ABSTRACT

YU, YAO. Performance Approximations and Controls to Queueing Networks. (Under the direction of Yunan Liu.)

This dissertation provides closed-form formulas to approximate the performance of general queueing models and effective control policies to reduce congestion in queueing networks. Queueing models are widely used in the analysis of service systems such as call centers and healthcare systems. For the smaller-scale system having exponential distributions, it is possible to obtain exact performance solutions. But when the scale is large or the system is endowed with more realistic stochastic features, it can be extremely difficult to develop closed-form performance formulas. Computer simulations can be used for studying system performance, but they are time-consuming and lack of structural insights. Hence, heavy-traffic limits, obtained from asymptotic analysis on queueing models, are widely used as efficient approximations.

To gain heavy-traffic limits, we need to justify the existence and uniqueness of heavy-traffic limits. Then, by appropriately scaling heavy-traffic limits, we can use to develop approximations for realistic queueing models. However, if approximations sometimes do not perform good, we refine approximations following some engineering principles to enhance the approximation accuracy.

In this dissertation, we study the heavy-traffic limits for various queueing models (networks) in Chapters 2-5. We start from establishing heavy-traffic limits for loss model ($GI/GI/n/0$) in Chapter 2. Due to the difficulty, we prove the existence of convergent sequences for $GI/GI/n/0$ in the asymptotic analysis but left the uniqueness and the analytical form of heavy-traffic limits unsolved. In Chapters 3 and 4, based on the well established heavy-traffic limits for time-varying queues with customer abandonments ($G_t/M/s_t + GI$), we find approximations sometimes become ineffective to approximate system performance. Therefore, we develop simple engineering approximation formulas for the overloaded steady-state multi-server queues ($GI/GI/n + GI$) and time-varying queues with customer abandonments ($G_t/GI/s_t + GI$). We then conduct extensive numerical experiments to test the effectiveness of new approximations across a wide range of model parameters. In Chapter 5, we observe that, in real service systems with dedicated servers, customers usually experience pre-arrival delays before actually arrive at the dedicated server. When the pre-arrival delay is close to zero, the well-studied joining-the-shortest queue (JSQ) policy has been proven effective. However, JSQ starts to degrade significantly as this delay grows. We propose new routing policies which make the queue asymptotically achieve resource pooling and develop the corresponding heavy-traffic limits. We also conduct computer simulations to verify the efficiency of the new policy.

Performance Approximations and Controls to Queueing Networks

by
Yao Yu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Industrial Engineering

Raleigh, North Carolina

2017

APPROVED BY:

Shuangchi He	Shu-Cherng Fang
James Wilson	Lorena Bociu Minor of Advisory Committee

Yunan Liu
Chair of Advisory Committee

DEDICATION

To my family, for their consistent love and supports.

BIOGRAPHY

Yao Yu was born and grew up in Changzhi, Shanxi Province, China, with parents and a sibling. After Receiving two Bachelor degrees (Industrial Safety and Finance) from University of Science and Technology Beijing (formerly known as Beijing Steel and Iron Institute) in 2011, Yao Yu become a graduate student in the Department of Industrial and Systems Engineering at North Carolina State University since then.

ACKNOWLEDGEMENTS

First and foremost, I am deeply thankful to my advisor, Dr. Yunan Liu, who inspired me in many ways. Before sitting in his class, I have no clue on (the toughness of) graduate studies; before running with him, I have no clue on how athletic a scholar could be; before I saw how he trained his table tennis skills, I have no clue on how professional and eager a "learner" could be; and before doing research with him, I have no clue on how knowledgeable and smart an advisor could be. He is too versatile to amaze me for finite times, and I hope I can reach his level at least in one aspect. Except setting an example for me, more importantly, he inspired my quest, paved the way for my research, and provided chances for me to experience various research topics and environments. This dissertation will be nowhere without his prompt, patient and professional guidance. The past five years is and will be the period that I deeply appreciate in my life.

Next, I would like to thank Dr. Ward Whitt, Dr. Shuangchi He and Dr. Junfei Huang for their knowledgeable teaching and generous helps. It was a honor and a valuable experience to work with Dr. Ward Whitt. From him, I learned how to establish things from the fundamental level. I was also impressed by his grand passion on research. It was also a very valuable experience to work with Dr. Shuangchi He, whose wisdom and insight on problems also amazed me for uncountably many times. It was also a valuable experience to work with Dr. Junfei Huang, whose serious altitudes in mathematics influenced me. I felt so fortunate to have chances to work under their guidance.

I also would like to thank my committee members Dr. Shu-Cherng Fang, Dr. James Wilson, Dr. Lorena Bociu, Dr. Shuangchi He for their valuable time and comments, which enrich my knowledge and help me refine this dissertation. I would like to thank Dr. Shu-Cherng Fang for countless great suggestions; I also would like to thank Dr. James Wilson for his precise suggestions to improve the methodology; I would like to thank Dr. Lorena Bociu for her great comments and her knowledgeable teaching; and I would like to thank Dr. Shuangchi He again for his great guidance and valuable time.

I would like to thank my fellow students: Qi An, Korhan Aras, Chi-Yi Chen, Xiaoling Guo, Beixiang He, Liam Huang, Yao-Huei Huang, Shan Jiang, Yu-Liang Lin, Jian Luo, Sha Luo, Kang-Ting Ma, Tiantian Nie, Xiaohu Qian, Ye Tian, Ziteng Wang, Xin Yan, Zinan Yi, Jiahua Zhang, Ling Zhang, Jing Zhou for being great friends. I would like to thank Yueqing Li, Jingyi Yang and Ming Zheng for sharing wonderful two year to me in one living room. I would like to thank Amy, Blake, Dai, Diana, Jack, Justin, Michael, Mindy, Paul, Robert, Sarah, Stewart, etc., for helping me overcome culture shocks since my first day in the US. In particular, I would like to thank Diana Williams for leaving us so many happy memories, before we can recall no more forever.

At last, I would like to thank my family for their consistent love and supports. Without them, I am nowhere.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

This dissertation is motivated by the growing needs to improve the performance of service systems, such as healthcare systems, call centers, judicial and penal systems, both front office and back-office operations in business systems and computing systems, see [1, 122, 36] and references therein. Realistic features of service systems include (i) non-Markovian probabilities structures [15], (ii) customer abandonments [20], (iii) time-varying arrivals and staffings [74] and (iv) network structures with various routing policies. All these stated features make service systems difficult to analyze exactly, therefore effective approximation techniques are needed. Approximation techniques based on heavy-traffic limits for queues systems have been proven useful since the 1930s. This dissertation aims to (i) advance methodology in heavy-traffic limits for steady-state queueing systems, (ii) provide useful engineering approximations for a general queueing system based on heavy-traffic limits, and (iii) develop asymptotically optimal routing policies for queueing networks.

1.1 Non-Markovian Queueing Models with Impatient Customers

1.1.1 Customer Abandonment

Customer abandonment is common in service systems: customers will often choose to leave if they have not entered service within a reasonable delay. For example, in a hospital emergency room, patients sometimes leave the hospital before being seen by doctors, known as the *left-without-being-seen* (LWBS) effect [122]. Meanwhile, a small number of abandonment results in significant changes in the system performance such as the customer waiting time and the waiting queue length [32, 20]. In addition, a higher level of abandonment usually indicates a lower grade of customer satisfaction. Hence the customer abandonment has been recognized as an important feature in service systems [32, 83, 123, 20].

For service systems with impatient customers, we focus on several key performance measures including the waiting queue length, waiting time, number of busy servers and total number of

customers in the system, probability of customer abandonment, and probability of delay. We next introduce the most basic queueing models: the Erlang models.

1.1.2 The Classic Erlang Models

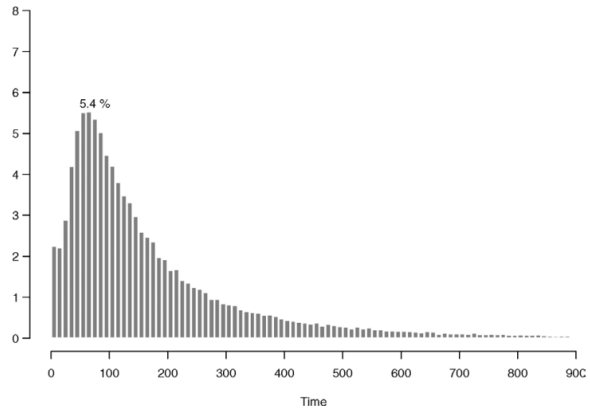
Erlang models assume exponentially distributed service times for mathematical convenience. The earliest queueing model model is the Erlang-C model [24, 25], referred to as $M/M/n/\infty$, where the first two "M"'s indicate that the interarrival times and service times are exponentially distributed, the "∞" means an infinity waiting capacity, and the "n" is the total number of servers. Customers are served in the order of their arrivals, i.e. *first-come, first-served* (FCFS). Customer service times are independent of interarrival times. Unlike Erlang-C, the Erlang-B model, or the $M/M/n/0$ loss model, has no waiting line. So customers arriving at an $M/M/n/0$ system will be blocked and lost if no server is idle. Both Erlang-C and the Erlang-B models are quite tractable but restrictive, because they do not consider the customer abandonment. Later in [32], the Erlang-A model, or the $M/M/n/0+M$ model, was developed, where the "$+M$" indicates that customer patience times are exponentially distributed. Erlang-A models can be regarded as a generalization of both Erlang-B and Erlang-C models. If the abandonment rate decreases to zero, Erlang-A degenerates to Erlang-C; if the abandonment rate increases to ∞, Erlang-A becomes Erlang-B, because customers, if ever delayed, will be blocked as if they abandon immediately.

1.1.3 Non-Makorvian Models

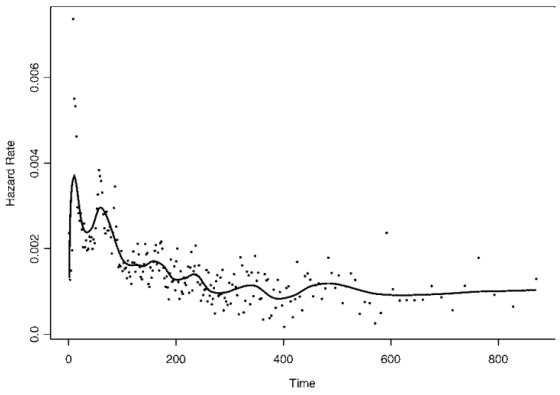
The Erlang models are easy to analyze because the system state (such as the number of customers in the system) is a birth-and-death process, which is a special case of continuous-time Markov chain (CTMC). However, real service systems are much more complex than Markovian models. For example, statistical analysis shows that customer service times in large-scale call centers are nearly lognormally distributed and the patience times can be far from exponentially distributed, as their hazard rates are not a constant, see Figures 1.1 in [15].

There is a significant number of papers that study non-Markovian models. Whitt [107] shows that the abandonment distribution beyond its mean has a significant impact upon queueing performance, so the $M/M/n+GI$ model (having a general abandonment-time distribution, i.e., the "$+GI$") does not resemble the corresponding Erlang-A model. Although the service time distribution beyond its mean plays a limited role when estimating mean performance measures (such as the mean waiting time) [117], we find that the service time distribution significantly affects other system performance measures (such as variances and tail probabilities of queue length, waiting time, etc.), see Chapter 3.

For steady-state $M/M/n+GI$ queues, closed-form expressions, though rather complicated, have been established for some key performance measures. But for the more general $M/GI/n+GI$



(a) Histogram of sevice time with mean=200, std=249



(b) Hazard rate of patience time

Figure 1.1: Estimated service and patience time from call centers in [15]

model, exact formula is not established yet. Due to the intractability of the system performance, researchers are motivated to seek approximations, many of which are based on heavy-traffic limits of queueing models, see Kingman [61, 63, 64] for the $GI/GI/1$ single-server model and later Iglehart and Whitt [50, 51, 108] for the $GI/GI/n$ multi-server queues. Halfin and Whitt later established a many-server heavy-traffic limit for the $GI/M/s$ model in [38], which initiated a sequence of groundbreaking works on many-server heavy-traffic limits.

On the other hand, Whitt [117] developed fluid models for the stationary $GI/GI/n + GI$ models, and Liu and Whitt [74, 75, 76] extended the stationary fluid model to the time-varying $G_t/GI/s_t + GI$ model, having time-varying arrivals (the G_t) and staffing (the s_t). Later in [77], Liu and Whitt proved that a fluid model is the many-server heavy-traffic limit for the corresponding queueing models by establishing *functional weak laws of large numbers* (FWLLN) results. The fluid model provides approximations for first order performance measures such as the mean number of customers in waiting, mean waiting time, etc. Meanwhile, Liu and Whitt in [77] also established a many-server heavy-traffic limits *funcational central limit theorem* (FCLT) results for the time-varying $G_t/M/s_t + GI$ queueing model. Their results can be used to approximate variances and distributions of the time-varying queues having exponential service times. Based on their theoretical work in [77], we establish Gaussian based approximations for the steady-state $GI/GI/n + GI$ queueing models and verify their accuracy across a wide range of system parameters in Chapter 3. In Chapter 4, we extend our results to the time-varying $G_t/M/n + GI$ and $G_t/GI/n + GI$ systems.

1.2 Routing to Improve Performance

Real service systems can often be modeled as queueing networks having a finite number of stations. In a hospital, for instance, patients are routed among different departments such as emergency departments, operating rooms, intensive care units, etc. Patient flows and optimal routing policies in hospitals have been widely studied [27, 79, 2, 42]. However, patient flows and routing outside a hospital have not yet drawn enough attention. For instance, many hospitals start to announce their delays online, so patients can pick a less congested hospital by comparing the announced online waiting times of all hospitals. However, the actual waiting time experienced by patients may be significantly different from the announced waiting time, because the information can quick become outdated after the commute time of patients. In 2003, a Canadian lady brought her eighty-year old mother to the emergency department in Alberta, Canada. Although the announced waiting time of the emergency department was fifteen to eighteen minutes, they had to wait five hours until a doctor was available. Consequently, her mother passed away. If the delay announcements were more accurate, they could have decided to visit another less-congested hospital [88].

Motivated by health care systems, in Chapter 5 we consider a queueing network having a finite number of single-server queues in parallel. Each customer will be routed to one queue according to certain routing policies. Before actually arriving at the destined queue, each customer will experience a random commute delay. We plan to device an optimal routing policy with the objective of minimizing the waiting times at all queues. Parallel queues are usually less effective than queueing models with one waiting line and a pooled service resource (i.e., number of servers) [22]. When the commute delay is close to 0, it has been proven in [95] that it is optimal to route customers using the *join-the-shortest-queue* (JSQ) rule, which successfully synchronizes the sizes of waiting lines at all queues when the system is in heavy traffic. Unfortunately, our simulation experiments show that JSQ performs poorly when the commute delay increases. We improve the system performance by developing a probabilistic routing policy and prove that it is asymptotically optimal as the traffic intensity approaches 1. In addition, we conduct simulation experiments to provide engineering confirmations. See Chapter 5 for more details.

1.3 Heavy-Traffic Limits of Queueing Models

Heavy-traffic limits of queueing systems have been proven useful in approximating performance functions in service systems. To explain the general framework of heavy-traffic limits, we consider a sequence of queueing systems indexed by a scaler n, and show that, as $n \to \infty$, the system performance functions converge in appropriate functional spaces to their corresponding limits. There are two types of heavy-traffic regimes: (i) the conventional heavy-traffic (CHT) regime and (ii) the many-server heavy-traffic (MSHT) regime. In the CHT setting, the number of servers is usually

kept fixed while the system is asymptotically in heavy traffic because the traffic intensity goes to 1. In the MSHT case, the traffic intensity is usually kept fixed while the number of servers goes to infinity. In the rest of this section, we first review the general background of weak convergence and discuss the general approach to prove weak convergence results (§ 1.3.1). We next review CHT and MSHT in §§ **??** and 1.3.2.

1.3.1 Weak Convergence and Tightness

Heavy-traffic limits are often obtained through weak convergence, also called convergence in distributions. The most commonly used functional space in queueing theory is space $\mathbb{D}$, which is the space of all right-continuous functions with limits from the left. For instance, both the queue length process and the workload process can be defined as random elements in space $\mathbb{D}$. In general, weak convergence is defined for a sequence of probability measures in a separable metric space space $\mathbb{P}$. The weak convergence of a sequence of random elements X_n in $\mathbb{D}$ means the convergence of a sequence of probability measures induced by X_n, $\{P_n, n \geq 1\}$, to a limiting probability measure P, if for all continuous and bounded functions f,

$$\lim_{n\to\infty} \int f \mathrm{d}P_n = \int f \mathrm{d}P,$$

and we denote the weak convergence $P_n \Rightarrow P$. See [8, 113] for more details.

Prohorov [90] developed the Prohorov metric d_π to measure the distance of two probability measures. Specifically, for any $P_1, P_2 \in \mathbb{P}$, the Prohorov metric is defined by

$$d_\pi(P_1, P_2) \equiv \inf \{\varepsilon > 0 : P_1(B) \leq P_2(B^\varepsilon) + \varepsilon, \text{ for all } B \subset \mathbb{D}\},$$

where B^ε is the open neighborhood containing B with "$\varepsilon-$more radius". Mathematically, it is defined as $B^\varepsilon \equiv \{Y \in \mathbb{D} : m(X, Y) < \varepsilon, \text{ for some } X \in B\}$. Here $m(\cdot, \cdot)$ is the metric measuring the distance between stochastic processes in space $\mathbb{D}$, and we will specify it later. See more details in [113].

On the other hand, the convergence of Prohorov metric $d_\pi(P_n, P) \to 0$ implies the weak convergence $P_n \Rightarrow P$. However, when specifying weak convergence, instead of studying the probability measures, we often discuss random elements (such as random variables, etc.). In view of heavy-traffic limits in queueing theory, we are mainly interested in stochastic processes in space $\mathbb{D}$ and space $\mathbb{C}$, which is the space of continuous functions. Unlike the uniform metric for space $\mathbb{C}$, elements in space $\mathbb{D}$ require weaker metrics, such as J_1, J_2 and M_1 metric, see [113] for details.

There are mainly two methods to show weak convergence: one is the *continuous mapping theorem* (CMT) and the other is the compactness approach. We review the basic version of CMT: if $X_n \Rightarrow X$ in space $\mathbb{D}$, and a function $f : \mathbb{D} \to \mathbb{D}$ is continuous, then $Y_n = f(X_n) \Rightarrow Y = f(X)$. See Theorems

3.4.1-3.4.4 in [113] for extensions and more details of CMT.

Though convenient, the CMT approach sometimes is difficult to use because convenient continuous functions are hard to find. So the compactness approach is usually regarded as a standard approach to prove weak convergences. The compactness approach requires *relative compactness*. A sequence of probability measures is relative compactness if and only if their closure is compact. Prohorov's theorem [90] provided a way to relate relative compactness to *tightness*.

Definition 1.3.1 (see §1.5 in [8]). *A subset A of probability measures in space $\mathbb{P}$ is tight if and only if for any ε, there exists a compact subset K_ε in the corresponding funcational space such that $P(K_\varepsilon) > 1 - \varepsilon$ for all $P \in A$.*

The Arzelà-Ascoli theorem facilitates the developments of various tightness criterion for stochastic processes in space $\mathbb{C}$ with uniform topology, and Billingsley [8] extended those results for space $\mathbb{D}$. We refer interested readers to [8, 113] for details.

In Chapter 2 of this dissertation, we prove the tightness of properly scaled steady-state number for the idle servers in the $GI/GI/n/0$ loss model by using the definition of tightness shown in Definition 1.3.1.

1.3.2 Heavy-Traffic Regimes

Different heavy-traffic limits are obtained in different heavy-traffic regimes, which depicts how queueing parameters changes as the system scale n (the index) increases. There are two kinds of heavy-traffic regimes: convention heavy-traffic regime (CHT) and many-server heavy-traffic regime (MSHT).

CHT In CHT, the number of customers and the arrival rate λ is fixed but the traffic intensity (the ratio of the arrival rate to the service rate) approaches to 1. Kingman [61, 63] studied the conventional heavy-traffic limit for the sequence of $GI/GI/1$ queue. In the nth $GI/GI/1$, there is only one server and customers follows FCFS. Customer interarrive times are i.i.d. $A_1, A_2, \ldots$ with rate λ and *squared coefficient variation* (SCV) c_λ^2. A new arrival will enter the service if there is any server idle, otherwise she will join the end of waiting queue. Their service times are i.i.d. $\{S_{n,i}, i \geq 1\}$, having rate μ_n and SCV c_s^2. As $n \to \infty$, we have the traffic intensity $\rho_n \equiv \frac{\lambda}{\mu_n} \to 1$. When the traffic intensity ρ_n approaches to 1, the number of customers in $GI/GI/1$ increases to ∞, but the steady-state waiting time of the n^{th} queue, denoted as W_n, is weakly convergent: $(1 - \rho_n)W_n \Rightarrow W$, where W is an exponential random variable having rate $2\mu/(c_\lambda^2 + c_s^2)$ (see [63] for more details).

Borovhov [10] and Iglehart and Whitt [50, 51] extended the heavy-traffic limits to $GI/GI/s$ multi-server queueing systems. In particular, they scale the time in the nth queue and let the traffic

intensity approach 1 in the specify way:

$$\sqrt{n}(1-\rho_n) \to \beta, \quad \text{as } n \to \infty,$$

where β could be any positive real number. In this regime, they showed the properly scaled total number of customers is weakly convergent to a diffusion process, i.e. $n^{-1/2}X_n(t) \Rightarrow \hat{X}(t)$ in space $\mathbb{D}$, given the initial condition holds $n^{-1/2}X_n(0) \Rightarrow \hat{X}(0)$. Here $\hat{X}(\cdot)$ is a reflected Brownian motion with drift $-\beta\mu$ and diffusion coefficient $\mu(c_\lambda^2 + c_s^2)$.

MSHT Besides CHT, another widely used heavy-traffic regime is MSHT. In MSHT, the service rate is fixed but the arrival rate λ_n and the number of servers s_n increases as the system index $n \to \infty$ (see [48, 49] and Borovkov [11]). The way in which λ_n and s_n increase as $n \to \infty$ can be arbitrary, and the most basic MSHT is to let them change proportionally, i.e., $s_n = n, \lambda_n = n\lambda$. However, in this setting, the traffic intensity $\rho_n = \frac{\lambda}{\mu} < 1$, so some important performance metric, such as probability of delay, is 0. But in real practice, the probability of delay is an important metric measuring the quality of service and is not 0. Hence a more delicate MSHT is required. Halfin and Whitt in their seminal work [38] developed a new MSHT regime for the $M/M/n$ Erlang-C model and extend it to the $GI/M/n$ model. The regime is later referred to as the *quality-efficiency-driven* (QED) regime or Halfin-Whitt regime. In Halfin-Whitt regime, the traffic intensity ρ_n approaches to 1 in a specific way as $n \to \infty$:

$$\sqrt{n}(1-\rho_n) \to \beta, \tag{1.1}$$

where the constant $\beta > 0$ is called the quality-of-server (QoS) coefficient. Halfin and Whitt [38] showed that the properly scaled number of waiting customers is weakly convergent to a diffusion process with piecewise drift term $m(x) = -\mu\beta\mathbf{1}_{\{x\geq 0\}} - \mu(x+\beta)\mathbf{1}_{\{x<0\}}$ and volatility 2μ. In Halfin-Whitt regime, the steady-state probability of delay is convergent to a nondegenerate probability $\alpha(\beta) \in (0,1)$ and it is dependent on the QoS coefficient β. Puhalskii and Reiman [91] extended Halfin-Whitt regime to $GI/PH/n$ and Jelenkovic et al [53] studied $GI/D/n$. Reed [93] and Puhalskii and Reed [92] later generalized to queues having generally distributed service times, i.e., $GI/GI/n$.

Furthermore, Garnett el all. discussed the heavy-traffic limits in Halfin-Whitt regime for queues with customer abandonment ($M/M/n+M$) in [32]. In the presence of abandonment (the $+M$), the queueing model is stable when the traffic intensity $\rho_n \geq 1$, thus the QoS coefficient β can be negative. Mandelbaum and Zeltyn [84] later introduced Halfin-Whitt regime for the $M/M/n+GI$ model, having generally distributed abandonment times. Mandelbaum and Momcilovic [82] proposed diffusion approximations for the *virtual waiting time* (VWT) process and queue length process for $GI/GI/n+GI$ queues in Halfin-Whitt regime. Then in [19], Dai et al. studied $GI/PH/n+GI$ model by extending results in Puhalskii and Reiman [91] and Whitt [115]. Later, Dai and He [20]

discovered the deterministic relationship between scaled abandonment process and queue length process in Halfin-Whitt regime. Kang and Ramanan [55] studied the functional law of large number (fluid) limits for $GI/GI/n$ queue by providing a measure-valued process recording the elapsed service time for each server at present and then Kaspi and Ramana [56] extended theirs results to $GI/GI/n+GI$ queues which allows customer abandonment. Meanwhile, Zhang [124] also justified FWLLN limits for $GI/GI/n+GI$ queues which proposed as fluid models in [117]. Later Kaspi and Ramanan [57] proved a functional central limit theorem of state process for $GI/GI/n$ queues in terms of elapsed service time process as well as waiting time process.

Recently, Atar [5] studied a another heavy-traffic regime called *nondegenerate-slowdown* regime (NDS) which is "in the middle of" the most basic MSHT and Halfin-Whitt regime. Here "nondegenerate-slowdown" means the ratio between customer's delay and her service time, named as slowdown, is not trivial but in the order of $O(1)$. In the NDS regime, the joint limits of the sojourn time and service time processes were shown to be an independent reflected Ornstein-Uhlenbeck process and a "white-noise" (zero-mean) process with a size-biased mixture of exponentials as the marginal. Ward and Glynn [102, 101] and Huang et al. [41] reviewed and explored unified FCLT limits and approximations combining conventional heavy-traffic and Halfin-Whitt regimes.

1.3.3 Heavy-Traffic Limits and Gaussian Approximations

Different heavy-traffic limits are obtained in different heavy-traffic regimes. For a sequence of queueing models in a heavy-traffic regimes, the performance measure, say the queue length Q_n, is indexed by system scaler n. The heavy-traffic limits are the weak convergence limits of performance measures under different scaling. We first define the LLN-scaled performance measure $\bar{Q}_n$ and the corresponding FWLLN limit Q.

$$\bar{Q}_n \equiv \frac{Q_n}{n} \Rightarrow Q,$$

where Q is deterministic. Then we define the CLT-scaled performance measures $\hat{Q}_n$ and its FCLT limit $\hat{Q}$.

$$\hat{Q}_n \equiv \frac{Q_n - nQ}{\sqrt{n}} \Rightarrow \hat{Q},$$

where $\hat{Q}$ is stochastic. Then, the heavy-traffic limits provides the approximations to various performance measure of a queueing system. For example,

$$Q_n(t) \approx nQ(t) + \sqrt{n}\hat{Q}(t) + o(\sqrt{n}). \tag{1.2}$$

In this dissertation, we aim to develop effective engineering approximations for key performance measures of the stationary $GI/GI/n+GI$ and the time-varying $G_t/GI/s_t+GI$ queues across a wide range of model parameters. Our approximations are based on MSHT limits established in Whitt

[117] and Liu and Whitt [73, 75, 74, 77]. Although the queueing models they considered have a time-varying feature, which is more general than stationary systems, the established heavy-traffic limits can be applied to stationary queueing models, by setting the time-varying model parameters to constants and letting time parameter $t \to \infty$. MSHT FWLLN limits are deterministic approximations that estimate the mean of a performance measure and FCLT limits provide stochastic refinements to characterize stochastic fluctuations around FWLLN limits. However, we find that the effectiveness of approximations deteriorates when the system switches between overloaded intervals (OL) and underloaded intervals (UL).

Based on those observations, in Chapter 3 we modified the MSHT approximations to *truncated-Guassian approximation* (TGA) so that good performance prevails in a wide range of parameters. We also extend TGA's even further to *truncated-Guassian approximation with GI service* (TGAG) so that they apply the $GI/GI/n + GI$ model having non-exponential service times. Extensive simulation works have been done to test the practical range of our approximations. In Chatper 4, we extend our techniques to time-varying $G_t/M/n + GI$ and more general $G_t/GI/n + GI$ queues, and conduct various experiments to test the robustness of approximations.

In Chapter 5, we prove that parallel single-server queue under the routing policy JSQ-REB (joining-shortest-queue with root-excess bias) coincides a pooled system in a heavy-traffic regime. Specifically, we can also approximate queue length dedicated system by the heavy-traffic limit of a $GI/GI/1$ pooled system, which has been shown to be a reflective Brownian motion [17, 113], namely

$$Q_{n,1}(t) \approx Q_{n,2}(t) \approx \frac{1}{2}\sqrt{n}RBM(t) + o(\sqrt{n}), \tag{1.3}$$

where $Q_{n,i}(t)$ is the queue length for the i^{th} single server in the parallel system and $RMB(t)$ is a reflective Brownian motion with negative drift $-\beta\mu$ and variabilities determined by interarrival times and service times.

1.4 Simulations

In this dissertation, an extensive amount of computer experiments simulating stochastic queueing systems have been conducted to evaluate the accuracy of our Gaussian approximations. All simulations are run in MATLAB.

1.4.1 Simulation Procedure

In this dissertation, we evaluated the accuracy of approximations for various steady-state queueing performance. In order to avoid biases caused by transient behavior starting from empty, we start the sampling after the queueing system approaches steady state in each run. For example, if T is total simulation budget, we collect the data from $0.5T$ and end it at $0.95T$. We cut off the interval

$[0.95T, T]$ in sampling in case arrival customers occurred in the final portion has no waiting time recorded yet.

For all performance measures that we consider, we in total run $R = 2,000$ runs, and in each run, we collect the data evenly along the time line $[0.5T, T]$. In particular, in the r^{th} simulation replication, $1 \leq r \leq R$, we periodically generate virtual arrivals at deterministic times $t_1, t_k, \ldots, t_N$ with $t_k \equiv k\Delta t$ and $\Delta t = 0.1$, $1 \leq k \leq N \equiv \lfloor 0.5T/\Delta t \rfloor$. The virtual customers arrive and abandon as if they are the real customers but they will not be removed from the queue if they abandon. In each run r, we sample the number of customers in queue (Q), the number customers in the service (B), and head-of-line waiting time (HWT) (W). In particular, we sample the continuous-time queue-length process and number of busy servers as well as the waiting time of the head-of-line customers at discrete time points $t_1, t_2, \ldots, t_{N_v}$, and record them as $Q_r(k)$, $B_r(k)$ and $W_r(k)$, while excluding the virtual arrivals. For potential waiting time, we collect the waiting time of virtual customers since they will not leave the system even though they are noted as abandoned, denoted as $V_r(k)$. For probabilities of delay and abandonment, we sample indicator variables, $\eta_r^d(k)$ and $\eta_r^a(k)$, and record 1 if the k^{th} virtual customer is delayed or abandoned and 0 otherwise.

1.4.2 Experiments Design

A wide range of experiments is conducted to analyze the efficiency of approximations under different conditions such as abandonment rate, the number of servers (or system scales), and traffic intensity.

Customer abandonment plays a crucial role in deciding performance of stationary $GI/GI/n + GI$ (and time-varying $G_t/GI/s_t + GI$). As stated before, when abandonment rate is approaching 0, the model becomes a $GI/GI/n$ ($G_t/GI/s_t$); when the rate goes to ∞, the model becomes $GI/GI/n/0$ ($G_t/GI/s_t/0$), since all delayed customers will abandon immediately. Hence we evaluate our approximations in a wide range of the abandonment rate.

In addition, all the approximations are based on MSHT limits so they should work well when the system scale is large, i.e. when the number of servers is large. However, we tested the effectiveness of our approximations in smaller systems by decreasing the number of servers from 100 to 20, 10, 5 and even 3.

In Chapter 3 and Chapter 4 we tested new approximations in system with particular loadings for which the performance measures become difficult to evaluate. Furthermore, in Chapter 5, we examine the asympotic optimality of the JSQ-REB policy for general parallel single-server queues with pre-arrival delays.

1.5 Organization

In Chapter 2, we prove the tightness of diffusion scaled number of idle servers in $GI/GI/n/0$ loss model in Halfin-Whitt regime. In this regime, as mentioned before, the traffic intensity is scaled by the number of servers n as $\sqrt{n}(1-\rho_n) \to \beta > 0$, while the server rate is fixed at 1 with no loss of generality. Little literature studies the loss model in Halfin-Whitt regime because no tractable approach is handy to analyze general $GI/GI/n/0$ model. In this chapter, we first construct an auxiliary system through which the number of idle servers is converted to the waiting queue length in a single-server queue, then inspired by Garmarnik and Goldberg [31], we showed the tightness of the CLT-scaled number in waiting for the single-server queue, and thus the tightness for original loss model.

In Chapter 3, we develop refined MSHT approximations for various key performance measures such as the probability of delay, mean and variance of queue length and waiting time, etc. In this chapter, we firstly obtain the steady state version of the corresponding MSHT fluid and FCLT limits based on MSHT for time-varying queueing models [77]. Then, we develop Gaussian-based performance approximation formulas for the $GI/M/n+GI$ model. We later extend our formulas to approximate the performance of models having GI service times by incorporating the SCV of the service times. Extensive simulations are conducted to verify the accuracy of our approximations for a wide range of model parameters.

In Chapter 4, we extend our results in Chapter 3 and develop approximation formulas of the performance measures for the time-varying $G_t/M/n+GI$ and $G_t/GI/n+GI$ models, such as the mean and variance of queue length and waiting time, etc. Our time-varying formulas are developed by truncating the MSHT fluid and FCLT limits for the time-varying queueing models developed in [77]. We again conduct extensive simulations to verify the robustness of our time-varying approximation formulas for a wide range of system parameters.

In Chapter 5, we study a queueing network with a finite number of single-server queues in parallel. Each customer, before arriving at a queue, has to experience a random commute delay. We first show that the commonly used JSQ policy can be disastrous to system performance and cause high system congestion. Therefore, we proposed a probabilistic JSQ policy, JSQ-REB. The JSQ-REB policy combines the JSQ policy with probabilistic routing, where the routing probabilities are fine-tuned based on the system excess capacity. We proved the asymptotic optimality of JSQ-REB policy by showing that the CLT-scaled queue length process is asymptotically equivalent to that of the corresponding pooled system. Simulations are conducted to provide engineering confirmations.

1.6 Notations and Acronyms

Notations and acronym used in this dissertation are summarized in Tables 1.1 and **??** below.

Table 1.1: Summary of notations and acronyms

Notations and acronums	Meaning
a.s.	almost surely, or with probability one
$B_n(\cdot)$	the number of busy servers in queue
$\mathbb{C}$	space of continuous functions
ccdf.	complementary cumulative distribution function
cdf.	cumulative distribution function
CL	critically loaded, or traffic intensity $\rho_n \to 1$
CHT	conventional heavy-traffic limit
CLT	central limit theorem
CMT	continuous mapping theorem
$\mathbb{D}$	space of right-continuous functions with left limits
DGA	direct Gaussian approximation
EAP	exogenous arrival process
ED	efficiency driven
FCFS	first come first served
FCLT	functional central limit theorem
fdd.	finite-dimensional distribution
FWLLN	functional weak law of large numbers
HWT	head-of-line waiting time
i.i.d	independent and identically distributed
JSQ	jointing-shortest-queue
JSQ-REB	jointing-shortest-queue with a root-excess bais
LLN	law of large numbers
MSHT	many-server heavy-traffic
MMPP	Markov Modulated Poisson process
OL	overloaded interval
P_n^a	probability of abandonment
P_n^d	time-averaged probability of delay
$P_n^{d,C}$	customer-averaged probability of delay
pdf.	probability density function
PWT	potential waiting time

Acronyms	Meaning
$\mathcal{Q}_n$	the $G/G/n$ many-server queue
$\mathcal{Q}_n^l$	the $G/G/n/0$ loss model
$\mathcal{Q}_n^\infty$	the $G/G/n\infty$ infinity-server queue
$Q_n(\cdot)$	the number waiting customers in queue
$Q(\cdot)$	FWLLN limit of $Q_n(\cdot)$
$\hat{Q}(\cdot)$	FCLT limit of $Q_n(\cdot)$
QD	quality-driven
QED	quality-and-efficiency driven, i.e. Halfin-Whitt regime
ρ	traffic intensity
$\mathbb{R}$	real line
$\mathbb{R}_+$	positive real line
SCV	squared coefficient of variation
TGA	truncatd Gaussian approximation
TGAG	truncatd Gaussian approximation adapted to GI service
UL	underloaded interval
VWT	virtual waiting time
$V_n(\cdot)$	potential waiting time for queueing system
$W_n(\cdot)$	waiting time of the head-of-line customers in queue
$X_n(\cdot)$	total number of customers in the queueing model
$\to$	almost surely convergence
$\Rightarrow$	weak convergence, or convergence in distribution

Chapter 2

Tightness of Steady-State $GI/GI/n/0$ in Haflin-Whitt regime

2.1 Introduction

Markovian queueing models are mathematically tractable, but statistical analysis of real service systems reveals the service times and abandonment times are far from exponentially distributed, see Figure 1.1. However, the exact analysis for general queueing models is almost infeasible. Hence, heavy-traffic limits are developed to approximate various system performance measures. The crux to prove heavy-traffic limits is to show the weak convergence of system performance measures. The standard proof procedure is first to show prove the existence of convergent subsequences, and then to verify the uniqueness of limits for all convergent sequences. In order to show the existence of the convergent subsequences, it is suffice to show the tightness defined in Definition 1.3.1. In this paper, we aim to prove the tightness of the sequence of system performance measures.

Motivated by a theoretical work in [31] on steady-state queue length for multiserver queueing systems ($GI/GI/n$) in Haflin-Whitt regime, we aim to show the tightness of steady-state number of customers for loss models in the same regime. In [31], Gamarnik and Goldberg proved the tightness of steady-state queue length in the system and the corresponding large-deviation exponent. However, they did not study the tightness related to system idleness, namely the number of idle servers in the service pool. Later Goldberg [33] showed the tightness for idleness of the $GI/M/n$ models having exponential service time. To extend the tightness for more general queues, we firstly study the tightness of loss model. Then, according to the fact that the number of idle servers in the loss model serve as an upper bound of that in various queueing models such as $GI/GI/n$, $GI/GI/n+GI$, etc. Thus, our tightness result for the loss model is a stepping stone for results of more general queueing models.

Meanwhile, there is a growing interest in practice use of loss models per se. Loss models have

been widely applied in healthcare, due to the limited resources, to control the flow (and overflow) of patients requiring intensive care [71], to model ambulance deployment [96] and to determine the required number of beds, rooms and rental vehicles [21, 3] and [66].

In this chapter, we mainly study the tightness of the number of idle servers for the steady-state loss model denoted as $GI/GI/n/0$, which paves the way for the heavy-traffic limits of steady-state $GI/GI/n/0$ loss queueing models. There has been a large body of literature on the analysis of steady-state performance measures on loss queueing models since early twentieth century, most of which mainly focuses on the analysis of probability of blocking (PoB). Erlang [25] studied the fundamental steady-state $M/M/n/0$ loss model in which both interarrival times and service times are exponentially distributed. The author proposed Erlang-B function to calculate PoB. Then in § 5.7.2 in [97], the steady-state number of busy servers in the $M/GI/n/0$ loss model was shown to be insensitive with respect to the service-time distribution beyond its mean. Therefore, a generalization of the service-time distribution is redundant. Later progress on generalizing Erlang's [25] work was made in by [18] and [100], by conducting the theoretical analysis for $GI/M/n/0$ loss models.

Notwithstanding, analytical solutions for further generalized loss model are almost intractable, and thus heavy-traffic limits are developed to estimate the system performance of loss models. Borovkov, in [11] and [12], provided rigorous basis for approximations of $GI/M/n/0$ heavily loaded loss model such as Hayward's approximations in [70]. In [104], Whitt gave a mathematical proof of Hayward's approximation for $GI/M/n/0$ and further proposed effective approximations for $GI/GI/n/0$ models having non-exponential service times. Whitt's work [114], based on state-dependent analysis in the previous one [107], proposed FCLT approximations for the $GI/GI/n/m_n$ model, where m_n is the capacity of waiting room which is allowed to increase as system scale n grows. Later in [116], Whitt gave the heavy-traffic limit for $GI/H_2/n/m_n$ model having specific hyper-2 exponentially distributed service times. In both works, the waiting room capacity m_n is required to increases as n grows, so they do not apply to the $GI/GI/n/0$ loss model. Recently, Li and Whitt [69] directly dealt with approximation techniques for the $GI/G/n/0$ loss model, by relaxing the independence assumption among service times. Their approach heavily draws on the so-called *peakedness*, that is the ratio of the variance to the mean of the steady-state number of the corresponding infinite-server model. However, the heavy-traffic limit for the loss model is still yet to established.

We focus on the loss model in Haflin-Whitt regime. Our result can help pave the way for future work on FCLT limits. The Halfin-Whitt regime, as defined in § 1.3.2, though originally analyzed in [26] and [52], and formally studied in [38]. In the work of [38], the regime for multiple parallel servers with Markovian service times denoted $GI/M/n$, is realized by scaling both the system scale n and the traffic intensity ρ_n in the way defined in (1.1). We can use that the traffic intensity ρ_n approaches 1 when n increases. In Haflin-Whitt regime, the probability of delay (PoD) is converging to a non-trivial function of β. The more general formula of the PoD limit was found in [116], extending the

regime to queues with a mixture of exponential distributions with a mass point at 0. Mandelbaum et al. [32] and [123] studied queues in Haflin-Whitt regime with customers abandonment, in which the limiting PoD is a function of $\beta \in \mathbb{R}$ as well as the ratio between abandonment rate and processing rate.

Our contributions are twofold. First, we show that the CLT-scaled steady-state number of idle servers for the $GI/GI/n/0$ loss model in Halfin-Whitt regime is tight, which is not proved before. But we point out that it remains an open problem to obtain non-trivial heavy-traffic limits, and the problem is beyond the scope of this dissertation. Another contribution is the auxiliary system we construct to prove the tightness. In the auxiliary system, the number of idle servers in the loss model equal to the number of waiting customers in the single-server queue. Such a feature enables us to study more queueing models such as $GI/GI/n+GI$ queues with impatient customers and queueing networks.

The rest of this chapter is organized as follows. We start § 2.2 to prove the tightness in the loss model. We converted the problem of the loss model to a problem for a single-server queue in § 2.2.1, by constructing a sequence of auxiliary systems, and we prove the tightness of for the single-server queue in § 2.2.2. We conclude this chapter in § 2.3.

2.2 Tightness for $GI/GI/n/0$

To describe a $GI/GI/n/0$ loss model, we first assume two independent random variables U's and V's having unit mean and bounded $(2+\varepsilon)$th moment, for some $\varepsilon > 0$. That is, we assume that $\mathrm{E}[A] = \mathrm{E}[S] = 1$ and $\mathrm{E}[A^{2+\varepsilon}] < \infty$, $\mathrm{E}[S^{2+\varepsilon}] < \infty$. We then consider a sequence of FCFS $GI/GI/n/0$ models under Halfin-Whitt regime, denoted as $\{\mathcal{Q}_1^l, \mathcal{Q}_2^l, \dots\}$. In the model $\mathcal{Q}_n^l$, interarrival times are distributed as U/λ_n, and service times are distributed as V, where the λ_n indicates interarrival rate and we will specify it later. We define the total number of customers in $\mathcal{Q}_n^l$ as $X_n^l(t)$. Initially, we assume there are n customers in the system being served; each has the residual service time distributed as $R(V)$ with ccdf.:

$$\mathbb{P}\left(R(V) > x\right) \equiv \frac{\int_x^\infty \mathbb{P}\left(V > y\right) \mathrm{d}y}{\mathrm{E}[V]} = \int_x^\infty \mathbb{P}\left(V > y\right) \mathrm{d}y. \tag{2.1}$$

In Haflin-Whitt regime, we fix that the Halfin-Whitt scalar at $\beta > 0$ and let the arrival rate $\lambda_n = n - \beta\sqrt{n} + o(\sqrt{n})$. We say $f(x) = o(x)$ if $f(x)/x \to 0$ as x increases. For convenience, we assign $\lambda_n = n - \beta\sqrt{n}$ in the remaining of this chapter.

Our main result is obtained under Assumption T_0 proposed in [31], and we adopt it here.

