general queueing system, waiting time of a customer and expected service rate. Using Brumelle's formula, an approximation for time-varying workload in the system can be written as,

$$E(Z_t) \approx \rho^*(t)E(W_t) + \rho^*(t)\Big(\frac{c_s^2+1}{2\ \mu}\Big),$$

where $\rho^*(t)$ is the time-varying traffic intensity, μ is the service rate and c_s^2 is the SCV of service process.

By approximating $E(Z_t)$ in the above equation using Z_t^* from (7.2.3), $E(W_t)$ becomes

$$E(W_t) \approx max\{0, \frac{Z_t^*}{\rho^*(t)} - \frac{c_s^2+1}{2\mu}\}. \tag{7.4.2}$$

The time-varying approximation of queue length and waiting time for each station can be obtained using (7.4.1) and (7.4.2).

7.4.1 Network Performance Measures

So far, only the approximations for the performance measures of a single queue have been discussed. Using these approximations, the total network performance can also be derived. Suppose that there are various routes for customers who enter the system. Customers with similar route can be grouped in to a class. Thus, routing is fixed for a specific class. Let m be the number of classes in a network. It is assumed that a customer's waiting time and service time are independent of their class.

Therefore, time-varying approximation for mean sojourn time $E(S_t^l)$ of customer class l is,

$$E(S_t^l) \approx \sum_{j=1}^{n_l} n_j^l (E(W_{t,j}) + \frac{1}{\mu_j}), \qquad l = 1, 2, ..., m, \tag{7.4.3}$$

where n_j^l is the number of times a customer in class l visits node j. Since the model considered here is a feed-forward queueing network, n_j^l is always one. $E(W_{t,j})$ is the time-varying approximation for waiting time at node j before the service begins and $1/\mu_j$ is the mean service time at node j. And assume that the service time of a customer does not depend on which class they belongs to.

7.5 Algorithm For Time-Varying Queueing Model

This section presents the general structure of the IDC-based algorithm for deriving time-varying approximations of network performance measures in a feed-forward queueing network model.

7.5.1 Algorithm

Input: Routing matrix P, External arrival rate function $\lambda(t)$, Service rate function $\mu_i, i = 1, 2, 3$.
Output: time-varying approximation for the system performance measures

Step 1: Obtain cumulative arrival rate ($\Lambda(t)$), cumulative service rate ($M(t)$) and time-varying traffic intensity ($\rho^*(t)$) for each station.
Step 2: Obtain time-varying approximation for traffic equations by (7.3.1).
Step 3: Solve the limiting variability equations using (7.3.9).
Step 4: Obtain IDCs of total arrival processes using (7.3.11)
Step 5: For a selected station, obtain IDW by (7.2.5).
Step 6: Obtain time-varying approximations for mean workload in each station by (7.2.9).
Step 7: Obtain time-varying approximations for mean queue length and waiting time at each station and the average sojourn time in different routes.

In step 3 using the fact that the weight function $w_i(t)$ tends to 1 as t tends to ∞ to solve the limiting variability equations. Step 4 uses the SCVs obtained from step 3 to get IDCs for the external arrival and service processes. Based on the rate functions and IDCs, IDWs and further approximations can be computed.

7.6 Numerical Study

In this section, a three-station feed-forward queueing network model with FCFS discipline is considered. The algorithm is implemented in this model, and the network performance is analysed using the time-varying approximations. The specific key performance measures such as queue length and waiting time of customers in each queue and average sojourn time in the system are analysed.

7.6.1 A Three-station Queueing Network Model

In this example, there are three stations, and customer feedback is not allowed. Here, after service from station 1, customers are split into two and move towards station 2 and station 3. Superposition of customers from station 2 and station 1 occurs in queue 3.

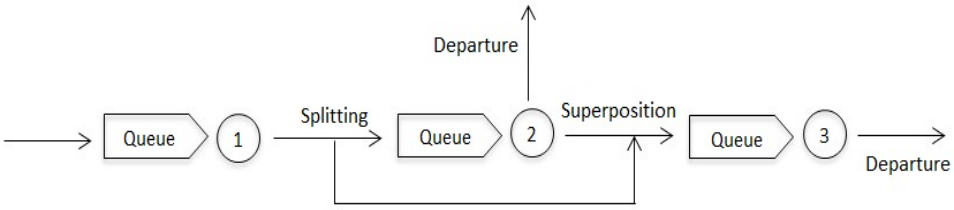


Figure 7.1: A three-station queueing network model

Let the transition probabilities be, $p_{12} = 0.6$, $p_{23} = 0.7$, $p_{13} = 0.4$. Customer departure from the network is possible from station 2 and station 3.

Here, we bring in the time-varying feature focusing on a sinusoidal function as the external arrival rate function. *i.e.,*

$$\lambda_{0,1}(t) = 1 + 0.8\ sin(\pi\gamma t),\ \gamma = 0.1,\ t \geq 0.$$

Then, the cumulative external arrival rate becomes

$$\Lambda_{0,1}(t) = \int_0^t \big(1 + 0.8\ sin(\pi\gamma u)\big) du.$$

Let us take the service rate functions at each node to be constant with respect to time, as $\mu_1 = 3$, $\mu_2 = 5$ and $\mu_3 = 7$. Corresponding cumulative service rates become, $M_1(t) = 3t$, $M_2(t) = 5t$ and $M_3(t) = 7t$. Cumulative arrival rate for each station can be obtained as $\Lambda_1(t) = \Lambda_{0,1}(t)$, $\Lambda_2(t) = M_1(t)\ p_{12}$ and $\Lambda_3(t) = M_1(t)\ p_{13} + M_2(t)\ p_{23}$.

Now considering four cases of this network. In each case, there is a set of values

for SCVs of external arrival process and service processes with the cases labelled as A, B, C, and D, as shown in Table 7.1.

Case	$c^2_{a,0,1}$	$c^2_{s,1}$	$c^2_{s,2}$	$c^2_{s,3}$
A	0.75	0.75	0.75	0.75
B	0.50	1.00	2.25	2.00
C	1.25	0.80	0.80	1.00
D	2.00	1.50	2.00	0.50

Table 7.1: Variability of external arrival distribution and service distributions of four different cases in the study.

By solving the limiting variability equations in (7.3.12) which involves the 15×15 matrix $E - K(\infty)$, we get the vector c^2 for 4 different cases. Then for each case, obtain IDC, $I_a(\Lambda_i(t))$ from (7.3.11) and hence IDW, $I_w(\Lambda_i(t))$ using (7.2.6). The value of b in (7.2.9) is chosen as $\sqrt{2}$ in order to provide a better approximation, see Whitt and You (2019). Then the time-varying approximation for mean workload can be obtained from (7.2.9).

Figure 7.2 illustrates time-varying approximation for mean waiting time in each queue for the cases considered in this study. Similarly, Figure 7.3 presents time-varying approximation for mean queue length in the stations. The mean queue length at station 1 is periodic in nature. Since station 3 has low traffic intensity, it has shorter mean waiting times and queue length. In stations with a higher SCV, the waiting time and queue are longer.