Assumption 1. *We assume that*

(a) both A_n and S_n are non-trivial r.v.'s, that is, neither Var(A) nor Var(S) is zero, and

(b) $\limsup_{t\to 0} t^{-1}\left(F(t)-F(0)\right) < \infty$ holds where $F(\cdot)$ is the cumulative distribution function for A_n and S_n, and

(c) the steady state of loss model $\mathcal{Q}_n^l$ exists, namely, $X_n^l(\infty) \equiv \lim_{t\to\infty} X_n^l(t)$ exists.

Remark 2.2.1. *We explain the intuition of Assumption 1. Part (a) indicates that interarrival times and service times are not constant but randon variables. Part (b) means the pdf's of interarrival times and service times, if exists, is finite at the origin. Such an assumption is not restrictive because all continuous distributions and discrete distributions having zero mass at origin satisfy it. Part (c) guarantees the our discussions on steady state is meaningful, and the sufficient conditions for Part (c) are discussed in [4].*

In each loss model $\left\{\mathcal{Q}_n^l, n=1,2,\ldots\right\}$, we obtain the CLT-scaled steady-state number of idle servers

$$\hat{X}_n^l \equiv \frac{\left(n - X_n^l(\infty)\right)}{\sqrt{n}},$$

and $\hat{X}_n^l$ is a random variable in $\mathbb{R}$. Based on Assumption 1, we give the main result.

Theorem 2.2.1. *The sequence of random variables $\left\{\hat{X}_n^l, n=1,2,\ldots\right\}$ is tight in $(\mathbb{R}, \mathscr{B})$,*

where $\mathscr{B}$ is the Borel σ-algebra on $\mathbb{R}$.

The proof consists two steps: first, we construct an auxiliary system; in the system, the number of idle servers for a loss model is equal to waiting queue length for a single-server queue. Second, we show tightness of the queue length for the single-server queue, by adopting techniques developed by Gamarnik and Goldberg in [31]. We prove the tightness according to Definition 1.3.1. Here, the compact subset K is choosed to be a closed interval $[0,x]$ for some finite x.

2.2.1 Step I: An Auxiliary System

In this section, we construct an auxiliary system or each $\mathcal{Q}_n^l$, denoted as $\mathcal{Q}_n^{l,au}$. As Figure 2.1 depicts, the auxiliary system is a closed network having two stations. Such a tandem structure is initially inspired by the model used in [65]. The first station, called Station 1, is a loss model having m_n i.i.d. servers with service time distributed as S; the second station is a single-server queue with the service time distributed as A_n. We set the number of servers in Station 1 to be m_n satisfying

$$m_n = \left\lceil \lambda_n \left(1 - \beta n^{-1/2}\right) \right\rceil < \lambda_n \tag{2.2}$$

to make sure that Station 2 is stable and stays in Haflin-Whitt regime. We explain why m_n is set so later. The number of customers in Station i in $\mathcal{Q}_n^{l,au}$ is defined as $X_{n,i}(t)$ for any time $t \geq 0$, and

their CLT-scaled process are

$$\hat{X}_{n,1}(t) \equiv \frac{m_n - X_{n,1}(t)}{\sqrt{n}} \quad \text{and} \quad \hat{X}_{n,2}(t) \equiv \frac{X_{n,2}(t) - 1}{\sqrt{n}}.$$

The two stations shown in Figure 2.1 are fed by each other: departures from Station 1 (2) are new arrivals to Station 2 (1). Moreoever, arrivals to Station 1 might be blocked since the station does not have waiting buffer. To keep system closed, we directly send those blocked customers to Station 2. We will discuss the benefits of this feature furthermore.

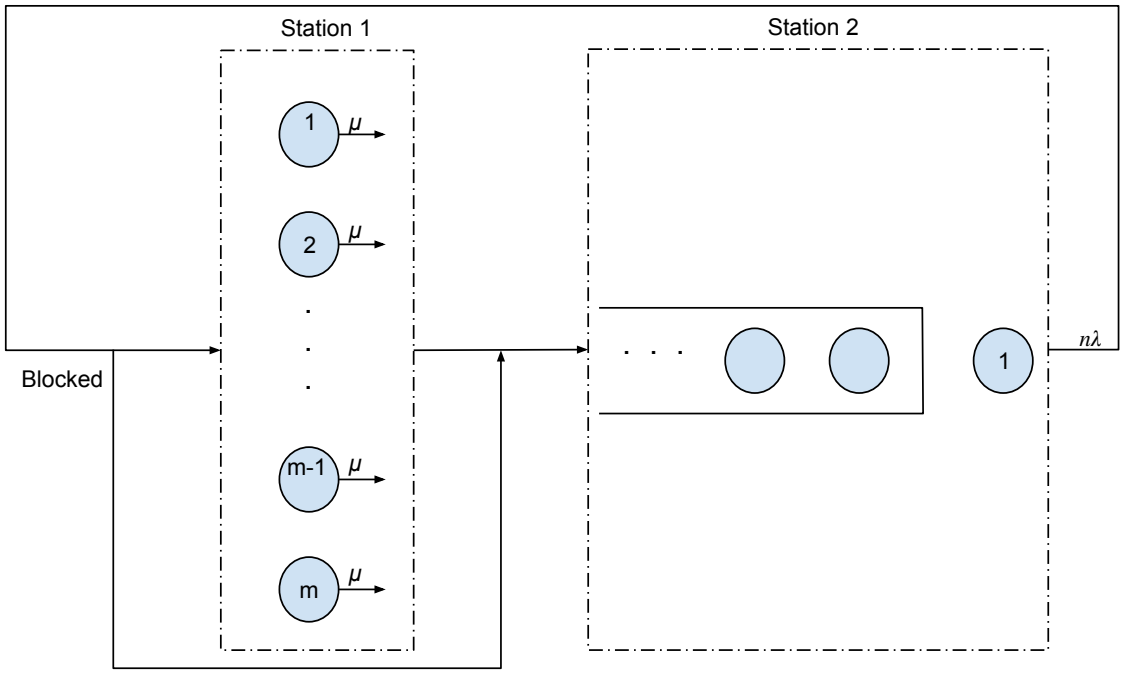


Figure 2.1: An auxiliary system

Initial Condition: In the nth auxiliary system, we assume there are (m_n+1) customers in the auxiliary system: m_n of them are in Station 1 and 1 customer is in Station 2. Customers in Stations 1 have i.i.d. residual service times distributed as $R(S)$, the same as those in $\mathcal{Q}_n^l$, and customers in Station 2 has residual service time distributed as $R(A_n)$. Both $R(S)$ and $R(A_n)$ are defined in (2.1). With such an initialization, the auxiliary system depicted in Figure 2.1 has three properties as follows.

(i) The total number of customers in the auxiliary system's is fixed at $m_n + 1$, i.e. $X_{n,1}(t) +$

$X_{n,2}(t) = m_n + 1$, for any $t \geq 0$.

(ii) Station 2 is always busy, i.e. $X_{n,2}(t) \geq 1$ for any $t \geq 0$.

(iii) The number of customers in Station 1 bounds that of the loss model $\mathcal{Q}_n^l$, i.e.

$$X_n^l(t) - (n - m_n) \leq X_{n,1}(t) \leq X_n^l(t), \tag{2.3}$$

for any $t \geq 0$.

Property (i) because the auxiliary system is a closed network. The total number of customers in the system is fixed at $m_n + 1$. Now, after a little transformation, we find that $m_n - X_{n,1}(t) = X_{n,2}(t) - 1$, holding for any $t \geq 0$. That is to say, the number of idle servers in Station 1 equals the number of waiting customers (waiting queue length) in Station 2 for any time.

Property (ii) comes from Property (i) and the structure of the auxiliary system. Because of the fact that the total number is $m_n + 1$ but Station 1 allows at most m_n customers in it, there has to be at least 1 customer in the Station 2, namely Station 2 is always busy. Hence, the departure process of Station 2 is an equilibrium renewal process $A_n(t)$ associated with r.v. A_n, as the initial service time of Station 2 is the residual service time $R(A_n)$. Given that the departure process of Station 2 is the arrival process of Station 1, the first station shares the same arrival process with those of the loss model $\mathcal{Q}_n^l$ we built before.

Property (iii) comes from the comparison between Station 1 and loss model. For the loss model , We separate $\mathcal{Q}_n^l$ into two service pools. The first pool has m_n servers whereas the second pool has $(n - m_n)$ servers. Customers will preferably choose the first pool, and they are served in the second pool only if no servers idle in the first pool. In such a setting, the number of customers in the second pool is at most $(n - m_n)$, hence the number of customers is the first pool is no less than $\big(X_n^l(t) - (n - m_n)\big)$. Therefore, the relationships in (2.3) are shown.

Given these properties of the auxiliary system, we give Lemma 2.2.1, in which the number of idle servers for a loss model (Station 1) is bounded by the waiting queue length of a single-server queue (Station 2).

Lemma 2.2.1. *For any $t \geq 0$, we have*

$$\hat{X}_n(t) \leq \hat{X}_n^l(t) \leq \hat{X}_n(t) + 2\beta \tag{2.4}$$

for any time $t > 0$, where $n - X_n^l(t) \geq 0$ is the number of idle servers in $\mathcal{Q}_n^l$.

Proof. For writing convenience, we drop the time t and write $a \geq b$ to mean $a(t) \geq b(t)$ for any time $t \geq 0$.

$$X_{n,2} - 1 = m_n - X_{n,1} \leq n - X_n^l \leq n - X_{n,1} = X_{n,2} - 1 + (n - m_n), \tag{2.5}$$

Scaling both inequalities by $n^{-1/2}$ provides us (2.4). □

Based on Lemma 2.2.1, to show the tightness of CLT-scaled number of idle servers for a loss model, it is equivalent to show the tightness of queue length for Station 2.

2.2.2 Step II: Tightness of Queue Length for Station 2

In this section, we prove the tightness of number in waiting in Station 2, i.e., the tightness of the sequence $\left\{n^{-1/2}\left(X_{n,2}(\infty)-1\right), n=1,2,\ldots\right\}$. Here is the overview for the proof. We first bound total contents in Station 2 by that for a $\left(\sum_{i=1}^{m_n} G_i\right)/G/1$ queue, where $\sum_{i=1}^{m_n} G_i$ means the arrival process is a superposition of m_n equilibrium renewal processes. Secondly, we discretize the queue length expression for $\left(\sum_{i=1}^{m_n} G_i\right)/G/1$ by adopting Gamarnik and Goldberg's approach [31]. Then, based on the time discretization approach, we extend the techniques in Szczotka's proof in [99] to prove our goal.

Since Station 2 is always busy, we can regard it as the single-server queue with autonomous service, in which customers arrive according to the sum of m_n departure processes from Station 1, see [31] and [113].

Let the sequence of interarrival time and service time on the i^{th} server be $\left\{R\left(U_1^{(i)}\right), U_2^{(i)}, U_3^{(i)}, \ldots\right\}$ and $\left\{R\left(V_1^{(i)}\right), V_2^{(i)}, V_3^{(i)}, \ldots\right\}$, respectively. We denote the departure process on the i^{th} server for Station 1 as $D_i(t)$. Meanwhile, we construct an equilibrium renewal process $S_i(t)$ based on the sequence of service time $S_i(t) \equiv \max\left\{n \geq 0 : \sum_{k=1}^{n} V_k^{(i)} < t\right\}$.

Using a coupling technique, we show that $D_i(t) \leq S_i(t)$ for all $t \geq 0$. Figure 2.2 provides an intuitive comparison between $D_i(t)$ and $S_i(t)$. As long as some jobs' interarrival time is large enough to left server idle (blank areas), the wasted utilization of server will result in fewer number finished jobs (denoted as $D_i(t)$) than the counting process $S_i(t)$, for any time t after black areas appear. A rigorous proof for $\left(D_i(t) \leq S_i(t) \text{ for any } t \geq 0\right)$ is given below.

Rigorously, we prove $D_i(t) \leq S_i(t)$ for any $t \geq 0$ as follows.

Proof of Inequalify. According to the definition of couting process, we have

$$D_i(t) \equiv \max\left\{n \geq 0 : \left(\sum_{k=1}^{n} U_k^{(i)} \vee \sum_{k=1}^{n-1} V_k^{(i)}\right) + V_k^{(i)} < t\right\}, \tag{2.6}$$

$$S_i(t) \equiv \max\left\{n \geq 0 : \sum_{k=1}^{n} V_k^{(i)} < t\right\}. \tag{2.7}$$

Since $\left(\sum_{k=1}^{n} U_k^{(i)} \vee \sum_{k=1}^{n-1} V_k^{(i)}\right) \geq \sum_{k=1}^{n-1} V_k^{(i)}$, by definitions (2.6) and (2.7), we have $S_i(t) \geq D_i(t)$ for any time $t > 0$. □

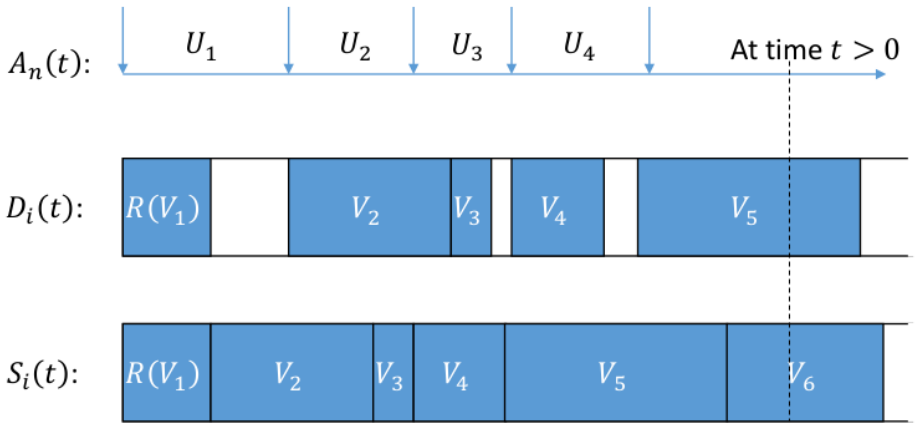


Figure 2.2: Comparison between $D_i(t)$ and $S_i(t)$ in terms of arrival times.

The queue length of Station 2 is given by that of a single-server queue with autonomous service.

$$X_{n,2}(\infty) - 1 = \sup_{t\geq 0}\left\{\sum_{i=1}^{m} D_i(t) - A_n(t)\right\} \leq \sup_{t\geq 0}\left\{\sum_{i=1}^{m} S_i(t) - A_n(t)\right\} \leq \sum_{i=1}^{m}\sup_{t\geq 0}\left\{S_i(t) - \frac{1}{m}A_n(t)\right\}, \tag{2.8}$$

where the first equality holds by mathematical induction (see Corollary 1 in [31]), the second inequality holds because that $D_i(t) \leq S_i(t)$ for all $t \geq 0$, and the third inequality holds by the fact that $\sup\{a+b\} \leq \sup\{a\} + \sup\{b\}$ for any $a, b \in [-\infty, \infty]$.

Gamarnik and Goldberg [31] developed an innovative method to discretize the right-hand-side term in (2.8) so that it can be further related to the waiting time of $GI/GI/1$ queue. Then we prove the equivalent problem by generalizing the proof techniques given in [99]. Their discretization approach is as follows: for each $i = 1, 2, \ldots, m_n$, the function $\sup\{S_i(t) - m^{-1}A_n(t)\}$ increases at the time when $S_i(t)$ jumps, namely at k_i^{th} renewal time $\sum_{l=1}^{k_i} V_l^{(i)}$ for all $k_i \in \mathbb{N}$. As $\sup\{\cdot\}$ is a continuous function, the right-hand-side term can be expressed as

$$\sum_{i=1}^{m}\sup_{t\geq 0}\left\{S_i(t) - \frac{1}{m}A_n(t)\right\} = \sum_{i=1}^{m}\sup_{k_i\geq 0}\left\{k_i - \frac{1}{m}A_n\left(\sum_{l=1}^{k_i} V_l^{(i)}\right)\right\}.$$

Hence, according to Definition 1.3.1, to show the sequence of random variables $\left\{\hat{X}_{n,2}(\infty), n \geq 1\right\}$ is tight, it is suffice to show that for any $\eta > 0$, there exists a compact subset $K_\eta = [0, x_\eta]$ such that

$$\mathbb{P}\left(\hat{X}_{n,2}(\infty) \in K_\eta\right) > 1 - \eta, \quad \text{for sufficient large } n.$$

In other words, we need to show that, for any $\eta > 0$ there exist $x > 0$ and an integer $n_0 \in \mathbb{N}$ such

that.

$$\mathbb{P}\left(n^{-1/2}\sum_{i=1}^{m}\sup_{k_i\geq 0}\left\{k_i - m_n^{-1}A_n\left(\sum_{l=1}^{k_i}V_l^{(i)}\right)\right\} > x\right) < \eta, \quad \text{for any } n \geq n_0.$$

For convenience, we define

$$\zeta_{n,k} = 1 - \frac{1}{m_n}\left(A_n\left(V_k + \sum_{l=1}^{k-1}V_l\right) - A_n\left(\sum_{l=1}^{k-1}V_l\right)\right), \tag{2.9}$$

where $\zeta_{n,k}$'s are i.i.d. with each other. By the inequality

$$\mathbb{P}\left(\sum_{i=1}^{m_n}X_i > x\right) \leq \sum_{i=1}^{m_n}\mathbb{P}\left(X_i > \frac{x}{m_n}\right), \tag{2.10}$$

we can prove Inequality (2.9) by showing

$$\mathbb{P}\left(n^{-1/2}\sup_{k\geq 0}\left\{\sum_{l=1}^{k}\zeta_{n,k}\right\} > x\right) < \frac{\eta}{m_n}. \tag{2.11}$$

To prove Inequality 2.11, we extend the proofs of Lemma 3-5, Theorem 5 in [31] and Theorem 1 in [99]. We focus on justify the condition that guarantees (2.11) according to [31].

Proof for (2.11). We begin the proof with three preliminary Lemmas. Lemmas 2.2.2-2.2.4 below all come from Gamarnik and Goldberg [31]. And we refer readers who are interested in proofs to Theorem 10.2 in [8], [119] (as they mentioned) and [31].

Lemma 2.2.2 (Lemma 3 in [31]). *Suppose $k < \infty$, $X_1, \ldots, X_k$ is a sequence of general random variables, $S_j = \sum_{i=1}^{j} X_i$ and $M_k = \max_{j\leq k}|S_j|$. Further suppose that there exist real numbers $r > 2$ and $\alpha > 0$, and a sequence of non-negative numbers $u_1, \ldots, u_k$ s.t. for all $0 \leq i \leq j \leq k$ and $x > 0$,*

$$\mathbb{P}\left(|S_j - S_i| \geq x\right) \leq x^{-4\alpha}\left(\sum_{i<l\leq j}u_l\right)^{\frac{r}{2}}.$$

Then there exists a finite constant $K_{\alpha,\beta}$, depending only on α and β, s.t. for all $x > 0$,

$$\mathbb{P}\left(M_k \geq x\right) \leq K_{\alpha,\beta}x^{-4\alpha}\left(\sum_{0<l\leq k}u_l\right)^{\frac{r}{2}}.$$

Lemma 2.2.3 (Lemma 4 in [31]). *For all $r \geq 2$, there exists $C_r < \infty$ (depending only on r) s.t. for all r.v. X satisfying $E[X] = 0$ and $E[|X|^r]$, if $\{X_i, i \geq 1\}$ is a sequence of i.i.d. random variables*

distributed as X, then for all $k \geq 1$,

$$E\left[\left|\sum_{i=1}^{k} X_i\right|^r\right] \leq C_r k^{\frac{r}{2}} E\left[|X|^r\right] \tag{2.12}$$

Lemma 2.2.4 (Lemma 5 in [31])**.** *Let X denote any non-negative random variable s.t. $E[X] = \mu^{-1} \in (0, \infty)$, and $E[X^r] < \infty$ for some $r \geq 2$. Let $\{Z_i^e(t), i \geq 1\}$ denote a set of i.i.d. equilibrium renewal processes with renewal distribution X. Then there exists $C_{X,r} < \infty$ (depending only on X and r) s.t. for all $n \geq 1$ and $t \geq 0$,*

$$E\left[\left|\sum_{i=1}^{n} Z_i^e(t) - \mu n t\right|\right] \leq C_{X,r}\left(1 + (nt)^{\frac{r}{2}}\right).$$

Now we begin to show (2.11). First, we notice that mean of $\zeta_{n,k}$ is negative for any finite k and the mean approaches to zero as n increases, because

$$\mathrm{E}\left[\zeta_{n,k}\right] = 1 - \frac{n - \beta\sqrt{n} + o(\sqrt{n})}{n - 2\beta\sqrt{n}} < 0, \quad \text{and} \quad \lim_{n\to\infty} \mathrm{E}\left[\zeta_{n,k}\right] = \lim_{n\to\infty} -\frac{\beta + o(1)}{\sqrt{n} - 2\beta} = 0. \tag{2.13}$$

Then we show (2.11) as follows.

$$\begin{aligned}
&\mathbb{P}\left(n^{-1/2}\sup_{k\geq 0}\sum_{l=1}^{k}\zeta_{n,l} > \frac{x}{m}\right) \leq \mathbb{P}\left(\sum_{l=1}^{k}\zeta_{n,l} > \frac{\sqrt{n}}{m}x, \text{ for some } k\right) \\
&= \mathbb{P}\left(\sum_{l=1}^{k}(\zeta_{n,l} - \mathrm{E}[\zeta_{n,l}]) > \frac{\sqrt{n}}{m}x + \frac{k\beta}{\sqrt{n} - 2\beta}, \text{ for some } k\right) \\
&\leq \sum_{j=0}^{\infty}\mathbb{P}\left(\sum_{l=1}^{k}(\zeta_{n,l} - \mathrm{E}[\zeta_{n,l}]) > \frac{\sqrt{n}}{m}x + \frac{2^j\beta}{\sqrt{n} - 2\beta}, \text{ for some } k \in [2^j, 2^{j+1}]\right) \\
&\leq \sum_{j=0}^{\infty}\mathbb{P}\left(\sum_{l=1}^{2^j}(\zeta_{n,l} - \mathrm{E}[\zeta_{n,l}]) + \max_{2^j\leq k\leq 2^{j+1}}\sum_{l=2^j}^{2^{j+1}}(\zeta_{n,l} - \mathrm{E}[\zeta_{n,l}]) > \frac{\sqrt{n}}{m}x + \frac{2^j\beta}{\sqrt{n} - 2\beta}\right) \\
&\leq \sum_{j=0}^{\infty}\mathbb{P}\left(\sum_{l=1}^{2^j}(\zeta_{n,l} - \mathrm{E}[\zeta_{n,l}]) > \frac{1}{2}\tilde{x}_{m,j}\right) + \sum_{j=0}^{\infty}\mathbb{P}\left(\max_{1\leq k\leq 2^j}\sum_{l=1}^{k}(\zeta_{n,l} - \mathrm{E}[\zeta_{n,l}]) > \frac{1}{2}\tilde{x}_{m,j}\right),
\end{aligned} \tag{2.14}$$

where the first inequality holds because that $\sum_{l=1}^{k}\zeta_{n,l} > 0$ happens only for some finite k due to (2.13), and the last inequality holds by (2.10). Here we define

$$\tilde{x}_{m,j} \equiv \frac{1}{2}\cdot\left(\frac{\sqrt{n}}{m}x + \frac{2^j\beta}{\sqrt{n} - 2\beta}\right). \tag{2.15}$$

For further convenience, we define $\bar{\zeta}_{n,l} = \zeta_{n,l} - \mathrm{E}[\zeta_{n,l}]$ and rewrite (2.14) as

$$\mathbb{P}\left(n^{-1/2}\sup_{k\geq 0}\sum_{l=1}^{k}\zeta_{n,l} > \frac{x}{m}\right) \leq \sum_{j=0}^{\infty}\mathbb{P}\left(\sum_{l=1}^{2^j}\bar{\zeta}_{n,l} > \tilde{x}_{m,j}\right) + \sum_{j=0}^{\infty}\mathbb{P}\left(\max_{1\leq k\leq 2^j}\sum_{l=1}^{k}\bar{\zeta}_{n,l} > \tilde{x}_{m,j}\right). \quad (2.16)$$

According to Markov inequality, Lemma 2.2.3 and Lemma 2.2.4, both terms at the right-hand-side are bounded as

$$\mathbb{P}\left(\sum_{l=1}^{2^j}\bar{\zeta}_{n,l} > \tilde{x}_{m,j}\right) \leq \frac{\mathrm{E}\left|\sum_{l=1}^{2^j}\bar{\zeta}_{n,l}\right|^r}{\tilde{x}_{m,j}^r} \leq C_{r,1}\left(\sqrt{2^j}\right)^r \mathrm{E}\left|\bar{\zeta}\right|^r \frac{1}{\tilde{x}_{m,j}^r}, \quad (2.17)$$

$$\mathbb{P}\left(\left|\sum_{l=i+1}^{j}\bar{\zeta}_{n,l}\right| > \tilde{x}_{m,j}^r\right) \leq \frac{\mathrm{E}\left|\sum_{l=i+1}^{j}\bar{\zeta}_{n,l}\right|}{\tilde{x}_{m,j}^r} \leq C_{r,2}\left(\sqrt{j-i}\right)^r \mathrm{E}\left|\bar{\zeta}\right|^r \frac{1}{\tilde{x}_{m,j}^r}. \quad (2.18)$$

Combining 2.18 and Lemma 2.2.2 and letting $\alpha = 1$ in Lemma 2.2.2, we have

$$\mathbb{P}\left(\max_{1\leq k\leq 2^j}\sum_{l=1}^{j}\bar{\zeta}_{n,l} > \tilde{x}_{m,j}^r\right) \leq K\left(\sqrt{2^j}\right)^r \mathrm{E}\left|\bar{\zeta}\right|^r \frac{1}{\tilde{x}_{m,j}^r}. \quad (2.19)$$

Define $M_r \equiv (C_{r,1} + K)\mathrm{E}|\bar{\zeta}|^r < \infty$, $r = 2 + \epsilon$ where $\epsilon > 0$, and plug (2.15), (2.17) and (2.19) into (2.16). Then we will have

$$\begin{aligned}
&\mathbb{P}\left(n^{-1/2}\sup_{k\geq 0}\sum_{l=1}^{k}X_l > \frac{x}{m}\right) \leq \sum_{j=1}^{\infty}\frac{M_r}{2}\cdot\frac{\sqrt{2^j}}{\tilde{x}_{m,j}^r} = \frac{M_r}{2}\sum_{j=1}^{\infty}\left(\frac{\sqrt{n}}{m\sqrt{2^j}}x + \frac{\sqrt{2^j}\beta}{\sqrt{n}-2\beta}\right)^{-(2+\varepsilon)}, \\
&\leq \frac{M_r}{2}\left(\sum_{j=0}^{n}\left(\frac{\sqrt{n}}{m\sqrt{2^j}}x\right)^{-(2+\varepsilon)} + \sum_{j=n+1}^{\infty}\left(\frac{\sqrt{2^j}\beta}{\sqrt{n}-2\beta}\right)^{-(2+\varepsilon)}\right) \\
&= \frac{M_r}{2}\left(\left(\frac{\sqrt{n}x}{m}\right)^{-(2+\varepsilon)}\sum_{j=0}^{n}2^{-(1+\varepsilon/2)j} + \left(\frac{\sqrt{2^{n+1}}\beta}{\sqrt{n}-2\beta}\right)^{-(2+\varepsilon)}\sum_{j=0}^{\infty}2^{-(1+\varepsilon/2)j}\right) \\
&\leq \frac{M_r}{2}\cdot\left[\left(\frac{m\sqrt{2^n}}{\sqrt{n}}x^{-1}\right)^{2+\varepsilon} + \left(\frac{\sqrt{n}/\beta - 2}{\sqrt{2^n}}\right)^{2+\varepsilon}\right]. \quad (2.20)
\end{aligned}$$

For any $\eta > 0$, we can choose n_0 satisfying $\left(\frac{\sqrt{n_0}/\beta-2}{\sqrt{2^{n_0}}}\right)^{2+\varepsilon} < \eta/(mM_r)$, and for any $n \geq n_0$, we can choose $x \geq m\sqrt{\frac{2^{n_0}}{n_0}}\cdot\sqrt[2+\varepsilon]{mM_r/\eta}$ so that the right hand side of (2.20) is bounded by η/m. □

The proved Inequality (2.11) imply that the sequence of $\left\{\hat{X}_{n,2}(\infty), n = 1, 2, \ldots\right\}$ is tight, which further implies that $\left\{\hat{X}_n^l(\infty), n = 1, 2, \ldots\right\}$ is tight by Lemma 2.2.1. This concludes the proof of

Theorem 2.2.1.

2.3 Conclusion

In this chapter, we studied tightness of steady-state CLT-scaled the number of idle servers for $GI/GI/n/0$ models in Haflin-Whitt regime. To prove the tightness for the loss model, we established an auxiliary system through which the number of idle server for the loss queue equals the number of waiting customers for a single-server queue. Then we show the tightness for that single-server queue by extending innovative proof methods proposed in [31] and [99]. The auxiliary system could be helpful on solving the tightness for other queueing systems such as $GI/GI/n + GI$ and other problems in queue networks.

Chapter 3

Approximations for Heavily-Loaded $G/GI/n+GI$ Queues

3.1 Introduction

In this chapter we develop and evaluate new engineering approximation formulas for heavily-loaded non-Markovian queueing systems with customer abandonment. Models that account for customer abandonment from queue due to customer impatience have generated substantial interest in recent years because of their application to call centers and other service systems. This has led to renewed interest in the Markovian $M/M/n+M$ Erlang-A model [32, 81, 115]. However, data analysis from service systems has also shown that the distributions of the service and patience times are often not nearly exponential [15]. Data analysis also has suggested that in some cases the arrival process might not be well modeled by a Poisson process; see [6, 44, 54, 59, 125].

Thus, we focus on the stationary $G/GI/n+GI$ queueing system, allowing a non-Poisson (and even non-renewal) stationary arrival process (the G), independent and identically distributed (i.i.d.) service times with a general distribution (the first GI), multiple (n) servers working in parallel, unlimited waiting space, customer abandonment according to i.i.d. patience times with a general distribution (the $+GI$) and the FCFS service discipline. We provide simple formulas to approximate the mean, variance and distribution of important steady-state performance measures, including the number of customers in the system, the number in queue and the waiting time (which we take to be the potential waiting time, i.e., the time a potential arrival at time t that is infinitely patient would have to wait before starting service). We also give formulas to approximate important steady-state probabilities, including probability of abandonment (P^a), the time-averaged probability of delay (P^d) and customer-averaged probability of delay ($P^{d,C}$). It is worthy to mention that time-averaged probability of delay equal to customer-averaged probability only in queues with Markovian arrivals, and such a feature is called *Poisson Arrivals See Time Average* (PASTA). In general queueing

systems having GI arrivals, they are not identical but the difference is little.

The stationary $G/GI/n + GI$ queueing model has several model elements. Even the basic $M/M/n+M$ Erlang-A model has four parameters: the arrival rate λ, the individual service rate μ, the number of servers n and the abandonment rate θ. Without loss of generality (by choosing units to measure time), we can let $\mu = 1$ so that there is the parameter triple (λ, n, θ) or equivalently (ρ, n, θ), with $\rho \equiv \lambda/n\mu \equiv \lambda/n$ being the traffic intensity. In addition, the stationary $G/GI/n+GI$ queueing model has a general service-time cdf G with mean μ^{-1}, a general patience-time cdf F with mean θ^{-1} and a general stationary arrival process with rate λ, allowing complex dependence among interarrival times. We develop performance approximations for that general $G/GI/n + GI$ model and we conduct extensive experiments to study how all its model elements affect (i) the performance of the system and (ii) the accuracy of the proposed approximations. We summarize our conclusions in §3.10.

3.1.1 A Basis in Many-Server Heavy-Traffic Limits

We primarily base these new engineering approximations on the MSHT FCLT for the time-varying non-Markovian $G_t/M/n_t + GI$ queueing model in [77], with Gaussian process limits, but we also draw on the MSHT FWLLN for the more general time-varying $G_t/GI/n_t + GI$ queueing model in [74, 75, 117], yielding deterministic fluid limits. Scaled fluid limits appear as centering terms in the FCLT and so provide the mean values of the resulting Gaussian approximations. These MSHT limits apply to the stationary $G/GI/n + GI$ model as a special case. In this chapter we investigate the quality of those approximations and develop heuristic refinements that are more effective.

For service systems, there is strong motivation for the time-varying feature, because service systems typically have arrival rates that vary strongly by the time of day. Nevertheless, what we do here still has significant relevance for service systems because stationary models often can be applied when the service times are relatively short, as in most call centers. With short service times, stationary models can be used in a nonstationary way via the pointwise stationary approximation [35, 105], as reviewed in [34]. Indeed, that is the common approach used to staff service systems in practice.

The MSHT limits indicate when the approximations should be effective. Because MSHT limits involve letting the arrival rate and number of servers grow without bound, we expect the approximations to be more effective with large scale. Hence, our base case has $n = 100$ servers, but we also find that the approximations are quite effective for smaller scale, e.g., $n = 20$, 10 and 5.

The MSHT limits for the time-varying non-Markovian $G_t/M/n_t + GI$ model in [77] are for systems with a continuous time-varying arrival-rate function that makes the system alternate between overloaded (OL) intervals, where we locally have $\rho > 1$, and underloaded (UL) intervals, where

we locally have $\rho < 1$, without being critically loaded (CL), where we locally have $\rho \approx 1$, over a positive interval. That limit yields one-sided approximations in each of the OL and UL intervls. Nevertheless, simulation experiments showed that the approximations are remarkably effective for difficult time-varying systems that are mostly not nearly critically loaded. However, for stationary models, these MSHT limits hold only for OL models with $\rho > 1$ or UL models with $\rho < 1$. Moreover, only the OL approximations are effective for the queue length and waiting time processes. For that reason, the approximations in this chapter are only for heavily loaded models. We do provide information about the corresponding (less useful) approximations for UL systems in §3.8 and in the appendix [78].

We investigate the direct application of the MSHT Gaussian process limit in [77], but we find that the direct approximations are ineffective when the traffic intensity ρ is near 1. Thus, a significant contribution here is to develop effective heuristic refinements. We conduct extensive simulation experiments investigating when these approximations are effective. We find that the effectiveness of the refined Gaussian approximations for heavily loaded models primarily depends on two parameters: the traffic intensity ρ and the abandonment rate θ. Assuming that the scale (which is characterized by the relevant number of servers n) is not too small, we find that the refined approximations are effective roughly for $\rho > 1.02$ and $\theta < 2.0$. Since our approximations tend to work better when the system is heavily loaded, the quality of the approximations tends to improve as ρ increases and as θ decreases. We also find that our approximations are relatively robust to the variability in the arrival process, service times and patience times, in a reasonable range.

Even though we focus on heavily loaded models with $\rho > 1$, as observed previously, e.g. [107, 117], these are practical cases that often occur in practice, because the abandonment always keeps the system stable. (For example, see the $M/M/n + M$ base case with $\rho = 1.05$ in Table 3.2. The steady-state mean number in system and abandonment probabilities are representative of what is often seen in practice.) For these cases, the proposed approximations not only cover general non-Markovian models, but the accuracy of the approximations and simplicity of the formulas makes the proposed approximations attractive alternatives even for the Markovian $M/M/n+M$ Erlang-A model, as in [32, 81]; e.g., for quick approximations of OL models it may not be necessary to solve any birth-and-death equations.

From the range of abandonment rates θ, it should be evident that we are only considering models with abandonment, and only at a typical level; we are not considering the $G/G/n/0$ loss model or the $G/G/n/\infty$ delay model, which already have been quite extensively studied, e.g., see [69, 114, 112] and references therein. While the proposed approximations should be useful, they are far from universal approximations; we are *not* claiming that the approximations apply to all multi-server queueing models. To appreciate the limitations, recall that the waiting time distribution in $M/M/n$ queue and $G/GI/n$ generalizations has an exponential tail [38], unlike the much more

rapidly decaying tails of our Gaussian approximations, which arise because of the abandonment.

The nice Gaussian approximations here can be understood as a generalization of exact results for the $M/M/n+M$ model that hold when $\theta = \mu$, where μ is the individual service rate. As discussed in §6 of [28], for that special parameter choice, the number in system has the same structure as in the associated $M/M/\infty$ infinite-server model, so that the steady-state distribution is exactly Poisson, which is approximately Gaussian provided that the arrival rate is not too small. Limit theorems supporting such Gaussian approximations for the $M/M/\infty$ model go back to [48] and there has been much work since then. The limits in [77] can be viewed as generalizations.

3.1.2 Overview of the Proposed Approach

We now give a brief overview of our approach. We start by applying the MSHT limits to generate a Gaussian approximation for the number of customers in the system. Unfolding the limit in the usual way, we obtain a direct approximation for the total number of customers in the system at time t, $X_n(t)$, as

$$X_n(t) \approx nX(t) + \sqrt{n}\hat{X}(t) \stackrel{\mathrm{d}}{=} \mathcal{N}(nX(t), n\sigma^2_{\hat{X}}(t)), \tag{3.1}$$

where X is the fluid approximation analyzed directly in [74] and obtained as a limit in the MSHT FWLLN in [75], while $\hat{X}$ is a zero-mean Gaussian process with variance $\sigma^2_{\hat{X}}(t) \equiv Var(\hat{X}(t))$ obtained from the MSHT FCLT in [77] (which assumes M service), and $\mathcal{N}(m, \sigma^2)$ denotes a Gaussian random variable with mean m and variance σ^2. To obtain associated approximations for the steady-state variable, denoted by $X_n(\infty)$, we take the direct approach and simply let $t \to \infty$ in (3.1) and obtain

$$X_n(\infty) \approx nX(\infty) + \sqrt{n}\hat{X}(\infty) \stackrel{\mathrm{d}}{=} \mathcal{N}(nX(\infty), n\sigma^2_{\hat{X}}(\infty)). \tag{3.2}$$

Of course, the associated number in queue satisfies $Q_n(t) \equiv (X_n(t) - n)^+$, where $(a)^+ \equiv \max\{a, 0\}$ and $\equiv$ denotes equality by definition, but if $\rho > 1$, then its MSHT limit holds without that truncation. If $\rho > 1$, then $X(\infty) > 1$ in (3.2), so that the direct MSHT limit for the number in queue yields

$$Q_n(\infty) \approx n(X(\infty) - 1) + \sqrt{n}\hat{X}(\infty), \tag{3.3}$$

without any truncation. That makes the approximation have $P(Q_n(\infty) < 0) > 0$, even though that is not possible. Similarly, the direct approximation allows $P(B_n(\infty) > n) > 0$, where $B_n(\infty)$ is the steady-state number in service, even though that is not possible. As a consequence, these direct MSHT approximations produce probability distributions in regions that cannot occur. Moreover, this defect can seriously degrade the approximations; e.g., see Table 3.2.

Because these direct Gaussian approximations are ineffective except when the system is significantly OL or UL, e.g., when $\rho > 1.2$, we investigate a simple refinement based on truncation that was proposed in (1.2) of [77], but never tested; indeed that is the purpose of the present chapter. As in (1.2) of [77], the first step is to make the obvious refinement for the queue length, letting

$$Q_n(\infty) \equiv (X_n(\infty) - n)^+ \approx (nX(\infty) + \sqrt{n}\hat{X}(\infty) - n)^+. \tag{3.4}$$

Similarly, we let the number in service be

$$B_n(\infty) \equiv X_n(\infty) \wedge n \approx (nX(\infty) + \sqrt{n}\hat{X}(\infty) \wedge n,$$

where $a \wedge b \equiv \min\{a, b\}$. We call these approximations based on the FCLT for the $G/M/n + GI$ model in [77] the *truncated Gaussian approximations* (TGAs); see §3.3 for the details. In this chapter we show that the approximations based on the MSHT limit plus this simple truncation refinement are remarkably effective for OL models.

3.1.3 Related Literature

There are two streams of important related work. The first stream is the literature on MSHT limits, starting with models without abandonment in [48, 38] and then continuing with stationary models with abandonment [20, 19, 32, 43, 55, 94], and then time-varying Markov models in [80] and non-Markovian models in [74, 75, 77]; we refer to those papers for further discussion of the MSHT literature.

The second stream is the literature on exact results and approximations for the $M/GI/n + GI$ model [7, 13, 14, 123]. The exact results for the $M/M/n + GI$ model in [123] led to further studies that provided better understanding, such as [83]. However, exact results for the $M/GI/n + GI$ model with non-exponential service remains an important open problem. Approximations for the $M/GI/n + GI$ model were developed in [107]. As we have confirmed in our simulation experiments, these approximations in [107] are remarkably effective, but they are substantially more complicated, requiring approximation by a state-dependent $M/M/s + M(n)$ model and then a numerical algorithm.

A main conclusion in [107, 117] was that the steady state performance of the $M/GI/n + GI$ model tends to be nearly insensitive to the service-time distribution beyond its mean. Our experiments confirm that conclusion for the performance measures considered, but not for all performance measures. In particular, our simulations show that the mean values of the steady-state queue length and waiting time have this near-insensitivity property, but the variance and distribution do not; they depend significantly on the service-time distribution beyond its mean. Moreover, the simulations show that the new approximations of these quantities for the $M/GI/n + GI$ model with a

non-exponential service times are effective; e.g., see Tables 3.7 and 3.8. Since the limit in [77] is only for the $G_t/M/s_t + GI$ model, with exponential service times, our approximation in this step is only heuristic. It does follow from a natural modification to account for GI service, but it remains to be better justified theoretically.

Refinements to the direct MSHT approximations for the time-varying Markovian $M_t/M/s_t + M$ model in [80] have also been developed by [86]. These Gaussian and skewness closure approximations are remarkably effective, even for critically loaded models, but so far they are restricted to Markovian models.

3.1.4 Organization of the Chapter

In §3.2 we give a brief review of the heavy-traffic limits which are the building blocks for our approximations. In §3.3 we develop the TGA formulas for the $GI/GI/n + GI$ queueing systems. In §§3.4-3.7 we present results of numerical examples to test the performance of the Gaussian approximations: In §3.4 we consider the Markovian $M/M/n + M$ model; in §3.5 we consider the extensions to non-M arrival processes and non-M patience distributions; in §3.6 we consider the extension to non-M service; and in §3.7 we consider examples with fewer servers and a corresponding lower arrival rate.

We provide additional supporting material in §§3.8-3.9. In §3.8 we give examples exposing the limitations of the approximations. In §3.8.1 and §3.8.2 we compare our approximations to previous approximations developed by Whitt [107] and Reed and Tezcan [94]; these tend to perform better in the QED regime. In §3.8.3 we expose the limitations of the approximations for UL models. In §3.9 we elaborate on the simulation methodology. We draw conclusions in §3.10. We present additional material in an appendix [78], which is available as supporting information on the journal and author web pages.

3.2 The Many-Server Heavy-Traffic Limits for Stationary Models

The MSHT limits are obtained by considering a sequence of stationary queueing models indexed by the integers n. In the n^{th} model, there are $s_n = \lceil ns \rceil$ servers and the arrival rate is $\lambda_n = n\lambda$, where λ is the base arrival rate. (The scaling factor becomes the number of servers when we let $s \equiv 1$.) The service times and patience times are unscaled; they are assumed to come from independent sequences of i.i.d. random variables with cumulative distribution functions (cdfs) G and F with means $E[S] = 1/\mu = 1$ and $E[A] = 1/\theta$ and finite second moments. Let f and $\bar{F}$ be the probability density function (pdf) and complementary cumulative distribution function (ccdf) of F; and let $\bar{G}$ be the ccdf of G. Thus the traffic intensity is $\rho_n = \lambda_n/s_n\mu = \lambda/s\mu \equiv \rho$ for all n.

The arrival process $N_n(t)$ is assumed to satisfy a FCLT

$$(N_n(t) - n\lambda t)/\sqrt{n} \Rightarrow c_\lambda \mathcal{B}(t) \quad \text{in } \mathbb{D} \quad \text{as } n \to \infty, \tag{3.5}$$

where $\mathcal{B}$ is a standard Brownian motion (BM), $c_\lambda > 0$ is a variability parameter, $\Rightarrow$ denotes weak convergence and $\mathbb{D}$ denotes the space of right-continuous functions that have left limits; see [8, 77, 113] for details.

Let $Q_n(t)$ and $B_n(t)$ to be the number of customers waiting in queue and in service at time t. Let $X_n(t) = Q_n(t) + B_n(t)$ be the total number in the system. Let $W_n(t)$ and $V_n(t)$ to be the head-of-line waiting time (the elapsed waiting time of the head-of-line customer if there is any) and the potential waiting time at time t (the waiting time of an infinitely patient arrival at time t if there were an arrival at that time). We next review the MSHT FWLLN and FCLT limits for the stationary model, which are the building blocks for our Gaussian approximations.

3.2.1 The FWLLN Yielding Fluid Limits for the $G/GI/n + GI$ Queue

Let the LLN-scaled processes be

$$\bar{X}_n(t) = \frac{X_n(t)}{n}, \quad \bar{Q}_n(t) = \frac{Q_n(t)}{n}, \quad \bar{B}_n(t) = \frac{B_n(t)}{n}, \quad \bar{W}_n(t) = W_n(t) \quad \text{and} \quad \bar{V}_n(t) = V_n(t).$$

The waiting times need no spatial scaling because the service-time and patience-time cdf's are not scaled. By Theorem 1 in [75], we have the joint convergence for the LLN-scaled functions

$$\left(\bar{X}_n, \bar{Q}_n, \bar{B}_n, \bar{W}_n, \bar{V}_n\right) \Rightarrow (X, Q, B, w, v), \quad \text{in } \mathbb{D}^5, \tag{3.6}$$

where the limit is a vector of deterministic functions, specified in [117, 74]. We next summarize the steady state performance.

Theorem 3.2.1 (Theorem 3.1 in [117] and Theorem 4.1 in [73])**.** *The stationary fluid model with capacity s arising as the MSHT FWLLN limit of stationary $G/GI/n+GI$ model with s_n servers has a steady state characterized by the deterministic vector $(b(\infty), q(\infty), B(\infty), Q(\infty), X(\infty), w(\infty), v(\infty))$ in $\mathbb{R}^7$, where $X(\infty) = B(\infty) + Q(\infty)$ and the other variables depend on the value of the traffic intensity $\rho \equiv \lambda/s\mu = \lambda/s$.*

(a) If $\rho \le 1$, then for $x \ge 0$,

$$B(\infty) = \int_0^\infty b(x)\,dx = s\rho, \quad b(x) = \lambda \bar{G}(x), \quad Q(\infty) = \int_0^\infty q(x)\,dx = w(\infty) = v(\infty) = q(x) = 0.$$

(b) If $\rho > 1$, then

$$B(\infty) = \int_0^\infty b(x)\,dx = s, \quad b(x) = s\mu\bar{G}(x) \quad \textit{for} \quad x \geq 0,$$
$$v(\infty) = w(\infty) = F^{-1}\left(1 - \frac{1}{\rho}\right), \quad q(x) = \lambda\bar{F}(x) \quad \textit{for} \quad 0 \leq x \leq w(\infty),$$
$$Q(\infty) = \int_0^\infty q(x)\,dx = \lambda\int_0^{w(\infty)} \bar{F}(x)dx.$$

3.2.2 The FCLT Yielding Gaussian limits for the $G/M/n + GI$ Queue

As in [77], we now restrict to the $G/M/n + GI$ model having M service. To provide the FCLT limits, let the CLT-scaled processes be

$$\hat{X}_n(t) = \frac{X_n(t) - nX(t)}{\sqrt{n}}, \quad \hat{Q}_n(t) = \frac{Q_n(t) - nQ(t)}{\sqrt{n}}, \quad \hat{B}_n(t) = \frac{B_n(t) - nB(t)}{\sqrt{n}},$$
$$\hat{W}_n(t) = \sqrt{n}\,(W_n(t) - w(t))\,, \quad \hat{V}_n(t) = \sqrt{n}\,(V_n(t) - v(t))\,, \tag{3.7}$$

where n is the number of servers, while $X(t), Q(t), B(t), w(t)$ and $v(t)$ are the deterministic limit functions in (4.3), i.e., the deterministic fluid functions in [74]. The FCLT follows directly from the FCLT for the time-varying $G_t/M/s_t + GI$ model (Theorems 4.2 and 5.1 in [77]) by simply letting $\lambda(t) = \lambda$ and $s(t) = s$.

Theorem 3.2.2 (MSHT FCLT limits for the $G/M/n + GI$ queues)**.** *Consider the sequence of $G/M/n + GI$ queueing models having s_n servers. Under regularity conditions in [77], including appropriate initial convergence at time* 0*,*

$$\left(\hat{X}_n, \hat{B}_n, \hat{Q}_n, \hat{W}_n, \hat{V}_n\right) \Rightarrow \left(\hat{X}, \hat{B}, \hat{Q}, \hat{W}, \hat{V}\right) \quad \textit{in } \mathbb{D}^5 \quad \textit{as } n \to \infty.$$

(a) When $\rho < 1$, i.e. in an underloaded (UL) interval, $\hat{Q} = \hat{W} = \hat{V} = 0$, and $\hat{B} \stackrel{\mathrm{d}}{=} \hat{X}$ satisfies the stochastic differential equation (SDE)

$$d\hat{X}(t) = -\mu\hat{X}(t)dt + \sqrt{c_\lambda^2\lambda + \mu X(t)}\,d\mathcal{B}(t),$$

where $\mathcal{B}$ is a standard Brownian motion, so that $\hat{X}(t)$ is a Gaussian process with

$$\sigma_X^2(t) \;\; = \;\; (c_\lambda^2 - 1)\int_0^t \bar{G}^2(t-u)\lambda du + \int_0^t \bar{G}(t-u)\lambda du.$$

(b) When $\rho > 1$, i.e. in an overloaded (OL) interval, $\hat{B}(t) = 0$, $\hat{Q}(t) = \hat{X}(t)$ where

$$\begin{aligned}
\hat{X}(t) &= \hat{X}(0)\bar{F}_w(t) + \sum_{i=1}^{3}\int_0^t K_i(t,u)\,d\mathcal{B}_i(u),\\
\hat{W}(t) &= \hat{W}(0)H(t,0) + \sum_{i=1}^{3}\int_0^t H(t,u)I_i(u)\,d\mathcal{B}_i(u),\\
\hat{V}(t) &= \frac{\hat{W}(t+v(t))}{1-\dot{w}(t+v(t))},
\end{aligned}$$

where $\mathcal{B}_1 \equiv \mathcal{B}_\lambda$, $\mathcal{B}_2 \equiv \mathcal{B}_s$ and $\mathcal{B}_3 \equiv \mathcal{B}_a$ are independent standard BMs, that are the FCLT limits of the scaled arrival process (the subscript "λ"), service process (the subscript "s") and abandonment process (the subscript "a"). Both $\hat{X}(t)$ and $\hat{W}(t)$ are zero mean Gaussian processes with variance process

$$\sigma_W^2(t) = \int_0^t H^2(t,u)I^2(u)\,du + H^2(t,0)\,Var\left(\hat{W}(0)\right), \tag{3.8}$$

$$\begin{aligned}
\sigma_X^2(t) &= \int_{t-w(t)}^{t} \lambda\,\bar{F}(t-u)\left((c_\lambda^2-1)\bar{F}(t-u)+1\right)du\\
&\quad +\lambda^2\bar{F}^2(w(t))\sigma_W^2(t) + Var\left(\hat{X}(0)\right)\cdot\left(\bar{F}_w(t)\right)^2.
\end{aligned} \tag{3.9}$$

Finally, $H(t,u)$, $I_i(t)$, $K_i(t)$ are deterministic analytic functions given by

$$\begin{aligned}
H(t,u) &= \exp\left\{\int_u^t h(v)\,dv\right\} = \exp\left\{\int_u^t -\frac{f(w(\infty))}{\bar{F}(w(\infty))}\,dv\right\} = e^{-\frac{f(w(\infty))}{\bar{F}(w(\infty))}(t-u)},\\
I_1^2(t) &= \frac{c_\lambda^2\bar{F}(w(u))b(u,0)}{\tilde{q}^2(u,w(u))} = \frac{c_\lambda^2\bar{F}(w(\infty))s\mu}{\lambda^2\bar{F}^2(w(\infty))} = c_\lambda^2\lambda^{-1},\\
I_2^2(t) &= \frac{b(u,0)}{\tilde{q}^2(u,w(u))} = \frac{s\mu}{\lambda^2\bar{F}^2(w(\infty))} = \rho/\lambda,\\
I_3^2(t) &= \frac{F(w(u))b(u,0)}{\tilde{q}^2(u,w(u))} = \frac{F(w(\infty))s\mu}{\lambda^2\bar{F}^2(w(\infty))} = (\rho-1)/\lambda,\\
I^2(t) &= I_1^2(t)+I_2^2(t)+I_3^2(t) = \frac{(c_\lambda^2-1)\rho^{-1}+2}{\mu s},\\
K_1(t,u) &= c_\lambda\bar{F}(t-u)\sqrt{\lambda(u)}\mathbb{1}_{\{t-w(t)u<t\}} + \tilde{q}(t,w(t))\sqrt{\lambda(u)}\bar{I}_1(L^{-1}(u))\mathbb{1}_{\{0\le u\le t-w(t)\}},\\
K_2(t,u) &= -\sqrt{b(t,0)-\dot{s}(t)}H(t,u),\\
K_3(t,u) &= -\sqrt{\lambda(u)F(t-u)\bar{F}(t-u)}\mathbb{1}_{\{t-w(t)\le u\le t\}}\\
&\quad + \tilde{q}(t,w(t))\sqrt{\lambda(u)}\bar{I}_3(L^{-1}(u))H(t,L^{-1}(u))\mathbb{1}_{\{0\le u\le t-w(t)\}},
\end{aligned}$$

where

$$\bar{I}_1(t) \;=\; \frac{c_\lambda \bar{F}(w(u))b(u,0)}{\tilde{q}(u,w(u))} = \frac{c_\lambda \bar{F}(w(\infty))s\mu}{\lambda \bar{F}(w(\infty))}, \quad \bar{I}_3(t) = -\frac{\sqrt{\bar{F}(w(u))b(u,0)}}{\tilde{q}(u,w(u))} = -\frac{\bar{F}(w(\infty))s\mu}{\lambda \bar{F}(w(\infty))},$$

$L(t) = t - w(\infty)$, $\lambda(t) = \lambda t$ and $\mathbb{1}_A$ is an indicator random variable of A.

We next provide steady-state distribution for the FCLT limits of the $G/M/n + GI$ model.

Theorem 3.2.3 (Steady-state of the MSHT FCLT limit of $G/M/n+GI$ queues)**.** *The steady-state of the Gaussian process arising in the MSHT FCLT limit for the sequence of $G/M/n + GI$ models with s_n servers in model n is given by the vector $(\hat{Q}(\infty), \hat{B}(\infty), \hat{X}(\infty), \hat{W}(\infty), \hat{V}(\infty))$ specified below:*

(a) If $\rho < 1$, then

$$\hat{Q}(\infty) = \hat{W}(\infty) = \hat{V}(\infty) = 0 \quad \textit{and} \quad \hat{B}(\infty) = \hat{X}(\infty) \stackrel{\mathrm{d}}{=} \mathcal{N}\left(0, \sigma_X^2\right),$$

where the variance is

$$\sigma_X^2 \equiv \lambda(c_\lambda^2 - 1)\int_0^\infty \bar{G}^2(u)\,du + \lambda \int_0^\infty \bar{G}(u)\,du. \tag{3.10}$$

(b) If $\rho > 1$, then

$$\hat{Q}(\infty) \stackrel{\mathrm{d}}{=} \hat{X}(\infty) \stackrel{\mathrm{d}}{=} \mathcal{N}\left(0, \sigma_X^2\right), \quad \hat{B}(\infty) = 0, \quad \hat{W}(\infty) \stackrel{\mathrm{d}}{=} \hat{V}(\infty) \stackrel{\mathrm{d}}{=} \mathcal{N}(0, \sigma_W^2),$$

where the variances are

$$\sigma_W^2 \equiv \frac{(c_\lambda^2 - 1) + 2\rho}{2\mu s \rho^2 f(w(\infty))} \quad \textit{and} \quad \sigma_X^2 \equiv (\mu s)^2 \sigma_W^2 + \lambda \int_0^{w(\infty)} \bar{F}(u)\left(1 + (c_\lambda^2 - 1)\bar{F}(u)\right) du, \tag{3.11}$$

with $w(\infty)$ being the steady-state fluid waiting time in the OL case, given in Theorem 4.1.1.

Proof. Because the limit process is a zero-mean Gaussian process, it suffices to show that the variances converge as $t \to \infty$. In particular, it suffices to show that $\sigma_W^2(t) \to \sigma_W^2$ and $\sigma_X^2(t) \to \sigma_X^2$ as $\to \infty$. That is easy to check in underloaded interval, thus we focus on the limit in the overloaded

interval.

$$\begin{aligned}
\sigma_W^2 &= \lim_{t\to\infty}\sigma_W^2(t) = \lim_{t\to\infty}\left(\int_0^t I^2(u)\cdot H^2(t,u)\mathrm{d}u + H^2(t,0)\mathrm{Var}(\hat{W}(0))\right)\\
&= \lim_{t\to\infty}\left(\frac{(c_\lambda^2-1)\rho^{-1}+2}{\mu s}\int_0^t e^{-\frac{2f(w(\infty))}{\bar{F}(w(\infty))}(t-u)}\mathrm{d}u + e^{-\frac{2f(w(\infty))}{\bar{F}(w(\infty))}t}\mathrm{Var}(\hat{W}(0))\right)\\
&= \frac{\bar{F}(w(\infty))}{2f(w(\infty))}\frac{(c_\lambda^2-1)\rho^{-1}+2}{\mu s} = \frac{(c_\lambda^2-1)+2\rho}{2\mu s\rho^2 f(w(\infty))},
\end{aligned}$$

where the last equality holds because $\bar{F}(w(\infty)) = 1/\rho$. Inserting it to (4.7), we get the remaining limit

$$\begin{aligned}
\sigma_X^2 \equiv \lim_{t\to\infty}\sigma_X^2(t) &= \lim_{t\to\infty}\mathrm{Var}\left(\hat{X}(0)\right)\cdot\left(\bar{F}_w(t)\right)^2 + \lim_{t\to\infty}\lambda^2\bar{F}^2(w(t))\sigma_W^2(t)\\
&\quad + \lim_{t\to\infty}\int_{t-w(t)}^t \lambda\,\bar{F}(t-s)\left((c_\lambda^2-1)\bar{F}(t-s)+1\right)\mathrm{d}s\\
&= (\lambda/\rho)^2\sigma_W^2 + \lim_{t\to\infty}\lambda\int_0^{w(t)}\bar{F}(u)\left((c_\lambda^2-1)\bar{F}(u)+1\right)\mathrm{d}u\\
&= (\mu s)^2\sigma_W^2 + \lambda\int_0^{w(\infty)}\bar{F}(u)\left((c_\lambda^2-1)\bar{F}(u)+1\right)\mathrm{d}u,
\end{aligned}$$

where $\lim_{t\to\infty}\bar{F}(w(t)) = \bar{F}(w(\infty)) = 1/\rho$, $\lim_{t\to\infty}\mathrm{Var}\left(\hat{X}(0)\right)\cdot\left(\bar{F}_w(t)\right)^2 = 0$ since $\bar{F}(t)\to 0$ as $t\to\infty$ but $\mathrm{Var}\left(\hat{X}(0)\right)$ is bounded.

Full convergence of the multivariate Gaussian distribution in $\mathbb{R}^5$ requires convergence of the pairwise covariances too. That also follows from the representation in Theorem 4.1.2, but we omit the details, because we will not be applying that. □

Remark 3.2.1 (Interchange of the two limits)**.** Our MSHT approximation is based on an iterated limit in which first $n\to\infty$ and then $t\to\infty$, but we need the iterated limit in the other order. As often is done in MSHT approximations, we are assuming without proof that a limit interchange is valid, in particular,

$$\lim_{t\to\infty}\hat{X}(t) = \lim_{t\to\infty}\lim_{n\to\infty}\hat{X}_n(t) = \lim_{n\to\infty}\lim_{t\to\infty}\hat{X}_n(t) = \lim_{n\to\infty}\hat{X}_n(\infty). \tag{3.12}$$

We conjecture that (3.12) holds for the $G/GI/n{+}GI$ model under mild regularity conditions, if any. For the more elementary $M/M/n$, $M/M/n+M$ and $M/M/n+GI$ models, that limit interchange was proved in [38, 32, 123].

Remark 3.2.2 (The underloaded case)**.** The MSHT limits above are relatively elementary in the underloaded case with $\rho<1$. In that case the results reduce to corresponding infinite-server results in [89] and references there, as shown for the $M/M/n$ model in [48].

3.3 The Steady-State Gaussian Approximations

In this section, we develop the Gaussian approximation formulas for the steady-state distribution of the stationary $G/GI/n{+}GI$ model. Henceforth, we focus on steady-state random variables and no longer discuss stochastic processes. Thus, for simplicity, we omit the time argument t, which would become ∞ in steady state; i.e., we let Q_n denote the steady-state queue length in model n with s_n servers and we let Q and $\hat{Q}$ denote the steady-state of the fluid and FCLT limits, respectively, and similarly for the other variables. (Previously, these were denoted by $Q_n(\infty)$, $Q(\infty)$ and $\hat{Q}(\infty)$.) Since the steady-state head-of-line waiting time W_n and potential waiting time V_n should coincide, as do the limits in Theorem 4.1.1 (b) and Theorem 3.2.3 (b), we henceforth use only the notion W_n, w and $\hat{W}$ for the waiting time (instead of $W_n(\infty)$, $V_n(\infty)$, $w(\infty)$, $v(\infty)$, $\hat{W}(\infty)$ and $\hat{V}(\infty)$).

We start in §3.3.1 with direct Gaussian approximations (DGA) and then turn to the basic truncation refinement TGA for the $G/M/n{+}GI$ model in §3.3.2. Afterwards, in §3.3.3 we generalize TGA from M service to GI service, which we refer to as TGA-G.

3.3.1 Direct Gaussian Approximations

The most straightforward performance approximation is a direct application of the steady-state of the deterministic fluid and the zero-mean Gaussian limits. We we use superscript "DGA" to denote these approximations, which are

$$X_n \approx X_n^{DGA} \equiv nX + \sqrt{n}\hat{X}, \tag{3.13}$$

$$B_n \approx B_n^{DGA} \equiv nB + \sqrt{n}\hat{B}, \quad Q_n^{DGA} \equiv nQ + \sqrt{n}\hat{Q}, \tag{3.14}$$

$$W_n \approx W_n^{DGA} \equiv w + \frac{1}{\sqrt{n}}\hat{W}, \tag{3.15}$$

where X, B, Q and w are given in Theorem 4.1.1 and $\hat{X}$, $\hat{B}$, $\hat{Q}$ and $\hat{W}$ are given in Theorem 3.2.3. (Recall that we have eliminated the infinite time argument from the steady-state quantities given in Theorems 4.1.1 and 3.2.3.)

In addition to approximating means and variances, defined in the obvious direct way, we have associated approximations for the P^d and P^a, namely,

$$P^d \approx P^{d,C} \approx P_D^{DGA}(n) \equiv \mathbb{P}\left(W_n^{DGA} > 0\right) \quad \text{and} \tag{3.16}$$

$$P^a \approx P_A^{DGA}(n) \equiv \mathbb{P}\left(W_n^{DGA} > A\right), \tag{3.17}$$

where A is a generic independent patience time.

3.3.2 Truncated Gaussian Approximations

It is natural to refine the DGA approximations by truncation because $Q_n = (X_n - s_n)^+$ and $B_n = (X_n \wedge s_n)^+$, where a^+ is $\max\{0, a\}$ and $a \wedge b = \min\{a, b\}$. Thus, our TGA approximations for Q_n and B_n are

$$\begin{aligned} Q_n^{TGA} &\equiv \left(X_n^{DGA} - ns\right)^+ = \sqrt{n}\sigma_X \left(\frac{\hat{X}}{\sigma_X} + \frac{\sqrt{n}(X-s)}{\sigma_X}\right)^+ \equiv \sqrt{n}\sigma_X \left(\mathcal{Z} \vee -a_X(n)\right) + n(X-s), \\ B_n^{TGA} &\equiv \left(X_n^{DGA} \wedge ns\right)^+ = \sqrt{n}\sigma_X \left((\mathcal{Z} \wedge -a_X(n)) + \frac{\sqrt{n}}{\sigma_X} X\right)^+, \end{aligned} \tag{3.18}$$

where $\mathcal{Z}$ is a standard Gaussian random variable, σ_X is given in (3.11), and

$$a_X(n) = \sqrt{n}(X-s)/\sigma_X. \tag{3.19}$$

In the same spirit, we truncate the DGAs for waiting times (3.15) to obtain their TGAs.

$$W_n^{TGA} \;=\; V_n^{TGA} = \left(W_n^{DGA}\right)^+ = \left(w + \frac{\hat{W}}{\sqrt{n}}\right)^+ = w + \frac{\sigma_W}{\sqrt{n}}\left(\mathcal{Z} \vee -a_W(n)\right), \tag{3.20}$$

where σ_W is given in (3.11) and

$$a_W(n) = \sqrt{n}w/\sigma_W. \tag{3.21}$$

The means, variances and distributions are then approximated in the obvious way. The formulas can be conveniently expressed as functions of the standard normal ccdf $\bar{\Phi}$ and pdf ϕ via

$$\begin{aligned} \mathrm{E}[V_n^{TGA}] &= \mathrm{E}[w + \frac{\sigma_W}{\sqrt{n}}\left(\mathcal{Z} \vee -a_W(n)\right)] = w\left(\Phi(a_W(n)) + \frac{\phi(a_W(n))}{a_W(n)}\right), \\ \mathrm{Var}(V_n^{TGA}) &= \mathrm{Var}(w + \frac{\sigma_W}{\sqrt{n}}\left(\mathcal{Z} \vee -a_W(n)\right)), \\ &= \frac{(a_W(n)\sigma_W)^2}{n}\left[\left(1 + \frac{1}{a_W^2(n)}\right)\bar{\Phi}(a_W(n)) - \frac{\phi(a_W(n))}{a_W(n)} - \left(\frac{\phi(a_W(n))}{a_W(n)} - \bar{\Phi}(a_W(n))\right)^2\right], \\ \mathrm{E}[Q_n^{TGA}] &= \mathrm{E}[n(X-s) + \frac{\sigma_X}{\sqrt{n}}\left(\mathcal{Z} \vee -a_X(n)\right)] = n(X-s)\left(\Phi(a_X(n)) + \frac{\phi(a_X(n))}{a_X(n)}\right), \\ \mathrm{Var}(Q_n^{TGA}) &= \mathrm{Var}(n(X-s) + \frac{\sigma_X}{\sqrt{n}}\left(\mathcal{Z} \vee -a_X(n)\right)) \\ &= n(a_X(n)\sigma_X)^2\left[\left(1 + \frac{1}{a_X^2(n)}\right)\bar{\Phi}(a_X(n)) - \frac{\phi(a_X(n))}{a_X(n)} - \left(\frac{\phi(a_X(n))}{a_X(n)} - \bar{\Phi}(a_X(n))\right)^2\right], \end{aligned}$$

where $a_W(n)$ and $a_X(n)$ are given in (3.21) and (3.19).

The P^d, $P^{d,C}$ and P^a are natural analogs of (3.16) and (3.17), i.e.,

$$\begin{aligned} P^d &\approx P^{d,c} \approx P_D^{TGA}(n) \equiv \mathbb{P}\left(W_n^{TGA} > 0\right) = \mathbb{P}\left(w\left(\frac{\mathcal{Z}}{a_W(n)} + 1\right) > 0\right) \\ &= \bar{\Phi}\left(-a_W(n)\right), \end{aligned} \tag{3.22}$$

$$\begin{aligned} P^a &\approx P_A^{TGA}(n) \equiv \mathbb{P}\left(W_n^{TGA} > A\right) = \mathbb{P}\left(w\left(\frac{\mathcal{Z}}{a_W(n)} + 1\right) > A\right) \\ &= \int_0^\infty \bar{\Phi}\left(a_W(n)\left(\frac{x}{w} - 1\right)\right) f(x)\mathrm{d}x, \end{aligned} \tag{3.23}$$

where f is the pdf of the patience time, as before.

Notice that several of the TGA approximations coincide with their DGA counterparts, i.e.,

$$(X_n^{TGA}, \mathrm{E}[X_n^{TGA}], \mathrm{Var}(X_n^{TGA}), P_A^{TGA}(n), P_D^{TGA}(n)) = (X_n^{DGA}, \mathrm{E}[X_n^{DGA}], \mathrm{Var}(X_n^{DGA}), P_A^{DGA}(n), P_D^{DGA}(n))$$

3.3.3 The Refinement for Non-Exponential Service: TGA-G

Recall that both DGAs and TGAs are developed based on Theorems 4.1.2 and 3.2.3 for the $G/M/n + GI$ queue. However, the form of the limit in 3.2.3 allows us to identify the impact of the service process. In particular, we see that the three sources of variability – the arrival process, service times and patience times – produce identifiable impacts on the limiting Gaussian processes through three independent Brownian motions. Thus, even though the impact of the service process is complicated, we can exploit this separability to see how to introduce a good candidate heuristic approximation.

First, for the OL model, where asymptotically all servers are busy all the time, we can act as if the servers are continuously busy. Because of the assumed exponential service time distribution, the approximating service-completion process is a Poisson process with rate $\mu s(t)$, which yields a Brownian FCLT limit, corresponding to the BM $\mathcal{B}_2 = \mathcal{B}_s$ in Theorem 4.1.2 (See Theorem 4.2 in [77] for details; also see (4.5) and (6.64) there). When service becomes GI, and when the servers are all busy, the service process is the superposition of i.i.d. renewal processes. Because the number of servers n grows in the limit, that superposition departure process is complicated, as discussed in §9.8 of [113]. However, one candidate approximation is to act as if n is fixed. Then the departure process again satisfies a FCLT with a Brownian motion limit, but with the prefactor c_s, just as if the service process were a single renewal process with c_s^2 being the squared coefficient of variation (SCV, variance divided by the square of the mean) of the time between renewals; see §9.4 of [113].

Consequently, in our heuristic approximation we capture the nonexponential service distribution by scaling the BM $\mathcal{B}_2$ by adding the prefactor c_s. That leads to replacing the "1" by "c_s" in the

numerator of (3.11), yielding

$$\sigma^2_{W_G} \quad = \quad \frac{(c_\lambda^2 - 1) + (c_s + 1)\rho}{2\mu s \rho^2 f(w)}. \tag{3.24}$$

By replacing σ_W by σ_{W_G} in (3.11), (3.18)–(3.20), we obtain TGA-G. It is significant that TGA-G reduces to TGA when service is M, for which $c_s^2 = 1$.

3.4 Evaluating the Gaussian Approximations for Markov Models

Since the Markovian $M/M/n + M$ model is relatively tractable, we primarily want to develop effective approximations for other non-Markov $G/GI/n+GI$ models. Nevertheless, it is convenient to start examining the proposed Gaussian approximations by making comparisons with exact numerical results for the $M/M/n + M$ model because numerical algorithms are readily available. A minimum requirement for our proposed approximations is that they perform well for this basic model.

Hence, we start evaluating the proposed approximations in this section by comparing with exact numerical results for $M/M/n+M$ model. For that purpose, we use the numerical algorithm in [107]. That algorithm was developed to treat the more general $M/GI/n+GI$ model by approximating it by the Markovian $M/M/n + M(n)$ model with state-dependent abandonment rate, but it applies to the $M/M/n + M$ model as a special case. That model includes a finite waiting room; we let it be so large that it does not affect the formulas.

Our base case has $n = 100$ servers, but we also examine smaller systems later. For the overloaded systems of primary interest, we considered a range of traffic intensities from $\rho = 1.5$ down to 1.001. The case $\rho = 1.5$ is so overloaded that there is little need for truncation, so that the TGA and DGA approximations nearly coincide, and the performance is very good; see Table A.1 in the appendix. We thus start by showing the experimental results for $\rho = 1.2$ with abandonment rates $0.1 \le \theta \le 4.0$ in Table 3.1, where $M(m)$ refers to an exponential distribution with mean m. For the lower three abandonments rates ($\theta \le 0.5$), the system is still highly heavily loaded. For these two cases in Table 3.1, TGA and DGA are quite close with all errors less than 1%. However, otherwise we see that TGA provides significant improvement.

We regard the case $\rho = 1.2$ as quite heavily loaded, much more in the ED heavy-traffic regime than the QED regime. Hence, we primarily focus on models with lower traffic intensities. For the traffic intensity, we regard $\rho = 1.05$ as our base case; it corresponds to levels often encountered in practice. Moreover, for $n = 100$ and $\rho = 1.05$, the system is operating in the more practical QED regime, which can be characterized by the quality-of-service (QoS) parameter $\beta \equiv (1-\rho)\sqrt{n} = -0.5$; see [38, 32]. Table 3.2 shows the main performance measures for six abandonment rates with $0.1 \le \theta \le 4.0$. Table 3.2 shows that DGA performs poorly in this case, but TGA provides dramatic

improvement, having all errors in the means $E[Q]$ and $E[W]$ and variances $Var(Q)$ and $Var(W)$ less than 10% for $\theta \leq 2.0$.

As indicated earlier, we find good performance for $\theta \leq 2.0$. For $\theta = 0.1$ and 0.25 with $(n, \rho) = (100, 1.05)$, the maximum percentage error among the means $\mathrm{E}[X]$, $\mathrm{E}[Q]$, $\mathrm{E}[W]$ and the probabilities P^d and P^a for TGA was 7%. On the other hand, Table A.4 shows that the performance of TGA degrades for $\theta = 4$ and 10, but then the high abandonment rate makes the system far from being overloaded; e.g., $\mathrm{E}[Q] = 1.47$ and $\mathrm{E}[W] = 0.017$ for $\theta = 4$.

Most of the rest of this chapter is devoted to showing that the good results in Table 3.2 extend to a wide class of models and parameters. We illustrate in Figures 3.1-3.3, which graphically show the performance for the $M/M/n + M$ model as a function of ρ and θ. Figures 3.1 and 3.2 show the performance as a function of ρ for two values of θ: 0.5 to 2.0. Figure 3.3 shows the performance as a function of θ for $\rho = 1.05$.

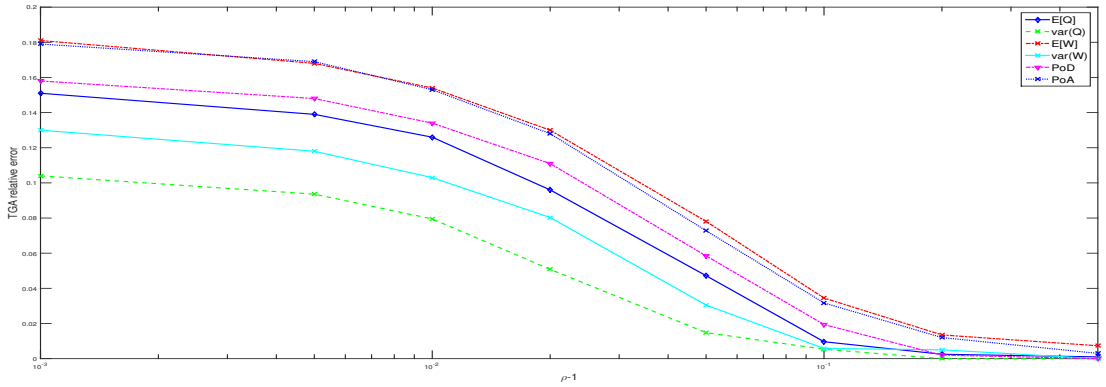


Figure 3.1: The relative error in the approximations of six performance measures as a function of the traffic intensity ρ for $1.001 \leq \rho \leq 1.500$ (with $\rho - 1$ in log scale) in the $M(1/100\rho)/M(1)/100 + M(2.0)$ model with abandonment rate $\theta = 0.5$.

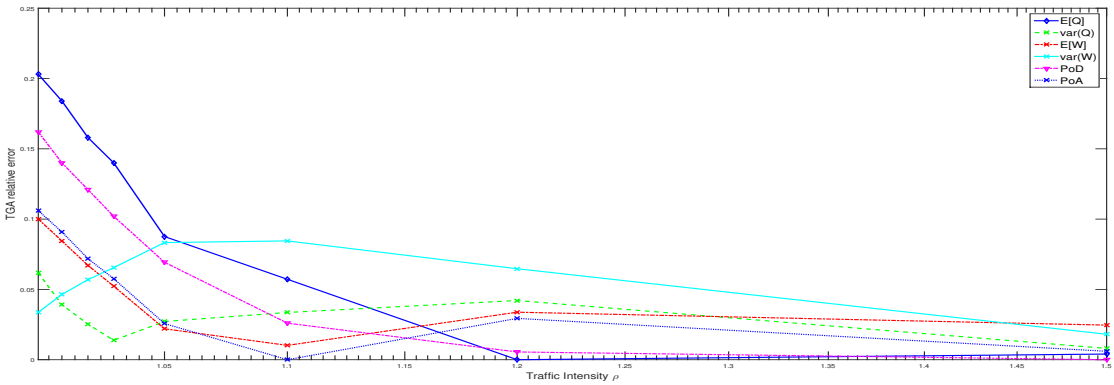


Figure 3.2: The relative error in the approximations of six performance measures as a function of the traffic intensity ρ for $1.001 \leq \rho \leq 1.500$ (with (with $\rho - 1$ in log scale) in the $M(1/100\rho)/M(1)/100 + M(0.5)$ model with abandonment rate $\theta = 2.0$.

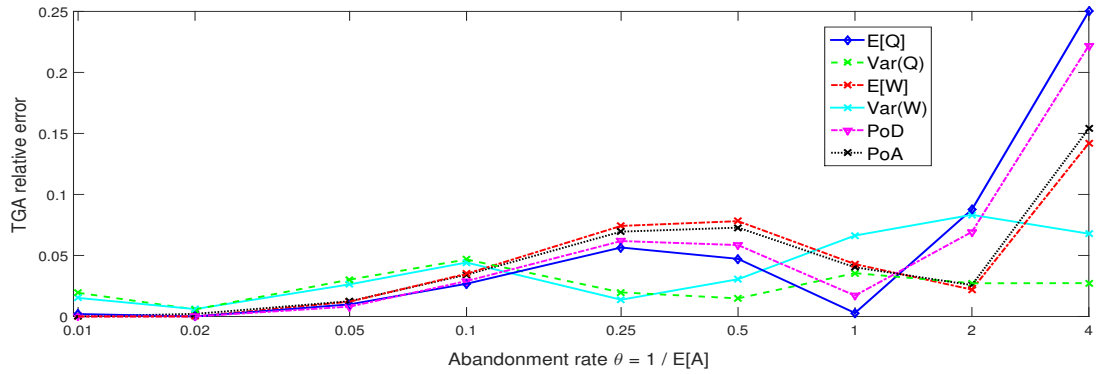


Figure 3.3: The relative error in the approximations of six performance measures as a function of the abandonment rate θ for $0.01 \leq \theta \leq 4.00$ (with θ in log scale) in the $M(1/105)/M(1)/100 + M(1/\theta)$ model having traffic intensity $\rho = 1.05$.

Table 3.1: A comparison of the TGA and DGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $\lambda = 100\rho$ and $\rho = 1.2$ for six values of θ, $0.10 \leq \theta \leq 4.00$.

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	3.00E+2	3.00E+2	same	1.80E+2	1.80E+2	same	1.40E+2	1.40E+2	same
rel. err.		0%			0%			0%	
Var(X)	1.20E+3	1.20E+3	same	4.80E+2	4.80E+2	same	2.39E+2	2.40E+2	same
rel. err.		0%			0%			0%	
E[Q]	2.00E+2	2.00E+2	2.00E+2	8.00E+1	7.99E+1	7.99E+1	4.00E+1	3.99E+1	3.99E+1
rel. err.		0%	0%		0%	0%		0%	0%
Var(Q)	1.20E+3	1.20E+3	1.20E+3	4.80E+2	4.80E+2	4.80E+2	2.38E+2	2.40E+2	2.38E+2
rel. err.		0%	0%		0%	0%		1%	0%
E[W]	1.83E+0	1.82E+0	1.82E+0	7.34E-1	7.29E-1	7.29E-1	3.70E-1	3.65E-1	3.65E-1
rel. err.		0%	0%		1%	1%		1%	1%
Var(W)	1.00E-1	1.00E-1	1.00E-1	4.00E-2	4.00E-2	4.00E-2	1.99E-2	2.00E-2	1.98E-2
rel. err.		0%	0%		0%	0%		1%	0%
P^d	1.00E+0	1.00E+0	same	1.00E+0	1.00E+0	same	9.97E-1	9.95E-1	same
rel. err.		0%			0%			0%	
P^a	1.67E-1	1.66E-1	same	1.67E-1	1.66E-1	same	1.67E-1	1.65E-1	same
rel. err.		0%			1%			1%	

	$\theta = 1$			$\theta = 2$			$\theta = 4$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	1.20E+2	1.20E+2	same	1.10E+2	1.10E+2	same	1.04E+2	1.05E+2	same
rel. err.		0%			0%			1%	
Var(X)	1.20E+2	1.20E+2	same	6.35E+1	5.99E+1	same	3.76E+1	2.99E+1	same
rel. err.		0%			6%			20%	
E[Q]	2.01E+1	1.98E+1	2.00E+1	1.02E+1	9.86E+0	1.02E+1	5.22E+0	4.92E+0	5.47E+0
rel. err.		1%	1%		3%	0%		6%	5%
Var(Q)	1.14E+2	1.20E+2	1.13E+2	5.22E+1	5.99E+1	5.00E+1	2.29E+1	2.99E+1	2.14E+1
rel. err.		5%	1%		15%	4%		31%	7%
E[W]	1.88E-1	1.82E-1	1.83E-1	9.76E-2	9.10E-2	9.43E-2	5.17E-2	4.60E-2	5.09E-2
rel. err.		3%	3%		7%	3%		11%	2%
Var(W)	9.61E-3	1.00E-2	9.40E-3	4.48E-3	5.00E-3	4.19E-3	2.04E-3	2.50E-3	1.81E-3
rel. err.		4%	2%		11%	7%		23%	11%
P^d	9.72E-1	9.66E-1	same	9.06E-1	9.01E-1	same	8.01E-1	8.21E-1	same
rel. err.		1%			1%			2%	
P^a	1.68E-1	1.64E-1	same	1.70E-1	1.65E-1	same	1.74E-1	1.73E-1	same
rel. err.		2%			3%			1%	

Table 3.2: A comparison of the TGA and DGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $\lambda = 100\rho$ and $\rho = 1.05$ for six values of θ, $0.10 \leq \theta \leq 4.00$.

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	1.52E+2	1.50E+2	same	1.22E+2	1.20E+2	same	1.11E+2	1.10E+2	same
rel. err.		1%			2%			1%	
Var(X)	9.25E+2	1.05E+3	same	3.47E+2	4.20E+2	same	1.81E+2	2.10E+2	same
rel. err.		16%			21%			16%	
E[Q]	5.22E+1	4.99E+1	5.08E+1	2.30E+1	1.99E+1	2.17E+1	1.27E+1	9.94E+0	1.21E+1
rel. err.		4%	2%		14%	6%		22%	5%
Var(Q)	8.99E+2	1.05E+3	9.41E+2	3.05E+2	4.20E+2	3.11E+2	1.35E+2	2.10E+2	1.33E+2
rel. err.		19%	7%		37%	2%		56%	1%
E[W]	5.14E-1	4.88E-1	4.96E-1	2.29E-1	1.95E-1	2.12E-1	1.28E-1	9.80E-2	1.18E-1
rel. err.		5%	3%		15%	8%		23%	7%
Var(W)	8.59E-2	1.00E-1	8.97E-2	2.93E-2	4.00E-2	2.97E-2	1.31E-2	2.00E-2	1.27E-2
rel. err.		19%	7%		36%	1%		53%	3%
P^d	9.67E-1	9.39E-1	same	8.90E-1	8.35E-1	same	8.03E-1	7.56E-1	same
rel. err.		3%			6%			6%	
P^a	4.97E-2	4.80E-2	same	5.47E-2	5.09E-2	same	6.04E-2	5.75E-2	same
rel. err.		4%			7%			5%	

	$\theta = 1$			$\theta = 2$			$\theta = 4$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	1.05E+2	1.05E+2	same	1.01E+2	1.02E+2	same	9.86E+1	1.01E+2	same
rel. err.		0%			1%			3%	
Var(X)	1.05E+2	1.05E+2	same	6.85E+1	5.24E+1	same	5.00E+1	2.61E+1	same
rel. err.		0%			23%			48%	
E[Q]	7.03E+0	4.92E+0	7.01E+0	3.88E+0	2.36E+0	4.22E+0	2.12E+0	1.13E+0	2.65E+0
rel. err.		30%	0%		40%	8%		47%	25%
Var(Q)	5.92E+1	1.05E+2	5.71E+1	2.57E+1	5.24E+1	2.50E+1	1.10E+1	2.61E+1	1.13E+1
rel. err.		77%	3%		103%	3%		139%	3%
E[W]	7.22E-2	4.90E-2	6.91E-2	4.09E-2	2.40E-2	4.18E-2	2.32E-2	1.20E-2	2.65E-2
rel. err.		32%	4%		42%	2%		48%	14%
Var(W)	5.88E-3	1.00E-2	5.49E-3	2.64E-3	5.00E-3	2.42E-3	1.18E-3	2.50E-3	1.10E-3
rel. err.		70%	7%		89%	9%		112%	6%
P^d	7.00E-1	6.88E-1	same	5.92E-1	6.33E-1	same	4.87E-1	5.95E-1	same
rel. err.		2%			6%			22%	
P^a	6.70E-2	6.43E-2	same	7.40E-2	7.59E-2	same	8.07E-2	9.31E-2	same
rel. err.		3%			3%			16%	

3.5 Evaluating the Approximations for $G/M/n + GI$ Models

We now start investigating the approximations for non-Markov models, first considering the $G/M/n + GI$ models for which the MSHT limits have been established in [77]. For these and all other non-Markov models, we will use simulation to estimate the exact values of the performance measures. We estimated all the performance measures using 2000 independent replications over the time interval $[0, 100]$, starting empty in each case. To ensure the system is nearly in steady state, we use the data after time $T' = 40$. To construct confidence intervals, on each sample path, we sample values for the performance measures (e.g., queue length and waiting times) at evenly-spaced time points in the interval $[40, 100]$. We describe the detailed simulation methodology in §3.9.

3.5.1 Non-Poisson Arrivals

Theorems 4.1.1-3.2.3 show that the MSHT limits depend on the non-Poisson arrival process only through the asymptotic variance c_λ^2 in the FCLT assumed for the arrival process in (4.2). For a renewal arrival process the parameter c_λ^2 coincides with the scv of an interarrival time, but not more generally.

Renewal processes. To illustrate a non-Markovian arrival process, we consider a non-Poisson renewal process. We let the interarrival-time distribution be lognormal, denoted by $LN(\lambda^{-1}, c_\lambda^2)$, where λ^{-1} is its mean, which is the reciprocal of the fixed arrival rate $\lambda = n\rho = 105$, and c_λ^2 denotes its scv. Recall that an $LN(\lambda^{-1}, c_\lambda^2)$ random variable is distributed as $e^{\hat{\mu}+\hat{\sigma}\mathcal{Z}}$, where $\mathcal{Z}$ is a standard Gaussian random variable, $\hat{\mu} = \log\left(\lambda^{-1}/\sqrt{1+c_\lambda^2}\right)$ and $\hat{\sigma} = \sqrt{\log\left(1+c_\lambda^2\right)}$.

Table 3.3 and Figure 3.4 show the experimental results for five values of c_λ^2 with $0.25 \le c_\lambda^2 \le 4.0$. Table 3.3 also shows the experimental results for the case of a Poisson (M) arrival process, which is interesting, because the $LN(\lambda^{-1}, 1)$ distribution is different from the corresponding $M(\lambda^{-1})$ exponential distribution, even though they have the same mean and variance.

Table 3.3 compares the Gaussian approximations to simulations for the $LN/M/n+M$ queueing model with $(n, \rho, \theta) = (100, 1.05, 0.5)$. First, Table 3.3 shows that the interarrival time distribution has a significant impact upon performance. For example, $E[Q]$ increases from 11.5 to 16.1 as c_λ^2 increases from 0.25 to 4.0. Moreover, by comparing the results for $M(\lambda^{-1})$ and $LN(\lambda^{-1}, 1)$, where the approximations coincide, we see that the interarrival-time distribution matters little beyond its mean and variance, just as predicted.