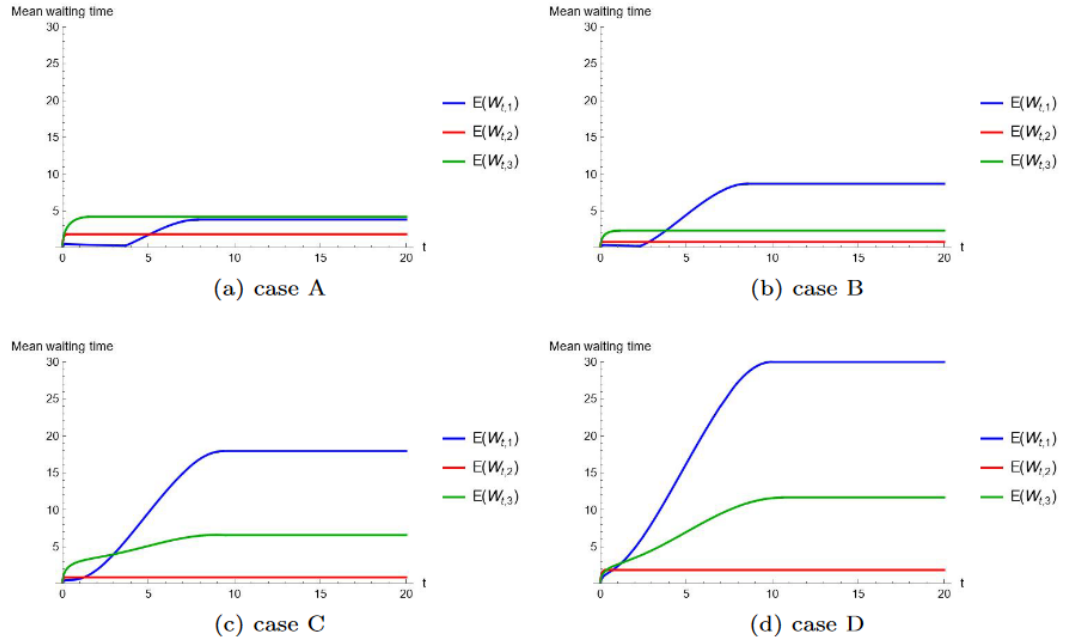


Figure 7.2: Mean waiting time for four different cases considered in this study

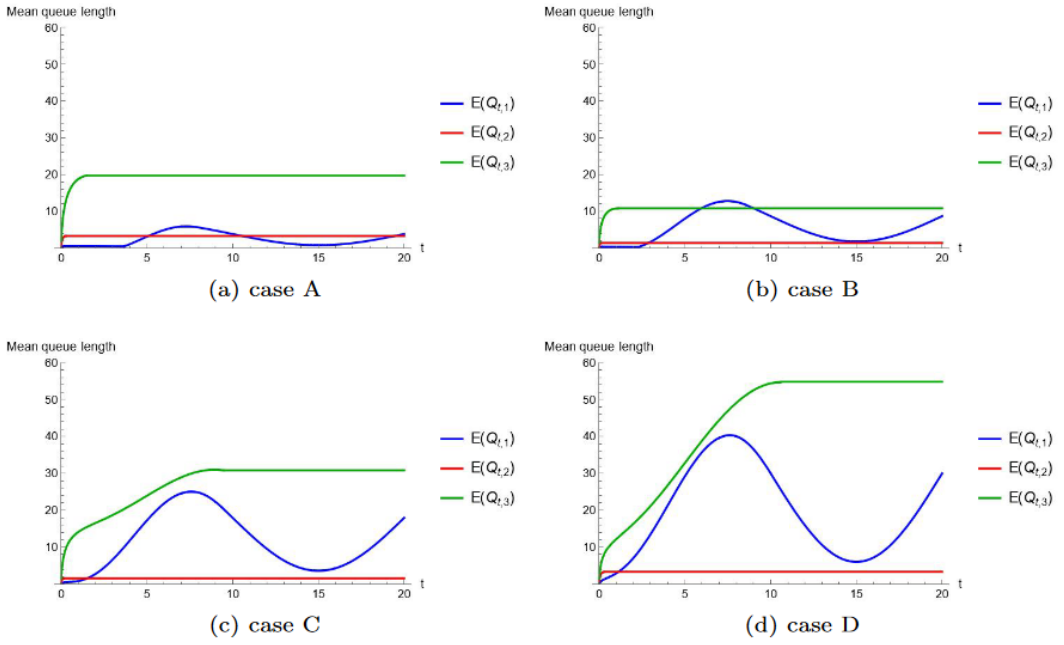


Figure 7.3: Mean queue length for four different cases considered in this study

As shown in the network model, customers have three different routes to choose from. (i) [1] $\rightarrow$ [2] $\rightarrow$ [3] (ii) [1] $\rightarrow$ [2] and (iii) [1] $\rightarrow$ [3] and the routes are

denoted by $r_i, \quad i = 1,\ 2,\ 3$. The average sojourn time for customers in each route is computed separately and it is depicted in Figure 7.4. As compared to routes 1 and 2, route 3 has the shortest sojourn time. Customers in route 1 experience the longest sojourn time among the four cases. The algorithm is implemented and produced figures using Wolfram Mathematica 12.3.

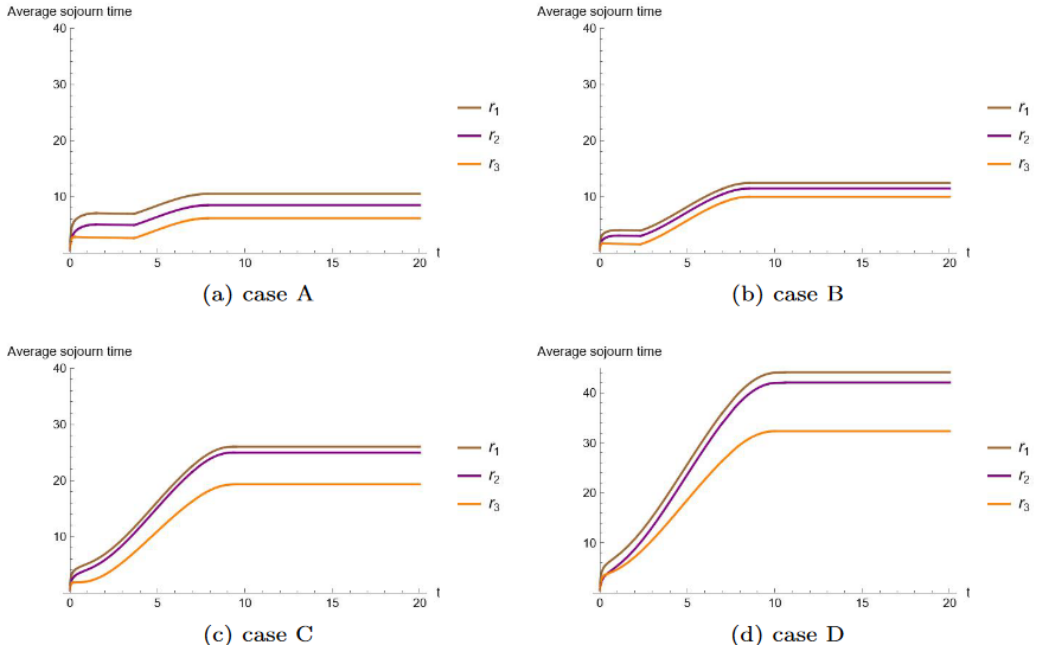


Figure 7.4: Sojourn time in three routes r_1, r_2 and r_3 for four different cases considered in this study

7.7 Summary of the Chapter

Queueing networks, assuming non-Markovian and non-stationary processes can be applied to various real-world service systems. In this chapter, a time-varying approximation algorithm is formulated based on indices of dispersion, for approximating performance measures in a feed-forward non-stationary general queueing network. The approach involves forward-time construction of workload process in non-Markovian time-varying queues which facilitate complex network approxima-

tions.