Second, Table 3.3 shows that, just like Table 3.2, TGA performs consistently well, whereas DGA does not. Figure 3.5 adds to the story by showing that the full distributions of the queue length Q_n and potential waiting time W_n are well approximated by TGA as well.

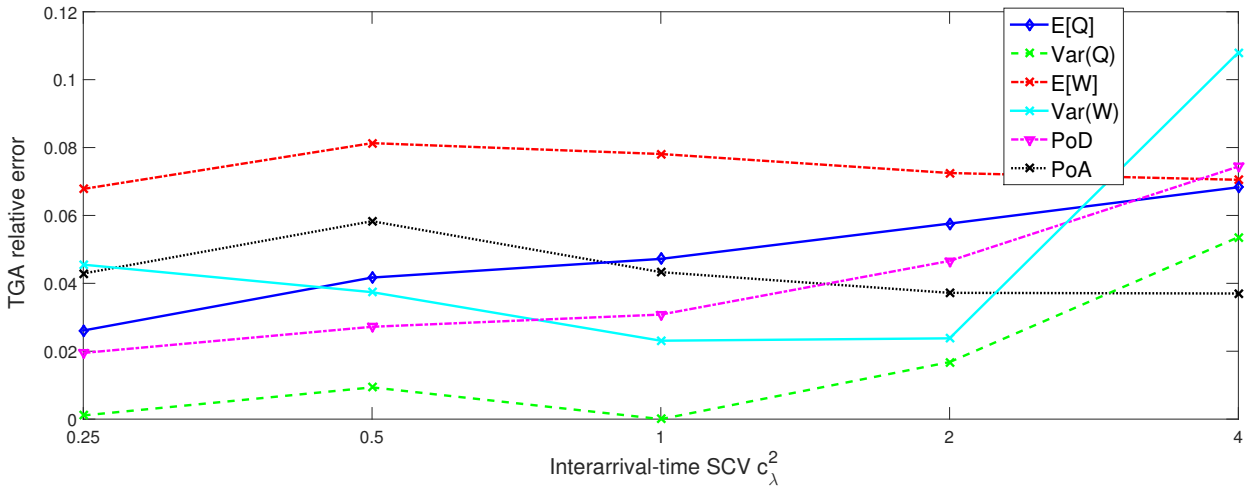


Figure 3.4: The relative error in the approximations of six performance measures as a function of the interarrival-time scv c_λ^2 for $0.25 \leq c_\lambda^2 \leq 4.0$ (with c_λ^2 in log scale) in the $LN(1/105, c_\lambda^2)/M(1)/100 + M(2)$ model.

Markov-Modulated Poisson processes (MMPP's) We next consider an alternative non-Poisson arrival process: Markov-modulated Poisson process (MMPP), which is a Poisson process having a random rate modulated by a continuous-time Markov chain (CTMC) $\{\Gamma(t), t \geq\}$, e.g., see [29]. Specifically, we can construct an MMPP by composition:

$$N_n(t) \equiv M\left(n\rho \int_0^t \alpha(\Gamma(u))\, du\right),$$

where M is a rate-1 Poisson process, and the random rate $\alpha(\Gamma(t)) = \alpha_i$ when the CTMC $\Gamma(t) = i$. In particular, we now consider an MMPP with an underlying CTMC $\{\Gamma(t), t \geq 0\}$ that is a birth-and-death process having three states 0, 1 and 2. Let CTMC-dependent arrival rate be $(\alpha_0, \alpha_1, \alpha_2) = (3, 1, 1/3)$. The long-run rate of the MMPP arrival process is

$$\lambda_n = n\rho\lambda^*, \quad \lambda^* \equiv \lim_{t\to\infty} t^{-1} \int_0^t \alpha(\Gamma(u))\, du = \sum_{j=0}^{2} \pi_j\, \alpha_j,$$

where $\pi \equiv (\pi_0, \pi_1, \pi_2)$ is the steady state distribution for the CTMC. We consider four sets of birth and deaths rates: (i) $\hat{\lambda}_0 = 20/81$, $\hat{\lambda}_1 = 5/27$, $\hat{\mu}_1 = \hat{\mu}_2 = 10/81$, (ii) $\hat{\lambda}_0 = 20/27$, $\hat{\lambda}_1 = 5/9$, $\hat{\mu}_1 = \hat{\mu}_2 = 10/27$, (iii) $\hat{\lambda}_0 = 20/9$, $\hat{\lambda}_1 = 5/3$, $\hat{\mu}_1 = \hat{\mu}_2 = 10/9$, and (iv) $\hat{\lambda}_0 = 40/9$, $\hat{\lambda}_1 = 10/3$, $\hat{\mu}_1 = \hat{\mu}_2 = 20/9$, which yield the same steady state $\pi = (1/6, 1/3, 1/2)$ and asymptotic rate $\lambda^* = 1$, but different asymptotic variability parameter: (i) $c_\lambda^2 = 10$, (ii) $c_\lambda^2 = 4$, (iii) $c_\lambda^2 = 2$ and (iv) $c_\lambda^2 = 1.5$,

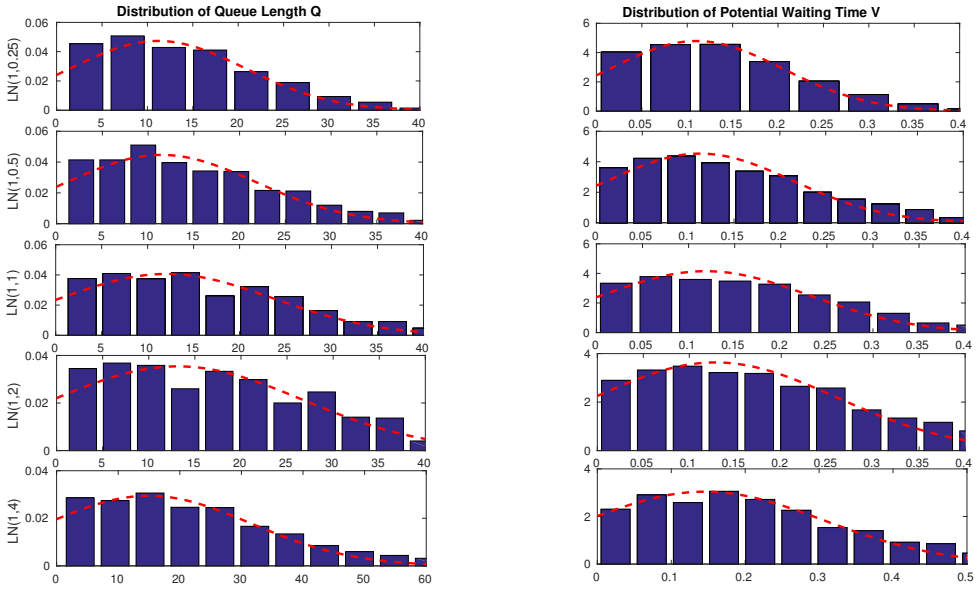


Figure 3.5: Simulation estimates (histograms) of the TGA approximating distributions of the steady-state queue length Q_n (left) and waiting time V_n (right) for the $LN(\lambda^{-1}, c_\lambda^2)/M(1)/100 + M(2)$ model with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$.

where

$$c_\lambda^2 = 1 + c_C^2, \quad \text{where} \quad c_C^2 \equiv 2 \sum_{j=1}^{2} \frac{1}{\hat{\lambda}_j \pi_j} \left(\sum_{i=1}^{j} \pi_i \left(\alpha_i - \lambda_C \right) \right)^2 .$$

See Proposition 1 in [111] and also see [39].

We denote by $MMPP(\lambda^{-1}, c_\lambda^2)$ our MMPP arrival process having rate λ and variability parameter c_λ^2. Table 3.4 compares TGA to the simulation results for $MMPP/M/n + M$ models with $\lambda = 105$, $n = 100$, $\mu = 1$, $\theta = 0.5$, and different variability parameters $c_\lambda^2 = 1.5, 2, 4$ and 10.

3.5.2 Non-Exponential Abandonment

It has been shown that the full abandonment distribution has a significant impact on the performance; e.g., see [107, 117]. We confirm that here when we study how our TGAs works in $M/M/n + GI$ models using different abandonment distributions, again using the lognormal distribution. Figure 3.6 and Tables 3.5 and 3.6 compare the DGAs and TGAs to the simulations of the $M/M/n + LN(2, c_{ab}^2)$ model with the same parameter triple $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$ we have been using, where scv of abandonment distribution c_{ab}^2 ranges from 0.25 to 4.0. Paralleling Table 3.3, we add a column for the results of Erlang-A models.

Figure 3.6 and Tables 3.5 and 3.6 show that TGA is again consistently quite accurate for

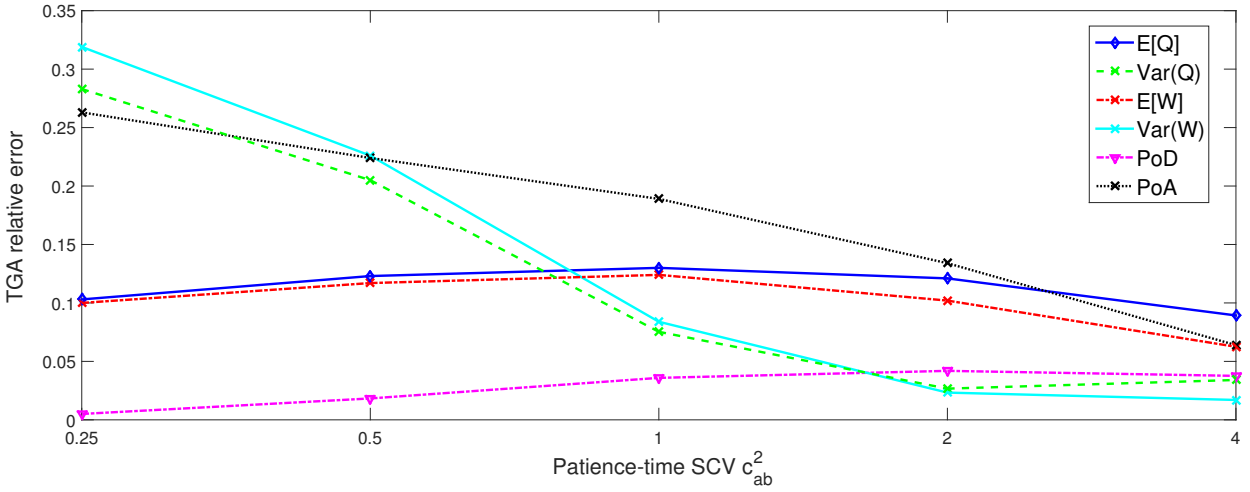


Figure 3.6: The relative errors in the approximations of six performance measures as a function of the patience-time scv c_{ab}^2 for $0.25 \leq c_{ab}^2 \leq 4.0$ (with c_{ab}^2 in log scale) in the $M(1/105)/M(1)/100 + LN(2, c_{ab}^2)$ model.

all the first-moment measures (mean and probability) in all cases, but the accuracy degrades to $20 - 30\%$ for the variances when c_{ab}^2 is low. Moreover, by comparing the results for the $M/M/n + LN(2, 1)$ and $M/M/n + M(2)$ models in Table 3.6, we can see that the patience distribution has a significant impact on the queueing system beyond its mean and variance, unlike the interarrival-time distribution. Approximations of distributions of Q_n and W_n are given in Figure 3.7.

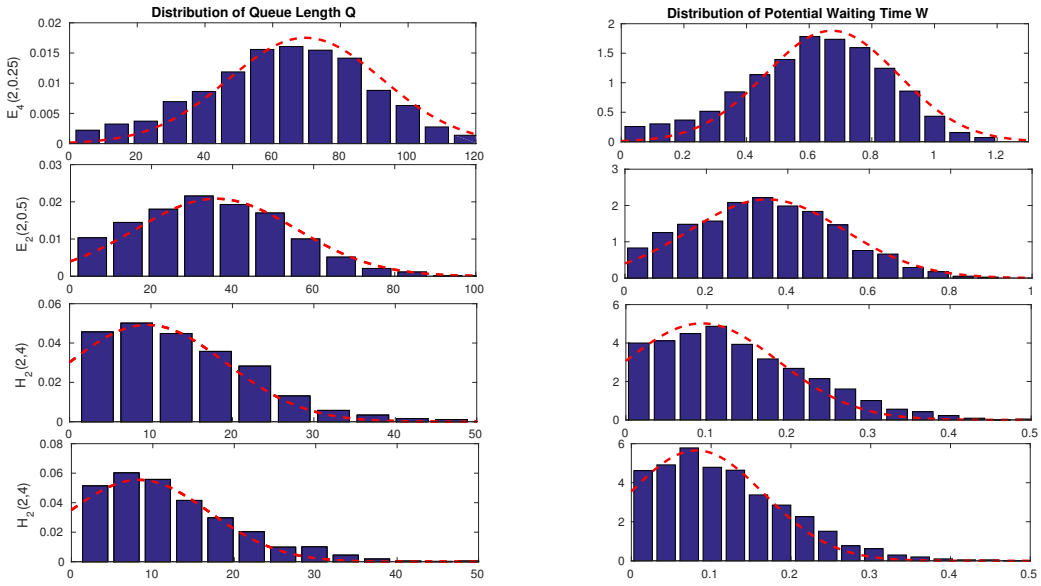


Figure 3.7: Simulation estimates (histograms) of the TGA approximating distributions of the steady-state queue length Q_n (left) and waiting time V_n (right) in the $M/M/100 + LN(\theta^{-1}, c_{ab}^2)$ model with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$.

Table 3.3: A comparison of the TGA and DGA approximations to simulation estimates in the $LN(\lambda^{-1}, c_\lambda^2)/M(1)/100 + M(2)$ model with i.i.d. lognormal interarrival times and $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$ for six values of the interarrival time scv c_λ^2, $0.25 \leq c_\lambda^2 \leq 4.00$.

	$c_\lambda^2 = 0.25$			$c_\lambda^2 = 0.5$			$c_\lambda^2 = 1$		
Perf.	Sim	DGA	TGA	Sim	DGA	TGA	Sim	DGA	TGA
E[X]	1.11E+2	1.10E+2	same	1.11E+2	1.10E+2	same	1.11E+2	1.10E+2	same
rel. err.	±1.16E-1	1%		±1.25E-1	1%		±1.44E-1	1%	
Var(X)	1.16E+2	1.31E+2	same	1.38E+2	1.57E+2	same	1.80E+2	2.10E+2	same
rel. err.	±2.62E+1	13%		±2.83E+1	14%		±3.27E+1	17%	
E[Q]	1.15E+1	9.94E+0	1.12E+1	1.20E+1	9.94E+0	1.15E+1	1.27E+1	9.94E+0	1.21E+1
rel. err.	±1.04E-1	14%	3%	±1.10E-1	17%	5%	±1.24E-1	22%	5%
Var(Q)	9.22E+1	1.31E+2	9.23E+1	1.07E+2	1.57E+2	1.06E+2	1.33E+2	2.10E+2	1.33E+2
rel. err.	±3.18E+0	42%	0%	±3.62E+0	48%	0%	±4.46E+0	58%	0%
E[W]	1.18E-1	9.80E-2	1.10E-1	1.23E-1	9.80E-2	1.13E-1	1.28E-1	9.80E-2	1.18E-1
rel. err.	±1.05E-3	17%	7%	±1.10E-3	20%	8%	±1.22E-3	23%	7%
Var(W)	9.46E-3	1.29E-2	9.03E-3	1.07E-2	1.52E-2	1.03E-2	1.30E-2	2.00E-2	1.27E-2
rel. err.	±1.05E-3	36%	5%	±1.10E-3	42%	4%	±1.22E-3	54%	2%
P^d	8.22E-1	8.06E-1	same	8.08E-1	7.86E-1	same	7.80E-1	7.56E-1	same
rel. err.	±3.02E-3	2%		±3.15E-3	3%		±3.31E-3	3%	
$P^{d,C}$	8.41E-1	8.06E-1	same	8.24E-1	7.86E-1	same	8.08E-1	7.56E-1	same
rel. err.	±3.00E-3	4%		±3.10E-3	5%		±3.20E-3	6%	
P^a	5.60E-2	5.36E-2	same	5.83E-2	5.49E-2	same	6.01E-2	5.75E-2	same
rel. err.	±6.45E-4	4%		±6.70E-4	6%		±6.99E-4	4%	

	$M, c_\lambda^2 = 1$			$c_\lambda^2 = 2$			$c_\lambda^2 = 4$		
Perf.	Exact	DGA	TGA	Sim	DGA	TGA	Sim	DGA	TGA
E[X]	1.11E+2	1.10E+2	same	1.12E+2	1.10E+2	same	1.13E+2	1.10E+2	same
rel. err.		1%		±1.69E-1	2%		±2.17E-1	3%	
Var(X)	1.81E+2	2.10E+2	same	2.60E+2	3.15E+2	same	4.08E+2	5.25E+2	same
rel. err.		16%		±3.88E+1	21%		±5.02E+1	29%	
E[Q]	1.27E+1	9.94E+0	1.21E+1	1.39E+1	9.94E+0	1.31E+1	1.61E+1	9.94E+0	1.50E+1
rel. err.		22%	5%	±1.40E-1	29%	6%	±1.73E-1	38%	7%
Var(Q)	1.35E+2	2.10E+2	1.33E+2	1.79E+2	3.15E+2	1.82E+2	2.61E+2	5.25E+2	2.75E+2
rel. err.		56%	1%	±5.81E+0	76%	2%	±8.52E+0	101%	6%
E[W]	1.28E-1	9.80E-2	1.18E-1	1.38E-1	9.80E-2	1.28E-1	1.56E-1	9.80E-2	1.45E-1
rel. err.		23%	7%	±1.36E-3	29%	7%	±1.63E-3	37%	7%
Var(W)	1.31E-2	2.00E-2	1.27E-2	1.68E-2	2.95E-2	1.72E-2	2.32E-2	4.86E-2	2.57E-2
rel. err.		53%	3%	±1.36E-3	76%	3%	±1.63E-3	110%	11%
P^d	8.03E-1	7.56E-1	same	7.51E-1	7.16E-1	same	7.26E-1	6.72E-1	same
rel. err.		6%		±3.46E-3	5%		±3.76E-3	8%	
$P^{d,C}$	8.03E-1	7.56E-1	same	7.81E-1	7.16E-1	same	7.70E-1	6.72E-1	same
rel. err.		6%		±3.50E-3	8%		±3.50E-3	7%	
P^a	6.04E-2	5.75E-2	same	6.46E-2	6.22E-2	same	7.29E-2	7.02E-2	same
rel. err.		5%		±7.69E-4	4%		±8.58E-4	4%	

Table 3.4: A comparison of the TGA approximations to simulation estimates in the $MMPP(\lambda^{-1}, c_\lambda^2)/M/100 + M(\theta^{-1})$ model with $(\lambda, \rho, \theta) = (105, 1.05, 0.5)$ for four values of the arrival process variability parameter c_λ^2 in (4.2), $1.5 \leq c_\lambda^2 \leq 10.0$.

	$c_\lambda^2 = 1.5$		$c_\lambda^2 = 2$		$c_\lambda^2 = 4$		$c_\lambda^2 = 10$	
Perf.	Sim	TGA	Sim	TGA	Sim	TGA	Sim	TGA
E[X]	1.12E+2	1.10E+2	1.12E+2	1.10E+2	1.13E+2	1.09E+2	1.15E+2	1.10E+2
rel. err.	±1.62E-1	2%	±1.77E-1	2%	±2.16E-1	3%	±3.20E-1	5%
Var(X)	2.22E+2	2.62E+2	2.65E+2	3.15E+2	4.30E+2	5.25E+2	9.19E+2	1.15E+3
rel. err.	±3.71E+1	18%	±4.08E+1	19%	±5.15E+1	22%	±8.05E+1	26%
E[Q]	1.34E+1	1.26E+1	1.38E+1	1.31E+1	1.60E+1	1.50E+1	2.06E+1	1.91E+1
rel. err.	±1.37E-1	6%	±1.48E-1	5%	±1.78E-1	7%	±2.55E-1	7%
Var(Q)	1.62E+2	1.58E+2	1.88E+2	1.82E+2	2.93E+2	2.75E+2	6.02E+2	5.35E+2
rel. err.	±5.46E+0	2%	±6.21E+0	3%	±9.55E+0	6%	±1.95E+1	11%
E[W]	1.34E-1	1.24E-1	1.37E-1	1.28E-1	1.55E-1	1.45E-1	1.91E-1	1.85E-1
rel. err.	±1.33E-3	8%	±1.43E-3	6%	±1.67E-3	6%	±2.26E-3	3%
Var(W)	1.53E-2	1.50E-2	1.74E-2	1.72E-2	2.56E-2	2.57E-2	4.68E-2	4.94E-2
rel. err.	±5.16E-4	2%	±5.76E-4	1%	±8.21E-4	0%	±1.50E-3	6%
P^d	7.83E-1	7.33E-1	7.61E-1	7.16E-1	7.22E-1	6.72E-1	6.61E-1	6.18E-1
rel. err.	±3.33E-3	6%	±3.50E-3	6%	±3.55E-3	7%	±4.00E-3	7%
$P^{d,C}$	7.85E-1	7.33E-1	7.76E-1	7.16E-1	7.44E-1	6.72E-1	7.20E-1	6.18E-1
rel. err.	±4.80E-3	7%	±4.90E-3	8%	±5.20E-3	10%	±5.40E-3	14%
P^a	6.32E-2	5.82E-2	6.44E-2	6.02E-2	7.15E-2	6.72E-2	8.61E-2	8.28E-2
rel. err.	±7.49E-4	8%	±7.86E-4	7%	±8.71E-4	6%	±1.10E-3	4%

Table 3.5: A comparison of the TGA and DGA approximations to simulation estimates in the $M/M/100 + LN(\theta^{-1}, c_{ab}^2)$ model with with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$ for four values of the patience scv c_{ab}^2, $0.25 \leq c_{ab}^2 \leq 4.00$.

	$LN(2, 0.25)$			$LN(2, 0.5)$		
Perf.	Sim	DGA	TGA	Sim	DGA	TGA
E[X]	1.77E+2	1.85E+2	same	1.52E+2	1.59E+2	same
rel. err.	±4.92E-1	4%		±3.94E-1	4%	
Var(X)	6.39E+2	4.54E+2	same	5.22E+2	4.04E+2	same
rel. err.	±1.63E+2	29%		±1.15E+2	23%	
E[Q]	7.67E+1	8.46E+1	8.46E+1	5.21E+1	5.85E+1	5.85E+1
rel. err.	±4.89E-1	10%	10%	±3.88E-1	12%	12%
Var(Q)	6.33E+2	4.54E+2	4.54E+2	5.07E+2	4.04E+2	4.03E+2
rel. err.	±6.55E+1	28%	28%	±3.72E+1	20%	21%
E[W]	7.39E-1	8.13E-1	8.13E-1	5.05E-1	5.64E-1	5.64E-1
rel. err.	±4.54E-3	10%	10%	±3.64E-3	12%	12%
Var(W)	5.42E-2	3.69E-2	3.69E-2	4.46E-2	3.45E-2	3.45E-2
rel. err.	±5.80E-3	32%	32%	±3.30E-3	22%	23%
P^d	9.95E-1	1.00E+0	same	9.81E-1	9.99E-1	same
rel. err.	±8.22E-4	1%		±1.47E-3	2%	
P^a	4.83E-2	6.10E-2	same	4.90E-2	6.00E-2	same
rel. err.	±8.35E-4	26%		±8.02E-4	22%	

	$LN(2, 2)$			$LN(2, 4)$		
Perf.	Sim	DGA	TGA	Sim	DGA	TGA
E[X]	1.18E+2	1.21E+2	same	1.10E+2	1.11E+2	same
rel. err.	±1.75E-1	2%		±1.13E-1	1%	
Var(X)	2.25E+2	2.23E+2	same	1.43E+2	1.42E+2	same
rel. err.	±4.14E+1	1%		±2.49E+1	0%	
E[Q]	1.90E+1	2.07E+1	2.13E+1	1.12E+1	1.10E+1	1.22E+1
rel. err.	±1.58E-1	9%	12%	±9.19E-2	1%	9%
Var(Q)	1.87E+2	2.23E+2	1.92E+2	9.96E+1	1.42E+2	1.03E+2
rel. err.	±6.78E+0	19%	3%	±2.74E+0	43%	3%
E[W]	1.87E-1	2.01E-1	2.06E-1	1.12E-1	1.08E-1	1.19E-1
rel. err.	±1.52E-3	7%	10%	±8.96E-4	3%	6%
Var(W)	1.72E-2	2.02E-2	1.76E-2	9.43E-3	1.31E-2	9.59E-3
rel. err.	±6.30E-4	17%	2%	±2.62E-4	39%	2%
P^d	8.84E-1	9.21E-1	same	7.97E-1	8.27E-1	same
rel. err.	±2.85E-3	4%		±3.07E-3	4%	
P^a	5.45E-2	6.18E-2	same	6.10E-2	6.49E-2	same
rel. err.	±7.34E-4	13%		±7.19E-4	6%	

Table 3.6: The impact of the abandonment distribution beyond its mean and variance: a performance comparison between $M(\lambda^{-1})/M/n + M(2)$ and $M(\lambda^{-1})/M/n + LN(2,1)$ models, where $(\lambda, \rho, n) = (105, 1.05, 100)$.

	$M(2)$			$LN(2,1)$		
Perf.	Exact	DGA	TGA	Sim	DGA	TGA
E[X]	1.05E+2	1.05E+2	same	1.32E+2	1.36E+2	same
rel. err.		0%		±2.79E-1	3%	
Var(X)	1.05E+2	1.05E+2	same	3.58E+2	3.18E+2	same
rel. err.		0%		±7.27E+1	11%	
E[Q]	7.03E+0	4.92E+0	7.01E+0	3.24E+1	3.65E+1	3.66E+1
rel. err.		30%	0%	±2.67E-1	13%	13%
Var(Q)	5.92E+1	1.05E+2	5.71E+1	3.32E+2	3.18E+2	3.07E+2
rel. err.		77%	3%	±1.74E+1	4%	8%
E[W]	7.22E-2	4.90E-2	6.91E-2	3.15E-1	3.53E-1	3.54E-1
rel. err.		32%	4%	±2.54E-3	12%	12%
Var(W)	5.88E-3	1.00E-2	5.49E-3	2.98E-2	2.82E-2	2.73E-2
rel. err.		70%	7%	±2.54E-3	6%	9%
P^d	7.00E-1	6.88E-1	same	9.48E-1	9.82E-1	same
rel. err.		2%		±2.19E-3	4%	
P^a	6.70E-2	6.43E-2	same	5.08E-2	6.04E-2	same
rel. err.		3%		±7.74E-4	19%	

3.6 Non-Exponential Service

3.6.1 Refined Gaussian Approximations for the $M/GI/n+M$ Model

We now evaluate the heuristic approximation TGA-G developed in §3.3.3. We let the service-time distribution be phase-type, denoted by PH with fixed mean $1/\mu = 1$ and scv c_s^2 ranging in $[0.25, 4]$. To be specific, for cases with $c_s^2 = 0.25, 0.5 < 1$, we used Erlang 4 (E_4) and Erlang 2 (E_2) distribution and for cases with $c_s^2 = 2, 4 > 1$, we used the two-phase hyperexponential distribution (H_2) with balanced means, see [109] for more details.

Table 3.7 compares the approximations for all three Gaussian approximations for the $M/PH/n+M$ model. Table 3.7 shows that TGA-G outperforms TGA, while both are far better than DGA. The new approximation TGA-G is especially important for the variances, which depend quite strongly on the service-time distribution, unlike the means. Figure 3.8 shows that the TGA-G approximation accuracy tends to be independent of the service-time distribution scv c_s^2, consistent with the observations in [107, 117]. Figure 3.9 shows that corresponding TGA-G approximations of the distributions remain good.

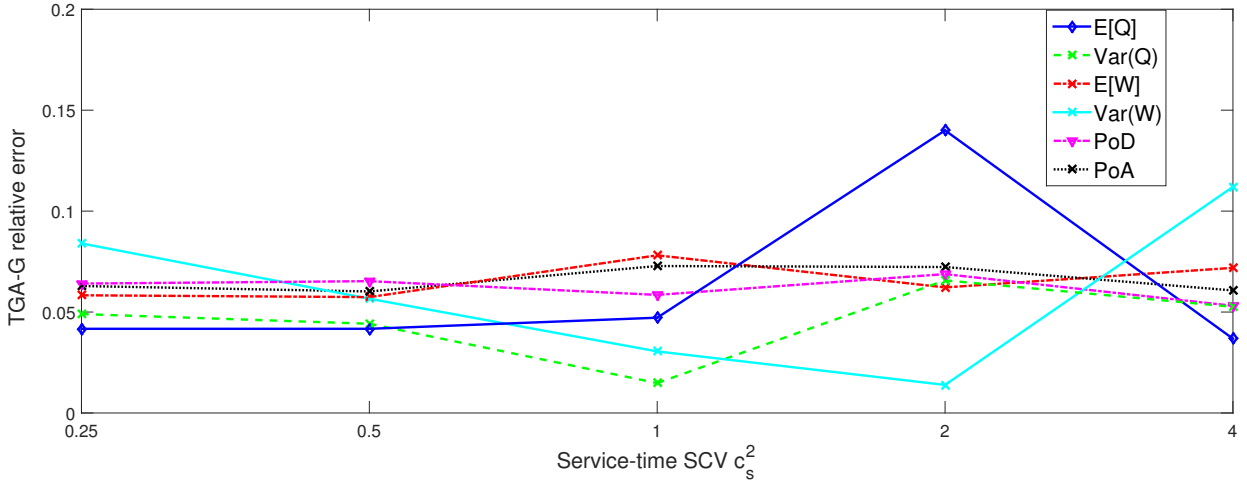


Figure 3.8: The relative error in the TGA-G approximations of six performance measures as a function of the service-time scv c_s^2 for $0.25 \leq c_s^2 \leq 4.0$ (with c_s^2 in log scale) in the $M(1/105)/PH(1, c_s^2)/100 + M(2)$ model

In Table 3.8 we consider the $M/H_2/n+M$ model, with hyperexponential service according to $H_2(1/\mu, c_s^2)$, which denotes mean $1/\mu = 1$ and SCV $c_s^2 = 2$, for a range of abandonment rates with $0.25 \leq \theta \leq 2.0$. We observe that the performance of TGA-G is acceptable. However, just as for the $M/M/n+M$ model, the approximation accuracy degrades when θ increases; see the appendix for

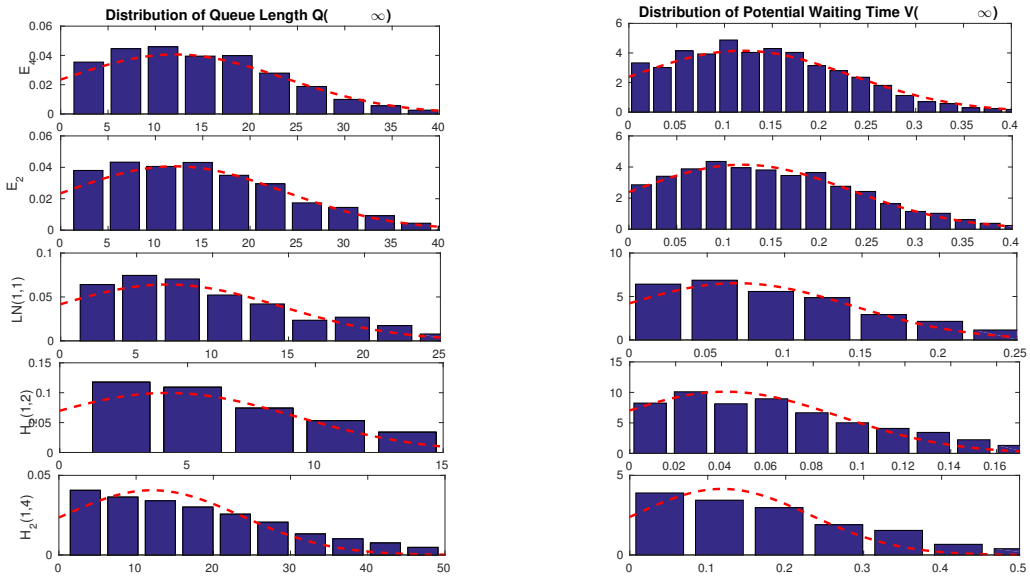


Figure 3.9: Simulation estimates (histograms) of the TGA-G approximating distributions of the steady-state queue length Q_n (left) and waiting time V_n (right) in the $M(105^{-1})/PH/100 + M(2)$ model for five service distributions with $0.25 \leq c_s^2 \leq 4.0$.

more examples.

3.6.2 The General $GI/GI/n + GI$ Model

We have also considered examples in which all three stochastic components of the $G/GI/n+GI$ model are non-exponential. We illustrate some of these now. Table 3.9 shows comparisons of TGA and TGA-G to simulation estimates for the $H_2/PH/n+H_2$ model, having a renewal arrival process with H_2 inter-arrival times, PH service times (in the settings of Table 3.7), $n = 100$ servers, and H_2 patience times. We fix the scv's for arrival and patience times at $c_\lambda^2 = c_{ab}^2 = 2$ and consider a range of service scv: $0.25 \leq c_s^2 \leq 4.0$.

Just as in Table 3.7, Table 3.9 shows that the mean values such as $E[Q]$ and the probabilities such as P^d are relatively insensitive to the service-time distribution beyond its mean; these entries differ little in the three cases. However, as before, we see differences in the variances. Table 3.9 shows that both TGA and TGA-G are effective for the mean values such as $E[Q]$ and the probabilities such as P^d, but TGA-G provides significant improvement for the variances. Additional results on other combinations of arrival processes, patience times and service times are given in Table A.3 in the appendix.

Table 3.7: A comparison of the TGA-G, TGA and DGA approximations to simulation estimates in the $M(\lambda^{-1})/PH(1, c_s^2)/100 + M(\theta^{-1})$ model with $(\lambda, \rho, \theta) = (100, 1.05, 0.5)$ for four different phase-type (Ph) service distributions characterized by their scv c_s^2, $0.25 \leq c_s^2 \leq 4.00$.

	$c_s^2 = 0.25$				$c_s^2 = 0.5$			
Perf. Meas.	Sim.	DGA	TGA	TGA-G	Sim.	DGA	TGA	TGA-G
E[X]	1.11E+2	1.10E+2	same	same	1.11E+2	1.10E+2	same	same
rel. err.	±1.21E-1	1%			±1.30E-1	1%		
Var(X)	1.34E+2	1.60E+2	same	same	1.50E+2	1.81E+2	same	same
rel. err.	±2.72E+1	19%			±2.94E+1	20%		
E[Q]	1.20E+1	9.94E+0	1.21E+1	1.15E+1	1.22E+1	9.94E+0	1.21E+1	1.17E+1
rel. err.	±1.06E-1	17%	1%	4%	±1.13E-1	19%	1%	4%
Var(Q)	1.02E+2	1.60E+2	1.33E+2	1.07E+2	1.13E+2	1.81E+2	1.33E+2	1.18E+2
rel. err.	±3.45E+0	57%	30%	6%	±3.83E+0	60%	18%	5%
E[W]	1.20E-1	9.80E-2	1.18E-1	1.13E-1	1.22E-1	9.80E-2	1.18E-1	1.15E-1
rel. err.	±1.01E-3	18%	1%	6%	±1.10E-3	20%	3%	6%
Var(W)	9.41E-3	1.50E-2	1.27E-2	1.02E-2	1.06E-2	1.71E-2	1.27E-2	1.12E-2
rel. err.	±3.20E-4	59%	35%	8%	±3.64E-4	60%	19%	6%
P^d	8.42E-1	7.88E-1	same	same	8.27E-1	7.73E-1	same	same
rel. err.	±2.67E-3	6%			±2.81E-3	6%		
P^a	5.72E-2	5.36E-2	same	same	5.81E-2	5.46E-2	same	same
rel. err.	±7.60E-4	6%			±7.81E-4	6%		

	$c_s^2 = 2$				$c_s^2 = 4$			
Perf. Meas.	Sim.	DGA	TGA	TGA-G	Sim.	DGA	TGA	TGA-G
E[X]	1.01E+2	1.02E+2	same	same	1.12E+2	1.10E+2	same	same
rel. err.	±6.90E-2	1%			±2.17E-1	2%		
Var(X)	7.32E+1	6.28E+1	same	same	2.57E+2	3.10E+2	same	same
rel. err.	±1.38E+1	14%			±5.03E+1	20%		
E[Q]	3.93E+0	2.36E+0	4.22E+0	4.48E+0	1.36E+1	9.94E+0	1.21E+1	1.31E+1
rel. err.	±3.68E-2	40%	7%	14%	±1.84E-1	27%	11%	4%
Var(Q)	2.74E+1	6.28E+1	2.50E+1	2.92E+1	1.90E+2	3.10E+2	1.33E+2	1.80E+2
rel. err.	±5.49E-1	129%	9%	7%	±7.74E+0	63%	30%	5%
E[W]	4.18E-2	2.40E-2	4.18E-2	4.44E-2	1.39E-1	9.80E-2	1.18E-1	1.29E-1
rel. err.	±3.85E-4	43%	0%	6%	±1.90E-3	29%	15%	7%
Var(W)	2.88E-3	6.03E-3	2.42E-3	2.84E-3	1.97E-2	3.00E-2	1.27E-2	1.75E-2
rel. err.	±5.93E-5	109%	16%	1%	±8.32E-4	53%	35%	11%
P^d	5.81E-1	6.21E-1	same	same	7.54E-1	7.14E-1	same	same
rel. err.	±3.36E-3	7%			±4.50E-3	5%		
P^a	7.47E-2	8.01E-2	same	same	6.43E-2	6.04E-2	same	same
rel. err.	±9.21E-4	7%			±1.06E-3	6%		

Table 3.8: A comparison of the TGA-G approximations to simulation estimates in the $M(\lambda^{-1})/H_2(1,2)/100 + M(\theta^{-1})$ model with $\lambda = 100$, $\rho = 1.05$ and five different abandonment rates θ, $0.1 \leq \theta \leq 2.0$.

	$\theta = 0.1$		$\theta = 0.25$		$\theta = 0.5$		$\theta = 1$		$\theta = 2$	
Perf.	Sim	TGA-G	Sim	TGA-G	Sim	TGA-G	Sim	TGA-G	Sim	TGA-G
E[X]	1.53E+2	1.50E+2	1.23E+2	1.20E+2	1.12E+2	1.10E+2	1.05E+2	1.05E+2	1.01E+2	1.02E+2
rel. err.	±8.30E-1	2%	±3.20E-1	2%	±1.75E-1	2%	±1.07E-1	0%	±6.90E-2	1%
Var(X)	1.24E+3	1.26E+3	4.31E+2	5.03E+2	2.15E+2	2.51E+2	1.18E+2	1.26E+2	7.32E+1	6.28E+1
rel. err.	±2.69E+2	2%	±8.23E+1	17%	±4.00E+1	17%	±2.27E+1	6%	±1.38E+1	14%
E[Q]	5.38E+1	5.12E+1	2.37E+1	2.22E+1	1.32E+1	1.25E+1	7.19E+0	7.35E+0	3.93E+0	4.48E+0
rel. err.	±8.16E-1	5%	±3.00E-1	6%	±1.49E-1	5%	±7.62E-2	2%	±3.68E-2	14%
Var(Q)	1.19E+3	1.09E+3	3.76E+2	3.57E+2	1.59E+2	1.53E+2	6.67E+1	6.62E+1	2.74E+1	2.92E+1
rel. err.	±1.05E+2	9%	±1.91E+1	5%	±5.78E+0	4%	±1.84E+0	1%	±5.49E-1	7%
E[W]	5.32E-1	5.01E-1	2.37E-1	2.18E-1	1.34E-1	1.23E-1	7.42E-2	7.26E-2	4.18E-2	4.44E-2
rel. err.	±8.06E-3	6%	±3.00E-3	8%	±1.51E-3	8%	±7.83E-4	2%	±3.85E-4	6%
Var(W)	1.16E-1	1.05E-1	3.71E-2	3.43E-2	1.60E-2	1.47E-2	6.81E-3	6.40E-3	2.88E-3	2.84E-3
rel. err.	±1.03E-2	10%	±1.90E-3	8%	±5.93E-4	8%	±1.94E-4	6%	±5.93E-5	1%
P^d	9.48E-1	9.20E-1	8.64E-1	8.13E-1	7.82E-1	7.36E-1	6.80E-1	6.72E-1	5.81E-1	6.21E-1
rel. err.	±2.84E-3	3%	±3.47E-3	6%	±3.72E-3	6%	±3.74E-3	1%	±3.36E-3	7%
P^a	5.12E-2	4.83E-2	5.60E-2	5.19E-2	6.29E-2	5.79E-2	6.87E-2	6.72E-2	7.47E-2	8.01E-2
rel. err.	±9.66E-4	6%	±9.03E-4	7%	±9.25E-4	8%	±9.48E-4	2%	±9.21E-4	7%

Table 3.9: $H_2(\lambda^{-1},2)/PH/100 + H_2(1/\theta,2)$ with $(\lambda,\rho,\theta) = (100, 1.05, 0.5)$.

	$c_s^2 = 0.25$			$c_s^2 = 0.5$		
Perf.	Sim	TGA	TGA-G	Sim	TGA	TGA-G
E[X]	1.09E+2	1.07E+2	1.07E+2	1.09E+2	1.07E+2	1.07E+2
rel. err.	±1.21E-1	1%	1%	±1.31E-1	1%	1%
Var(X)	1.84E+2	2.37E+2	2.00E+2	1.95E+2	2.37E+2	2.15E+2
rel. err.	±2.64E+1	29%	8%	±2.89E+1	22%	11%
E[Q]	1.07E+1	1.05E+1	1.01E+1	1.09E+1	1.05E+1	1.03E+1
rel. err.	±9.32E-2	2%	6%	±1.02E-1	3%	5%
Var(Q)	1.12E+2	1.29E+2	1.12E+2	1.21E+2	1.29E+2	1.19E+2
rel. err.	±3.01E+0	15%	0%	±3.44E+0	7%	1%
E[W]	1.06E-1	1.03E-1	9.84E-2	1.07E-1	1.03E-1	1.00E-1
rel. err.	±8.77E-4	2%	7%	±9.71E-4	4%	7%
Var(W)	1.00E-2	1.22E-2	1.05E-2	1.10E-2	1.22E-2	1.12E-2
rel. err.	±2.65E-4	22%	5%	±3.09E-4	12%	3%
P^d	7.52E-1	6.88E-1	7.04E-1	7.42E-1	6.88E-1	6.97E-1
rel. err.	±3.05E-3	9%	6%	±3.18E-3	7%	6%
$P^{d,C}$	7.67E-1	6.88E-1	7.04E-1	7.56E-1	6.88E-1	6.97E-1
rel. err.	±3.04E-3	9%	6%	±3.10E-3	7%	6%
P^a	6.54E-2	6.34E-2	6.09E-2	6.66E-2	6.34E-2	6.19E-2
rel. err.	±8.09E-4	3%	7%	±8.67E-4	5%	7%

	$c_s^2 = 2$			$c_s^2 = 4$		
Perf.	Sim	TGA	TGA-G	Sim	TGA	TGA-G
E[X]	1.09E+2	1.07E+2	1.07E+2	1.09E+2	1.07E+2	1.07E+2
rel. err.	±1.64E-1	1%	1%	±1.95E-1	1%	1%
Var(X)	2.30E+2	2.37E+2	2.69E+2	2.50E+2	2.37E+2	3.13E+2
rel. err.	±3.66E+1	3%	17%	±4.36E+1	5%	25%
E[Q]	1.13E+1	1.05E+1	1.09E+1	1.13E+1	1.05E+1	1.14E+1
rel. err.	±1.29E-1	7%	3%	±1.52E-1	7%	0%
Var(Q)	1.47E+2	1.29E+2	1.43E+2	1.61E+2	1.29E+2	1.62E+2
rel. err.	±4.69E+0	12%	3%	±5.75E+0	20%	0%
E[W]	1.13E-1	1.03E-1	1.07E-1	1.14E-1	1.03E-1	1.11E-1
rel. err.	±1.28E-3	9%	5%	±1.54E-3	10%	2%
Var(W)	1.42E-2	1.22E-2	1.36E-2	1.60E-2	1.22E-2	1.55E-2
rel. err.	±4.58E-4	14%	4%	±5.93E-4	23%	3%
P^d	7.11E-1	6.88E-1	6.77E-1	6.94E-1	6.88E-1	6.64E-1
rel. err.	±3.88E-3	3%	5%	±4.51E-3	1%	4%
$P^{d,C}$	7.25E-1	6.88E-1	7.08E-1	7.56E-1	6.88E-1	6.97E-1
rel. err.	±3.80E-3	9%	6%	±4.54E-3	7%	6%
P^a	6.80E-2	6.34E-2	6.53E-2	6.92E-2	6.34E-2	6.78E-2
rel. err.	±9.75E-4	7%	4%	±1.10E-3	8%	2%

3.7 Smaller Scale: Lower Arrival Rates and Fewer Servers

Since the MSHT limits involve a sequence of queueing systems with increasing scale, the MSHT approximations DGA and TGA-G should perform better as the scale increases. Thus, we considered the base case with $n = 100$, because it is large but also small enough to be of practical value. However, we also want to apply the approximations to even smaller scale systems. Thus, in this section, we examine the effectiveness of DGA, TGA and TGA-G for smaller systems.