Decomposition approximations for OQNs can be challenging due to the interdependence of the arrival process on the flow of customers through the network. By utilizing the system of equations for the index of dispersion (variance-time function) and considering network operations such as departure, superposition, and splitting, the interdependence in OQNs can be addressed. Measures like mean queue length, mean waiting time, and mean sojourn time are crucial for evaluating the performance of a queueing system. The time-varying approximation of these measures is highly influenced by the variability in service processes and external arrival processes. The numerical study provides an understanding of how variability in these processes impacts network performance measures. The developed algorithm effectively discerns the time-dependent behaviour of an OQN with feedback restricted queues.

CHAPTER 8

FINAL CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

8.1 Introduction

Queueing models are essential tools for analysing and optimizing service systems where waiting line plays a critical role. However, most large-scale service systems in real life are influenced by time-varying conditions, such as non-stationary arrival rates, service rates, and other factors that impact overall system performance. Analysing time-dependent behaviour in such non-stationary queueing systems is more challenging than studying their steady-state behaviour. This study mainly focused on transient behaviour of non-stationary queueing systems with time-varying arrival and/service rates. Developed transient distributional laws related to virtual workload in single server queues as well as tandem queueing networks with finite and infinite capacities. When it comes to general non-stationary queueing networks instead of tandem queueing networks, analytical and approximation methods are of-

ten required to effectively examine the time-dependent behaviour. So time-varying approximations for performance measures in such queueing networks are also developed in order to understand the dynamic behaviour. Python and Mathematica are the computational tools utilised in this study.

8.2 Summary of the Thesis

Chapter 1 provided the contextualization of the study. It discussed the basic characteristics of a queueing system and introduced relevant notations. The chapter also covered the concept of a network of queues, its classifications, and network operations. Additionally, it presented the structural outline of the study.

In Chapter 2, some analytical issues related to complex queueing networks are discussed. It includes a discussion on non-product form networks and their analytical complexities, as well as approximation methods and simulation techniques. Other sources of complexity, such as blocking, feedback, and non-stationarity, are also considered. A study of applications of queueing networks is also presented.

In Chapter 3, some basic concepts of time-varying queueing systems and the non-homogeneous Poisson process are discussed. Inferential aspects of the intensity function of non-homogeneous Poisson process, specifically the maximum likelihood estimation method, are provided. Simulation and real case studies are conducted based on the maximum likelihood estimation, and the estimates of the intensity function are illustrated. Additionally, a brief literature survey is conducted, highlighting some analytical issues related to the analysis of time-varying queues.

In Chapter 4, the transient distributional relationship between the average workload and the customer's waiting time for a single-server non-Markovian queue with time-varying arrival and service rates is derived. Data analysis is conducted on

simulated call centre data using Python, and presented the results.

In Chapter 5, the transient laws for single-server queues discussed in Chapter 4 are extended to a k-station tandem network. Furthermore, an algorithm for analysing transient performance measures in a k-station tandem queueing network is developed, and a numerical study is conducted based on the algorithm. The numerical study supported the effectiveness of the algorithm, and the results provided insights into the transient behaviour of tandem networks, specifically in bottleneck scenarios.

Chapter 6 dealt with tandem queueing networks with capacity constraints that influence the flow of entities through the system. It compared two blocking mechanisms, blocking-after-service (BAS) and blocking-before-service (BBS), based on a two-station tandem network with finite capacity on the intermediate queue, and developed transient performance measures for the system under both mechanisms. Using a numerical approach, it presented how these mechanisms influence the time-varying number of patients and virtual workload in the system. The results demonstrated that the BAS mechanism slightly outperforms the BBS mechanism in reducing unwanted congestion.

In Chapter 7, general feed-forward open queueing networks with time-varying arrival rates are considered, and a time-varying approximation algorithm for their performance measures is developed. Mean queue length, mean waiting time, and average sojourn time in a feed-forward open queueing network can be effectively approximated using the algorithm. A numerical study is performed on a three-station feed-forward open queueing network by computing the time-varying approximation for performance measures and investigating the influence of variability in arrival and service processes on the time-varying approximations.

8.3 Limitations of the Study

While this research offers important contributions to the understanding of non-stationary queueing models, it is important to recognise its limitations. In this study, transient analysis is limited to tandem queueing networks and single-server systems, as analysing the transient behaviour of more complex queueing systems significantly increases the complexity.

One limitation of this study lies in the validation of the theoretical frameworks presented. Real data on non-stationary queueing setups, specifically arrival times, service starting times, and service ending times, are extremely difficult to obtain. As a result, real data could not be produced for the study. Therefore, synthetic data were used to validate the results. Simulated data from a call centre were generated and analysed. However, time-varying arrival data alone is relatively easier to obtain. To study the inferential aspects, real data on bus arrival times were used.

In the case of networks of non-stationary queues, simulation techniques are challenging, since the behaviour of the system may not be predictable at each time interval. To capture these dynamics during simulation, complex modelling and adjustments are required. Moreover, simulation of a non-stationary network may require running simulations over different time intervals with frequent updates, which can result in greater computational costs and longer simulation times. As a result, numerical studies were used in this study to validate theoretical results in tandem queueing networks and to approximate performance measures in general queueing networks.

In spite of these limitations, the study makes a unique contribution by exploring transient measures in non-stationary queueing models, an area where limited research exists. Identifying these limitations provides a basis for future research.

8.4 Recommendations for Future Research

Time-varying queueing networks are convenient for modelling extensive service systems such as healthcare systems, transportation systems and telecommunication. However, time-dependent performance analysis of queueing systems with time-varying queues has not received due importance. This study dealt with a similar scenario and tries to provide further directions for future research. Our primary focus is to further explore transient performance measures in time-varying queueing systems. This includes extending the transient distributional laws related to virtual workload in single-server queues to multi-server queue systems. Additionally, some aspects involve the implementation of network operations such as feedback and blocking within the queueing models examined in this study. It has already studied the impact of blocking on a two-station tandem network of single-server queues with capacity restriction on intermediate queue. This can be generalized to a more complex tandem networks. Feedback operation in time-varying tandem queueing networks is another intriguing area, offering more practical applications. Furthermore, extending the results to different queueing disciplines are some interesting avenue for future research. In addition, the approximation algorithm for time-varying general single-server queueing networks discussed in the sixth chapter can be extended to multi-server queueing networks, incorporating network operations such as feedback and blocking. Statistical inferential problems associated with these queueing models will also form an interesting theme for future research.

List of Publications

1. Anjale Ramesh and M. Manoharan (2025) Transient Behaviour of Time-varying Tandem Queueing Networks. *OPSEARCH.* 62 (1), 104–118. doi: 10.1007/s12597-024-00790-0.

2. Anjale Ramesh and M. Manoharan (2025) Aspects of Blocking on Time-Varying Tandem Queueing Network. *Croatian Operational Research Review.* 16 (2), doi: 10.17535/crorr.2025.0016.

3. Anjale Ramesh and M. Manoharan (2025) ML Estimation of Intensity Function in Non-homogeneous Poisson Processes. *Operations Research Forum.* 6 (49), doi: 10.1007/s43069-025-00447-8.