In order to set the parameters for these smaller systems, it is good to exploit the MSHT limits. When the system is in the QED regime, we know that the scaling factor n (number of servers) and the traffic intensity ρ_n should roughly satisfies the relation

$$\sqrt{n}\,(1-\rho_n) \approx \beta, \quad -\infty < \beta < \infty, \tag{3.25}$$

where the β is the QoS factor. Since $\beta = \sqrt{n}(1-\rho_n) = -0.5$ when $\rho_n = 1.05$ and $n = 100$ as in previous tables, we now fix β at -0.5 as we change n. Note that this increases the traffic intensity as ρ decreases.

Table 3.10 shows the results for an $H_2/H_2/n + H_2$ model for three values of n: $20, 10, 5$, which correspond to traffic intensities $\rho_n = 1.11, 1.16, 1.22$. Table 3.10 shows that TGA-G remains effective for systems with fewer servers. Figure 3.10 shows that accuracy consistently degrades as the scale decreases, but remains reasonable. Figure 3.11 shows that TGA-G can be used to approximate

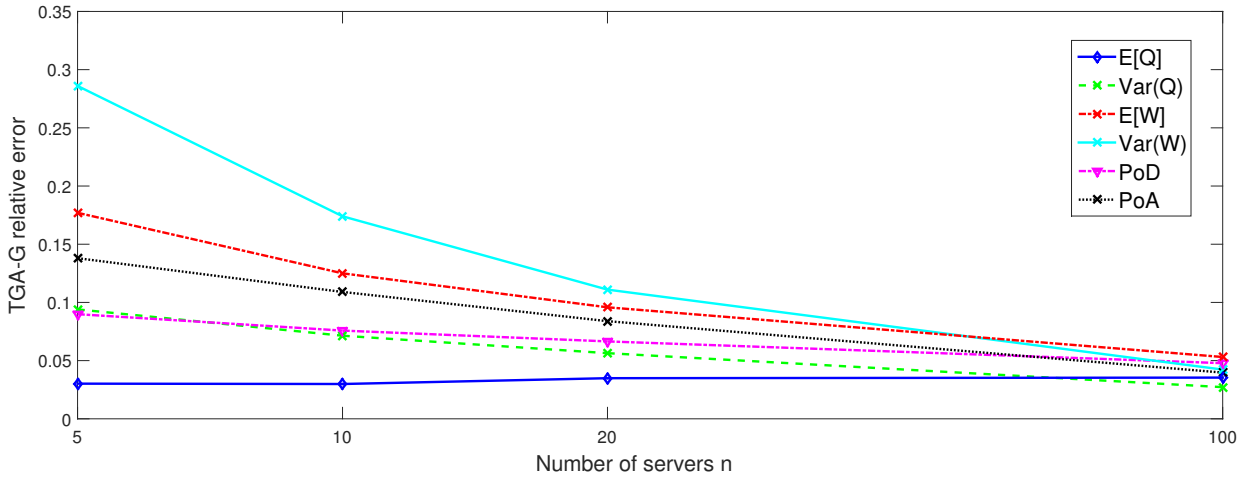


Figure 3.10: Simulation estimates of the relative errors in the approximations for six performance measures as a function of the number of servers, n, for $n = 5, 10, 20, 100$ (with n in log scale) in the $H_2(1/n\rho, 2)/H_2(1, 2)/n + H_2(2, 2)$ model.

Table 3.10: Smaller scale: a comparison of the TGA-G and DGA approximations to simulation estimates in the $H_2(\lambda^{-1})/H_2(1,2)/n + H_2(\theta^{-1},2)$ model with three pairs (n,ρ) with $\rho = 1-\beta/\sqrt{n}$, $\lambda = n\rho$, $\beta = -0.5$ and $n = 20$, 10 and 5

	$n=20$			$n=10$			$n=5$		
Perf.	Sim	DGA	TGA-G	Sim	DGA	TGA-G	Sim	DGA	TGA-G
E[X]	2.40E+1	2.34E+1	2.34E+1	1.29E+1	1.24E+1	1.24E+1	7.08E+0	6.69E+0	6.69E+0
rel. err.	±7.46E-2	3%	3%	±5.38E-2	4%	4%	±3.96E-2	5%	5%
Var(X)	4.86E+1	5.06E+1	5.69E+1	2.52E+1	2.65E+1	2.97E+1	1.32E+1	1.41E+1	1.57E+1
rel. err.	±3.84E+0	4%	17%	±1.55E+0	5%	18%	±6.61E-1	6%	19%
E[Q]	5.16E+0	4.82E+0	4.98E+0	3.68E+0	3.46E+0	3.57E+0	2.65E+0	2.49E+0	2.57E+0
rel. err.	±5.93E-2	7%	4%	±4.31E-2	6%	3%	±3.21E-2	6%	3%
Var(Q)	3.19E+1	2.74E+1	3.01E+1	1.68E+1	1.42E+1	1.56E+1	9.04E+0	7.50E+0	8.19E+0
rel. err.	±1.03E+0	14%	6%	±5.46E-1	15%	7%	±3.01E-1	17%	9%
E[W]	2.61E-1	2.28E-1	2.36E-1	3.77E-1	3.18E-1	3.30E-1	5.60E-1	4.44E-1	4.61E-1
rel. err.	±2.96E-3	13%	10%	±4.38E-3	16%	13%	±6.90E-3	21%	18%
Var(W)	7.54E-2	6.01E-2	6.70E-2	1.61E-1	1.19E-1	1.33E-1	3.67E-1	2.34E-1	2.62E-1
rel. err.	±2.55E-3	20%	11%	±5.72E-3	26%	18%	±1.45E-2	36%	29%
P^d	7.22E-1	6.85E-1	6.74E-1	7.26E-1	6.82E-1	6.71E-1	7.34E-1	6.79E-1	6.68E-1
rel. err.	±3.74E-3	5%	7%	±3.75E-3	6%	8%	±3.74E-3	7%	9%
$P^{d,C}$	7.45E-1	6.85E-1	6.74E-1	7.62E-1	6.82E-1	6.71E-1	7.81E-1	6.79E-1	6.68E-1
rel. err.	±3.65E-3	8%	10%	±3.62E-3	10%	12%	±3.47E-3	13%	14%
P^a	1.43E-1	1.28E-1	1.31E-1	1.92E-1	1.67E-1	1.71E-1	2.54E-1	2.15E-1	2.19E-1
rel. err.	1.72E-3	11%	9%	2.15E-3	13%	11%	±2.70E-3	16%	14%

the distributions of key performance measures. Figure 3.11 shows that the simulated histograms for the $H_2/M/n + H_2$ model are well approximated by the pdfs of the corresponding Gaussian approximations, for $n = 100, 50, 20$ and 5. Additional simulation results appear in the appendix.

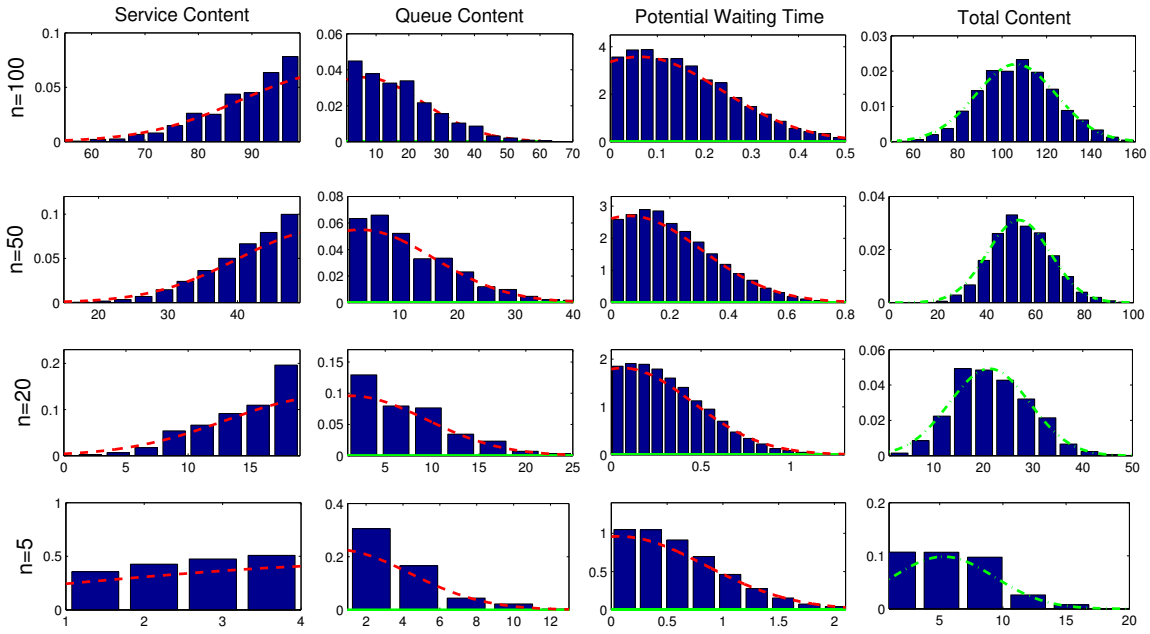


Figure 3.11: Simulation estimates (histograms) of the TGA-G approximating distributions for B_n, Q_n, V_n and X_n in the $H_2(\lambda^{-1}, 2)/M/n + H_2(\theta^{-1}, 2)$ model with $\beta = -0.5$ and four values of n, $5 \leq n \leq 100$.

3.8 Limitations of the Proposed Approximations

We first compare our TGA-G approximation to previous approximations in [107] and [94]. Then we show that the performance is not good for UL models.

3.8.1 Comparison with Approximations in [107]

A numerical approximation algorithm for the $M/GI/n+GI$ model was developed and evaluated in [107]. It was based on an application of an exact analysis of an associated state-dependent Markov $M/M/n+M(n)$ queue, after approximating the GI abandonment by an appropriate state-dependent $M(n)$ abandonment mechanism. (The GI service was simply approximated by M.) That numerical procedure has the advantage that it applies to all loadings (underloaded, critically loaded and overloaded), but it is much more computationally intensive. That approximation was shown to be quite effective. A shortcoming of [107] that we address here is that it does not describe the impact of non-M arrival processes and service times.

We now compare our new TGA-G approximation to the approximation developed in [107] by comparing to the displayed results in Tables 6 and 7 of [107], which are also for $n = 100$, but for the relatively light loading $\rho = 1.02$, which is at the edge of the range of effectiveness for TGA-G.

As expected for this relatively light loading, Tables 3.11 and 3.12 show that the engineering approximations in [107], labeled as *Eng. Approx. (W05)*, are more accurate overall. However, TGA-G is better for D service.

First, Table 3.11 shows that TGA-G performs reasonably well for the Erlang E_2 patience distribution, except for the variance $\mathrm{Var}(Q)$, even if not as accurate as Table 7 of [107]. In fact, TGA-G performs better for D service. On the other hand, Table 3.12 shows that the performance of TGA-G degrades significantly for the $LN(1,1)$ patience distribution. Nevertheless, the TGA-G (=DGA) approximation for $\mathrm{E}[X]$ remains good.

In summary, we have seen in previous sections that our proposed TGA-G approximation for heavily-loaded $G/GI/n+GI$ model is remarkably effective for a wide class of models. Nevertheless, there are limitations, as exposed by Table 3.12. Lack of accuracy is most likely as the loading decreases toward critical loading. That breakdown is likely to occur sooner (for higher ρ) if the component model elements deviate more from M. The examples show difficulties for low loading ($\rho = 1.02$) and non-M patience distributions.

Table 3.11: Comparison of the TGA-G and DGA approximations with the $M/M/n + M(n)$ approximation in [107] and simulation estimates for the $M(102^{-1})/GI(1, c_s^2)/100/200 + E_2$ model with four service-time cdf's with a range of scv's: $0.0 \leq c_s^2 \leq 4.0$.

Serv. Dist.	$D, c_s^2 = 0$				$E_2, c_s^2 = 0.5$			
Perf.	Sim	Eng. Approx. (W05)	DGA	TGAG	Sim	Eng. Approx. (W05)	DGA	TGAG
$\mathbb{P}(W = 0)$	1.80E-1	2.50E-1	1.05E-1	same	2.17E-1	2.50E-1	2.08E-1	same
rel. err.	±1.30E-3	28%	42%		±2.10E-3	13%	4%	
P^a	3.09E-2	3.81E-2	2.86E-2	same	3.51E-2	3.81E-2	3.83E-2	same
rel. err.	±1.70E-4	19%	7%		±2.90E-4	8%	9%	
E[Q]	1.11E+1	1.14E+1	1.08E+1	1.13E+1	1.15E+1	1.14E+1	1.08E+1	1.24E+1
rel. err.	±4.20E-2	3%	3%	2%	±7.50E-2	1%	6%	8%
Var(Q)	8.93E+1	1.22E+2	7.80E+1	6.42E+1	1.12E+2	1.22E+2	1.79E+2	1.22E+2
rel. err.	±4.00E-1	27%	12%	28%	±7.10E-1	8%	60%	9%
E[X]	1.10E+2	1.10E+2	1.11E+2	same	1.10E+2	1.10E+2	1.11E+2	same
rel. err.	±4.90E-2	0%	1%		±9.20E-2	0%	1%	

Serv. Dist.	$M, c_s^2 = 1$				$LN(1, 1)$			
Perf.	Sim	Eng. Approx. (W05)	DGA	TGAG	Sim	Eng. Approx. (W05)	DGA	TGAG
$\mathbb{P}(W = 0)$	2.46E-1	2.50E-1	2.33E-1	same	2.33E-1	2.50E-1	2.33E-1	same
rel. err.	±2.00E-3	2%	5%		±2.10E-3	7%	0%	
P^a	3.78E-2	3.81E-2	4.18E-2	same	3.70E-2	3.81E-2	4.18E-2	same
rel. err.	±3.20E-4	1%	11%		±2.70E-4	3%	13%	
E[Q]	1.18E+1	1.14E+1	1.08E+1	1.29E+1	1.17E+1	1.14E+1	1.08E+1	1.29E+1
rel. err.	±7.50E-2	3%	8%	9%	±6.30E-2	3%	8%	10%
Var(Q)	1.29E+2	2.20E+2	1.43E+2	1.80E+2	1.23E+2	1.22E+2	2.20E+2	1.43E+2
rel. err.	±9.40E-1	70%	11%	39%	±7.20E-1	1%	79%	16%
E[X]	1.10E+2	1.10E+2	1.11E+2	same	1.10E+2	1.10E+2	1.11E+2	same
rel. err.	±9.10E-2	0%	1%		±7.20E-1	0%	1%	

Table 3.12: Comparison of the TGA-G and DGA approximations with the $M/M/n+M(n)$ approximation in [107] and simulation estimates for the $M(102^{-1})/GI(1, c_s^2)/100/200 + LN(1, 1)$ model with four service-time cdf's with a range of scv's: $0.0 \leq c_s^2 \leq 4.0$.

Serv. Dist.	$E_2, c_s^2 = 0.5$				$M, c_s^2 = 1$			
Perf.	Sim	Eng. Approx. (W05)	DGA	TGA-G	Sim	Eng. Approx. (W05)	DGA	TGA-G
$\mathbb{P}(W = 0)$	2.11E-1	2.47E-1	1.76E-1	same	2.42E-1	2.47E-1	1.95E-1	same
rel. err.	±1.30E-3	15%	17%		±2.60E-3	2%	20%	
P^a	3.48E-2	3.79E-2	5.13E-2	same	3.76E-2	3.79E-2	5.51E-2	same
rel. err.	±2.10E-4	8%	47%		±3.20E-4	1%	47%	
E[Q]	1.14E+1	1.10E+1	1.29E+1	1.43E+1	1.14E+1	1.10E+1	1.29E+1	1.46E+1
rel. err.	±3.90E-2	3%	13%	25%	±7.10E-2	4%	13%	27%
Var(Q)	1.03E+2	1.07E+2	1.99E+2	1.43E+2	1.16E+2	1.07E+2	2.31E+2	1.61E+2
rel. err.	±3.90E-1	4%	94%	40%	±4.60E-1	8%	100%	39%
E[X]	1.10E+2	1.09E+2	1.13E+2	1.13E+2	1.10E+2	1.09E+2	1.13E+2	same
rel. err.	±5.30E-2	1%	3%		±9.20E-2	0%	3%	

Serv. Dist.	$LN(1, 1)$				$LN(1, 4)$			
Perf.	Sim	Eng. Approx. (W05)	DGA	TGA-G	Sim	Eng. Approx. (W05)	DGA	TGA-G
$\mathbb{P}(W = 0)$	2.29E-1	2.47E-1	1.95E-1	same	2.11E-1	2.47E-1	2.41E-1	same
rel. err.	±1.50E-3	7%	15%		±1.30E-3	15%	14%	
P^a	3.66E-2	3.79E-2	5.51E-2	same	3.48E-2	3.79E-2	6.66E-2	same
rel. err.	±2.40E-4	3%	51%		±2.10E-4	8%	91%	
E[Q]	1.14E+1	1.10E+1	1.29E+1	1.46E+1	1.14E+1	1.10E+1	1.29E+1	1.55E+1
rel. err.	±5.10E-2	4%	13%	27%	±3.90E-2	3%	13%	36%
Var(Q)	1.11E+2	1.07E+2	2.31E+2	1.61E+2	1.03E+2	1.07E+2	3.40E+2	2.16E+2
rel. err.	±4.30E-1	3%	109%	45%	±3.90E-1	4%	231%	111%
E[X]	1.10E+2	1.09E+2	1.13E+2	same	1.10E+2	1.09E+2	1.13E+2	same
rel. err.	±6.20E-2	1%	3%		±5.30E-2	1%	3%	

3.8.2 Comparison with Approximations in [94]

A MSHT limit was established for the $GI/M/n + GI$ model in [94] and its corresponding numerical approximation was developed and evaluated. The MSHT limit in [94] was in the QED regime, whereas ours in Theorem 3.2.3 in [77] is in the ED regime. Nevertheless, it is natural to compare the two candidate approximations for systems with $\rho > 1$, because any such system can be regarded as one in a sequence of systems satisfying a QED limit or an ED limit.

Before we present numerical comparison results, we first summarize the differences between TGA and [94]. First, just like [77], the MSHT limit in [94] is for the model with exponential service times, and the resulting approximation in [94] is limited to that case. Second, the formulas in [94] require the knowledge of the entire patience-time cdf F or, equivalently, of its hazard-rate function $h(t) \equiv f(t)/\bar{F}(t)$. In contrast, the ED fluid limit used in TGA depends on the patience-time cdf F only on the value of its inverse F^{-1} at the limiting fluid abandonment rate $(\rho - 1)/\rho$. The TGA fluid head-of-line waiting time $w(\infty)$ is determined by $w(\infty) = F^{-1}((\rho - 1)/\rho)$. Third, the approximation formulas in [94] are rather complicated, requiring computation of the triple integrals in (11), (12) and (13) in [94]. In contrast, the TGA approximations are easier to compute. Finally, the hazard-rate scaling in the MSHT limit in [94] makes that approximation especially effective for models in which the patience-time hazard-rate function changes rapidly near the origin.

The examples in [94] considered three different patience distributions. The first two are hyper-exponential (H_2) distributions with parameter triples (p, θ_1, θ_2). The first is a regular H_2 distribution with $(p, \theta_1, \theta_2) = (0.5, 1, 2)$. The second is a more extreme H_2 distribution with $(p, \theta_1, \theta_2) = (0.9, 1, 200)$, which changes rapidly near the origin. The third is another relatively extreme hazard-rate function that increases rapidly near the origin. In [94] the new approximations were compared to exact values and approximations "Z&M" in [123].

We now compare our new TGA approximation to the approximation developed in [94] by comparing our results to the displayed results in [94] for the cases with $\rho > 1$. These appear in Tables 3, 6 and 9 of [94]. As in [94], we consider a range of scale n from $n = 10$ to $n = 500$, with the traffic intensity ρ and number of server n satisfying the relation (3.25) with $\beta = -1$, which are: $n = 10$ ($\rho = 1.31$), $n = 50$ ($\rho = 1.14$), $n = 100$ ($\rho = 1.10$), $n = 200$ ($\rho = 1.07$), and $n = 500$ ($\rho = 1.022$). (The last case is at the edge of the range of effectiveness for TGA.)

Table 3.13 compares TGA to the hazard-rate scaling (HRS) approximation in [94] and the exact and approximate values in [123] for the first H_2 patience distribution with parameter $(p, \theta_1, \theta_2) = (0.5, 1, 2)$. For this "easy" example, Table 3.13 shows that all methods are effective for large scale, but TGA is consistently appreciably better for the mean $E[V]$, especially for small scale.

However, Tables 3.14 and 3.15 show that the performance of TGA degrades significantly for the last two more challenging patience distributions, while the HRS approximation in [94] continues to be effective. However, even for these challenging examples, TGA performs better than HRS in

Table 3.13: A comparison of TGA and DGA approximations for the probability of delay P^d, the mean wait $E[V]$ and the probability of abandonment P^a to the results in Table 3 of [94] for the $M/M/n+H_2$ model with $\rho = 1+1\sqrt{n}$, $\lambda = n\rho\mu$, $\mu = 1$ and H_2 patience with $(p, \theta_1, \theta_2) = (0.5, 1, 2)$. The exact results and approximations Z&M come in [123].

	P^d				E[V]				P^a			
n	Exact	Z&M	HRS	TGA	Exact	Z&M	HRS	TGA	Exact	Z&M	HRS	TGA
10	0.7952	0.7408	0.7806	0.7921	11.3858	12.8341	14.6407	10.1028	0.2755	0.3209	0.2679	0.2448
	-	7%	2%	0%	-	13%	29%	11%	-	16%	3%	11%
50	0.7839	0.7586	0.7777	0.7982	5.6986	6.0298	6.4212	5.4381	0.1403	0.1507	0.1384	0.1399
	-	3%	1%	2%	-	6%	13%	5%	-	7%	1%	0%
100	0.7814	0.7633	0.7771	0.7927	4.1484	4.4716	4.32	4.0355	0.1026	0.108	0.1016	0.103
	-	2%	1%	1%	-	8%	4%	3%	-	5%	1%	0%
200	0.7796	0.7667	0.7766	0.7913	2.9963	3.0844	3.1859	2.9561	0.0743	0.0771	0.0738	0.0747
	-	2%	0%	2%	-	3%	6%	1%	-	4%	1%	1%
500	0.7781	0.7698	0.7762	0.7937	1.932	1.9681	2.0093	1.9307	0.0481	0.0492	0.0478	0.0489
	-	1%	0%	2%	-	2%	4%	0%	-	2%	1%	2%

[94] and Z&M in [123] in the case of $n = 10$ (small scale). Indeed, for $n = 10$, the TGA values are reasonable, whereas Z&M is off by a factor of 10.5 for the mean $E[V]$ in Table 3.14 and HRS is off by a factor of 15.0 for the mean $E[V]$ in Table 3.15. The is partly explained by the scaling, because for $n = 10$ the QED scaling in (3.25) makes $\rho = 1.31$ even though $\beta = -0.5$.

We conclude that all these approximations have advantages, which at least partly depends on whether the QED or ED regime is most natural.

Table 3.14: A comparison of TGA and DGA approximations for the probability of delay P^d, the mean wait $E[V]$ and the probability of abandonment P^a to the results in Table 6 of [94] for the $M/M/n + H_2$ model with $\rho = 1 + 1\sqrt{n}$, $\lambda = n\rho\mu$, $\mu = 1$ and H_2 patience with $(p, \theta_1, \theta_2) = (0.9, 1, 200)$.

	P^d				E[V]				P^a			
n	Exact	Z&M	HRS	TGA	Exact	Z&M	HRS	TGA	Exact	Z&M	HRS	TGA
10	0.7909	0.3139	0.7821	0.7462	11.8884	1.1349	16.2358	9.9089	0.2763	0.4268	0.2677	0.2584
	-	60%	1%	6%	-	90%	37%	17%	-	54%	3%	6%
50	0.7158	0.3242	0.709	0.6014	4.5527	0.5277	5.405	3.4147	0.1455	0.1838	0.1428	0.1211
	-	55%	1%	16%	-	88%	19%	25%	-	26%	2%	17%
100	0.6663	0.327	0.6587	0.5789	2.7067	0.377	3.0844	1.2267	0.1087	0.1313	0.1072	0.0761
	-	51%	1%	13%	-	86%	14%	55%	-	21%	1%	30%
200	0.6063	0.329	0.5979	0.584	1.4942	0.2686	1.6453	0.6486	0.0808	0.0936	0.08	0.0613
	-	46%	1%	4%	-	82%	10%	57%	-	16%	1%	24%
500	0.5213	0.3309	0.5127	0.5858	0.6201	0.1711	0.6577	0.3495	0.0541	0.0596	0.0538	0.0495
	-	37%	2%	12%	-	72%	6%	44%	-	10%	1%	9%

Table 3.15: A comparison of TGA and DGA approximations for the probability of delay P^d, the mean wait $E[V]$ and the probability of abandonment P^a to the results in Table 9 of [94] for the $M/M/n + GI$ model with $\rho = 1 + 1\sqrt{n}$, $\lambda = n\rho\mu$, $\mu = 1$ and increasing patience-time hazard rate.

	P^d				E[V]				P^a			
n	Exact	Z&M	HRS	TGA	Exact	Z&M	HRS	TGA	Exact	Z&M	HRS	TGA
10	0.4454	0.7408	0.2913	0.6324	0.8557	1.1349	12.8341	1.1201	0.3312	0.3209	0.3297	0.4072
	-	66%	35%	42%	-	33%	1400%	31%	-	3%	0%	23%
50	0.4041	0.7586	0.3559	0.6647	0.5277	6.4212	0.565	0.9411	0.1679	0.1507	0.1658	0.2854
	-	88%	12%	64%	-	1117%	7%	78%	-	10%	1%	70%
100	0.4152	0.7633	0.3859	0.6783	0.4786	4.4716	0.4877	0.8368	0.1215	0.108	0.1202	0.2254
	-	84%	7%	63%	-	834%	2%	75%	-	11%	1%	86%
200	0.4351	0.7667	0.4169	0.6917	0.4107	3.0844	0.4191	0.7245	0.087	0.0771	0.0863	0.1692
	-	76%	4%	59%	-	651%	2%	76%	-	11%	1%	94%
500	0.4688	0.7698	0.4588	0.7089	0.3347	1.9681	0.3404	0.5779	0.0553	0.0492	0.055	0.1084
	-	64%	2%	51%	-	488%	2%	73%	-	11%	1%	96%

3.8.3 Underloaded Models

In this section we present the performance of the underloaded $M/M/n+M$ in QED regime, where β defined in (3.25) is fixed at 0.5. During UL intervals, it is easy to check that P^d and P^a are 0 in the MSHT limits. The fluid and FCLT limit of waiting time in underloaded intervals implies that $w(\infty) = v(\infty) = 0$ and $\sigma_W = 0$. In order to improve the performance, we replace $w(\infty) = 0$ by $\mathrm{E}\left[W^{TGA}\right] = \mathrm{E}\left[Q^{TGA}\right]/\mu$ and $\sigma_W = 0$ by $\mathrm{Var}\left(W^{TGA}\right) = \mathrm{Var}\left(Q^{TGA}\right) + \mathrm{E}\left[Q^{TGA}\right]$, then design $P_D^{TGA}(n)$ and $P_A^{TGA}(n)$ as

$$\begin{aligned} P_D^{TGA}(n) &= \bar{\Phi}\left(-a'_W(n)\right), \\ P_A^{TGA}(n) &= \int_0^\infty \bar{\Phi}\left(a'_W(n)\left(\frac{x}{w}-1\right)\right) f(x)\mathrm{d}x, \end{aligned} \tag{3.26}$$

where $a'_W(n) = \sqrt{n}\mathrm{E}[W^{TGA}]/\mathrm{Var}\left(\hat{W}^{TGA}\right)$.

The main idea of (3.26) is to use $W^{TGA} = \sum_{i=1}^{Q^{TGA}} S_i$ to replace the zero waiting time; here S_i is the processing time of the i^{th} customers. We omit the impact of abandonments since according to the numerical results on underloaded intervals, the probability of abandonments are about 10^{-2}. To verify (3.26), it is suffice to prove that

$$\mathrm{E}\left[W^{TGA}\right] = \mathrm{E}\left[Q^{TGA}\right]/\mu, \quad \text{and} \quad \mathrm{Var}\left(W^{TGA}\right) = \mathrm{Var}\left(Q^{TGA}\right) + \mathrm{E}\left[Q^{TGA}\right].$$

Proof. It is obvious for the expression of $\mathrm{E}[W^{TGA}]$ so we only focus on $\mathrm{Var}\left(W^{TGA}\right)$ here.

$$\begin{aligned} \mathrm{E}\left[\left(W^{TGA}\right)^2\right] &= \mathrm{E}\left[\sum_{i=1}^{Q^{TGA}} S_i^2 + 2\sum_{j<i}^{Q^{TGA}} S_i S_j\right] = \mathrm{E}\left[\mathrm{E}\left[\sum_{i=1}^{Q^{TGA}} S_i^2 + 2\sum_{j<i}^{Q^{TGA}} S_i S_j \middle| Q^{TGA}\right]\right] \\ &= \mathrm{E}\left[2Q^{TGA} + Q^{TGA}(Q^{TGA}-1)\right] = \mathrm{E}\left[Q^{TGA}\right] + \mathrm{E}\left[\left(Q^{TGA}\right)^2\right], \end{aligned}$$

which implies the expression of $\mathrm{Var}(W^{TGA})$. □

Table 3.16 shows the performance of an underloaded $M/M/n+M$ model with $\rho = 0.95$. In particular, with parameters $\lambda = 100$, $\rho = 0.95$ and $0.1 \le \theta \le 4.0$. Table 3.16 shows good performance of TGA for the means of X_n and B_n in all cases and for the variances of X_n and B_n with $0.5 \le \theta \le 2.0$, but poor performance otherwise.

Table 3.16: The performance for underloaded models: a comparison between simulation estimates and exact numerical values for the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $n = 100, \rho = 0.95$ and $0.1 \leq \theta \leq 2$

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	1.00E+2	9.50E+1	same	9.80E+1	9.50E+1	same	9.64E+1	9.50E+1	same
rel. err.		5%			3%			1%	
Var(X)	2.18E+2	9.50E+1	same	1.56E+2	9.50E+1	same	1.20E+2	9.50E+1	same
rel. err.		56%			39%			21%	
E[B]	9.44E+1	9.50E+1	9.31E+1	9.40E+1	9.50E+1	9.31E+1	9.36E+1	9.50E+1	9.31E+1
rel. err.		1%	1%		1%	1%		2%	0%
Var(B)	4.91E+1	9.50E+1	5.31E+1	5.03E+1	9.50E+1	5.31E+1	5.11E+1	9.50E+1	5.31E+1
rel. err.		94%	8%		89%	6%		86%	4%
E[Q]	5.78E+0	0.00E+0	1.89E+0	4.06E+0	0.00E+0	1.89E+0	2.88E+0	0.00E+0	1.89E+0
rel. err.		100%	67%		100%	53%		100%	34%
Var(Q)	1.04E+2	0.00E+0	1.59E+1	5.67E+1	0.00E+0	1.59E+1	3.22E+1	0.00E+0	1.59E+1
rel. err.		100%	85%		100%	72%		100%	51%
E[V]	6.16E-2	0.00E+0	1.89E-2	4.37E-2	0.00E+0	1.89E-2	3.14E-2	0.00E+0	1.89E-2
rel. err.		100%	69%		100%	57%		100%	40%
Var(V)	1.13E-2	0.00E+0	1.78E-3	6.24E-3	0.00E+0	1.78E-3	3.60E-3	0.00E+0	1.78E-3
rel. err.		100%	84%		100%	72%		100%	51%
P^d	4.49E-1	0.00E+0	6.73E-1	4.06E-1	0.00E+0	6.73E-1	3.64E-1	0.00E+0	6.73E-1
rel. err.		100%	50%		100%	66%		100%	85%
P^a	6.09E-3	NaN	1.88E-3	1.07E-2	NaN	4.70E-3	1.51E-2	NaN	9.38E-3
rel. err.		NaN	69%		NaN	56%		NaN	38%

	$\theta = 1$			$\theta = 2$			$\theta = 4$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	9.50E+1	9.50E+1	same	9.38E+1	9.50E+1	same	9.28E+1	9.50E+1	same
rel. err.		0%			1%			2%	
Var(X)	9.50E+1	9.50E+1	same	7.78E+1	9.50E+1	same	6.64E+1	9.50E+1	same
rel. err.		0%			22%			43%	
E[B]	9.31E+1	9.50E+1	9.31E+1	9.26E+1	9.50E+1	9.31E+1	9.21E+1	9.50E+1	9.31E+1
rel. err.		2%	0%		3%	1%		3%	1%
Var(B)	5.16E+1	9.50E+1	5.31E+1	5.16E+1	9.50E+1	5.31E+1	5.12E+1	9.50E+1	5.31E+1
rel. err.		84%	3%		84%	3%		86%	4%
E[Q]	1.92E+0	0.00E+0	1.89E+0	1.21E+0	0.00E+0	1.89E+0	7.23E-1	0.00E+0	1.89E+0
rel. err.		100%	1%		100%	57%		100%	161%
Var(Q)	1.69E+1	0.00E+0	1.59E+1	8.28E+0	0.00E+0	1.59E+1	3.82E+0	0.00E+0	1.59E+1
rel. err.		100%	6%		100%	92%		100%	315%
E[V]	2.13E-2	0.00E+0	1.89E-2	1.38E-2	0.00E+0	1.89E-2	8.60E-3	0.00E+0	1.89E-2
rel. err.		100%	11%		100%	37%		100%	120%
Var(V)	1.93E-3	0.00E+0	1.78E-3	9.79E-4	0.00E+0	1.78E-3	4.76E-4	0.00E+0	1.78E-3
rel. err.		100%	8%		100%	81%		100%	273%
P^d	3.17E-1	0.00E+0	6.73E-1	2.68E-1	0.00E+0	6.73E-1	2.21E-1	0.00E+0	6.73E-1
rel. err.		100%	112%		100%	151%		100%	205%
P^a	2.02E-2	NaN	1.87E-2	2.54E-2	NaN	3.70E-2	3.04E-2	NaN	7.26E-2
rel. err.		NaN	7%		NaN	46%		NaN	139%

3.9 Simulation Methodology

We used simulation to estimate the exact values for all non-Markovian models. We now provide extra details about our simulation methodology. For $n = 100$, we estimated all the performance measures using 2000 independent replications over the time interval $[0, T]$ with $T = 100$, starting empty in each case. To have statistical precision for all the steady-state estimates, we need to ensure, first, that the system has approximately reached the steady state before sampling, and, second, that enough sampled data are collected to give reasonable accuracy, which we judge by using 95% confidence intervals. We discuss these issues in turn.

3.9.1 From Transient to Steady State

To avoid bias caused by the initial transient starting empty, we do not collect data from an initial portion of each run. We stop sampling at time $0.95T = 95$ We also eliminate a final portion so that we can observe the waiting times experienced by all arrivals in the main measurement interval; i.e., to avoid abnormal zeroes in sampled potential waiting times, which results from that some virtual customers' (potential) waiting times not being sampled at the end of simulation at $T = 100$. In particular, for $n = 100$ we use the data in $[40, 95]$ from each run over $[0, 100]$ to estimate the steady-state performance functions.

To illustrate the initial transient and when it tends to disappear, we show an example, using the $H_2(105^{-1}, 2)/H_2(1, 2)/100 + H_2(2, 2)$ model. Figure 3.12 shows plots of the transient (time-dependent) mean and variance functions for the queue length. Figure 3.12 shows that the performance is close to steady state after time 20. To be safe, we use the data in $[40, 95]$ to estimate the steady-state performance functions.

3.9.2 The Sampling Procedure

To determine the potential waiting times at time t, which are for an arrival with unlimited patience that would arrive at time t (the usual virtual waiting time, modified to include unlimited patience), we generate virtual customers that do not affect the other customers. In particular, in the r^{th} simulation replication, $1 \le r \le R$, we periodically generate virtual arrivals at deterministic times $t_1, t_k, \ldots, t_{N_v}$ with $t_k \equiv k\Delta t$ and $\Delta t = 0.1$, $1 \le k \le N_v \equiv \lfloor T/\Delta t \rfloor$. The virtual customers have the same waiting time distribution. They abandon as if they are the real customers but they will not be removed from the queue if they abandon. They still wait in queue until their turn to enter service so that we can record their virtual waiting time as potential waiting times. We do not allow them to enter service, so these virtual customers do not affect the system dynamics. We use indicator variables $\eta^a_{r,k}$ and $\eta^d_{r,k}$to record if the virtual arrival at t_k on the r^{th} path abandons and

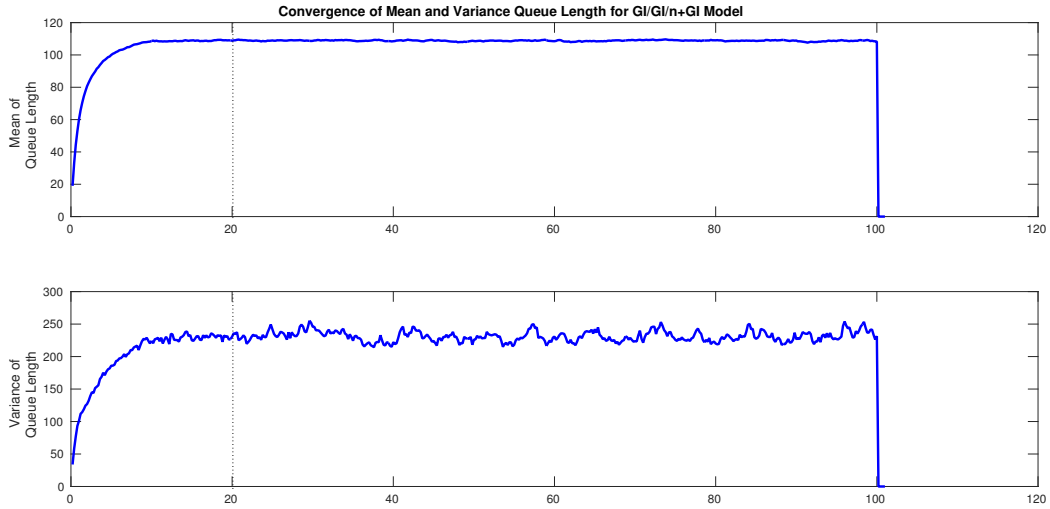


Figure 3.12: Simulation estimates of the mean and variance of the queue length as a function of time in the $H_2/H_2/n + H_2$ model to show the approach to steady state.

is delayed, namely

$$\eta^d_{r,k} = \begin{cases} 1, & \text{if } V_r(k) > 0, \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad \eta^a_{r,k} = \begin{cases} 1, & \text{if } V_r(k) > A_k, \\ 0, & \text{otherwise} \end{cases}, \tag{3.27}$$

where A_k is the patience time of the k^{th} virtual arrival, $V_r(k) \equiv E_r(k) - A_r(k)$ records the potential waiting time at time t_k, $A_r(k) \equiv k \cdot \Delta t$ and $E_r(k)$ are the times at which the k^{th} virtual customer arrives and enters service. We also use indicators variables $\eta^{d,C}_{r,k}$'s to denote if the k^{th} real (non-virtual) customer is delayed once it arrives on the r^{th} replication.

For the number of customers in queue and service at each time, we sample the continuous-time queue-length process and number of busy servers at discrete time points $t_1, t_2, \ldots, t_{N_v}$, denoted by $Q_r(k)$ and $B_r(k)$. Here we make sure to exclude the virtual arrivals.

3.9.3 Constructing Confidence Intervals

All estimates of target performance measures and corresponding confidence intervals are based on assuming i.i.d. samples, which is justified because we take a single estimate from each of the $R = 2000$ independent samples.

To illustrate how we construct these estimators, we use the queue-length process Q for an example. On the r^{th} path, we sample values for the queue length at $N = 551$ evenly-spaced time points in the interval $[0.4T, 0.95T] = [40, 95]$, denoted by $Q_{r,1}, \ldots, Q_{r,N}$. To construct the confidence

intervals for E[Q] and E[Q^2], first, for each replication $r = 1, 2, \ldots, R$, we let

$$\widetilde{Q}_r \equiv \frac{1}{N}\sum_{l=1}^{N} Q_{r,l} \quad \text{and} \quad \widetilde{Q}_r^{(2)} \equiv \frac{1}{N}\sum_{l=1}^{N} (Q_{r,l})^2. \tag{3.28}$$

Even though the random variables being averaged in (3.28) are typically dependent, these are valid estimators for the true mean E[Q] and second moment E[Q^2]. Experience shows that the average of these $N = 551$ values has lower variance than a single observation from the end of the run.

To get the overall estimators of E[Q] and second moment E[Q^2], and their CI's, we use the R independent samples $\widetilde{Q}_1, \ldots, \widetilde{Q}_R$ ($\widetilde{Q}_1^{(2)}, \ldots, \widetilde{Q}_R^{(2)}$) to compute the sample mean and sample variance of the queue length and its second moment in the usual way, i.e.,

$$\bar{Q}(R) \equiv \frac{1}{R}\sum_{r=1}^{R} \widetilde{Q}_r \quad \text{and} \quad S_Q^2(R) \equiv \frac{1}{R-1}\sum_{r=1}^{R} \left(\widetilde{Q}_r - \bar{Q}(R)\right)^2, \tag{3.29}$$

$$\bar{Q}^{(2)}(R) \equiv \frac{1}{R}\sum_{r=1}^{R} \widetilde{Q}_r^{(2)} \quad \text{and} \quad S_{Q^{(2)}}^2(R) \equiv \frac{1}{R-1}\sum_{r=1}^{R} \left(\widetilde{Q}_r^{(2)} - \bar{Q}^{(2)}(R)\right)^2. \tag{3.30}$$

The random variables $\bar{Q}(R)$ and $\bar{Q}^{(2)}(R)$ in (3.29) our our final estimators for the true mean E[Q] and second moment E[Q^2].

As usual, the $(1 - 100\alpha\%)$-confidence intervals for the mean and second moment of the queue length are

$$\left[\bar{Q}(R) - z_{\alpha/2}\sqrt{\frac{S_Q^2(R)}{R}}, \bar{Q}(R) + z_{\alpha/2}\sqrt{\frac{S_Q^2(R)}{R}}\right], \quad \text{and} \tag{3.31}$$

$$\left[\bar{Q}^{(2)}(R) - z_{\alpha/2}\sqrt{\frac{S_{Q^{(2)}}^2(R)}{R}}, \bar{Q}^{(2)}(R) + z_{\alpha/2}\sqrt{\frac{S_{Q^{(2)}}^2(R)}{R}}\right], \tag{3.32}$$

where z_α is the α-percentile of the standard Gaussian distribution. Since we use 95% CI's, we use $\alpha = 0.025$.

Since $Var(Q) = E[Q^2] - (E[Q])^2$, we estimate the variance by

$$\bar{V}(R) \equiv \bar{Q}^{(2)}(R) - \left(\bar{Q}(R)\right)^2. \tag{3.33}$$

We then approximate the CI halfwidth of the the variance by the CI halfwidth of the second

moment. We thus roughly estimate the CI of the variance as

$$\left[\bar{V}(R) - z_{\alpha/2}\sqrt{\frac{S^2_{Q^{(2)}}(R)}{R}}, \bar{V}(R) + z_{\alpha/2}\sqrt{\frac{S^2_{Q^{(2)}}(R)}{R}}\right]. \tag{3.34}$$

We discuss this approximation further with the numerical example below.

For the probability of abandonment (similar procedure for the time-averaged probability of delay), we sample values for the indicator at $N = 551$ evenly-spaced time points in the interval $[0.4T, 0.95T]$ on the r^{th} run, denoted by $\eta^a_{r,1}, \ldots, \eta^a_{r,N}$, and we let $\widetilde{P}^a_r \equiv (1/N)\sum_{l=1}^{N} \eta^a_{r,l}$, for $r = 1, 2, \ldots, R$. The $(1 - 100\alpha\%)$-confidence interval is

$$\left[\bar{P}^a(R) - z_{\alpha/2}\sqrt{\frac{S^2_a(R)}{R}}, \bar{P}^a(R) + z_{\alpha/2}\sqrt{\frac{S^2_a(R)}{R}}\right],$$

where

$$\bar{P}^a(R) \equiv \frac{1}{R}\sum_{r=1}^{R} \widetilde{P}^a_r \quad \text{and} \quad S^2_a(R) \equiv \frac{1}{R-1}\sum_{r=1}^{R}\left(\widetilde{P}^a_r - \bar{P}^a(R)\right)^2.$$

For the probability of customer-averaged delay, we sample values for the indicator at N_r time points, where N_r is the number of arrivals occurred in the interval $[0.5T, T]$ on the r^{th} replication, denoted by $\eta^{d,C}_{r,1}, \ldots, \eta^{d,C}_{r,N_r}$, and we let $\tilde{P}^{d,C}_r = (1/N_r)\sum_{n=1}^{N_r} \eta^{d,C}_{r,n}$ for $r = 1, 2, \ldots, R$. The $(1-100\alpha\%)$-confidence interval is

$$\left[\bar{P}^{d,C}(R) - z_{\alpha/2}\sqrt{\frac{S^2_{d,C}(R)}{R}}, \bar{P}^{d,C}(R) + z_{\alpha/2}\sqrt{\frac{S^2_{d,C}(R)}{R}}\right],$$

where

$$\bar{P}^{d,C}(R) = \frac{1}{R}\sum_{r=1}^{R} \tilde{P}^{d,C}_r, \quad \text{and} \quad S^2_{d,C} = \frac{1}{R-1}\sum_{r=1}^{R}\left(\tilde{P}^{d,C} - \bar{P}^{d,C}(R)\right)^2.$$

To substantiate our procedures and verify that we obtain adequate statistical precision, we compare the estimated performance measures of $M/M/n+M$ model to corresponding exact solutions, which are calculated by the same algorithms of [107]. The Table 3.17 shows that the procedures are sound and the the statistical precision is adequate.

We use Table 3.17 to elaborate on the approximate CI for the variance. To do so, we focus on the queue length in the case $\theta = 0.1$. Notice that the CI for the mean is 30.0 ± 0.52, so that the relative halfwidth for the mean is 1.7%. However, by squaring the upper and lower limits,

Table 3.17: A comparison between simulation estimates and exact numerical values for the $M(102^{-1})/M(1)/100 + M(\theta^{-1})$ model, confirming the validity of both algorithms

	$\theta = 0.1$		$\theta = 0.25$		$\theta = 0.5$		$\theta = 1$		$\theta = 2$	
Perf. Meas.	Exact	Sim.	Exact	Sim.	Exact	Sim.	Exact	Sim.	Exact	Sim.
E[X]	1.52E+2	1.52E+2	1.22E+2	1.22E+2	1.11E+2	1.11E+2	1.05E+2	1.05E+2	1.01E+2	1.01E+2
rel. err.		±6.75E-1		±2.78E-1		±1.46E-1		±8.47E-2		±5.61E-2
Var(X)	9.25E+2	9.07E+2	3.47E+2	3.46E+2	1.81E+2	1.80E+2	1.05E+2	1.05E+2	6.85E+1	6.81E+1
rel. err.		±2.12E+2		±7.03E+1		±3.32E+1		±1.79E+1		±1.12E+1
E[Q]	5.22E+1	5.20E+1	2.30E+1	2.31E+1	1.27E+1	1.27E+1	7.03E+0	7.01E+0	3.88E+0	3.90E+0
rel. err.		±6.67E-1		±2.62E-1		±1.26E-1		±6.15E-2		±3.06E-2
Var(Q)	8.99E+2	8.82E+2	3.05E+2	3.06E+2	1.35E+2	1.34E+2	5.92E+1	5.91E+1	2.57E+1	2.58E+1
rel. err.		±7.88E+1		±1.55E+1		±4.53E+0		±1.41E+0		±4.55E-1
E[W]	5.14E-1	5.12E-1	2.29E-1	2.30E-1	1.28E-1	1.28E-1	7.22E-2	7.20E-2	4.09E-2	4.11E-2
rel. err.		±6.52E-3		±2.57E-3		±1.24E-3		±6.16E-4		±3.13E-4
Var(W)	8.59E-2	8.41E-2	2.93E-2	2.94E-2	1.31E-2	1.31E-2	5.88E-3	5.88E-3	2.64E-3	2.64E-3
rel. err.		±7.58E-3		±1.50E-3		±4.45E-4		±1.42E-4		±4.69E-5
P^d	9.67E-1	9.67E-1	8.90E-1	8.91E-1	8.03E-1	8.03E-1	7.00E-1	6.99E-1	5.92E-1	5.95E-1
rel. err.		±1.93E-3		±2.91E-3		±3.17E-3		±3.11E-3		±2.81E-3
P^a	4.97E-2	4.98E-2	5.47E-2	5.50E-2	6.04E-2	6.07E-2	6.70E-2	6.65E-2	7.40E-2	7.40E-2
rel. err.		±8.36E-4		±8.51E-4		±8.43E-4		±8.29E-4		±8.34E-4

we see that a rough symmetric CI for $(E[Q])^2$ is 900 ± 30, using the gap at the upper limit, so that the relative halfwidth for the square of the mean is 3.3%. Our direct estimate of the second moment is $613 + (30)^2 = 1513$ and our direct estimate of its CI is 1513 ± 43.6, so that the relative halfwidth is 2.8%. Our approximation thus estimates the CI of the variance as 613 ± 43.6, so that the approximate relative halfwidth is 7.1%, which we judge to be conservative.

3.10 Conclusion

In this chapter we have developed and evaluated approximations for the key steady-state performance measures in the heavily loaded stationary $G/GI/n + GI$ model. These approximations are remarkably simple and easy to implement, even though they have a relatively complicated basis in MSHT limits. As can be seen from §3.3.2, the basic truncated Gaussian approximations (TGA) for $\rho > 1$ can be expressed in terms of (i) the deterministic fluid approximating pair $(w, Q) \equiv (w(\infty), Q(\infty))$ in Theorem 4.1.1 (b), (ii) the limiting variance pair $(\sigma_W, \sigma_X) \equiv (\sigma_W(\infty), \sigma_X(\infty))$ in Theorem 3.2.3 (b) and (iii) the standard (mean-0, variance-1) Gaussian cdf Φ and pdf ϕ. The TGA-G refinement in §3.3.3 only requires an adjustment to σ_W using the service-time scv c_s^2. The approximate probability of abandonment P^a in (3.23) also requires a numerical integration with the full patience pdf f. Most of these approximate performance measures are easier to compute than corresponding quantities in the Markov $M/M/n + M$ Erlang-A model; see [81, 107]. For that model, the advantage of such simple approximations was previously emphasized by [32, 115, 81].

The basis for these simple approximations is the collection of MSHT limits for the $G/GI/n+GI$ model in the efficiency-driven (ED) regime in [74, 75, 77, 117], reviewed in §3.2. These MSHT limits explain why the approximations require that the scale (number of servers) n and the load (traffic intensity) ρ be suitably large, and that the abandonment rate be not too high.

It is significant that the approximations go beyond a direct application of the MSHT limits. We developed the approximations in §3.3. After presenting the direct DGA Gaussian approximations in §3.3.1, we applied truncation to obtain the refined TGA approximations in §3.3.2 and then subsequently, in §3.3.3, we heuristically modified the MSHT limit in [77] for non-exponential GI service to obtain the final TGA-G approximations, which coincide with TGA for exponential service times.

In §§3.4-3.7 we report results of extensive simulations studying the approximations. These experiments show that, for large scale with $n = 100$, the approximations are effective for a significant range of the traffic intensity (ρ) and the abandonment rate (θ) parameters, roughly for $\rho > 1.02$ and $\theta < 2.0$. After first comparing the approximations to exact numerical results for the Markov $M/M/n + M$ model in §3.4, we carefully examined the impact of non-Markov elements for the arrival process (including a non-renewal MMPP example) and the patience distribution in §3.5 and for the service time distribution in §§3.6.1 and 3.6.2. In §3.7 we showed that these approximations also remain effective for smaller scale, assuming that the remaining parameters are adjusted appropriately. In §3.9 we described the simulation methodology. Additional details are provided in an online appendix.

3.10.1 The Impact of Model Features on System Performance

It is well known that congestion tends to be increasing in λ and decreasing in μ, n and θ. The way that congestion depends on these basic parameters is quite well understood. We refer to [118] for a careful sensitivity study of performance in the Erlang-A model, which shows that performance is quite sensitive to small percentage changes in the arrival rate λ or the service rate μ (and thus the traffic intensity ρ), but is relatively insensitive to small changes in the abandonment rate θ.

The MSHT limits help expose the structure as well, as discussed for Markov models in [38, 32, 115]. First, the MSHT scaling shows that, for n not too small, performance depends on n and ρ primarily through the single parameter $(1-\rho)\sqrt{n} \equiv \beta$, with $\beta < 0$ corresponding to the overloaded case. Second, §4 of [115] shows for the $M/M/n + M$ model that performance depends on n and θ primarily through n/θ for large values of that ratio.

The fluid limits for the general $G/GI/n+GI$ model in Theorem 4.1.1 are very useful for exposing the primary impact of model elements upon performance. Theorem 4.1.1 shows that performance primarily depends on the arrival process only via its rate and on the service-time distribution only via its mean. In contrast, when $\rho > 1$, the fluid limit depends on the full patience-time cdf F,

but then only on the value of its inverse F^{-1} at the limiting fluid abandonment rate $(\rho - 1)/\rho$. the head-of-line waiting time $w(\infty)$ is determined by $w(\infty) = F^{-1}((\rho - 1)/\rho)$. From a practical engineering perspective, we see that this characteristic of the patience cdf is critical, not the mean or variance. Table 3.6 illustrated that the patience cdf matters beyond its mean and variance. The fluid approximation shows that, if the full patience cdf F gets larger in stochastic order, i.e., if the function $\bar{F}$ increases so that customers are more patient, then congestion should increase, just as in the $M/M/n + M$ model.

Moreover, As we have observed in §3.5, the general stationary G arrival process only affects the FCLT limit in Theorem 3.2.3 though the asymptotic variance parameter c_λ^2 appearing in the assumed FCLT for the arrival process in (4.2). That was illustrated in Table 3.3 when we displayed results for models that have $M(1)$ and $LN(1, 1)$ arrival processes. Theorem 3.2.3 shows that congestion tends to increase appproximately proportional to c_λ^2. In contrast, the patience-time cdf F affects the FCLT limit in a complicated way.

A main conclusion in [107, 117] was that the steady state performance of the $M/GI/n + GI$ model tends to be nearly insensitive to the service-time distribution beyond its mean. Our experiments confirm that conclusion for the main performance measures considered, e.g., for the mean values of the steady-state queue length and waiting time, but we show that the variance and full distribution depend significantly on the service-time distribution beyond its mean. Moreover, our refined TGA-G approximation successfully captures that effect.

Because Theorem 3.2.3 is for the $G/M/s + GI$ model, we do not directly see the impact of the service-time distribution, but the fluid limits suggests that it is not so great. Our heuristic refinement TGA-G depends on the service-time distribution only through its first two moments. In particular, the variance $\sigma_{W_G}^2$ in §3.3.3 is increasing in c_s^2. Because the approximation is quite effective, we conclude that the scv c_s^2 captures much of the impact.

3.10.2 The Impact of Model Features on the Accuracy of the Approximations

Much of this chapter has been devoted to carefully examining the accuracy of the proposed engineering approximations. First, for the $M/M/n + M$ model, Tables 3.1 and 3.2 and Figures 3.1-3.3 show that the accuracy of the approximation is quite good for large scale, which we take to be $n = 100$, provided that the model is reasonably overloaded (OL), as specified by our OL condition $\rho > 1.02$ and $\theta < 2.0$. Given that performance depends on n and ρ for n not too small primarily through the single parameter $(1 - \rho)\sqrt{n} \equiv \beta$, with $\beta < 0$ corresponding to the overloaded case, it is noteworthy that for $n = 100$ our OL condition $\rho > 1.02$ corresponds to $\beta < -0.2$. For large n,

Clearly, the system ceases to be OL as ρ decreases toward critical loading (CL), characterized by $\rho = 1$, and as θ increases above 2.0, as shown in Figures 3.1 and 3.3. The degradation of performance

for $\theta = 4$ and 10 is shown in Figure A.4 in the appendix. These figures show that, for $\rho > 1$, a high abandonment rate θ degrades approximation accuracy the most.

Tables 3.3 and 3.4 and Figure 3.4 show that the accuracy is quite insensitive to the arrival process variability as characterized by c_λ^2. Table 3.5 and Figure 3.6 show that the same is true for the variability of the patience distribution as characterized by c_{ab}^2 to a large extent, but there is some degradation as c_{ab}^2 gets small.

3.10.3 Directions for Future Research

There are many good directions for future research. First, exact performance measures are still needed for the $M/GI/n+GI$ model and more general models with GI service. Second, MSHT limits are still needed for the $G/GI/n+GI$ models with abandonment and GI service, going beyond the established MSHT fluid limits. More generally, It remains to provide better theoretical justification for the TGA-G approximation with non-exponential service-time distributions in §3.3.3 and/or even better simple approximations, if possible. It also remains to develop effective approximations for other ranges of the parameters. It remains to find and study new approximations such as those recently proposed in [43]; they seem to effectively cope with rapidly changing patience hazard rate functions caused by delay annnouncements over time.

Chapter 4

Approximations for Time-Varying $G_t/GI/s_t + GI$ Queues

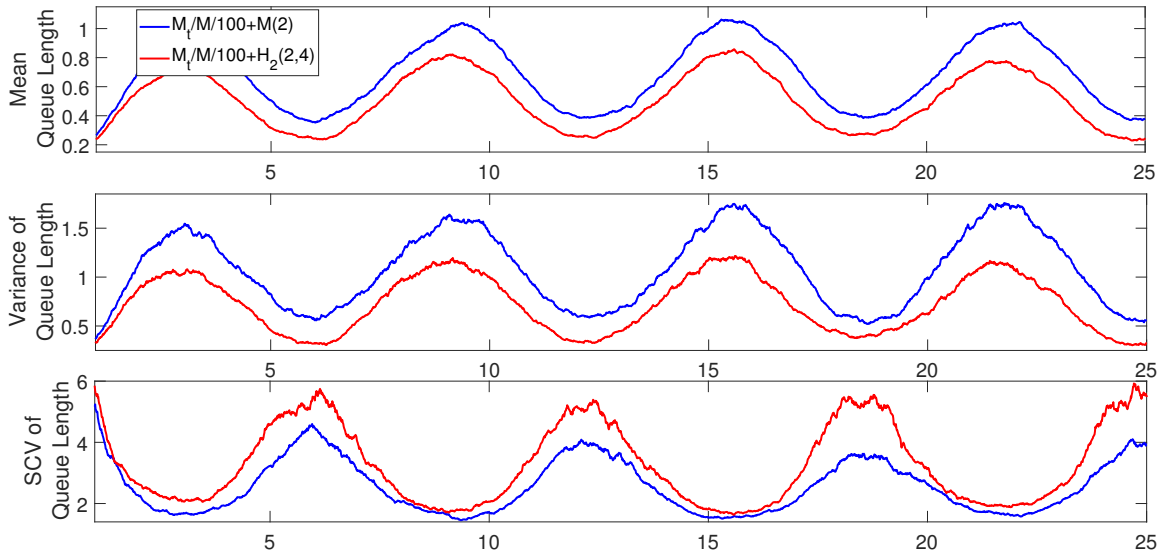
4.1 Introductions

This chapter is a sequel to Chapter 3, in which we build new engineering approximation formulas for steady-state heavily-loaded non-Markovian queueing systems with customers abandonment, denoted as the $G/GI/s + GI$, allowing non-Poisson arrival process (the first G), independent and identically distributed (i.i.d.) service times (the second GI) and customers patience times (the last GI). By extending Chapter 3, we establish engineering approximations for time-varying $G_t/GI/s_t + GI$ queueing, where the customer arrival rate varies as time evolves (the first G_t) and a time-dependent staffing the number of servers is allowed (the s_t).

Models accounting for time-varying features attract substantial interests in recent years because of their application to call centers, healthcare systems and other service systems, see [1] and [122]. It is then natural to use nonhomeogeneous Poisson process (NHPP) to simulate the time-varyinig arrival processes (see [68, 67]). However, statistical analysis of a teaching hospital in South Korea reveals that the estimated variance-to-mean ratio of arrival processes is far from that of a NHPP (see [60, 72] for more details). Therefore, to validly simulate the time-varying arrivals, we need a more general stochastic process: the nonstaionary non-Poisson processes (NNPP). With time-varying arrivals, system alternatively experiences UL (underloaded intervals as previously defined) and OL (overloaded intervals as previously defined), and such alternating behavior complicates analysis of system performance.

According to empirical studies on service systems in [16], service times and customers' patience times are usually not exponentially distributed, see Figure 1.1 in [16]. Therefore, in this chapter, we are motivated to characterize performance measures for the time-varying $G_t/GI/s_t + GI$ model exhibiting generally distributed service time and abandonment (customers' patience) times. We

Figure 4.1: Comparisions between queue lengths in Markovian $M_t/M/s+GI$ model and general $G_t/GI/s+GI$ model



provide simple formulas to approximate the time-depedent mean functions and variance functions of performance measures, including the total number of customers in the system, the number in queue, the number in service and waiting time.

Based on Markovian $M_t/M/s_t+M$ model, Massey and Pender [85] established Gaussian approximations by solving differential equations. Despite the tractability of the Markovian models, we observe that approximating a $G_t/GI/s_t+GI$ model by its associated $M_t/M/s_t+M$ model with the same mean service and abandonment times can be ineffective. Figure 4.1, for example, compares the simulated mean and variance functions of queue lengths of a Markovian $M_t/M/s+M$ model and a $G_t/GI/s+GI$ model, sharing the same arrival rate function $\lambda(t)=0.6\sin t+1$, system scaler n=100, constant number of servers $s(t)=1$, exponential service time with mean 1, lognormal inter-arrival times with SCV $c_\lambda^2=1$, and lognormal abandonment times with rate $\theta=0.5$ and SCV c_a^2.

From that figure, we can see significant performance gaps between two queues. Therefore, simply using Markovian models approximate the $G_t/GI/s+GI$ queue might lead to serious errors and poor decisions. Our work is to develop better approximations for those general queues.

The time-varying $G_t/GI/s_t+GI$ is characterized by several model parameters: the time-dependent arrival rate function $\lambda(t)$ and constant SCV c_λ^2, a general service time cdf G with mean $1/\mu$ and SCV c_s^2, a general patience time cdf F with mean $1/\theta$ and SCV c_a^2. Firstly, we explain what the G_t arrival with time-varying rate function $(\lambda(t))$ and constant SCV c_λ^2 means in our chapter. The G_t arrival process, denoted as $A(t)$, is a process associated with inter-arrival times. The rate

function $\lambda(t)$ provides the mean value of the process $A(t)$, i.e. $A(t)/t \to \Lambda(t)/t = (1/t)\int_0^t \lambda(s)\mathrm{d}s$, and the SCV c_λ^2 gives the variance of inter-arrival times. In contrast to renewal process, the inter-arrival times are mutually independently with each other but they are NOT i.i.d. random variables. Each inter-arrival time has a unique mean value depended on time t whereas they share the same level of variation c_λ^2. To generate a G_t arrival process, we apply the algorithm developed in [**?** 74]. The intuition idea is to use the identity feature of (equilibirum) stationary (delayed) renewal process that is, for any time t, we have $\mathrm{E}[A(t)] = \Lambda(t)$. Hence, based on a rate-1 stationary renewal process (SRP) with variance c_λ^2, denoted by $N(t)$, we generate the desired time-varying arrival process $A(t) = N(\Lambda(t))$, by scaling the time of the rate-1 SRP $\{N(t), t \geq\}$. Secondly, without loss of generality, we assume the mean service time $1/\mu = 1$.

Under the aforementioned setting, we aim to establish new performance approximations for the $G_t/GI/s_t+GI$ model and conduct extensive numerical experiments to study the impacts of system parameters on system performance and the robustness of the proposed approximations. We next review the MSHT limits.

4.1.1 Many-Server Heavy-Traffic Limits of $G_t/M/s_t + GI$ Model

The MSHT limits are obtained by considering a sequence of time-varying queueing models indexed by integers n. In the n^{th} model, both the number of servers and the arrival rate are scaled by n from the $G_t/GI/s_t+GI$ model, i.e. $s_n(t) = ns(t)$, and $\lambda_n(t) = n\lambda(t)$. Other system parameters such as distributions of service times and abandonment times are keep the same.

The arrival process $N_n(t)$ is assumed to satisfy a FWLLN

$$\frac{N_n(t)}{n} \Rightarrow \Lambda(t) = \int_0^t \lambda(t)\mathrm{d}t, \quad \text{in } \mathbb{D} \text{as } n \to \infty, \tag{4.1}$$

and FCLT

$$\left(N_n(t) - n\lambda\, t\right)/\sqrt{n} \Rightarrow c_\lambda \mathcal{B}(t) \quad \text{in } \mathbb{D} \quad \text{as } n \to \infty, \tag{4.2}$$

where $\mathcal{B}$ is a standard Brownian motion (BM), $c_\lambda > 0$ is a variability parameter, $\Rightarrow$ denotes weak convergence and $\mathbb{D}$ denotes, as defined in previous chapters, the space of right-continuous functions that have left limits; see [9, 77, 113] for details.

Let $Q_n(t)$ and $B_n(t)$ to be the number of customers waiting in queue and in service at time t. Let $X_n(t) = Q_n(t) + B_n(t)$ be the total number in the system. Let $W_n(t)$ and $V_n(t)$ to be the head-of-line waiting time (the elapsed waiting time of the head-of-line customer if there is any) and the potential waiting time at time t (the waiting time of an infinitely patient arrival at time t if there were an arrival at that time). We next review the MSHT FWLLN and FCLT limits for the time-varying queueing model, which are the building blocks for our Gaussian approximations.

Define the LLN-scaled processes be

$$\bar{X}_n(t) = \frac{X_n(t)}{n}, \quad \bar{Q}_n(t) = \frac{Q_n(t)}{n}, \quad \bar{B}_n(t) = \frac{B_n(t)}{n}, \quad \bar{W}_n(t) = W_n(t) \quad \text{and} \quad \bar{V}_n(t) = V_n(t).$$

The waiting times need no spatial scaling because the service-time and patience-time cdf's are not scaled. By Theorem 1 in [75], we have the joint convergence for the LLN-scaled functions

$$\left(\bar{X}_n, \bar{Q}_n, \bar{B}_n, \bar{W}_n, \bar{V}_n\right) \quad \to \quad (X, Q, B, w, v), \quad \text{in } \mathbb{D}^5, \tag{4.3}$$

where the limit is a vector of deterministic functions specified in [117, 74]. We give those limit functions as below.

Theorem 4.1.1 (Theorem 1 in [75])**.** *The fluid model with capacity s_t arising as the MSHT FWLLN limit of time-varying $G_t/GI/s_t + GI$ model with $s(t)$ servers can be by the deterministic functions $(B(t), Q(t), X(t), w(t))$ in $\mathbb{D}^5$. Here $X(t) = B(t) + Q(t)$ and the other functions are computed as follow.*

(a) If the system is in UL,

$$X(t) = B(t) = X(0)\bar{G}(t) + \int_0^t \lambda(s)\bar{G}(t-s)\,ds \quad \textit{and} \quad Q(t) = w(t) = 0.$$

(b) If the system is in UL, the service content $B(t) = s$, and, by Theorem 3 in [74], the fluid head-of-line waiting time $w(t)$ is the unique solution to the ODE

$$w'(t) = 1 - \frac{b(t,0)}{\lambda(t - w(t))\bar{F}(w(t))},$$

where $b(t,0)$ satisfies the fixed point equation

$$b(t,0) = s'(t) + \int_0^\infty \frac{b(0,y)g(t+y)}{\bar{G}(y)}\,dy + \int_0^t b(t-x,0)g(x)\,dx. \tag{4.4}$$

The number of customers waiting in the system, $Q(t)$, is defined

$$Q(t) = \int_0^{w(t)} \lambda(t-s)\bar{F}(s)\,ds, \tag{4.5}$$

and thus $X(t) = B(t) + Q(t)$.

We now review the FCLT limits. As in [77], we now restrict to the $G_t/M/s_t + GI$ model having

exponential service (the M). To provide the FCLT limits, let the CLT-scaled processes be

$$\hat{X}_n(t) = \frac{X_n(t) - nX(t)}{\sqrt{n}}, \quad \hat{Q}_n(t) = \frac{Q_n(t) - nQ(t)}{\sqrt{n}}, \quad \hat{B}_n(t) = \frac{B_n(t) - nB(t)}{\sqrt{n}},$$
$$\hat{W}_n(t) = \sqrt{n}\,(W_n(t) - w(t))\,, \quad \hat{V}_n(t) = \sqrt{n}\,(V_n(t) - v(t))\,, \tag{4.6}$$

where n is again the system scaler, while $X(t), Q(t), B(t), w(t)$ and $v(t)$ are the deterministic limit functions in (4.3), i.e., the deterministic fluid functions in [74].

Theorem 4.1.2 (MSHT FCLT limits for the $G_t/M/n + GI$ queues)**.** *Consider the sequence of $G_t/M/s_t + GI$ queueing models having $s(t)$ servers. Under regularity conditions in [77], including appropriate initial convergence at time* 0*,*

$$\left(\hat{X}_n, \hat{B}_n, \hat{Q}_n, \hat{W}_n, \hat{V}_n\right) \Rightarrow \left(\hat{X}, \hat{B}, \hat{Q}, \hat{W}, \hat{V}\right) \quad \text{in } \mathbb{D}^5 \quad \text{as } n \to \infty.$$

(a) If the system is in UL interval, we have $\hat{Q} = \hat{W} = \hat{V} = 0$, and $\hat{B} \stackrel{\mathrm{d}}{=} \hat{X}$ satisfies the stochastic differential equation (SDE)

$$d\hat{X}(t) = -\mu\hat{X}(t)dt + \sqrt{c_\lambda^2\lambda + \mu X(t)}\,d\mathcal{B}(t),$$

where $\mathcal{B}$ is a standard Brownian motion, so that $\hat{X}(t)$ is a Gaussian process with

$$\sigma_X^2(t) \;=\; (c_\lambda^2 - 1)\int_0^t \bar{G}^2(t-u)\lambda du + \int_0^t \bar{G}(t-u)\lambda du.$$

(b) If the system is in OL interval, we have $\hat{B}(t) = 0$, $\hat{Q}(t) = \hat{X}(t)$ where

$$\begin{aligned}
\hat{X}(t) &= \hat{X}(0)\bar{F}_w(t) + \sum_{i=1}^{3}\int_0^t K_i(t,u)\,d\mathcal{B}_i(u),\\
\hat{W}(t) &= \hat{W}(0)H(t,0) + \sum_{i=1}^{3}\int_0^t H(t,u)I_i(u)\,d\mathcal{B}_i(u),\\
\hat{V}(t) &= \frac{\hat{W}(t+v(t))}{1-\dot{w}(t+v(t))},
\end{aligned}$$

where $\mathcal{B}_1 \equiv \mathcal{B}_\lambda$, $\mathcal{B}_2 \equiv \mathcal{B}_s$ and $\mathcal{B}_3 \equiv \mathcal{B}_a$ are independent standard BMs, that are the FCLT limits of the scaled arrival process (the subscript "λ"), service process (the subscript "s") and abandonment process (the subscript "a"). Both $\hat{X}(t)$ and $\hat{W}(t)$ are zero mean Gaussian processes with variance

process

$$\begin{aligned}\sigma_W^2(t) &= \int_0^t H^2(t,u)I^2(u)\,du + H^2(t,0)\,Var\left(\hat{W}(0)\right),\\ \sigma_X^2(t) &= \int_{t-w(t)}^t \lambda\,\bar{F}(t-u)\left((c_\lambda^2-1)\bar{F}(t-u)+1\right)du\\ &\quad +\lambda^2\bar{F}^2(w(t))\sigma_W^2(t) + Var\left(\hat{X}(0)\right)\cdot\left(\bar{F}_w(t)\right)^2.\end{aligned}$$

Here, $H(t,u)$, $I_i(t)$, $K_i(t)$ are deterministic analytic functions given by

$$\begin{aligned}H(t,u) &= \exp\left\{\int_u^t h(v)dv\right\},\\ I_1^2(u) &= \frac{c_\lambda^2\bar{F}(w(u))b(u,0)}{\tilde{q}^2(u,w(u))},\\ I_2^2(u) &= \frac{b(u,0)}{\tilde{q}^2(u,w(u))},\\ I_3^2(u) &= \frac{F(w(u))b(u,0)}{\tilde{q}^2(u,w(u))},\\ I^2(t) &= I_1^2(t)+I_2^2(t)+I_3^2(t) = \frac{b(u,0)\left[1+(c_\lambda^2-1)\bar{F}(w(u))\right]}{\tilde{q}^2(u,w(u))},\\ K_1(t,u) &= c_\lambda\bar{F}(t-u)\sqrt{\lambda(u)}\mathbb{1}_{\{t-w(t)u<t\}} + \tilde{q}(t,w(t))\sqrt{\lambda(u)}\bar{I}_1(L^{-1}(u))\mathbb{1}_{\{0\le u\le t-w(t)\}},\\ K_2(t,u) &= -\sqrt{b(t,0)-\dot{s}(t)}H(t,u),\\ K_3(t,u) &= -\sqrt{\lambda(u)F(t-u)\bar{F}(t-u)}\mathbb{1}_{\{t-w(t)\le u\le t\}}\\ &\quad + \tilde{q}(t,w(t))\sqrt{\lambda(u)}\bar{I}_3(L^{-1}(u))H(t,L^{-1}(u))\mathbb{1}_{\{0\le u\le t-w(t)\}},\end{aligned}$$

where

$$\bar{I}_1(t) = \frac{c_\lambda\bar{F}(w(u))b(u,0)}{\tilde{q}(u,w(u))},\quad \bar{I}_3(t) = -\frac{\sqrt{\bar{F}(w(u))b(u,0)}}{\tilde{q}(u,w(u))}\quad \textit{and,}$$

$L(t)=t-w(t)$, $\lambda(t)=\lambda t$ and $\mathbb{1}_A$ is an indicator random variable of A.

4.1.2 Contributions and Organization

We establish new engineering approximations based on the MSHT FWLLN and FCLT for the time-varying $G_t/M/s_t+GI$ queueing model proved in [77]. The FWLLN and FCLT limits, on the one hand, immediately provides DGA, and on the other hand, indicate when DGA should be effective. Because MSHT limits are obtained in either UL or OL by letting system scaler n increase without bound, we expect DGA perform well for large-scale system staying in either UL or OL,

repectively. In this chapter, we verified our expectations by numerical experiments: DGA indeed become inefficient for smaller systems such as the system scaler $n = 50, 10, 5$, etc, or systems at the time near switching points.

Given the ineffective approximations that DGA provide, we here develop effective heuristic refinements, TGA and TGAG, for $G_t/M/s_t+GI$ and $G_t/GI/s_t+GI$ queueing systems. We conduct extensive numerical studies to investigate when these approximations are effective. We find that the effectiveness of refined Gaussian approximations, TGA and TGAG, depend on abandonment rate θ, stochastic variability of service times captured by SCV's c_s^2, and system scaler n. In particular, when abandonment rate is not too big, say $\theta \leq 2$, and service times are have reasonable fluctuations, say $c_s^2 \in (0.25, 4)$, and system scale is not too small, say $n \geq 3$, then TGA and TGAG are effective.

For abandonment rates θ, since we only consider queueing models with abandonment, we here only allow abandonment rates range at a typical level. We do not consider too big (small) abandonment rates in which case the system behaves more like loss models (infinity server queues).

This chapter is organized as follows. In § 4.2, we explain our simulation methodology and measures to evaluate the accuracy of various approximations. In § 4.3, we establish the new approximations: DGA, TGA and TGAG. We then investigate the applicable range of TGA in § 4.3.2 for time-varying $G_t/M/s_t+GI$ from three aspects: the time variability of arrival rate functions, the stochastic fluctuation of arrival processes, and the fluctuation of customers' abandonment times. In § 4.4, we consider $G_t/M/s_t+GI$ models, extended $G_t/GI/s_t+GI$ models, and smaller systems. We draw conclusions in § 4.5.

4.2 Simulation Methodology and Performance Measures

In this section, we indicate how we use simulations to estimate the exact values for time-varying queueing models, and how we evaluate the performance of our candidate approximations.

In all numerical experiments, we run in total $R = 2,000$ independent replications over the time interval $[0, T]$ with $T = 24$ and $T = 48$. In the r^{th} path, we periodically generate virtual arrivals as system observer at time $\Delta t, 2\Delta t, \ldots, N_r\Delta t$, where $\Delta t = 0.01$ and $N_r \equiv \lfloor T/\Delta t \rfloor$. Once being generated, the kth virtual observer collects the current number of customers in queue and service, denoted as $Q_r(k), B_r(k)$, and the head-of-line waiting time, $W_r(k)$. Then she joins the waiting queue with a pseudo patience time, after which she is labelled as "abandoned" but still stays in the system until her enter the service. Instead of entering the service, the observer will leave the system, and her virtual waiting time is sampled as the potential waiting time denoted as $V_r(k)$, the waiting time assuming customers have infinite patience.

We sample continuous-time queue-length process, number of busy server process, head-of-line waiting time process and potential waiting time process by $Q_r(k), B_r(k), W_r(k)$ and $V_r(k)$'s at discrete time points. Then, to construct estimators of actual mean and variance processes, we take

average of mean and variance processes among all R replications at each time point $k\Delta t$. Taking an example of queue-length process Q, we use $\bar{Q}(k)$, $\bar{Q}^{(2)}(k)$, and $V_Q(k)$ to denote the estimator of mean, second moment, and variance of the queue length at time $k\Delta t$.

$$\bar{Q}(k) = \frac{1}{R}\sum_{r=1}^{R} Q_r(k), \quad \text{and} \quad \bar{Q}^{(2)}(k) = \frac{1}{R}\sum_{r=1}^{R} Q_r^2(k), \quad \text{and} \quad V_Q(k) = \bar{Q}^{(2)}(k) - \left(\bar{Q}(k)\right)^2 \tag{4.7}$$

Similarly, we have the estimators of mean and variance processes number of busy servers B, head-of-line waiting time W and potential waiting time V.

We evaluate the performance of our candidate approximations by *average squared error* (ASE) defined in [45, 47, 46]. Again taking an example of queue-length process Q, we write

$$\mathrm{ASE}\left(\bar{Q}\right) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{\bar{Q}(k)}{n} - \mathrm{E}\left[e_i(k)\right]\right)^2 \quad \text{and} \quad \mathrm{ASE}\left(V_Q\right) = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{V_Q(k)}{n} - \mathrm{Var}\left[e_i(k)\right]\right)^2 \tag{4.8}$$

where $e_i(k)$ is the candidate approximations at time $k\Delta t$, either DGA or TGA, n is the system scaler, and $N = \min\{N_r, r = 1, \ldots, R\}$ is the number of sampled customers. We scale the actual estimations by system scalar n so that the average squared error is comparable across systems of different scales. It is worth mentioning that, due to the way we obtained heavy-traffic limits in § 4.1.1, the ASE of waiting time process, including both head-of-line (W) and potential waiting time (V), is different to the queue length process or number in service process.

$$\mathrm{ASE}\left(\tilde{W}\right) = \frac{1}{N}\sum_{k=1}^{N}\left(\tilde{W}(k) - \mathrm{E}\left[e_i(k)\right]\right)^2 \quad \text{and} \quad \mathrm{ASE}\left(V_W\right) = \frac{1}{N}\sum_{k=1}^{N}\left(nV_W(k) - \mathrm{Var}\left[e(k)\right]\right)^2,$$

where $e(k)$ is the corresponding candidate approximation, being either DGA or TGA.

As explained in [47], the ASE estimates the weighted mean squared error (WMSE), defined as follows.

$$\begin{aligned}
\mathrm{WMSE}(\bar{Q}) &= \frac{1}{\Lambda(T)}\int_0^T \lambda(t)\left(\frac{\mathrm{E}[Q_n(t)]}{n} - \mathrm{E}[e(t)]\right)\mathrm{d}t,\\
\mathrm{WMSE}^2(V_Q) &= \frac{1}{\Lambda(T)}\int_0^T \lambda(t)\left(\frac{\mathrm{Var}(Q_n(t))}{n} - \mathrm{Var}(e(t))\right)\mathrm{d}t,\\
\mathrm{WMSE}(\bar{W}) &= \frac{1}{\Lambda(T)}\int_0^T \lambda(t)\left(\mathrm{E}[W_n(t)] - \mathrm{E}[e(t)]\right)\mathrm{d}t,\\
\mathrm{WMSE}^2(V_W) &= \frac{1}{\Lambda(T)}\int_0^T \lambda(t)\left(n\mathrm{Var}(W_n(t)) - \mathrm{Var}(e(t))\right)\mathrm{d}t,
\end{aligned}$$

where $\Lambda(T) = \int_0^T \lambda(t)\mathrm{d}t$ is the mean total number of arrivals during time $[0, T]$ and $e(t)$ is the

candidate approximation, being either DGA or TGA. To measure the effectiveness of DGA and TGA relative to the average of system actual estimators, we adopt *relative root average squared error* (RRASE) defined as Eq. (5) in [47].

$$\mathrm{RRASE}(\bar{Q}) \equiv \frac{\sqrt{\mathrm{ASE}}(\bar{Q})}{(1/N)\sum_{k=1}^{N}\bar{Q}(k)/n} \quad \text{and} \quad \mathrm{RRASE}(V_Q) \equiv \frac{\sqrt{\mathrm{ASE}}}{(1/N)\sum_{k=1}^{N}V_Q(k)/n}, \tag{4.9}$$

$$\mathrm{RRASE}(\bar{W}) \equiv \frac{\sqrt{\mathrm{ASE}}}{(1/N)\sum_{k=1}^{N}\bar{W}(k)} \quad \text{and} \quad \mathrm{RRASE}(\bar{W}) \equiv \frac{\sqrt{\mathrm{ASE}}}{(1/N)\sum_{k=1}^{N}nV_W(k)}.. \tag{4.10}$$

Again, the RRASE of number in service process (B) coincides with (4.9) whereas the RRASE of potential waiting time (V) coincides with (4.10).

4.3 Gaussian Approximations for Time-Varying Models

In this section, we develop effective Gaussian approximation formulas for the time-varying $G_t/M/n+GI$ model. The MSHT limit developed in [74, 77], as reviewed in § 4.1.1, immediately provides DGA for various stochastic process measuring queueing performance. However, effectiveness of DGA drastically decays as the system switches from a underloaded regime to an overloaded regime, or vice versa, which motivates us to refine approximations. We name the newly developed Gaussian approximations by TGA, because a technique of truncation is adopted in our refinements.

We start in § 4.3.1 by reviewing DGA and then turn into the truncation refinement (i.e. TGA) for the $G_t/M/n+GI$ model in § 4.3.2. We then heurstically generalize TGA to the $G_t/GI/s+GI$ model having non-exponential service time distributions, which we denote them by TGAG.

4.3.1 Direct Gaussian Approximations

Many-server heavy-traffic limits in Theorems 4.1.1 and 4.1.2 directly provides Gaussian approximations for time-varying $G_t/M/s_t+GI$ models, which consist of deterministic fluid limit and the zero-mean Gaussian limits. We denote these approximations by using the superscript "DGA".

$$\begin{aligned}
\frac{X_n(t)}{n} &\approx X^{DGA}(t) \equiv X(t) + \frac{1}{\sqrt{n}}\hat{X}(t),\\
\frac{B_n(t)}{n} &\approx B^{DGA}(t) \equiv B(t) + \frac{1}{\sqrt{n}}\hat{B}(t),\\
\frac{Q_n(t)}{n} &\approx Q^{DGA}(t) \equiv Q(t) + \frac{1}{\sqrt{n}}\hat{Q}(t),\\
W_n(t) &\approx W^{DGA}(t) \equiv w(t) + \frac{1}{\sqrt{n}}\hat{W}(t),
\end{aligned}$$

where X, B, Q, w are given in Theorem 4.1.1 and $\hat{X}, \hat{B}, \hat{Q}$ and $\hat{W}$ in Theorem 4.1.2.

Though DGA are convenient to implement, they can only used for estimating large-scale queueing system staying away from switching points. That is to say, when the system approaches UL from OL, and vice versa, or when the system has fewer number servers, say $n = 10$, the performance of DGA is poor. See Figure 2 in [77] for DGA's bad performance in small system. Hence, we are motivated to develop refined approximations—TGA.

4.3.2 Truncated Gaussian Approximations

The effectiveness of DGA degenerates as system switches from UL to OL, or the other way around. It is because that the MSHT limits, which are the essential components of DGA, are obtained in UL or OL separately, and limits in different regimes do not coincide with each other. Therefore, when system is at the time near switching points, DGA fails to accurately estimate the system performance.

In order to refine DGA, we trace back the reason of the failure of DGA. In MSHT, when the system is switching from OL to UL, the queue length (and waiting times, etc.) mean function, according to MSHT in Theorem 4.1.1, approaches to zero. Suppose at time t, the FCLT limit $\hat{Q}(t)$ is negative and the mean $Q(t)$ is small enough, then the overall DGA Q^{DGA} could drop to negative, contradicting the nonnegativity of queue lengths. In other words, the distributions of DGA does not coincide with the that of real system measures, especially when $Q(t)$ (and other mean functions) or the system scaler n is so small that the corresponding DGA $Q^{DGA} = Q(t) + n^{-1/2}\hat{Q}(t)$ becomes negative. Figure 4.2 compares the distributions of DGA and the simulated queue length. In the left figure, when mean function is high, the Gaussian distribution of DGA has little difference from the real queue length distribution (the shaded area). However, in the right figure, when mean function is close to zero, the distribution of DGA significantly differs from that of practical queue lengths, so DGA fails.