4. Anjale Ramesh and M. Manoharan (2024) Time-varying Approximation for Performance Measures in Generalized Queueing Networks. (Communicated to *Yugoslav Journal of Operations Research.*)

Presentations in Conferences

1. “ML Estimation of Intensity Function in Non-Homogeneous Poisson Processes", International Seminar on ‘Applied Statistics and Data Analytics’ organised by Department of Statistics, University of Calicut, during 04 - 06 March, 2025.

2. “Transient Performance Measures in Time-Varying Tandem Queueing Networks", International Conference on ‘Statistical Sciences and Stochastic Modelling’ organised by Department of Statistics, University of Calicut, during 16 - 17 February, 2023.

3. “Approximation Methods for the Steady-State Performance Measures in Generalized Queueing Networks", International Conference on Statistics for Twenty-first Century-2021 (ICSTC - 2021) jointly organised by International Statistics Fraternity (ISF) and Department of Statistics, University of Kerala during 15 - 19 December, 2021.

Bibliography

[1] Afolalu, S., Ikumapayi, O., Abdulkareem, A., Emetere, M., and Adejumo, O. (2021). A short review on queuing theory as a deterministic tool in sustainable telecommunication system. *Materials Today: Proceedings*, 44:2884–2888.

[2] Akyildiz, I. F. (1988). Mean value analysis for blocking queueing networks. *IEEE Transactions on Software Engineering*, 14(4):418–428.

[3] Altiok, T. and Perros, H. G. (1987). Approximate analysis of arbitrary configurations of open queueing networks with blocking. *Annals of Operations Research*, 9(1):481–509.

[4] Armony, M., Israelit, S., Mandelbaum, A., Marmor, Y. N., Tseytlin, Y., and Yom-Tov, G. B. (2015). On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic systems*, 5(1):146–194.

[5] Ashour, S. and Jha, R. (1973). Numerical transient-state solutions of queueing systems. *SIMULATION*, 21(4):117–122.

[6] Avi-Itzhak, B. (1965). A sequence of service stations with arbitrary input and regular service times. *Management Science*, 11(5):565–571.

[7] Avi-Itzhak, B. and Halfin, S. (1993). Servers in tandem with communication

and manufacturing blocking. *Journal of Applied Probability*, 30(2):429–437.

[8] Avi-Itzhak, B. and Levy, H. (1995). A sequence of servers with arbitrary input and regular service times revisited: In memory of micha yadin. *Management Science*, 41(6):1039–1047.

[9] Avi-Itzhak, B. and Yadin, M. (1965). A sequence of two servers with no intermediate queue. *Management Science*, 11(5):553–564.

[10] Baccelli, F. and Brémaud, P. (2012). *Palm probabilities and stationary queues*, volume 41. Springer Science & Business Media.

[11] Badrinath, S. and Balakrishnan, H. (2017). Control of a non-stationary tandem queue model of the airport surface. In *2017 American Control Conference (ACC)*, pages 655–661.

[12] Bailey, N. T. J. (1954). A continuous time treatment of a simple queue using generating functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16(2):288–291.

[13] Balsamo, S. (2000). *Product Form Queueing Networks*, pages 377–401. Springer Berlin Heidelberg, Berlin, Heidelberg.

[14] Balsamo, S. (2011). Queueing networks with blocking: Analysis, solution algorithms and properties. *Network Performance Engineering: A Handbook on Convergent Multi-Service Networks and Next Generation Internet*, pages 233–257.

[15] Balsamo, S. and Personè, V. d. N. (1994). A survey of product form queueing networks with blocking and their equivalences. *Annals of Operations research*, 48:31–61.

[16] Balsamo, S., Rossi, G.-L. D., and Marin, A. (2015). Applying BCMP multi–class queueing networks for the performance evaluation of hierarchical and modular software systems. *International Journal of Computer Aided Engineering and Technology*, 7(2):145–157.

[17] Bandi, C., Bertsimas, D., and Youssef, N. (2015). Robust queueing theory. *Operations Research*, 63(3):676–700.

[18] Baskett, F., Chandy, K. M., Muntz, R. R., and Palacios, F. G. (1975). Open, closed, and mixed networks of queues with different classes of customers. *J. ACM*, 22(2):248–260.

[19] Bertsimas, D. and Mourtzinou, G. (1997). Transient laws of non-stationary queueing systems and their applications. *Queueing Syst. Theory Appl.*, 25(1/4):115–155.

[20] Bhat, U. N. (1968). *Elements of Applied Stochastic Processes*. Wiley, New York.

[21] Bigot, J., Gadat, S., Klein, T., and Marteau, C. (2013). Intensity estimation of non-homogeneous Poisson processes from shifted trajectories. *Electronic Journal of Statistics*, 7:881 – 931.

[22] Blanchet, J. and Chen, X. (2019). Perfect sampling of generalized jackson networks. *Mathematics of Operations Research*, 44(2):693–714.

[23] Bolch, G., Greiner, S., De Meer, H., and Trivedi, K. S. (2006). *Queueing networks and Markov chains: modeling and performance evaluation with computer science applications*. John Wiley & Sons.

[24] Bose, S. K. (2013). *An introduction to queueing systems*. Springer Science & Business Media.

[25] Boucherie, R. J. and Van Dijk, N. M. (2010). *Queueing networks: a fundamental approach*, volume 154. Springer Science & Business Media.

[26] Bramson, M. and Dai, J. (2001). Heavy Traffic Limits for Some Queueing Networks. *The Annals of Applied Probability*, 11(1):49 – 90.

[27] Brandt, A., Franken, P., and Lisek, B. (1990). *Stationary stochastic models*, volume 78. Walter de Gruyter GmbH & Co KG.

[28] Brumelle, S. L. (1971). On the relation between customer and time averages in queues. *Journal of Applied Probability*, 8(3):508–520.

[29] Burke, P. J. (1956). The output of a queuing system. *Operations Research*, 4(6):699–704.

[30] Chandy, K. M., Herzog, U., and Woo, L. (1975). Parametric analysis of queuing networks. *IBM Journal of Research and Development*, 19(1):36–42.

[31] Chen, H. and Yao, D. D. (2001). *Fundamentals of Queueing Networks*. Springer, New York.

[32] Choudhury, G. L., Lucantoni, D. M., and Whitt, W. (1997). Numerical solution of piecewise-stationary $M_t/G_t/1$ queues. *Operations Research*, 45(3):451–463.

[33] Cox, D. R. (1955). The analysis of non-markovian stochastic processes by the inclusion of supplementary variables. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):433–441.

[34] Crommelin, C. D. (1932). Delay probability formulae when the holding times are constant. *P.O. Elec. Engrs. J.*, 25:41–50.

[35] Crommelin, C. D. (1934). Delay probability formulae. *P.O. Elec. Engrs. J.*, 26:266–274.

[36] Dai, J. G. and Harrison, J. M. (1992). Reflected Brownian Motion in an Orthant: Numerical Methods for Steady-State Analysis. *The Annals of Applied Probability*, 2(1):65 – 86.

[37] Dai, J. G., Nguyen, V., and Reiman, M. I. (1994). Sequential Bottleneck Decomposition: An Approximation Method for Generalized Jackson Networks. *Operations Research*, 42(1):119–136.

[38] Dai, J. G. and Shi, P. (2017). A Two-Time-Scale Approach to Time-Varying Queues in Hospital Inpatient Flow Management. *Operations Research*, 65(2):514–536.

[39] Dallery, Y. and Frein, Y. (1993). On decomposition methods for tandem queueing networks with blocking. *Operations research*, 41(2):386–399.