It is then natural to refine the DGA through truncation because both the queue length and the number of busy servers can both be regarded as truncated processes of the total number of customers in the system $(X_n(t))$. Hence our TGA for queue length and number of busy servers processes are

$$Q_n^{TGA}(t) = \left(X_n^{DGA}(t) - s_n(t)\right)^+ \quad \text{and} \quad B_n^{TGA}(t) = X_n^{DGA}(t) \wedge s_n(t).$$

where $a^+ \equiv \max\{a, 0\}$ and $a \wedge b \equiv \min\{a, b\}$. Similarly, we truncate DGA for waiting times to obtain TGA, because the non-negativity of waiting time cannot be guaranteed in DGA when the system approaches UL from OL.

$$W_n^{TGA}(t) = \left(w(t) + \frac{\hat{W}}{\sqrt{n}}\right)^+ \quad \text{and} \quad V_n^{TGA}(t) = \left(v(t) + \frac{\hat{V}}{\sqrt{n}}\right)^+. \tag{4.11}$$

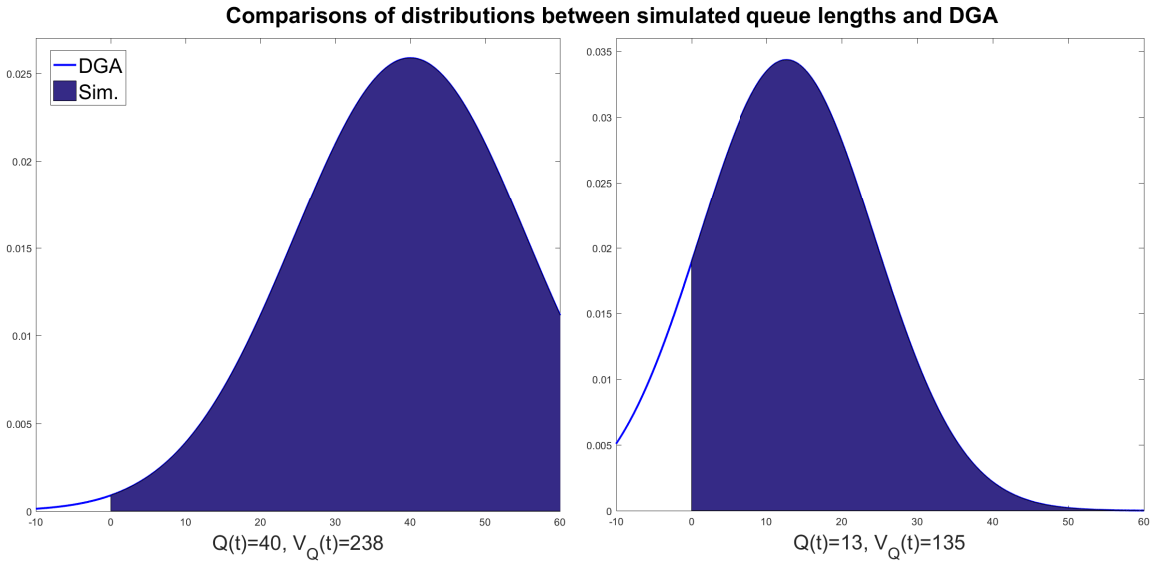


Figure 4.2: Comparisons of distributions between simulated queue lengths and DGA

Moreover, when system is in UL or switching from UL to OL, DGA of the waiting time is zero whereas that of the queue length is not, we can further refined the DGA of waiting time according to that queue length as follows.

When the system is in UL, we have

$$W_n^{TGA}(t) = \left(\sum_{i=1}^{Q^{TGA}(t)} \frac{1}{\mu s_n(t)} \right)^+.$$

4.3.3 Truncated Gaussian Approximations for Non-Exponential Service: TGAG

In this section, we move our focus to time-varying models with GI-service, the $G_t/GI/s+GI$ model. Recall that DGA and TGA are based on Theorem 4.1.2, which gives MSHT limits for time-varying $G_t/M/n+GI$ models having convenient exponential service times. However, through tracing back the formulas of the limit in Theorem 4.1.2, we are able to heuristically identify the impact of service time with the second model information—SCV.

In § 3.6.1, we refined TGA for stationary $GI/M/n+GI$ model to adapt to more general $GI/GI/n+GI$ model. For time-varying queueing model, following the similar idea, we target the service distribution component in the function of $I^2(t)$ defined in Theorem 4.1.2 and replace the simplified 1 by coefficient of variation of service time c_s.

$$I^2(t) = \frac{b(t,0)}{\tilde{q}(t,w(t))} \cdot \frac{\mathbf{c_s} + (c_\lambda^2 - 1)\bar{F}(w(t))}{\tilde{q}(t,w(t))}.$$

Then, based on the newly adapted $I^2(t)$ fuction, we update all approximations and name is TGA.

4.4 Evaluating approximations for $G_t/M/s+GI$ Models

Figure 4.3 below shows how TGA (dotted line) and DGA (dashed line) estimate system performance (solid line) of $H_2(t,2)/M(1)/100+H_2(2,2)$ model. The arrival rate function is sinusoidal: $\lambda(t)=1+0.6$, the staffing function is assumed to be constant $s(t)=1$ for convenience, the system scaler $n=100$, the service times are exponentially distributed with rate $\mu=1$, the abandonment times are hyper-2 exponentially distributed with rate $\theta=0.5$ and SCV $c_a^2=2$. The interarrival times are sampled from hyper-2 exponentially distributed i.i.d. with SCV $c_\lambda^2=2$.

In Figure 4.3, both DGA and TGA accurately estimate the system's mean (first moment) information. However, DGA performs poorly when estimating the variance (second moment) of system values, especially at switching point, where system changes its status from UL to OL, and vice versa. The poor performance of DGA is resulted from the aforementioned cut-off effects when system alternates between different regimes. The newly developed truncation techniques are designed to compensate such effects. After the truncation, TGA (in dotted line) significant improves the accuracy when estimating system's both mean and variance values.

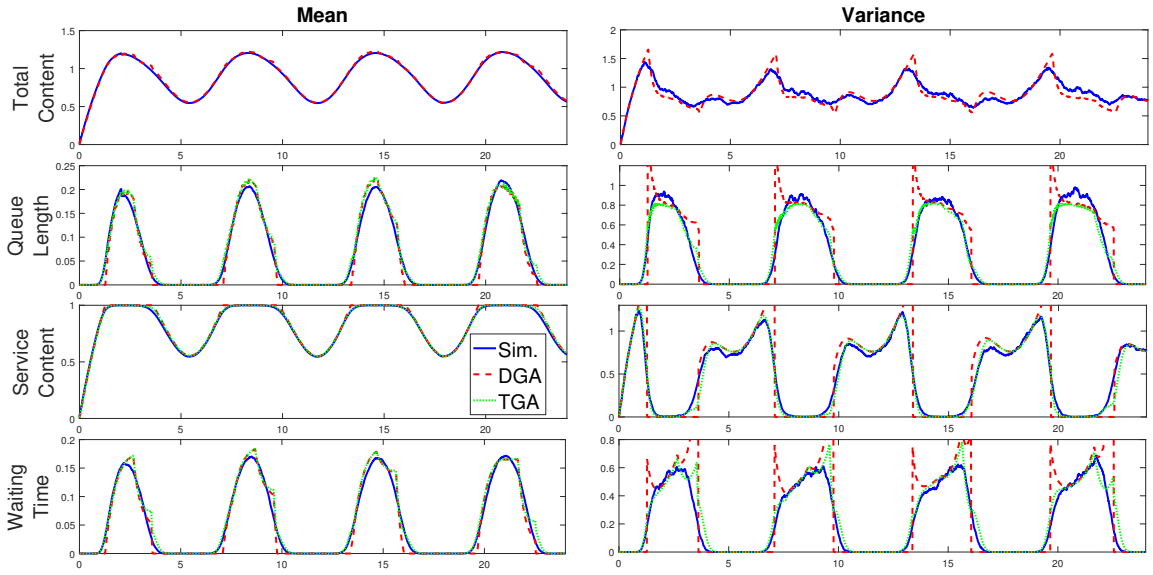


Figure 4.3: Comparison of the TGA and DGA approximations with the $H_2(t)/M(1)/100+H_2(2,2)$ model having arrival rate $\lambda(t)=60\sin t+100$ and SCV $c_\lambda^2=2$.

However, since refinements proposed above are in engineering sense, theoretical sensitivity anal-

ysis is hardly explorable. Hence, we conduct enormous numerical experiments to test the robustness of newly constructed TGA and TGAG. We start evaluations with $G_t/M/s + GI$ models, as our theoretical foundation, MSHT limits in [77], are derived in that setting.

4.4.1 Robustness in Abandonment Rates

We investigate the sensitivity of TGA with respect to abandonment rates, as abandonment significantly affects system performance. In particular, the system behaves like a loss queueing model (an infinity server queue) when abandonment rate approaches to infinite (zero). To study the impacts of abandonment rates, we choose our base model as $H_2(t, 2)/M(1)/100 + H_2(2, 2)$.

Table 4.1 evaluates the performance of DGA and TGA for a wide range of abandonment rates. Firstly, DGA is inefficient in estimating variance values of the system, whereas TGA significantly improves the accuracy and almost all RRASE is controlled below 20%, except for the variance of service content when abandonment rate is as low as $\theta = 0.25$. Notice that, when abandonment rate is close to 0, the whole system behaves more like $G_t/M/n$ loss queue. On the other hand, when abandonment rate is big, the system approaches to a loss model without buffer. Both cases are outside of scope of the available MSHT results in [77].

4.4.2 Robustness in the Time Variability of the Arrival Process

In real systems, arrival rates can be periodic functions with periods that depend on the operational time scale (e.g. a day, a week, etc.). In this section, we consider a sinusoidal arrival rate function because any periodic functions can be approximated by the sum of simple sinusoidal functions.

In Table 4.2, we estimate the RRASE of DGA and TGA when changing the relative amplitude of arrival rate function from $c = 0.2$ to 0.8. We find that, though the effectiveness of DGA drastically differs when changing the arrival rate functions, TGA's performance is stably good.

4.4.3 Robustness in the Stochastic Flutuation of the Arrival Process

In this section, we evaluate the effectiveness of DGA and TGA under different probabilistic distributions of arrival process. Though the theoretical base, the MSHT limits, are developed for $G_t/M/s_t + GI$ model allowing customers' arrival and abandonment to be non-exponentially distributed, the systems we are testing DGA and TGA for is in finite assumption. Therefore, the MSHT limits cannot precisely measure systems values, and it is required to conduct numerical simulations to compare the effectiveness of DGA and TGA on different levels of variations.

Indicated in Theorems 4.1.1 and 4.1.2, the G_t arrival processes affects the system values through its rate function and SCV c_λ^2. In § 4.4.2, we tested the DGA and TGA under different settings of arrival rate functions. So we now test the robustness of approximations in terms of the variation

Table 4.1: Evaluation of DGA and TGA on $H_2(\lambda^{-1}(t), 2)/M(1)/100 + H_2(\theta^{-1}, 2)$, where $\lambda(t) = 100 + 60\sin t$

	$\theta = 0.25$				$\theta = 0.5$			
	ASE		RRASE		ASE		RRASE	
Perf.	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA
E[X]	4.4E-4	same	1.8%	same	2.0E-4	same	1.4%	same
Var(X)	2.8E-2	same	5.5%	same	1.2E-2	same	5.7%	same
E[Q]	8.2E-4	3.4E-4	11.1%	7.2%	3.7E-4	1.4E-4	10.9%	6.7%
Var(Q)	3.2E-1	2.6E-2	25.0%	7.1%	1.1E-1	1.5E-2	25.9%	9.4%
E[B]	2.8E-4	4.6E-5	1.9%	0.8%	2.4E-4	3.8E-5	1.8%	0.7%
Var(B)	2.6E-1	1.5E-2	89.0%	21.2%	8.4E-2	1.5E-3	55.6%	7.3%
E[W]	1.6E-3	2.4E-4	16.5%	6.5%	5.6E-4	1.4E-4	15.0%	7.5%
Var(W)	9.2E+0	5.1E-3	137.7%	3.2%	1.1E+0	3.0E-4	94.8%	1.6%
	$\theta = 1$				$\theta = 2$			
	ASE		RRASE		ASE		RRASE	
Perf.	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA
E[X]	1.9E-4	same	1.5%	same	2.8E-4	same	1.9%	same
Var(X)	5.3E-3	same	5.9%	same	8.3E-3	same	10.3%	same
E[Q]	2.0E-4	8.8E-5	13.2%	8.8%	1.3E-4	8.8E-5	19.3%	15.8%
Var(Q)	5.5E-2	5.6E-3	35.8%	11.5%	3.2E-2	3.7E-3	58.1%	19.6%
E[B]	2.5E-4	6.9E-5	1.9%	1.0%	3.0E-4	1.2E-4	2.1%	1.4%
Var(B)	4.5E-2	1.5E-3	43.0%	7.8%	3.8E-2	3.6E-3	39.1%	11.9%
E[W]	2.1E-4	1.1E-4	15.7%	11.2%	1.1E-4	8.8E-5	21.0%	18.5%
Var(W)	1.4E-1	5.9E-4	77.4%	5.1%	2.9E-2	4.8E-4	79.1%	10.3%

level of arrival process, measured by SCV c_λ^2, in Table 4.3. However, different from those in previous numerical experiments adopting the simple hyper-2 exponential distributions, the interarrival times are lognormal distributed having a heavier tail than phase-type distributions, including the hyper-2. In order to know the impact of interarrival distribution other than first two moments (rate function and SCV), we add the case when arrival process is non-homeogeneous Poission process (NHPP). It can be seen from Table 4.3 that distributions other than mean and variance hardly affects the system performance, as the corresponding MSHT limits show. More importantly, Table 4.3 reveals that our TGA is rather robust when the variational level of arrival times vary, e.g. $c_\lambda^2 \in (0.25, 4)$.

Table 4.2: Evaluation of DGA and TGA on $H_2(\lambda^{-1}(t), 2)/M(1)/100 + H_2(2, 2)$, where $\lambda(t) = 1 + c \sin t$, and $c = 0.2, 0.4, 0.6$ and 0.8

	$c = 0.2$				$c = 0.4$			
	ASE		RRASE		ASE		RRASE	
Perf.	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA
E[X]	7.7E-5	same	0.9%	same	1.1E-4	same	1.0%	same
Var(X)	1.9E-2	same	7.7%	same	6.5E-3	same	4.3%	same
E[Q]	7.0E-4	3.0E-5	36.3%	7.6%	3.8E-4	5.7E-5	16.4%	6.4%
Var(Q)	4.1E-1	6.2E-3	77.4%	9.5%	1.8E-1	6.1E-3	39.1%	7.2%
E[B]	5.6E-4	2.8E-5	2.6%	0.6%	3.2E-4	3.2E-5	2.0%	0.6%
Var(B)	2.5E-1	5.3E-3	88.5%	12.8%	1.3E-1	2.2E-3	62.5%	8.2%
E[W]	7.3E-4	2.7E-5	38.1%	7.2%	4.6E-4	5.8E-5	19.5%	6.9%
Var(W)	5.4E-1	2.4E-6	95.0%	0.2%	5.6E-1	4.2E-4	79.4%	2.2%
	$c = 0.6$				$c = 0.8$			
	ASE		RRASE		ASE		RRASE	
Perf.	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA
E[X]	2.0E-4	same	1.4%	same	3.0E-4	same	1.6%	same
Var(X)	1.2E-2	same	5.7%	same	1.7E-2	same	6.6%	same
E[Q]	3.7E-4	1.4E-4	10.9%	6.7%	4.0E-4	2.2E-4	8.5%	6.3%
Var(Q)	1.1E-1	1.5E-2	25.9%	9.4%	8.9E-2	2.1E-2	20.7%	10.0%
E[B]	2.4E-4	3.8E-5	1.8%	0.7%	2.1E-4	5.9E-5	1.7%	0.9%
Var(B)	8.4E-2	1.5E-3	55.6%	7.3%	6.7E-2	1.9E-3	56.8%	9.6%
E[W]	5.6E-4	1.4E-4	15.0%	7.5%	7.5E-4	2.7E-4	13.4%	8.1%
Var(W)	1.1E+0	3.0E-4	94.8%	1.6%	3.8E+0	1.3E-2	166.6%	9.6%

4.4.4 Robustness in the Variability of Abandonment Times

It has been shown that abandonment distribution significantly affects on system performance; for example, see [107, 117]. We confirm the impacts here using lognormal distributions. Table 4.4 compares the DGA and TGA to the simulation of the $LN(t)/M/100 + LN(2, c_a^2)$ model with the same parameter $\lambda(t) = 60 \sin t + 100$, abandonment rate $\theta = 0.5$ and SCV of abandonment distribution c_a^2 ranging from 0.25 to 10. Paralleling Table 4.3, Table 4.4 shows that, comparing to DGA, TGA consistently improve the accuracy estimating all the system values, but the accuracy degrades when the SCV of abandonment c_a^2 is low; for example, the RRASE of variance of number in service (Var(B)) is 18.4% when $c_a^2 = 0.25$, much higher than other cases.

Table 4.3: Evaluation of DGA and TGA on $LN(\lambda^{-1}(t), c_\lambda^2)/M(1)/100 + H_2(2,2)$, where $\lambda(t) = 1 + 0.6 \sin t$

	$c_\lambda^2 = 0.25$				$c_\lambda^2 = 0.5$				M_t Arrival			
	ASE		RRASE		ASE		RRASE		ASE		RRASE	
Perf.	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA
E[X]	4.6E-6	same	0.2%	same	8.2E-6	same	0.3%	same	1.6E-4	same	1.2%	same
Var(X)	1.7E-5	same	0.5%	same	8.9E-5	same	1.0%	same	3.8E-4	same	1.5%	same
E[Q]	3.2E-5	8.4E-6	3.3%	1.7%	3.9E-5	8.5E-6	3.6%	1.7%	8.0E-5	2.0E-5	5.1%	2.6%
Var(Q)	8.4E-4	9.1E-6	5.6%	0.6%	1.5E-3	1.3E-5	6.1%	0.6%	4.2E-4	1.2E-3	2.4%	4.0%
E[B]	1.2E-5	5.8E-7	0.4%	0.1%	1.1E-5	3.7E-9	0.4%	0.0%	1.3E-5	6.4E-5	0.4%	0.9%
Var(B)	1.6E-4	5.7E-7	5.4%	0.3%	3.2E-4	8.1E-6	6.5%	1.0%	1.2E-3	1.6E-4	9.8%	3.6%
E[W]	4.9E-5	8.7E-6	4.5%	1.9%	5.4E-5	7.0E-6	4.8%	1.7%	8.4E-5	9.6E-6	5.8%	2.0%
Var(W)	2.9E-3	2.7E-6	9.8%	0.3%	7.4E-3	3.6E-4	14.1%	3.1%	9.6E-3	3.6E-5	12.6%	0.8%
	$c_\lambda^2 = 1$				$c_\lambda^2 = 2$				$c_\lambda^2 = 4$			
	ASE		RRASE		ASE		RRASE		ASE		RRASE	
Perf.	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA
E[X]	1.0E-5	same	0.3%	same	1.3E-5	same	0.3%	same	6.2E-5	same	0.8%	same
Var(X)	7.7E-5	same	0.7%	same	2.2E-5	same	0.2%	same	1.4E-2	same	3.8%	same
E[Q]	6.3E-5	1.2E-5	4.6%	2.0%	9.4E-5	8.9E-6	5.5%	1.7%	3.0E-4	3.3E-5	9.4%	3.2%
Var(Q)	1.5E-3	2.9E-4	4.6%	2.0%	8.3E-3	3.9E-4	7.3%	1.6%	9.3E-2	9.1E-4	15.7%	1.6%
E[B]	2.3E-5	1.2E-7	0.6%	0.0%	3.8E-5	3.1E-7	0.7%	0.1%	8.8E-5	4.3E-6	1.1%	0.2%
Var(B)	7.0E-4	2.4E-5	7.4%	1.4%	2.8E-3	3.4E-4	10.3%	3.6%	1.4E-2	3.7E-3	14.7%	7.5%
E[W]	7.5E-5	6.8E-6	5.5%	1.7%	9.6E-5	1.2E-6	6.2%	0.7%	2.3E-4	1.3E-6	9.3%	0.7%
Var(W)	1.2E-2	3.2E-4	14.3%	2.3%	4.2E-2	3.2E-3	19.6%	5.4%	2.4E-1	4.4E-2	32.8%	14.0%

4.4.5 Robustness in the Variability of Service Times

We now evaluate TGAG, the heuristic approximation TGAG built in § 4.3.3 for $G_t/GI/s + GI$ model with non-exponential service time. We let the service time distribution be phase-type, denoted by PH with mean $\mu = 1$ and SCV c_s^2 ranging from 0.25 to 4. To be specific, for cases with $c_s^2 = 0.25$ and 0.5, we used Erlang 4 (E_2) and Erlang 2 (E_2) distribution, and for cases with $c_s^2 = 2$ and 4, we use two-phase hyperexponential distribution (H_2) with balanced means, see [109] for more details.

Figure 4.4 and Table 4.5 compare the effectiveness of DGA, TGA and TGAG for general time-varying queues. Figure 4.4 shows TGAG outperforms TGA especially when estimating the variance of queue length, and Table 4.5 shows the improvement is consistent in a wide range of c_s^2 (from 0.25 to 10). However, though both TGA and TGAG are far better than DGA, the two

Table 4.4: Evaluation of DGA and TGA on $LN(\lambda^{-1}(t), 2)/M(1)/100 + LN(2, c_a^2)$, where $\lambda(t) = 1 + 0.6 \sin t$

	$c_a^2 = 0.25$				$c_a^2 = 0.5$				$c_a^2 = 1$			
	ASE		RRASE		ASE		RRASE		ASE		RRASE	
E[X]	1.3E-3	same	2.6%	same	5.1E-4	same	1.8%	same	3.3E-4	same	1.5%	same
Var(X)	1.9E-1	same	12.6%	same	3.4E-2	same	7.2%	same	1.5E-2	same	6.4%	same
E[Q]	1.6E-3	1.2E-3	9.3%	8.1%	6.2E-4	4.5E-4	6.9%	5.9%	4.4E-4	2.8E-4	7.1%	5.7%
Var(Q)	7.0E-1	8.1E-2	29.6%	10.1%	2.3E-1	1.8E-2	23.4%	6.6%	1.0E-1	9.8E-3	22.3%	6.9%
E[B]	3.9E-4	2.4E-5	2.1%	0.5%	2.8E-4	2.0E-5	1.8%	0.5%	1.9E-4	2.1E-5	1.6%	0.5%
Var(B)	4.4E-1	2.6E-2	176.5%	42.8%	1.5E-1	4.5E-3	104.2%	18.0%	6.9E-2	1.6E-3	73.7%	11.2%
E[W]	1.7E-3	8.2E-4	10.2%	7.0%	8.6E-4	4.8E-4	8.7%	6.5%	6.1E-4	3.0E-4	9.2%	6.5%
Var(W)	2.0E+1	8.7E-2	154.3%	10.1%	7.6E+0	2.6E-2	141.8%	8.3%	2.9E+0	9.7E-3	131.1%	7.6%
	$c_a^2 = 2$				$c_a^2 = 4$				$c_a^2 = 10$			
	ASE		RRASE		ASE		RRASE		ASE		RRASE	
Perf.	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA
E[X]	2.2E-4	same	1.4%	same	1.8E-4	same	1.3%	same	1.6E-4	same	1.4%	same
Var(X)	7.5E-3	same	5.9%	same	5.7E-3	same	6.6%	same	1.8E-3	same	5.1%	same
E[Q]	3.0E-4	1.8E-4	7.7%	5.9%	2.0E-4	1.1E-4	8.7%	6.4%	1.5E-4	8.0E-5	11.6%	8.5%
Var(Q)	5.5E-2	6.1E-3	22.6%	7.5%	3.6E-2	5.0E-3	26.0%	9.7%	1.9E-2	1.7E-3	30.8%	9.3%
E[B]	1.5E-4	2.4E-5	1.4%	0.6%	1.6E-4	4.3E-5	1.5%	0.8%	1.6E-4	6.1E-5	1.5%	0.9%
Var(B)	3.8E-2	8.8E-4	56.0%	8.5%	2.3E-2	7.5E-4	44.3%	7.9%	1.5E-2	6.8E-4	36.6%	7.7%
E[W]	4.1E-4	2.0E-4	10.1%	7.0%	2.7E-4	1.4E-4	11.5%	8.1%	1.8E-4	9.3E-5	14.5%	10.5%
Var(W)	8.4E-1	1.4E-3	103.3%	4.2%	2.4E-1	4.9E-5	80.2%	1.2%	4.5E-2	3.5E-6	57.4%	0.5%

new approximations cannot measure the variance of service content (Var(B)) accurately. Except the Var(B), TGAG performs good when the SCV of service time is not too extreme; for example, $0.5 \leq c_s^2 \leq 4$.

4.4.6 Robustness in the System Scale

Since the MSHT limits involve a sequence of queuing systems with increasing scale, the MSHT approximations DGA and TGA should perform better as the total number of servers increases. Thus, we considered the base case with n = 100, because it is large but small enough to have practical meaning. However, we also want to apply the approximations to even smaller scale systems. Thus, in this section, we examine the effectiveness of DGA, and TGA for smaller systems.

Table 4.6 shows the results for an $LN(t)/M/n + LN(2, 2)$ model with arrival rate function $\lambda(t) = 60 \sin t + 100$ and arrival SCV $c_\lambda^2 = 1$. The number of servers is graduate decreased from base value $n = 100$ to 50, 20, 10, 5 and 3. Table 4.6 shows that TGA is still effective for systems with fewer servers. Though the accuracy degrades as scale decreases, TGA performs remains reasonable when the number of total servers is $n = 5$.

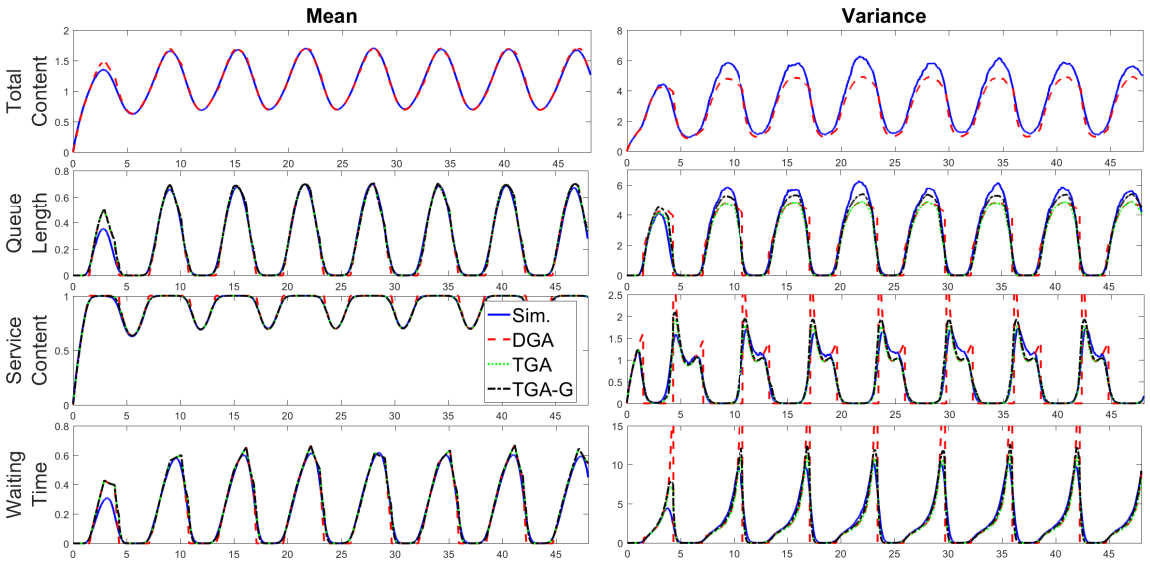


Figure 4.4: Comparison of the TGAG and DGA approximations with the $H2(t)/H_2(1,2)/100 + H_2(2,2)$ model with arrival rate $\lambda(t) = 60 \sin t + 100$.

4.5 Conclusions

In this article, we have developed and evaluated approximations for time-varying $G_t/GI/n+GI$ model. These approximations are remarkably simple and easy to implement, even though they have a relatively complicated basis in MSHT limits developed in [77]. These MSHT limits explain why the approximations require that the scale (number of servers) n and that the abandonment rate θ be not too high.

It is significant that the approximations go beyond a direct application of the MSHT limits. We developed the approximations in § 4.3.3. After presenting the direct DGA Gaussian approximations in § 4.3.1, we applied truncation to obtain the refined TGA approximations in § 4.3.2 and then subsequently, in § 4.3.3, we heuristically modified the MSHT limit in [77] for non-exponential GI service to obtain the final TGAG approximations, which coincide with TGA for exponential service times.

In § 4.4, we report results of extensive simulations studying the approximations and robustness of TGA and TGAG. In SS 4.4.1 and 4.4.2, these experiments show that, for large scale with $n = 100$, the approximations are robust under various arrival rate functions, which implies that effectiveness of TGA in real time-varying system, as general arrival rate can be approximated by finite sum of different sinusoidal functions. Moreover, approximations are efficient for a significant range of the abandonment rate (θ) parameters, roughly for < 2.0.

After examined the first moment system feature (arrival rate $\lambda(t)$ and abandonment rate θ),

we investigate the impact of non-Markov elements for the arrival process in § 4.4.3 and patience distribution in § 4.4.4, and for service distribution in § 4.4.5. According to numerical experiments therein, the arrival and abandonment SCV c_λ^2 and c_a^2 play rather limited role on affecting the effectiveness of approximations. In fact, the TGA performs consistently good when c_λ^2 and c_a^2 varies from 0.25 to 4. Given that the theoretic MSHT are developed for queues with exponential service time ($c_s^2 = 1$), our heuristically refined approximations, TGAG, is shown to be accurate when the SCV of service time does not deviate from one two much, say $0.5 \leq c_s^2 \leq 4$.

Table 4.5: Evaluation of DGA and TGA on $H_2(t)/PH(1, c_s^2)/100 + H_2(2, 2)$, where $\lambda(t) = 100 + 60 \sin t$

	$c_s^2 = 0.25$						$c_s^2 = 0.5$					
		ASE			RRASE			ASE			RRASE	
Perf.	DGA	TGA	TGAG	DGA	TGA	TGAG	DGA	TGA	TGAG	DGA	TGA	TGAG
E[X]	1.7E-3	same	same	3.4%	same	same	1.2E-3	same	same	2.9%	same	same
Var(X)	7.0E-1	same	same	30.4%	same	same	2.2E-1	same	same	15.9%	same	same
E[Q]	2.0E-3	1.4E-3	1.5E-3	13.8%	11.9%	12.0%	1.5E-3	1.0E-3	1.1E-3	12.7%	10.5%	10.6%
Var(Q)	8.9E-1	4.0E-1	8.6E-2	46.1%	31.0%	14.4%	4.9E-1	1.1E-1	2.8E-2	31.6%	15.3%	7.6%
E[B]	2.5E-4	1.4E-4	1.2E-4	1.8%	1.3%	1.2%	2.4E-4	6.4E-5	5.5E-5	1.7%	0.9%	0.8%
Var(B)	4.5E-1	9.2E-2	3.3E-2	120.0%	54.0%	32.3%	3.3E-1	3.7E-2	1.8E-2	103.0%	34.4%	24.1%
E[W]	1.8E-3	8.0E-4	8.1E-4	14.6%	9.8%	9.9%	1.6E-3	5.6E-4	5.7E-4	14.3%	8.6%	8.6%
Var(W)	1.1E+1	5.4E-1	1.8E-1	192.6%	43.5%	25.1%	9.1E+0	2.0E-1	7.5E-2	158.7%	23.7%	14.4%
	$c_s^2 = 1$						$c_s^2 = 2$					
		ASE			RRASE			ASE			RRASE	
Perf.	DGA	TGA	TGAG	DGA	TGA	TGAG	DGA	TGA	TGAG	DGA	TGA	TGAG
E[X]	7.8E-4	same	same	2.3%	same	same	3.3E-4	same	same	1.5%	same	same
Var(X)	3.2E-2	same	same	5.6%	same	same	3.5E-1	same	same	16.7%	same	same
E[Q]	1.2E-3	6.7E-4	6.7E-4	11.6%	8.8%	8.8%	6.5E-4	2.9E-4	2.9E-4	9.3%	6.2%	6.2%
Var(Q)	2.7E-1	3.2E-2	3.2E-2	21.0%	7.2%	7.2%	4.3E-1	3.1E-1	8.6E-2	24.0%	20.1%	10.7%
E[B]	2.2E-4	4.0E-5	4.0E-5	1.6%	0.7%	0.7%	3.1E-4	1.1E-5	1.1E-5	1.9%	0.4%	0.4%
Var(B)	2.3E-1	1.5E-2	1.5E-2	89.0%	23.2%	23.2%	1.8E-1	1.4E-2	1.8E-2	80.2%	22.2%	25.2%
E[W]	1.5E-3	3.6E-4	3.6E-4	14.8%	7.2%	7.2%	1.3E-3	4.2E-4	4.6E-4	14.4%	8.2%	8.6%
Var(W)	7.8E+0	4.1E-3	4.1E-3	127.9%	2.9%	2.9%	7.5E+0	1.0E-1	5.7E-3	108.3%	12.6%	3.0%
	$c_s^2 = 4$						$c_s^2 = 10$					
		ASE			RRASE			ASE			RRASE	
Perf.	DGA	TGA	TGAG	DGA	TGA	TGAG	DGA	TGA	TGAG	DGA	TGA	TGAG
E[X]	9.7E-4	same	same	2.6%	same	same	4.0E-3	same	same	5.5%	same	same
Var(X)	1.4E+0	same	same	30.4%	same	same	3.6E+0	same	same	43.4%	same	same
E[Q]	1.1E-3	8.2E-4	8.7E-4	12.3%	10.8%	11.1%	3.3E-3	3.2E-3	3.4E-3	23.5%	23.1%	24.1%
Var(Q)	1.2E+0	1.1E+0	1.7E-1	35.1%	34.2%	13.5%	2.7E+0	2.6E+0	2.0E-1	47.8%	47.4%	13.1%
E[B]	4.4E-4	4.3E-5	1.6E-5	2.3%	0.7%	0.4%	7.7E-4	2.3E-4	9.2E-5	3.0%	1.7%	1.0%
Var(B)	1.8E-1	2.8E-2	3.0E-2	80.8%	32.3%	33.4%	2.0E-1	5.2E-2	6.6E-2	84.5%	43.4%	48.8%
E[W]	1.9E-3	1.4E-3	1.7E-3	18.1%	15.9%	17.1%	4.6E-3	4.7E-3	5.6E-3	31.5%	32.1%	35.0%
Var(W)	8.4E+0	5.0E-1	1.4E-2	100.2%	24.4%	4.2%	1.0E+1	1.1E+0	2.5E-2	101.9%	33.5%	5.0%

Table 4.6: Evaluation of DGA, TGA and TGAG in smaller systems: $LN(t)/M/n+LN(2,2)$ where $c_\lambda^2 = 1$ and $n = 50, 20, 10, 5$ and 3

	$n = 100$				$n = 50$				$n = 20$			
	ASE		RRASE		ASE		RRASE		ASE		RRASE	
Perf.	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA
E[X]	2.2E-4	same	1.4%	same	2.1E-4	same	1.3%	same	4.7E-4	same	2.0%	same
Var(X)	7.5E-3	same	5.9%	same	1.1E-2	same	7.3%	same	2.0E-2	same	9.7%	same
E[Q]	3.0E-4	1.8E-4	7.7%	5.9%	4.9E-4	1.4E-4	9.6%	5.1%	1.9E-3	2.6E-4	17.6%	6.5%
Var(Q)	5.5E-2	6.1E-3	22.6%	7.5%	8.1E-2	6.0E-3	28.8%	7.9%	1.3E-1	7.7E-3	38.4%	9.3%
E[B]	1.5E-4	2.4E-5	1.4%	0.6%	3.7E-4	3.2E-5	2.2%	0.7%	1.1E-3	5.7E-5	3.9%	0.9%
Var(B)	3.8E-2	8.8E-4	56.0%	8.5%	5.1E-2	1.2E-3	66.5%	10.3%	8.0E-2	3.5E-3	90.3%	18.7%
E[W]	4.1E-4	2.0E-4	10.1%	7.0%	8.6E-4	1.9E-4	14.2%	6.8%	3.6E-3	4.0E-4	25.6%	8.6%
Var(W)	8.4E-1	1.4E-3	103.3%	4.2%	9.9E-1	2.1E-3	115.9%	5.4%	1.3E+0	1.1E-3	134.8%	4.1%
	$n = 10$				$n = 5$				$n = 3$			
	ASE		RRASE		ASE		RRASE		ASE		RRASE	
Perf.	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA	DGA	TGA
E[X]	6.0E-4	same	2.2%	same	1.6E-3	same	3.5%	same	3.3E-3	same	5.0%	same
Var(X)	2.6E-2	same	11.5%	same	4.4E-2	same	15.4%	same	6.3E-2	same	19.3%	same
E[Q]	4.6E-3	2.4E-4	24.9%	5.7%	1.4E-2	4.2E-4	36.1%	6.3%	2.7E-2	8.2E-4	44.6%	7.8%
Var(Q)	2.1E-1	6.9E-3	53.6%	9.7%	3.3E-1	7.7E-3	73.6%	11.3%	4.3E-1	8.6E-3	90.3%	12.8%
E[B]	3.1E-3	1.5E-4	6.6%	1.5%	7.6E-3	6.5E-4	10.8%	3.2%	1.5E-2	1.8E-3	15.6%	5.5%
Var(B)	1.1E-1	7.1E-3	114.3%	28.8%	1.5E-1	1.5E-2	144.4%	45.8%	1.8E-1	2.3E-2	167.2%	60.7%
E[W]	9.7E-3	8.4E-4	36.6%	10.8%	3.4E-2	3.8E-3	52.3%	17.5%	8.2E-2	1.2E-2	63.2%	24.0%
Var(W)	1.5E+0	1.0E-3	155.3%	4.0%	1.9E+0	1.9E-3	161.5%	5.1%	2.1E+0	1.2E-2	163.4%	12.0%

Chapter 5

Optimal Routing to Remote Queues

5.1 Introduction

In real practice, customers usually have to experience pre-arrival delay such as commute times before they enter the service systems. The impact of the pre-arrival delay on the service systems is significant. For instance, in a healthcare system, it takes some time for customers on the way toward hospitals or servers.

In parallel queues with dedicated servers, customers are required to make decisions on which server to visit before entering the queue. Based on available real-time system status including queue lengths or workloads, customers usually select the shortest queue. For example, customers always choose the line with fewest customers waiting in a supermarket. Indeed, it has been proved that to *join the shortest queue* (JSQ) is an "optimal" strategy when system is heavily loaded. However, if a customer, after selecting a server, experience a significant amount of delay before actually arriving at the targeted queue, the waiting time might become dramatically longer than expected. In 2003, a Canadian lady brought her 80-year-old mother to the emergency department in Alberta, Canada. Before they left home, the website said the waiting time was fifteen to eighteen minutes, however, they ended up waiting for more than four hours until a doctor was available [88].

Two main factors lead to the inaccurate emergency departments (EDs) wait times. Firstly, incorrect forecasts of the remaining workload (i.e., the time until a physician becomes available) is made based on incomplete information. EDs compute their anticipated wait times by multiplying the current queue length by the average treatment time. Unfortunately, sometimes treatment time could be extremely long thus the average treatment time shed little light on how long new customers have to wait. Consequently, some patients may need much longer treatments and thus the actual wait time of newcomers could be significantly extended.

Another critical issue is that practitioners who make wait time announcements usually fail to take into account the possible changes of system states during a patients commute time. The

number of ED patients and the severity of their conditions can vary widely from minute to minute. As a result, the announcement of delay at a particular time of the day, even if accurate, may quickly become outdated. This delay announcement may be useful for a patient who lives nearby, but inaccurate for those with a longer commute to the hospital. Consider another example: at 10am, announced wait time is 20 minutes at ED A and 60 minutes at ED B. There are 20 patients who are choosing to visit one of the two EDs, and make decisions based on online wait times. All patients decide to visit ED A with the shorter wait time. Assuming a one-hour average commute time to ED A, the actual wait time of ED A at 11am can suddenly bump up to several hours due to burst arrivals. At the same time, ED B may become idle at 11am due to lack of arrivals. From the viewpoint of the centralized health care network, this issue also causes highly unbalanced workloads at different EDs, and thus leads to system inefficiency.

Another example comes from computation systems such as networks of workstations or mirrored sites in the Internet [87]. In the distributed systems, if some computing units are idle while others are heavily loaded, system performance falls drastically. Therefore, the load balancer, or called the router, is used to improve the system performance by assigning jobs to different hosts (servers), aiming at achieving balanced loads among all servers. Routing rules is to designed to make use of real-time information, but, in real practice, the router records the status of server (either idle or busy) periodically so that the information could be outdated.

In this chapter, we aim to study the impacts of pre-arrival delays. We study a modification of parallel single server queues, which shows the feature of pre-arrival delays. In the modified queueing models, customers firstly choose a server to visit; after the pre-arrival delay, they arrive at the server, waiting in the line until the server is idle and then depart from the system. We first show that the commonly used JSQ policy is ineffective, yielding a large delay when the pre-arrival delay is large. Next, we develop a new policy that refines JSQ; we show that our new policy can eliminate the negative impact of the pre-arrival delay and significantly reduce customer waiting times.

5.1.1 Related Literature

Problems on customer routing in parallel queueing systems have been studied since 1950s, see Haight [37] and Kingman [62], in which authors analyzed the routing problem with differential-difference equation and generating functions. When choosing an efficient routing policy for parallel systems, JSQ is usually considered as a good candidate, as long as system states are available.

Winston [121] showed that JSQ is optimal to control admissions for parallel Markovian systems, by constructing a discrete-time version of queues, where the arrivals occurs at integer time point according to Bernoulli distribution. Later Weber [103] relaxed the Markovian assumption on arrival processes and exponentiality of service time but assuming the service time has a non-decreasing hazard rate function. Foschini and Salz [30], instead of applying discrete process adopted in [121],

proved the asymptotic optimality of JSQ using FCLT approximation.

However, when the traffic intensity is not close to 1, JSQ becomes ineffective. Whitt [110] provided several counterexamples from lightly loaded queues, and proposed several improved policies, though optimality is proved. However, Houck [40] numerically showed that the JSQ is close to optimal.

When the system is in the heavy-traffic regimem, i.e. traffic intensity ρ approach to 1, Reiman [95] firstly showed the asymptotic optimality of JSQ for a general parallel queueing model, where arrival processes are renewal processes and service times are generally distributed. In Reiman's work [95], if customers are routed to the shortest queue, then, as $\rho \to 1$, system state space collapse (SSC) occurs: the difference of queue lengths between two queues converges to zero. According to SSC, the author showed that the limiting stochastic process of the parallel system is equivalent to that of a pooled $GI/GI/n$ system, in which one waiting line is served by n parallel servers. In the pooled system, the service utilization is maximized because the resource pooling prevents the system from having a idle server and a waiting customer at the same time, ensuring that the non-idling constraints is always satisfied. The total queue lengths and waiting times are minimized, implying that JSQ is asymptotically optimal for parallel queueing systems. Whitt [106] reiterated that, in the heavy traffic regime, the multi-server pooled system behaves a "big" server having aggregated service rate with unlimited buffer (waiting room). Therefore, the FCLT limit of the parallel queueing system is a reflected Brownian motion.

Kelly and Laws [58], in Markovian models, considered a threshold strategy, which also achieves asymptotic optimality. The threshold policy works as follows: when one queue length is smaller than a threshold while the other is larger than that threshold, the majority of customers are routed to the shorter queue.

5.1.2 Ineffectiveness of JSQ in Prsence of Pre-Arrival Delays

JSQ achieves asymptotic optimality in dedicated-server models without pre-arrival delay. However, when the system exhibits a pre-arrival delay, JSQ becomes inefficient. In figure 5.1, we compares the queue lenghs of Markovian model by changing the pre-arrival delay. The interarrival times and service times are exponentially distributed with rate $\lambda = 2\rho = 1.98$ and $\mu = 1$, where $\rho = 0.99$ is the traffic intensity. Figure 5.1 illustrates the degradation of JSQ policy as pre-arrival delays increases. When we increase the pre-arrival delay from 0 to 50, 100, the number of waiting customers increases from 100 to 175 and 260. This phenomenon implies that the system performance experienced a significant congestion even adopting JSQ policy when pre-arrival delays are not negligible.

The pre-arrival delay negatively affects the system performance in two ways. First, due to the delay, the system information that is available to customers for routing decisions becomes

outdated. In addition, the outdated information recommends other customers to make similar mistaken judgments, misleading customers to wrong selections and causing largely imbalanced workloads in the system. When servers have large workload differentiations (e.g some congested whereas others idle), the system is inefficient and customers experience unexpected waiting times.