[40] Das, D., Pasupathy, K. S., Storlie, C. B., and Sir, M. Y. (2019). Functional regression-based monitoring of quality of service in hospital emergency departments. *IISE Transactions*, 51(9):1012–1024.

[41] D'Avignon, G. R. and Disney, R. L. (1977). Queues with instantaneous feedback. *Management Science*, 24(2):168–180.

[42] de Bruin, A. M., van Rossum, A. C., Visser, M. C., and Koole, G. M. (2007). Modeling the emergency cardiac in-patient flow: an application of queuing theory. *Health Care Management Science*, 10(2):125–137.

[43] Defraeye, M. and Nieuwenhuyse, I. V. (2011). Setting Staffing Levels in an Emergency Department: Opportunities and Limitations of Station-

ary Queueing Models. *Competition and Regulation in Network Industries*, 0(1):73–101.

[44] Defraeye, M. and Van Nieuwenhuyse, I. (2016). Staffing and scheduling under nonstationary demand for service: A literature review. *Omega*, 58:4–25.

[45] Disney, R. L. and Konig, D. (1985). Queueing networks: A survey of their random processes. *SIAM Review*, 27(3):335–403.

[46] Drazek, L. C. (2013). Intensity estimation for poisson processes. Master's thesis, The University of Leeds.

[47] Dudewicz, E. J. and Mishra, S. (1988). *Modern mathematical statistics*. John Wiley & Sons, Inc.

[48] Dudin, A., Kazimirsky, A., Klimenok, V., Breuer, L., and Krieger, U. (2005). The queueing model MAP| PH| 1| N with feedback operating in a markovian random environment. *Austrian Journal of Statistics*, 34(2):101–110.

[49] Ebert, A., Dutta, R., Mengersen, K., Mira, A., Ruggeri, F., and Wu, P. (2019). Likelihood-free parameter estimation for dynamic queueing networks: case study of passenger flow in an international airport terminal.

[50] Efrosinin, D., Vishnevsky, V., and Stepanova, N. (2024). A machine-learning approach to queue length estimation using tagged customers emission. In Vishnevskiy, V. M., Samouylov, K. E., and Kozyrev, D. V., editors, *Distributed Computer and Communication Networks: Control, Computation, Communications*, pages 265–276, Cham. Springer Nature Switzerland.

[51] Erlang, A. K. (1909). The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B*, 20:33–39.

[52] Feller, W. (1949). Fluctuation theory of recurrent events. *Transactions of the American Mathematical Society*, 67:98–119.

[53] Fendick, K. W. and Whitt, W. (1989). Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proceedings of the IEEE*, 77(1):171–194.

[54] Filipowicz, B. and Kwiecień, J. (2008). Queueing systems and networks. models and applications. *Bulletin of the polish academy of sciences technical sciences*, pages 379–390.

[55] Finch, P. D. (1959). Cyclic queues with feedback. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(1):153–157.

[56] Flottemesch, T. J., Gordon, B. D., and Jones, S. S. (2007). Advanced statistics: developing a formal model of emergency department census and defining operational efficiency. *Academic Emergency Medicine*, 14(9):799–809.

[57] Foley, R. D. and Disney, R. L. (1983). Queues with delayed feedback. *Advances in Applied Probability*, 15(1):162–182.

[58] Fralix, B. H. and Riaño, G. (2010). A new look at transient versions of little's law, and m/g/1 preemptive last-come-first-served queues. *Journal of Applied Probability*, 47(2):459–473.

[59] Frein, Y. and Dallery, Y. (1989). Analysis of cyclic queueing networks with finite buffers and blocking before service. *Performance Evaluation*, 10(3):197–210.

[60] Fry, T. C. (1928). *Probability and Its Engineering Uses.* The Bell Telephone Laboratories series, 12 th Edition, Macmillan.

[61] Galliher, H. P. and Wheeler, R. C. (1958). Nonstationary Queuing Probabilities for Landing Congestion of Aircraft. *Operations Research*, 6(2):264–275.

[62] Gangadhar, N. D. and Kadambi, G. R. (2022). Delay distributions in discrete time multiclass tandem communication network models. *International journal of electrical and computer engineering systems*, 13(6):417–425.

[63] Gerhardt, I. and Nelson, B. L. (2009). Approximating performance and traffic flow in nonstationary tandem networks of markovian queues.

[64] Gerum, P. C. L. and Baykal-Gürsoy, M. (2022). How incidents impact congestion on roadways: A queuing network approach. *EURO Journal on Transportation and Logistics*, 11:100067.

[65] Giambene, G. (2014). *Queuing theory and telecommunications*, volume 585. Springer.

[66] Glynn, P. W. and Iglehart, D. L. (1988). Simulation methods for queues: An overview. *Queueing systems*, 3(3):221–255.

[67] Gordon, W. J. (1967). Closed queueing systems with exponential servers. *Operation Research*, 15(2):145–155.

[68] Green, L., Kolesar, P., and Svoronos, A. (1991). Some effects of nonstationarity on multiserver markovian queueing systems. *Operations Research*, 39(3):502–511.

[69] Green, L. V. and Kolesar, P. J. (1991). The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37:84–97.

[70] Halfin, S. and Whitt, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Oper. Res.*, 29:567–588.

[71] Harrison, J. M. (1973). The heavy traffic approximation for single server queues in series. *Journal of Applied Probability*, 10(3):613–629.

[72] Harrison, J. M. and Lemoine, A. J. (1977). Limit theorems for periodic queues. *Journal of Applied Probability*, 14(3):566–576.

[73] Harrison, J. M. and Nguyen, V. (1990). The QNET method for two-moment analysis of open queueing networks. *Queueing systems*, 6:1–32.

[74] Harrison, J. M. and Reiman, M. I. (1981). Reflected brownian motion on an orthant. *The Annals of Probability*, 9(2):302–308.

[75] Harrison, J. M. and Williams, R. J. (1987). Multidimensional reflected brownian motions having exponential stationary distributions. *The Annals of Probability*, 15(1):115–137.

[76] Hasofer, A. M. (1964). On the single-server queue with non-homogeneous poisson input and general service time. *Journal of Applied Probability*, 1(2):369–384.

[77] Heyman, D. P. and Whitt, W. (1984). The asymptotic behavior of queues with time-varying arrival rates. *Journal of Applied Probability*, 21(1):143–156.

[78] Higasa, K., Sekine, K., and Itoh, E. (2023). Effectiveness of aircraft inter-arrival control in upstream traffic flow via a combined tandem fluid queue model and integer programming approach. *IEEE Access*, 11:15252–15270.

[79] Iglehart, D. L. and Whitt, W. (1970a). Multiple channel queues in heavy traffic. i. *Advances in Applied Probability*, 2(1):150–177.

[80] Iglehart, D. L. and Whitt, W. (1970b). Multiple channel queues in heavy traffic. ii: sequences, networks, and batches. *Advances in Applied Probability*, 2(2):355–369.

[81] Iglesias, R., Rossi, F., Zhang, R., and Pavone, M. (2019). A BCMP network approach to modeling and controlling autonomous mobility-on-demand systems. *The International Journal of Robotics Research*, 38(2-3):357–374.

[82] Izotova, A. and Valiullin, A. (2021). Comparison of poisson process and machine learning algorithms approach for credit card fraud detection. *Procedia Computer Science*, 186:721–726. 14th International Symposium "Intelligent Systems.

[83] Jackson, J. R. (1957). Networks of waiting lines. *Operations Research*, 5(4):518–521.

[84] Jackson, J. R. (1963). Jobshop-like queueing systems. *Management Science*, 10(1):131–142.

[85] Jackson, R. R. P. (1954). Queueing systems with phase type service. *OR*, 5(4):109–120.

[86] Jafarnejad Ghomi, E., Rahmani, A. M., and Qader, N. N. (2019). Applying queue theory for modeling of cloud computing: A systematic review. *Concurrency and Computation: Practice and Experience*, 31(17):e5186.

[87] Jagerman, D. L. (1975). Nonstationary blocking in telephone traffic. *Bell System Technical Journal*, 54(3):625–661.

[88] Jensen II, J. M. (2022). Bayesian estimation of the intensity function of a non-homogeneous poisson process. Master's thesis, Jacksonville State University.

[89] Keller, J. B. (1982). Time-dependent queues. *SIAM Review*, 24(4):401–412.

[90] Kelly, F. P. (1976). Networks of queues. *Advances in Applied Probability*, 8(2):416–432.

[91] Kendall, D. (1951). Some problems in the theory of queues. *Journal of the royal statistical society series b-methodological*, 13:151–173.

[92] Kendall, D. G. (1953). Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. *The Annals of Mathematical Statistics*, 24(3):338 – 354.

[93] Kim, S.-H., Vel, P., Whitt, W., and Cha, W. C. (2015). Poisson and non-poisson properties in appointment-generated arrival processes. *Oper. Res. Lett.*, 43(3):247–253.

[94] Kim, S.-H. and Whitt, W. (2013a). Estimating waiting times with the time-varying little's law. *Probability in the Engineering and Informational Sciences*, 27(4):471–506.

[95] Kim, S.-H. and Whitt, W. (2013b). Statistical analysis with little's law. *Operations Research*, 61(4):1030–1045.

[96] Kingman, J. (1969). Markov population processes. *Journal of Applied Probability*, 6(1):1–18.

[97] Knessl, C. and Yang, Y. P. (2002). An exact solution for an M(t)/M(t)/1 queue with time-dependent arrivals and service. *Queueing Systems*, 40(3):233–245.

[98] Kolesar, P. J., Rider, K. L., Crabill, T. B., and Walker, W. E. (1975). A queuing-linear programming approach to scheduling police patrol cars. *Operations Research*, 23(6):1045–1062.

[99] Konstantopoulos, P. and Walrand, J. (1989). Stationary and stability of fork-join networks. *Journal of Applied Probability*, 26(3):604–614.

[100] Koole, G. and Mandelbaum, A. (2002). Queueing models of call centers: An introduction. *Annals of Operations Research*, 113:41–59.

[101] Koopman, B. O. (1972). Air-terminal queues under time-dependent conditions. *Operations Research*, 20(6):1089–1114.

[102] Kouvatsos, D. D. and Awan, I.-U. (1998). Mem for arbitrary closed queueing networks with rs-blocking and multiple job classes. *Annals of Operations Research*, 79(0):231–269.

[103] Kouvatsos, D. D. and Xenios, N. P. (1989). Mem for arbitrary queueing networks with multiple general servers and repetitive-service blocking. *Performance evaluation*, 10(3):169–195.

[104] Kuehn, P. (1979). Approximate analysis of general queuing networks by decomposition. *IEEE Transactions on communications*, 27(1):113–126.

[105] Kurzhanskiy, A. A. and Varaiya, P. (2010). Active traffic management on road networks: a macroscopic approach. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1928):4607–4626.

[106] Kyritsis, A. I. and Deriaz, M. (2019). A machine learning approach to waiting time prediction in queueing scenarios. In *2019 Second International Conference on Artificial Intelligence for Industries (AI4I)*, pages 17–21.

[107] Lakatos, L., Szeidl, L., and Telek, M. (2013). *Introduction to queueing systems with telecommunication applications*, volume 388. Springer.

[108] Lee, H., Bouhchouch, A., Dallery, Y., and Frein, Y. (1998). Performance evaluation of open queueing networks with arbitrary configuration and finite buffers. *Annals of Operations Research*, 79(0):181–206.

[109] Lemoine, A. J. (1981). On queues with periodic poisson input. *Journal of Applied Probability*, 18(4):889–900.

[110] Leung, K., Massey, W., and Whitt, W. (1994). Traffic models for wireless communication networks. *IEEE Journal on Selected Areas in Communications*, 12(8):1353–1364.

[111] Li, C., Okamura, H., and Dohi, T. (2019). Parameter estimation of $M_t/M/1/K$ queueing systems with utilization data. *IEEE Access*, 7:42664–42671.

[112] Lindley, D. V. (1952). The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48(2):277–289.

[113] Little, J. D. C. (1961). A proof for the queuing formula: $l = \lambda w$. *Operations Research*, 9(3):383–387.

[114] Ma, N. and Whitt, W. (2019). Minimizing the maximum expected waiting time in a periodic single-server queue with a service-rate control. *Stochastic Systems*, 9(3):261–290.

[115] Mandelbaum, A. and Massey, W. A. (1995). Strong approximations for time-dependent queues. *Mathematics of Operations Research*, 20(1):33–64.

[116] Mandelbaum, A. and Zeltyn, S. (2007). Service engineering in action: the palm/erlang-a queue, with applications to call centers. In *Advances in services innovations*, pages 17–45. Springer.

[117] Mandelbaum, M. and Avi-Itzhak, B. (1968). Introduction to queueing with splitting and matching. *Isr. J. Technol.*, 6:376–382.

[118] Marie, R. A. (1979). An approximate analytical method for general queueing networks. *IEEE Transactions on Software Engineering*, (5):530–538.

[119] Massey, W. A. (1985). Asymptotic analysis of the time dependent M/M/1 queue. *Mathematics of Operations Research*, 10(2):305–327.

[120] Massey, W. A. and Whitt, W. (1994). An analysis of the modified offered-load approximation for the nonstationary erlang loss model. *The Annals of applied probability*, pages 1145–1160.

[121] McCarthy, M. L., Zeger, S. L., Ding, R., Aronsky, D., Hoot, N. R., and Kelen, G. D. (2008). The challenge of predicting demand for emergency department services. *Academic Emergency Medicine*, 15(4):337–346.

[122] Medhi, J. (2003). *Stochastic Models in Queueing Theory*. Elsevier.

[123] Meerkov, S. M. and Yan, C.-B. (2016). Production lead time in serial lines: Evaluation, analysis, and control. *IEEE Transactions on Automation Science and Engineering*, 13(2):663–675.

[124] Melamed, B. (1979). Characterizations of poisson traffic streams in jackson queueing networks. *Advances in Applied probability*, 11(2):422–438.

[125] Melikov, A. and Aliyeva, S. (2019). Refined approximate algorithm for steady-state probabilities of the large scale queuing systems with instantaneous and delayed feedback. In *Information Technologies and Mathematical Modelling. Queueing Theory and Applications: 18th International Conference, ITMM 2019, Named after AF Terpugov, Saratov, Russia, June 26–30, 2019, Revised Selected Papers 18*, pages 188–201. Springer.

[126] Melikov, A., Ponomarenko, L., and Rustamov, A. (2016a). Hierarchical space merging algorithm for the analysis of open tandem queueing networks. *Cybernetics and Systems Analysis*, 52:867–877.

[127] Melikov, A., Zadiranova, L., and Moiseev, A. (2016b). Two asymptotic conditions in queue with mmpp arrivals and feedback. In *Distributed Computer and Communication Networks: 19th International Conference, DCCN 2016, Moscow, Russia, November 21-25, 2016, Revised Selected Papers 19*, pages 231–240. Springer.

[128] Meyn, S. (2008). *Control techniques for complex networks*. Cambridge University Press, UK.

[129] Mizuno, S., Komiyama, Y., and Ohba, H. (2024). Proposal for optimizing number of servers in closed BCMP queueing network. *International Journal of Data Science and Analytics*, pages 1–10.

[130] Molina, E. C. (1927). Application of the theory of probability to telephone trunking problems. *Bell System Technical Journal*, 6(3):461–494.

[131] Nasr, W. W. and Taaffe, M. R. (2013). Fitting the pht/mt/s/c time-dependent departure process for use in tandem queueing networks. *INFORMS Journal on Computing*, 25(4):758–773.

[132] Nazarov, A., Melikov, A., Pavlova, E., Aliyeva, S., and Ponomarenko, L. (2021). Analyzing an m| m| n queueing system with feedback by the method of asymptotic analysis. *Cybernetics and Systems Analysis*, 57:57–65.

[133] Neely, M. J., Modiano, E., and Rohrs, C. E. (2003). Dynamic power allocation and routing for time varying wireless networks. In *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428)*, volume 1, pages 745–755. IEEE.

[134] Newell, G. F. (1965). Approximation methods for queues with application to the fixed-cycle traffic light. *Siam Review*, 7(2):223–240.

[135] Newell, G. F. (1968). Queues with time-dependent arrival rates: Ii. the maximum queue and the return to equilibrium. *Journal of Applied Probability*, 5(3):579–590.

[136] Newell, G. F. (1982). *Applications of Queueing Theory*. Springer Netherlands.

[137] Ngailo, T., Shaban, N., Reuder, J., Rutalebwa, E., and Mugume, I. (2016). Non homogeneous poisson process modelling of seasonal extreme rainfall events in tanzania. *Int. J. Sci. Res*, 5(10):1858–1868.

[138] Odoni, A. R. and Roth, E. (1983). An empirical investigation of the transient

behavior of stationary queueing systems. *Operations Research*, 31(3):432–455.

[139] Olu, O. T. (2019). Application of queuing theory to a bank's automated teller machine (atm) service optimization. *Mathematics Letters*, 5(1):8–12.

[140] Onvural, R. O. and Perros, H. G. (1989). Approximate throughput analysis of cyclic queueing networks with finite buffers. *IEEE Transactions on Software Engineering*, 15(6):800–808.

[141] Palm, C. (1937). *Inhomogeneous Telephone Traffic in Full-availability Goups*. Telefonaktiebolaget LM Ericsson.

[142] Palm, C. (1938). Analysis of the erlang traffic formula for busy-signal arrangements. *Ericsson Technics*, 5(9):39–58.

[143] Palm, C. (1947). Waiting times when traffic has variable mean intensity. *Ericsson Review*, 24:102–108.

[144] Palm, C. (1988). *Intensity variations in telephone traffic*. North-Holland.

[145] Pan, Y., Xu, Z., Guang, J., Chen, X., Dai, J., Wang, C., Zhang, X., Sun, J., Shi, P., Ding, Y., Wu, S., Yang, K., and Pan, H. (2021). A high-fidelity, machine-learning enhanced queueing network simulation model for hospital ultrasound operations. In *2021 Winter Simulation Conference (WSC)*, pages 1–12.

[146] Pekoz, E. A. and Joglekar, N. (2002). Poisson traffic flow in a general feedback queue. *Journal of Applied Probability*, 39(3):630–630.

[147] Pender, J. (2015). An analysis of nonstationary coupled queues. *Telecommunication Systems*, 61(4):823–838.

[148] Pender, J. (2017). Sampling the Functional Kolmogorov Forward Equations for Nonstationary Queueing Networks. *INFORMS Journal on Computing*, 29(1):1–17.

[149] Perros, H. and Snyder, P. (1989). A computationally efficient approximation algorithm for feed-forward open queueing networks with blocking. *Performance Evaluation*, 9(3):217–224.

[150] Perros, H. G. (1994). *Queueing networks with blocking*. Oxford University Press, Inc.

[151] Perros, H. G. and Altiok, T. (1986). Approximate analysis of open networks of queues with blocking: Tandem configurations. *IEEE transactions on software engineering*, (3):450–461.

[152] Pollaczek, F. (1930). Über eine aufgabe der wahrscheinlichkeitstheorie. i: Mitteilung aus dem telegraphentechnischen reichsamt. *Mathematische Zeitschrift*, 32(1):64–100.

[153] Prabhu, N. U. (1967). Transient behaviour of a tandem queue. *Management Science*, 13(9):631–639.

[154] Pujolle, G. and Ai, W. (1986). A solution for multiserver and multiclass open queueing networks. *INFOR: Information Systems and Operational Research*, 24(3):221–230.

[155] Rama Murthy M., S., Rao K., S., V, R., and Rao P., S. (2018). Transient analysis of k-node tandem queuing model with load dependent service rates. *International Journal of Engineering & Technology*, 7(3.31):141.

[156] Ran, B. and Boyce, D. (2012). *Dynamic urban transportation network models: theory and implications for intelligent vehicle-highway systems*, volume 417. Springer Science & Business Media.

[157] Rao, K. S. and Aparajitha, J. D. (2019). On two node tandem queueing model with time dependent service rates. *International Journal of System Assurance Engineering and Management*, 10(1):19–34.

[158] Ratneswaran, S. and Thayasivam, U. (2023a). Extracting potential travel time information from raw gps data and evaluating the performance of public transit - a case study in kandy, sri lanka. In *2023 3rd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, pages 1–7. IEEE.

[159] Ratneswaran, S. and Thayasivam, U. (2023b). An improved bus travel time prediction using multi-model ensemble approach for heterogeneous traffic conditions. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 2410–2415. IEEE.

[160] Reiman, M. I. (1984). Open queueing networks in heavy traffic. *Mathematics of Operations Research*, 9(3):441–458.

[161] Reiman, M. I. (1990). Asymptotically exact decomposition approximations for open queueing networks. *Operations research letters*, 9(6):363–370.

[162] Robertazzi, T. G. (2000). *Computer networks and systems: queueing theory and performance evaluation*. Springer Science & Business Media.

[163] Rolski, T. (1989). Queues with nonstationary inputs. *Queueing Systems*, 5:113–129.

[164] Ross, S. M. (2014). *Introduction to probability models*. Academic press, New York.

[165] Ross, S. M. (2022). *Simulation*. Academic press, Elsevier.

[166] Rosti, E., Schiavoni, F., and Serazzi, G. (1997). Queueing network models with two classes of customers. In *Proceedings Fifth International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems*, MASCOT-97, pages 229–234. IEEE Comput. Soc. Press.

[167] Rothkopf, M. H. and Oren, S. S. (1979). A closure approximation for the nonstationary m/m/s queue. *Management Science*, 25(6):522–534.

[168] Saaty, T. L. (1961). *Elements of queueing theory: with applications*, volume 34203. McGraw-Hill New York.

[169] Sahner, R., Trivedi, K. S., and Puliafito, A. (1996). *Product-Form Queueing Networks*, pages 85–102. Springer US, Boston, MA.

[170] Schwarz, J. A., Selinka, G., and Stolletz, R. (2016). Performance analysis of time-dependent queueing systems: Survey and classification. *Omega*, 63(C):170–189.

[171] Seo, D.-W. and Lee, H. (2011). Stationary waiting times in m-node tandem queues with production blocking. *IEEE Transactions on Automatic Control*, 56(4):958–961.

[172] Seo, D.-W., Lee, H.-C., and Ko, S.-S. (2008). Stationary waiting times in m-node tandem queues with communication blocking. *Management Science and Financial Engineering*, 14(1):23–34.

[173] Shakkottai, S., Srikant, R., and Stolyar, A. L. (2004). Pathwise optimality of the exponential scheduling rule for wireless channels. *Advances in Applied Probability*, 36(4):1021–1045.

[174] Shi, P., Chou, M. C., Dai, J. G., Ding, D., and Sim, J. (2016). Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science*, 62(1):1–28.

[175] Shortle, J. F., Thompson, J. M., Gross, D., and Harris, C. M. (2018). *Fundamentals of queueing theory*, volume 399. John Wiley & Sons.

[176] Sigman, K. (1990). The stability of open queueing networks. *Stochastic Processes and their Applications*, 35(1):11–25.

[177] Simonetta Balsamo, Vittoria Nitto Personé, R. O. (2001). *Analysis of Queueing Networks with Blocking*. Springer New York, NY.

[178] Sinu Lal, T., Krishnamoorthy, A., Joshua, V., and Vishnevsky, V. (2020). A two-stage tandem queue with specialist servers. *Applied Probability and Stochastic Processes*, pages 335–353.

[179] Slimacek, V. and Lindqvist, B. H. (2016). Nonhomogeneous poisson process with nonparametric frailty. *Reliability Engineering & System Safety*, 149:14–23.

[180] Sun, B., Jiang, Y., Wu, Y., Ye, Q., and Tsang, D. H. (2022). Performance analysis of mobile cloud computing with bursty demand: A tandem queue model. *IEEE Transactions on Vehicular Technology*, 71(9):9951–9966.

[181] Suri, R. and Diehl, G. W. (1984). A new'building block'for performance evaluation of queueing networks with finite buffers. *ACM SIGMETRICS Performance Evaluation Review*, 12(3):134–142.

[182] Syski, R., Eades, T., and Morris, H. (1960). Introduction to congestion theory in telephone systems. *(No Title)*.

[183] Taaffe, M. R. and Gordon, M. C. (1982). Approximating non-stationary queueing systems. Technical report, Institute of Electrical and Electronics Engineers (IEEE).

[184] Taaffe, M. R. and Ong, K. (1984). Approximating time-dependent non-exponential queues. Technical report, Institute of Electrical and Electronics Engineers (IEEE).

[185] Takacs, L. (1963). A single-server queue with feedback. *Bell system Technical journal*, 42(2):505–519.

[186] Takács, L. (1977). A queuing model with feedback. *RAIRO-Operations Research*, 11(4):345–354.

[187] Takahashi, Y., Miyahara, H., and Hasegawa, T. (1980). An approximation method for open restricted queueing networks. *Operations Research*, 28(3):594–602.

[188] Van Do, T. (2010). An efficient computation algorithm for a multiserver feedback retrial queue with a large queueing capacity. *Applied Mathematical Modelling*, 34(8):2272–2278.

[189] Varki, E. and Dowdy, L. W. (1996). Analysis of balanced fork-join queueing networks. In *Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '96, page 232–241, New York, NY, USA. Association for Computing Machinery.

[190] Veatch, M. H. and Wein, L. M. (1994). Optimal Control of a Two-Station Tandem Production/Inventory System. *Operations Research*, 42(2):337–350.

[191] Vishnevsky, V. and Gorbunova, A. V. (2022). Application of machine learning methods to solving problems of queuing theory. In Dudin, A., Nazarov, A., and Moiseev, A., editors, *Information Technologies and Mathematical Modelling. Queueing Theory and Applications*, pages 304–316, Cham. Springer International Publishing.

[192] Wang, C. (2017). Study on toll plaza design based on M/M/1 queue theory. In *2017 7th International Conference on Education, Management, Computer and Society (EMCS 2017)*, pages 755–759. Atlantis Press.

[193] Wein, L. M. (1989). Capacity allocation in generalized jackson networks. *Operations Research Letters*, 8(3):143–146.

[194] Whitt, W. (1983). The queueing network analyzer. *The bell system technical journal*, 62(9):2779–2815.

[195] Whitt, W. (1985). Queues with superposition arrival processes in heavy traffic. *Stochastic Processes and their Applications*, 21(1):81–91.

[196] Whitt, W. (1991). The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science*, 37(3):307–314.

[197] Whitt, W. (1994). Towards better multi-class parametric-decomposition approximations for open queueing networks. *Annals of Operations Research*, 48:221–248.

[198] Whitt, W. (2002). Stochastic-process limits: an introduction to stochastic-process limits and their application to queues. *Space*, 500:391–426.

[199] Whitt, W. (2014). Heavy-traffic limits for queues with periodic arrival processes. *Operations Research Letters*, 42(6–7):458–461.

[200] Whitt, W. (2015). Stabilizing performance in a single-server queue with time-varying arrival rate. *Queueing Systems*, 81:341–378.

[201] Whitt, W. (2018). Time-varying queues. *Queueing models and service management*, 1(2).

[202] Whitt, W. and You, W. (2018a). Heavy-traffic limit of the GI/GI/1 stationary departure process and its variance function. *Stochastic Systems*, 8(2):143–165.

[203] Whitt, W. and You, W. (2018b). Using robust queueing to expose the impact of dependence in single-server queues. *Operations Research*, 66(1):184–199.

[204] Whitt, W. and You, W. (2019). Time-varying robust queueing. *Operations Research*, 67(6):1766–1782.

[205] Whitt, W. and You, W. (2022). A robust queueing network analyzer based on indices of dispersion. *Naval Research Logistics (NRL)*, 69(1):36–56.

[206] Xiao, H. and Zhang, G. (2010). The queuing theory application in bank service optimization. In *2010 International Conference on Logistics Systems and Intelligent Management (ICLSIM)*, volume 2, pages 1097–1100. IEEE.

[207] Xiong, Y., Murdoch, D. J., and Stanford, D. A. (2015). Perfect sampling of a single-server queue with periodic poisson arrivals. *Queueing Systems*, 80(1–2):15–33.

[208] Yom-Tov, G. B. and Mandelbaum, A. (2014). Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299.

[209] Zhang, J. and Coyle, E. (1991). The transient solution of time-dependent M/M/1 queues. *IEEE Transactions on Information Theory*, 37(6):1690–1696.

[210] Zhao, M. and Xie, M. (1996). On maximum likelihood estimation for a general non-homogeneous poisson process. *Scandinavian Journal of Statistics*, 23(4):597–607.

[211] Ziya, S. (2008). On the relationships among traffic load, capacity, and throughput for the M/M/1/m, M/G/1/m-PS, and M/G/c/c queues. *IEEE Transactions on Automatic Control*, 53(11):2696–2701.

[212] Zychlinski, N., Mandelbaum, A., and Momčilović, P. (2018a). Time-varying tandem queues with blocking: modeling, analysis, and operational insights via fluid models with reflection. *Queueing Systems*, 89(1–2):15–47.

[213] Zychlinski, N., Momčilović, P., and Mandelbaum, A. (2018b). Time-varying many-server finite-queues in tandem: Comparing blocking mechanisms via fluid models. *Operations Research Letters*, 46(5):492–499.