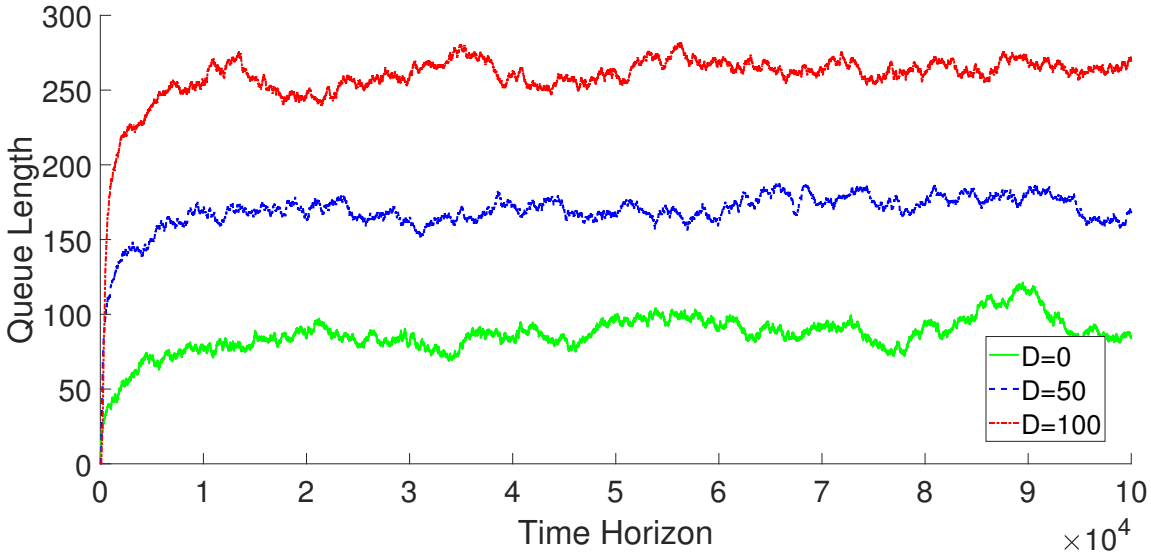


Figure 5.1: Number of waiting customers with different pre-arrival Delays.

To improve our understanding of this situation, we revisit the aforementioned example. Rather than looking at the overall queue length, we now focus on the lengths of two parallel queues. Figure 5.2 depicts the dynamics of each server with pre-arrival delay $D = 100$. It can be seen that the workloads are highly unbalanced; we observe a bouncing effect: when one queue length is high, the other becomes low.

Remark 5.1.1. *In Figure 5.2, the period of each alternation is about double of the pre-arrival delay. It is because that system needs the whole pre-arrival delay amount of time to realize the "shortest" queue is over-routed and it takes another whole pre-arrival delay amount of time to compensate the effect.*

5.1.3 Our Contributions

In this chapter, we study queueing models having pre-arrival delays. Countering to intuition, we discovered that the server currently having the fewest customer or lowest workload does not necessarily offer the shortest wait time. To minimize the overall waiting time, we develop a new

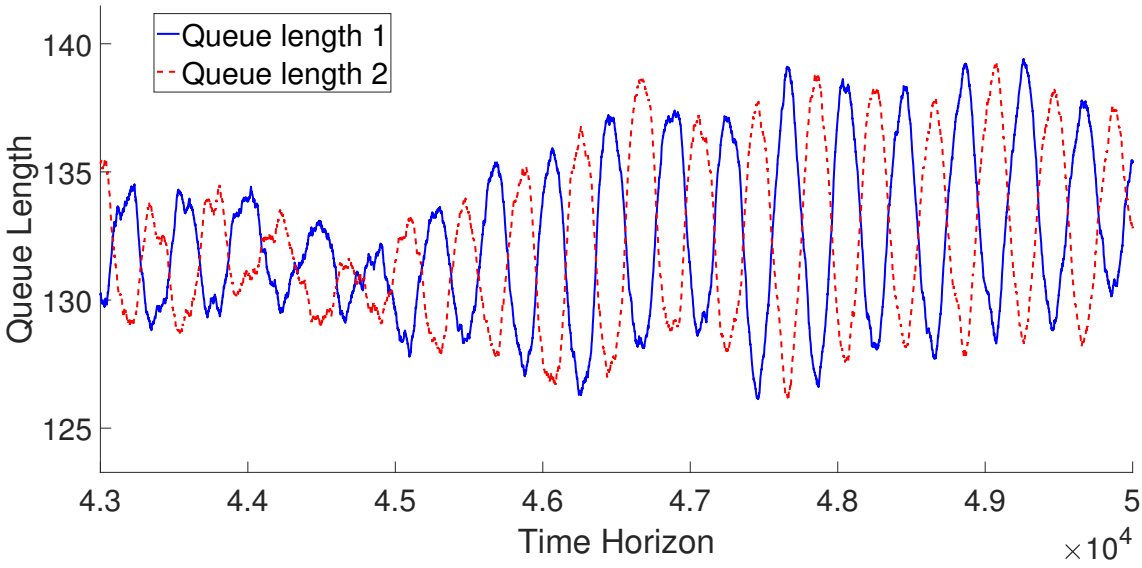


Figure 5.2: Bouncing effect in the dynamic of steady-state queue lengths for remote queue

routing policy and prove its asymptotic optimality.

The remaining of this chapter is organized as follows. In § 5.2, we establish a queueing model to capture the pre-arrival delays. We gain insights of the the degradation of JSQ through the numerical experiments and propose new routing policy to improve the system performance in § 5.3. In fact, we prove the asymptotic optimality of the new policy. Then we extend our routing policy to queues models with more general features in § 5.4. In § 5.5, we conduct numerical examples to show our policy is effective.

5.2 The Remote Queueing Model

To study the impacts of pre-arrival delays, we establish a queueing model with two parallel dedicated servers, in which customers experience pre-arrival delays after selecting a server to visit. No jockeying is allowed in the system and customers in the same queue are served following FCFS and their service times are i.i.d. random variables.

We call such a queueing model as a *remote queue* and give the Kendall's notation $(GI + D)/(GI/1)^2/\Pi$. The first GI with superscript D means the arrival process of customer demand is renewal process, and D indicates the pre-arrival delay. There are 2 queues in parallel, in which service times are i.i.d. distributed, which is denoted as $(GI/1)$. The last Π means the system routing policy adopted in the system. Figure 5.3 gives an illustration of the remote queue.

In the remote queue, customer demand is generated according to a renewal process $\{N(t), t \geq 0\}$. The inter-demand times are i.i.d. random variables with rate $\lambda < 1$ and SCV c_λ^2. Once a demand

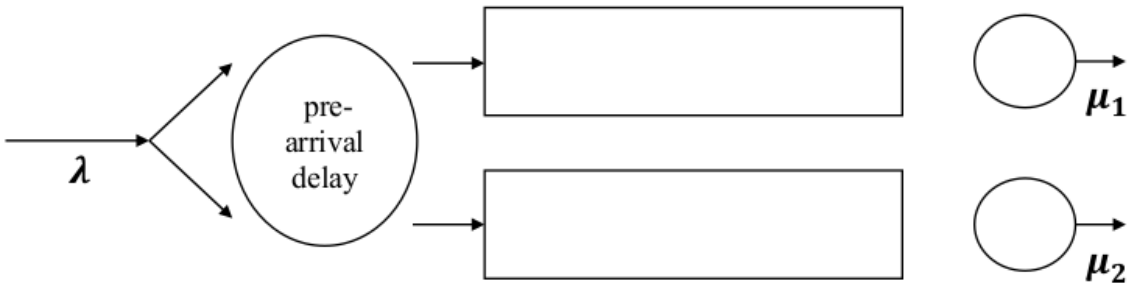


Figure 5.3: A remote queue with two single-server queues

is generated, the customer is routed to a specific server based on policy Π. After a constant pre-arrival delay D, she arrives at the dedicated server with unlimited waiting room. The customer service times are i.i.d. random variables with rate μ_i and SCV $c_{s,i}$, where $i = 1, 2$ indicating the visited server. Define $Q_i(t)$ as the number of total customers in queue i of the remote queue and $\rho = \lambda/(\mu_1 + \mu_2)$ as the traffic intensity of the remote queue. At time 0, the remote queue is assumed to be empty, i.e. $Q_1(0) = Q_2(0) = 0$.

We specify other system performance metrics as follows. In the remote queue, $N_i(t)$ is the number of customers that have been routed to Server i by time t. The process $A_i(t)$ is the arrival processes of Server i; it is obvious that $A_i(t) = N_i(t-d)$. We let $B_i(t)$ and $W_i(t)$ be the accumulated busy time and workloads of Server i by time t. Assuming the consecutive service time in Server i are $V_1^{(1)}, V_2^{(1)}, \ldots$, we have

$$B_i(t) + W_i(t) \equiv \sum_{k=1}^{A_i(t)} V_k^{(i)}.$$

Furthermore, we construct the associated renewal process $S_i(t)$ with respect to $\left\{V_k^{(i)}, k \geq 0\right\}$. It is known that $S_i(B_i(t))$ is the departure process of Server i in remote queue due to simple process time change. The queue length processes are denoted as $Q_i(t)$.

In this chapter, we focus on the remote queues in Haflin-Whitt regime. We construct a sequence of remote queues, denoted as $\mathcal{Q}_1^r, \mathcal{Q}_2^r, \ldots$, where $\mathcal{Q}_1^r$ is defined above. In $\mathcal{Q}_n^r$, system performance metrics are denoted with subscript n. For example, $Q_{n,i}(t)$ is the queue length process at Server i in the nth remote queue. The arrival rate λ_n and pre-arrival delay D_n is scaled in Assumption 2, whereas other system parameters $c_{n,\lambda}^2$ and $(\mu_{n,i}, c_{n,s,i}^2)$'s are fixed at the same values of $\mathcal{Q}_1^r$. Namely, $c_{n,\lambda}^2 = c_\lambda^2$, $(\mu_{n,i}, c_{n,s,i}^2) = (\mu_i, c_{s,i}^2)$.

Our analysis is conducted under Assumption 2, ensuring the sequence $\{\mathcal{Q}_1^r, \mathcal{Q}_2^r, \ldots\}$ is in CHT.

Assumption 2. *In each $\mathcal{Q}_n^r$, we have.*

$$Q_{n,i}(t) = Q_i(nt), \quad Q_{n,i}(0) = Q_i(0) = 0, \quad \textit{and} \quad \hat{Q}_{n,i}(t) = \frac{Q_{n,i}}{\sqrt{n}},$$

where the CLT-scaled queue length process $\hat{Q}_{n,i}(t)$ is defined by convention. As n increases,

$$\sqrt{n}\,(1-\rho_n) = \beta \quad \textit{and} \qquad D_n = D\sqrt{n},$$

where the traffic intensity $\rho_n = \lambda_n$ and $\beta > 0$ is the CHT scalar depicting the level of congestion of queueing system.

Remark 5.2.1. *In terms of the queue length, the relation $Q_{n,i}(t) = Q_i(nt)$ implies that the clock in $\mathcal{Q}_n^r$ is n times faster than the original $\mathcal{Q}_1^r$. Therefore, though the pre-arrival delay is $D_n = D\sqrt{n}$ recorded in the accelerated clock, the pre-arrival delay is in the order of $n^{-1/2}$ in the limit process, after the clock in tuned back. Hence, as $n \to \infty$, the pre-arrival delay in the time clock of limit process is $\frac{D_n}{n} = Dn^{-1/2}$.*

Remark 5.2.2. *In the nth remote queue, we assume the pre-arrival delay $D_n = D\sqrt{n}$ because the delay can be comparable to the waiting time in the system in such a scaling, since the waiting time W_n is in the same order of $\sqrt{n}$. When D_n is much smaller than W_n, then customers only cares about the waiting time so they will only join the shortest queue; when D_n is much larger than W_n, then customers only go to the nearest server without any selection. Hence, pracitically, it is meaningful only if the pre-arrival delay is comparable to the waiting time. However, theoretically, the pre-arrival delay could be scaled as $D_n = Dn^{3/4-\varepsilon}$ for any $\varepsilon > 0$ and our result will not change.*

Assumption 2 holds throughout this chapter unless statement and it provides the framework for our analysis. Later in § 5.4 the pre-arrival delay D is relaxed to any discretely distributed random variables on a finite support random.

5.3 JSQ with a Root-Excess Bias

As we have showed that JSQ is ineffective in remote queues, in this section, we develop a more effective routing policy. Specifically, we plan to refine JSQ by routing not all but some customers to the shortest queue. We first conduct simulation experiments to gain insights of this idea. Then we propose a *JSQ with a root-excess bias* (JSQ-REB) policy and prove its asymptotic optimality as $n \to 1$.

5.3.1 A Probabilistic JSQ

JSQ is ineffective in the face of pre-arrival delays, because JSQ routes too many customers to one queue misled by the outdated information. Hence, we consider a probabilistic routing policy for remote queues: with probability p, a new customer goes to the shortest queue; with probability $(1-p)$, she visit the longer queue. We test the system performance with different routing probability p through simulations. We simulate three Markovian $(M+D)/(M/1)^2$ remote queues with different pairs of traffic intensity and pre-arrival delay (ρ, D)'s: $(0.95, 20)$, $(0.99, 100)$ and $(0.995, 200)$. We let the routing probability p range from 0.5 to 1 and present results in Figure 5.4. We observe that, in all cases, the system queue length achieves the smallest value at time when the routing probability $p \in (0.5, 1)$. Furthermore, the optimal p's (corresponding the shortest queue lengths) in three subplots approach to 0.5 as the traffic intensity ρ increases from 0.95 to 0.995. So the optimal routing probability p is a decreasing function with respect to traffic intensity ρ.

In each sub plot, ther eare two benchmark examples. When the routing probability $p = 1$, the system sends all customers to the shorter queue, so JSQ is applied. We denote such a system with $p = 1$ as $(M + D)/(M/1)^2$/JSQ. when routing probability $p = 0.5$, the system sends customers to queues regardless of system status. Due to the Markovian property, the system behaves like two independent $M/M/1$ queue. We denote it as $(M + D)/(M/1)^2/(p = 0.5)$. It is known that two independent $M/M/1$ cannot compete with a $M/M/2$.

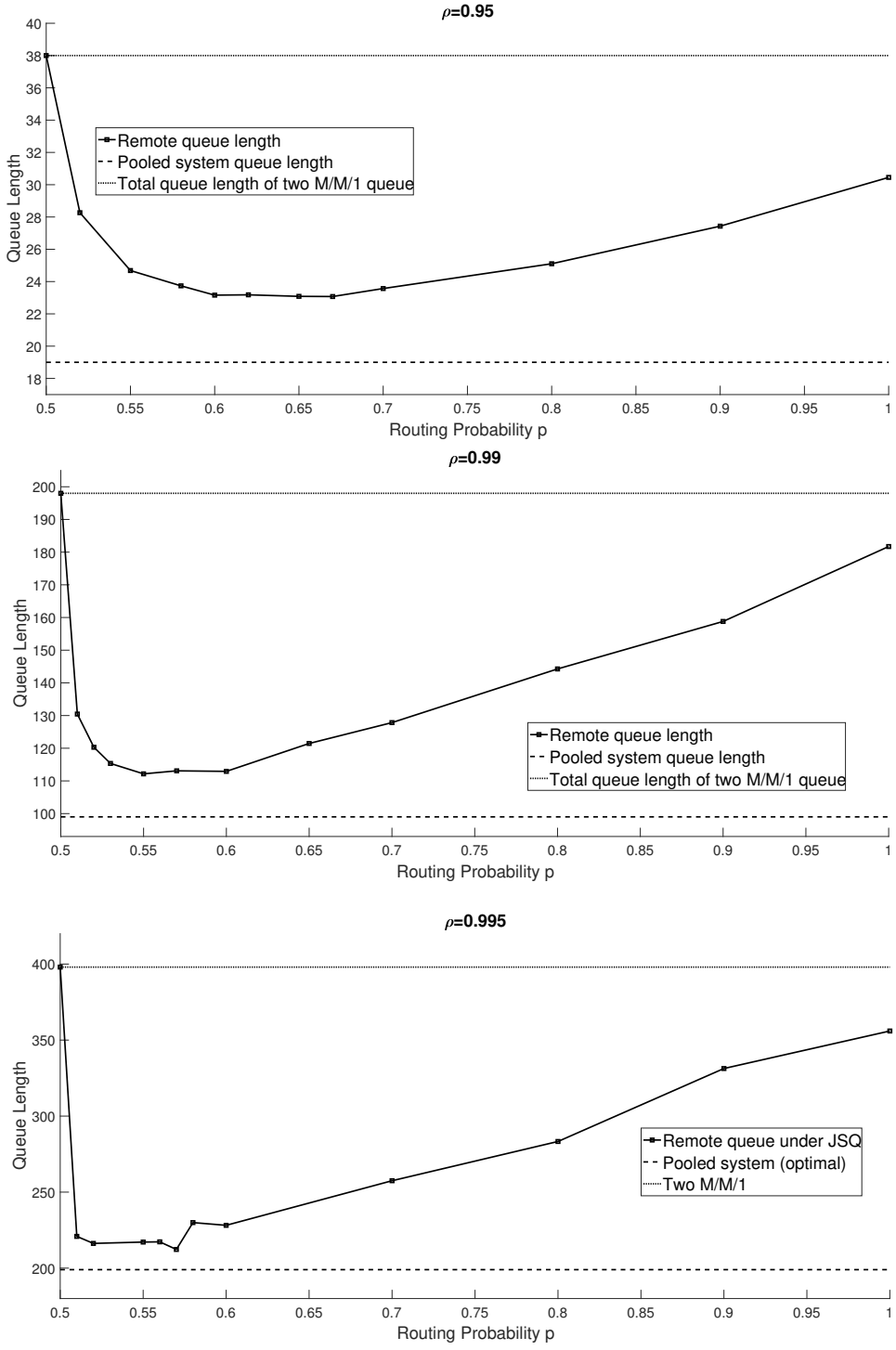


Figure 5.4: Steady-state queue lengths for remote queue having different routing probabilities

5.3.2 JSQ-REB and Its Asymptotic Optimality

Motivated by observations in § 5.3.2, we provide an asymptotically optimal routing policy for remote queues satisfying Assumption 2. According to observations in § 5.3.1, we give our new policy.

JSQ-REB. *In $\mathcal{Q}_n^r$, a new customer is routed to the shortest queue with probability*

$$p_{n,i} = \frac{\mu_i}{\mu} + \theta\sqrt{1-\rho_n}, \tag{5.1}$$

where i is the label of the shortest server, μ is the total output rate $\mu \equiv \mu_1 + \mu_2$, n is the system scaler, and $\theta > 0$ is a finite satisfying $p_{n,i} < 1$.

Now we give our main result.

Theorem 5.3.1. *Under Assumption 2, if the sequence of remote queues adopts JSQ-REB* (5.1)*, then we have*

$$\left(\hat{Q}_{n,1}, \hat{Q}_{n,2}\right) \Rightarrow (\mu_1, \mu_2)\,\hat{Q} \quad in \quad \mathbb{D}^2[0,T],$$

where $T > 0$ is finite and $\hat{Q}$ is a reflected Brownian motion with drift $-\beta\lambda$ and variance $\sqrt{1 + \sum_{i=1}^{2} \frac{\mu_i^2}{\mu^2}(c_\lambda^2 + c_{s,i}^2 - 1)}$.

Remark 5.3.1. *The parameter θ in* (5.1) *can be arbitrary positive number as long as $p_{n,i} < 1$. However, as $n \to \infty$, $1 - \rho_n = \frac{\beta}{\sqrt{n}} \to 0$, so θ can be any large numbers. Furthermore, Eq.* (5.1) *coinciding with our aforementioned observations since $p_n \to \frac{1}{2}$ as n increases.*

The structure of proof of Theorem 5.3.1 is inspired by [95]. We first show the asymptotic synchronization of CLT-scaled workloads of two queues in the remote queue. In particular, $\left|\mu_1^{-1}\hat{Q}_{n,1}(t) - \mu_2^{-1}\hat{Q}_{n,2}(t)\right|$, converges to 0 as $n \to \infty$ for any time t. Then we prove the resourcing pooling is achieved, by showing CLT-scaled total queue length process converges to a reflected Brownian motion (RBM), which is a FCLT-limit queue length process for the pooled $GI/GI/1$ model shareing the same interarrival and service time distributions. The two steps are given in Propositions 5.3.1 and 5.3.2.

The workload synchronization of the two servers is given in stronger mode, as shown below.

Proposition 5.3.1. *For the sequence of remote queues satisfying Assumption 2, we have*

$$\sup_{0\le t\le T}\left|\frac{\hat{Q}_{n,1}(t)}{\mu_1} - \frac{\hat{Q}_{n,2}(t)}{\mu_2}\right| \to 0, \quad as\ n \to \infty. \tag{5.2}$$

Before we prove Proposition 5.3.1, we need Lemmas 5.3.1 and 5.3.2 shown below. The proofs of Lemmas are rendered to § 5.6.

Lemma 5.3.1. *Let $N = \{N(t) : t \ge 0\}$ be a renewal process whose inter-renewal times have a finite rate μ and a finite coefficient of variation $c = \mu\sigma$. If we define a sequence of time scaled*

process $\{N_n(t), t \geq 0\}_{n=1}^{\infty}$ *in the fashion of* $N_n(t) = N(nt)$*, and then for any* $\alpha > 0$ *and* $\varepsilon > 0$*, we have*

$$\Delta_n(t, a, N) = \sqrt{n}\left(\frac{N_n(t + an^{-1/4}) - N_n(t)}{n}\right) - a\mu n^{1/4} \Rightarrow 0, \tag{5.3}$$

Lemma 5.3.2. *Sample* $N(\cdot)$ *defined in Lemma 5.3.1 into two processes,* $N_1(\cdot)$ *and* $N_2(\cdot)$*, as*

$$N_1(t) = \sum_{k=1}^{N(t)} \mathbf{1}\{U_k < 1/2 + \gamma_n\}, \quad \textit{and} \quad N_2(t) = N(t) - N_1(t),$$

where U_k*'s are i.i.d. uniform random variables supported on* $[0, 1]$ *and* $\gamma_n \downarrow 0$ *as* n *increases. Now we have, for any* $a > 0$*,* $t > 0$ *and small* $\varepsilon > 0$*,*

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{0\leq t\leq T} \left|\Delta_n(t, a, N_1) - \Delta_n(t, a, N_2) - 2a\mu\gamma_n n^{1/4}\right| > \varepsilon\right) = 0, \tag{5.4}$$

Proof. We prove Proposition 5.3.1 with reductio ad absurdum. For writing convenience, we define $N(s, t) \equiv N(t) - N(s+)$ for any stochastic process $\{N(t), t \geq 0\}$. Given Assumption 2 holds, we assume the gap between workloads in two servers can be nontrivial. That is to say, for any $\varepsilon > 0$, there exists a sequence of natural numbers $\{n_l, l = 1, 2, \ldots\}$ and $\delta > 0$ such that

$$\mathbb{P}\left(\sup_{0\leq t\leq T} \left|\frac{\hat{Q}_{n_l,1}(t)}{\mu_1} - \frac{\hat{Q}_{n_l,2}(t)}{\mu_2}\right| > \varepsilon\right) > \delta, \tag{5.5}$$

where $T > 0$ is a finite time point. For convenience, we use n to represent n_l and define

$$\hat{L}_n(t) = \frac{\hat{Q}_{n,1}(t)}{\mu_1} - \frac{\hat{Q}_{n,2}(t)}{\mu_2}, \quad \text{and} \quad L_n(t) = \sqrt{n}\hat{L}_n(t) = \frac{Q_{n,1}(t)}{\mu_1} - \frac{Q_{n,2}(t)}{\mu_2}.$$

Define two subordinate sample spaces

$$\Omega_n^+ \equiv \left\{\omega : \sup_{0\leq t\leq T} \tilde{L_{\omega n}}(t) > \varepsilon\right\} \quad \text{and} \quad \Omega_n^- \equiv \left\{\omega : \sup_{0\leq t\leq T} \tilde{L_{\omega n}}(t) < -\varepsilon\right\}.$$

Due to the symmetry of the gap of workloads between two servers, we focus on Ω_n^+. Later it can be seen that subspace Ω_n^- bears the same analysis.

In Ω_n^+, given $\sup_{0\leq t\leq T} \hat{L}_n(t) > \varepsilon$, there exists two time points t_1 and t_2 such that

$$t_2 \equiv t_{n,2} = \inf\left\{t > 0 : \sup_{0\leq t\leq T} \hat{L}_n(t) > \varepsilon\right\} \quad \text{and} \quad t_1 \equiv t_{n,1} = \sup\left\{t < t_2 : \sup_{0\leq t\leq T} \hat{L}_n(t) < \frac{\varepsilon}{2}\right\}.$$

Intuitively, t_2 is when $\hat{L}(t)$ upcrosses ε for the first time and t_2 is $\hat{L}(t)$ upcrosses $\varepsilon/2$ for the last

time before it reaches ε. Obviously, we have $\hat{L}_n(t_2) - \hat{L}_n(t_1) > \varepsilon/2$.

Now we argue the assumption in two situations to negate $\hat{L}_n(t_2) - \hat{L}_n(t_1) > \varepsilon/2$.

1). If $t_2 - t_1 < an^{1/4}$ holds for any constant $a > 0$, then we have $\hat{L}_n(t_2 - an^{-1/4}, t_2) > \varepsilon/2$ hold for any $\varepsilon > 0$. However, we have

$$\begin{aligned}
\hat{L}_n(t_2 - an^{-1/4}, t_2) &\equiv n^{-1/2}\left(\mu_1^{-1}Q_{n,1}(t_2 - an^{-1/4}, t_2) - \mu_2^{-1}Q_{n,1}(t_2 - an^{-1/4}, t_2)\right) \\
&= n^{-1/2}\left(\mu_1^{-1}\left(A_{n,1}(t_2 - an^{-1/4}, t_2) - S_1\left(B_{n,1}(t_2 - an^{-1/4}, t_2)\right)\right)\right. \\
&\quad \left.-\mu_2^{-1}\left(A_{n,2}(t_2 - an^{-1/4}, t_2) - S_2\left(B_{n,2}(t_2 - an^{-1/4}, t_2)\right)\right)\right) \\
&\le \mu_2^{-1}\sup_{t\in(t_1,t_2]}(\Delta_n(t, a, S_2)) - \mu_1^{-1}\inf_{t\in(t_1,t_2]}(\Delta_n(t, a, S_1)) \\
&\quad + \mu_1^{-1}\sup_{t\in(t_1,t_2]}(\Delta_n(t, a, A_1)) - \mu_2^{-1}\inf_{t\in(t_1,t_2]}(\Delta_n(t, a, A_2)) \\
&\quad + an^{1/4}\left(\frac{\lambda_{n,1}}{\mu_1} - \frac{\lambda_{n,2}}{\mu_2}\right) \\
&\Rightarrow 0 + \lim_{n\to\infty} an^{1/4}\left(\frac{\lambda_{n,1}}{\mu_1} - \frac{\lambda_{n,2}}{\mu_2}\right) = a\theta\beta^{1/2}\left(\frac{1}{\mu_2} + \frac{1}{\mu_1}\right),
\end{aligned}$$

where the convergence holds by Lemmas 5.3.1 and 5.3.2, and the last equality holds by the fact that

$$\frac{\lambda_{n,1}}{\mu_1} - \frac{\lambda_{n,2}}{\mu_2} \le \frac{\mu_1 + \theta\sqrt{1-\rho_n}}{\mu_1} - \frac{\mu_2 - \theta\sqrt{1-\rho_n}}{\mu_2} = \theta\sqrt{1-\rho_n}\left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right). \tag{5.6}$$

By choosing $a < \frac{\varepsilon}{2\theta\beta^{1/2}}\left(\frac{1}{\mu_2} + \frac{1}{\mu_1}\right)^{-1}$, we have $\hat{L}_n(t_2 - an^{-1/4}, t_2) < \varepsilon/2$ for any sufficient large n, which resulted in a contradiction. Thus, $t_2 - t_1 < an^{1/4}$ is not true.

2). If $t_2 - t_1 \ge an^{-1/4}$ for some $a > 0$, then we have $0 < \hat{L}_n(t_2 - t_1) \le \hat{L}_n(t_2 - an^{-1/4}) \le \varepsilon/2$. Recalling Remark 5.2.1 that the pre-arrival delay in the tuned clock of nth remote queue is $Dn^{-1/2}$, we have $t_2 - t_1 \ge an^{-1/4} > 2Dn^{-1/2}$ for sufficiently large n. Hence, the interval $[t_2 - an^{-1/4}, t_2]$ is long enough to guarantee the system keeps sending more customers to Server 2 that has the shortest waiting queue during this interval. Namely, we have $\lambda_{n,1} = \left(\frac{\mu_1}{\mu} - \theta\sqrt{1-\rho_n}\right)\lambda_n$ and $\lambda_{n,2} =$

$\left(\frac{\mu_2}{\mu} + \theta\sqrt{1-\rho_n}\right)\lambda_n$ according to JSQ-REB policy.

$$\begin{aligned}\hat{L}_n(t_2 - an^{-1/4}, t_2) \leq & \sup_{t\in(t_1,t_2]}(\Delta_n(t,a,S_2)) - \inf_{t\in(t_1,t_2]}(\Delta_n(t,a,S_1)) \\ & + \sup_{t\in(t_1,t_2]}(\Delta_n(t,a,A_1)) - \inf_{t\in(t_1,t_2]}(\Delta_n(t,a,A_2)) \\ & + an^{1/4}\left(\frac{\lambda_{n,1}}{\mu_1} - \frac{\lambda_{n,2}}{\mu_2}\right) \\ \Rightarrow & -a\beta^{3/2}\left(\frac{1}{\mu_1} + \frac{1}{\mu_2}\right) < 0,\end{aligned}$$

which is in stark contrasting to $\hat{L}_n(t_2 - an^{-1/4}, t_2) > 0$. Therefore, $t_2 - t_2 \geq an^{-1/4}$ is impossible, either.

In summary, because $(t_2 - t_1)$ can be neither greater than or equal to $an^{-1/4}$ nor less than it for some $a > 0$, the only way to conciliate the apparent disagreement is to deny the existence of such t_1 and t_2, which implies that Propositioin 5.3.1 holds. □

Given synchronization of workloads between two servers, we make a further step to on the pooling effect, that is we prove that our system performance of remote queue weakly converges to that of a pooled system shown in Proposition 5.3.2.

Proposition 5.3.2. *For the sequence of remote queues satisfying Assumption 2, if JSQ-REB policy is adopted, then we have*

$$\sup_{0\leq t\leq T}\left|\hat{Q}_{n,1}(t) + \hat{Q}_{n,2}(t) - \hat{Q}_n(t)\right| \to 0, \quad \text{as } n \to \infty, \tag{5.7}$$

where $\hat{Q}_n(t)$ is the CLT-scaled process of $GI/GI/1$ having arrival process characterized by (λ_n, c_λ^2) and service time distributions characterized by $(\mu_1, \mu_2, c_{s,1}^2, c_{s,2}^2)$. To be specific, the service time is distributed as $(\mu_i, c_{s,i}^2)$ with probability $\frac{\mu_i}{\mu}$.

Equipped with Proposition 5.3.1 that tells us two (averaged) workloads in either server of the remote queue are balanced, we now show the resource-pooling effect of the remote queue in the FCLT sense. To this end, we firstly coupled a pooled system from the remote queue, and then show the total queue length of the remote queue, $Q_{n,1}(t) + Q_{n,2}(t)$, equals to that of the pooled system in the FCLT sense. At last, we prove that such a coupled queue has the FCLT limit with the pooled $GI/GI/1$ system.

Before we prove Proposition 5.3.2, we first give a preliminary results in Lemma 5.3.3.

Lemma 5.3.3. *From renewal processes $N_n(t)$ defined before, we sample a counting process $N_{n,i}(t)$*

with probability $P_i(t)$ such that

$$P_i(t) = \begin{cases} p_{n,i} = \mu_i + \theta\sqrt{1-\rho_n} & \text{if } N_{n,i}(t) < N_{n,j}(t) \\ 1 - p_{n,j} = \mu_i - \theta\sqrt{1-\rho_n} & \text{if } N_{n,i}(t) > N_{n,j}(t) \end{cases},$$

where $i, j = 1, 2$ and $i \neq j$, $p_{n,i}$ is defined in (5.1)*. Then we have the weak convergence*

$$\left(\frac{N_{n,1}(t) - n\mu_1\lambda_n t}{\sqrt{n}}, \frac{N_{n,2}(t) - n\mu_2\lambda_n t}{\sqrt{n}}\right) \Rightarrow \left(\sqrt{\mu_1 + \mu_1^2\left(c_\lambda^2 - 1\right)}\mathcal{B}(t), \sqrt{\mu_2 + \mu_2\left(c_\lambda^2 - 1\right)}\mathcal{B}(t)\right), \tag{5.8}$$

where $\mathcal{B}(t)$ is a standard Brownian motion.

We give the proof of Lemma 5.3.3 in § 5.6.2.

Proof. We first construct a pooled queue, by coupling the service time and arrival processes from the remote queue $\mathcal{Q}_n^r$.

We collect the consecutive service times from both servers in remote queue to the pooled system, $\{V_1^{(l)}, V_2^{(l)}, \ldots\}$, where $l = 1, 2$, scale them with rate μ_l respectively, to unify their mean to be $\mu^{-1} = 1$. For these mean-1 random variables from two sequences, we permute them in accordance with the order of their appearances of in the system. Then we copy the permuted sequence of random variables to a pooled single server $G/G^*/1$ queue, having the arrival process as the remote queue. The second notation G^* implies that the service time are independent mean-1 random variables with mixed SCVs $c_{s,1}^2$ and $c_{s,2}^2$. Mathematically, if we use $\{V_1^*, V_2^*, \ldots\}$ to represent the consecutive service time for $G/G^*/1$ queue, then we have:

$$V_k^* = \begin{cases} \frac{\mu_1}{\mu} V_i^{(1)}, & \text{if the } k^{th} \text{ entering-service customer is the } i^{th} \\ & \text{entering-service customer of server 1 in remote queue,} \\ \frac{\mu_2}{\mu} V_i^{(2)}, & \text{if the } k^{th} \text{ entering-service customer is the } j^{th} \\ & \text{entering-service customer of server 2 in remote queue,}. \end{cases} \tag{5.9}$$

See Figure 5.5 for an intuition illustration of the coupling. Let $Q_n^*(t)$ be the total number of customers in the coupled $\cdot/G^*/1$ queue and the pooled queue is initially empty, i.e. $Q_n^*(0) = 0$.

Next, we show that the gap of scaled workloads between $G/G^*/1$ queue and the remote queue. Namely, we want to show that

$$\sup_{0 \leq t \leq T} n^{-1/2} \left|\mu W_n^*(t) - \mu_1 W_{n,1}(t) - \mu_2 W_{n,2}(t)\right| \Rightarrow 0; \tag{5.10}$$

Because the coupled $G/G^*/1$ queue share the same arrival processes $A_n(t) = N_n\left(t - Dn^{1/2}\right)$,

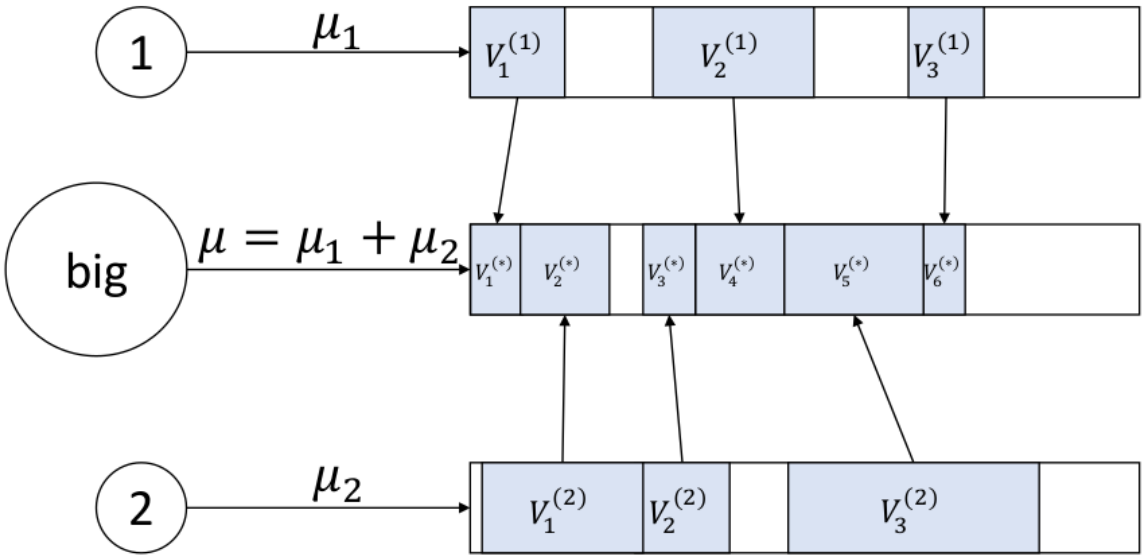


Figure 5.5: The illustration of coupling

we have

$$\mu W_n^*(t) + \mu B_n^*(t) = \sum_{k=1}^{A_n(t)} \mu V_k^* = \sum_{i=1}^{A_{n,1}(t)} \mu_1 V_i^{(1)} + \sum_{j=1}^{A_{n,2}(t)} \mu_2 V_i^{(2)} = \sum_{l=1}^{2} \mu_l \left(W_{n,l}(t) + B_{n,l}(t) \right).$$

Hence we define a new function

$$f_n(t) \equiv \mu_1 W_{n,1}(t) + \mu_2 W_{n,2}(t) - \mu W_n^*(t) = \mu B_n^*(t) - \mu_1 B_{n,1}(t) - \mu_2 B_{n,2}(t),$$

and want to show that $\sup_{0 \le t \le T} |f_n(t)| \to 0$.

Given the fact that busy times either increase along the time at the rate of 1 or keep flat, $f_n(t)$ can increase at the rate of either $1, \mu_2$ or μ_1, and can decrease at the rate of either μ_1 or μ_2. However, as long as the server of $G/G^*/1$ is busy, the minuend of $f_n(t)$, $B_n^*(t)$ is increasing at the rate of 1 and the subtrahend cannot grow faster than that, so $f_n(t)$ is always non-decreasing when $G/G^*/1$ queue is busy. In other words, $f_n(t)$ can decrease only when $G/G^*/1$ is empty. Let $C_n^*(t) \in \{0, 1\}$ and $C_n(t) \in \{0, 1, 2\}$ be the number of busy server(s) in $G/G^*/1$ queue and the remote queue $\mathcal{Q}_n^r$ at time $t \in [0, T]$, respectively. We can interpret the strictly increasing and decreasing intervals of

$f_n(t)$ as follows.

$$I_{\uparrow,n}(t) \equiv \left\{0 \le s \le t : \frac{\mathrm{d}f_n(s)}{\mathrm{d}s} > 0\right\} = \{0 \le s \le t : N_n^*(s) = 1, N_n(s) \le 1\},$$
$$I_{\downarrow,n}(t) \equiv \left\{0 \le s \le t : \frac{\mathrm{d}f_n(s)}{\mathrm{d}s} < 0\right\} = \{0 \le s \le t : N_n^*(s) = 0, N_n(s) \ge 1\},$$

where $I_{\uparrow,n}(t)$ and $I_{\downarrow,n}(t)$ are strictly increasing and decreasing intervals in $[0,t]$, respectively.

At any time $t \in I_{\downarrow,n}(T)$, as explained before, the $G/G^*/1$ queue is empty, whereas the remote queue is yet to finish all. However, only after remote queue finished all arrivals can the sum of its scaled busy time, $\mu_1 B_{n,1}(t) + \mu_2 B_{n,2}(t)$ matches the scaled busy time of $G/G^*/1$ queue $\mu B_n^*(t)$. Therefore, we have $\inf_{t\in I_{\downarrow,n}(T)} f_n(t) \ge 0$ is non-negative. Then by continuity of function $f_n(t)$, we write

$$\inf_{0\le t\le T} f_n(t) = \inf_{t\in I_{\downarrow,n}(T)} f_n(t) \ge 0.$$

which simplifies our targets of to $\sup_{0\le t\le T} f_n(t) < 0$ (the absolute value is dropped).

Now we focus on the interval $I_{\uparrow,n}(T)$ for where $T > 0$ is finite. For any $t \in I_{\uparrow,n}(T)$, there is at least one server being idle, i.e. $Q_{n,1}(t) \cdot Q_{n,2}(t) = 0$. So without loss of generality, we assume $Q_{n,1}(t) > 0, Q_{n,2}(t) = 0$. Starting from the left-hand side of our target,

$$\begin{aligned}
\sup_{0\le t\le T} \frac{1}{\sqrt{n}} f_n(t) &= \sup_{t\in I_{\uparrow,n}(T)} \frac{1}{\sqrt{n}} \left(\mu_1 W_{n,1}(t) + \mu_2 W_{n,2}(t) - \mu W_n^*(t)\right) \\
&= \sup_{t\in I_{\uparrow,n}(T)} \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{Q_{n,1}(t)} \mu_1 V_j^{(1)} - \sum_{k=1}^{Q_n^*(t)} \mu V_k^* \right) \le \sup_{t\in I_{\uparrow,n}(T)} \frac{1}{\sqrt{n}} \left(\sum_{j=1}^{|Q_{n,1}(t)-Q_{n,2}(t)|} \mu_i V_j^{(i)} \right) \\
&= \sup_{t\in I_{\uparrow,n}(T)} \left\{ \frac{1}{|Q_{n,1}(t) - Q_{n,2}(t)|} \sum_{j=1}^{|Q_{n,1}(t)-Q_{n,2}(t)|} V_j^{(i)} \cdot \left| \mu_1 \hat{Q}_{n,1}(t) - \mu_2 \hat{Q}_{n,2}(t) \right| \right\} \\
&\Rightarrow 0, \quad \text{as } n \uparrow \infty.
\end{aligned} \tag{5.11}$$

which implies (5.10). Therefore, Proposition 5.3.2 holds.

Lastly, we need to show that FCLT limit of queue length process for the coupled $G/G^*/1$ model is reflective Brownian motion.

At first, we define $B_{n,i}^*(t)$ as the accumulated busy time by time t when the service time is distributed according to parameters $(\mu, c_{s,i}^2)$. In the coupled $G/G^*/1$ queue, if a customer's service time is distributed according to parameters $(\mu, c_{s,i}^2)$, then she is served at Server i in the original remote queue $\mathcal{Q}_n^r$. Under the JSQ-REB policy, she can be routed to Server i with either probability

$p_{n,i} = \mu_i + \theta\sqrt{1-\rho_n}$ or the other probability $1 - p_{n,j} = mu_i - \theta\sqrt{1-\rho_n}$ $(j \neq i)$.

$$\frac{1}{\sqrt{n}}B^*_{n,i} = \frac{1}{\sqrt{n}}\left(\sum_{k=1}^{A_{n,i}(t)}\frac{\mu_i}{\mu}V^i_k - \sum_{k=1}^{Q^*_{n,i}(t)}\frac{\mu_i}{\mu}V^i_k\right) \tag{5.12}$$

where $B^*_{n,1}$ and $B^*_{n,2}$ partitions the total busy time B^*_n at any time $t \in [0,T]$.

Now we can rewrite queue length of a single server queue, $G/G^*/1$, by solving Skohorad problem, we have

$$Q^*_n(t) = Y_n(t) + \sup_{0\leq s\leq t}(Y_n(s))^+,$$

where

$$Y_n(t) \equiv \frac{1}{\mu}\sum_{i=1}^{2}\left((A_{n,i}(t) - \mu_i\lambda_n nt) - \left(S_i(B^*_{n,i}(t)) - \mu B^*_{n,i}(t)\right)\right) + (\frac{\lambda_n}{\mu} - 1)t. \tag{5.13}$$

Since $S_i(B^*_{n,i}(t))$ is renewal and $A_{n,i}(t) = N_{n,i}(t-D)$ have the same FCLT limit as $N_{n,i}(t)$, we only need to show the FCLT limit of $N_{n,i}(t)$.

Now combining Lemma 5.3.3, Eq. (5.8), and that $A_{n,i}(t) = N_{n,i}(t-D)$, we have

$$\frac{Y_n(t)}{\sqrt{n}} \Rightarrow -\beta t + \sqrt{1 + \sum_{i=1}^{2}\left(\frac{\mu_i}{\mu}\right)^2(c^2_\lambda + c^2_{s,i} - 1)}\mathcal{B}(t), \quad \text{as } n \to \infty$$

where $\mathcal{B}(t)$ is a Brownian motion. □

5.4 Extensions of JSQ-REB

In this section, we extend our main results to more general remote queue models. In § 5.3, the remote queue has a deterministic pre-arrival delay and two dedicated servers. We now relax these two assumptions. In § 5.4.1, we assume pre-arrival delay are discretetly distributed on a finite support; in § 5.4.2, we generalize to networks having m dedicated servers.

5.4.1 Random Pre-Arrival Delays

Previously, we assume the pre-arrival delay is a constant. Such an assumption gives the following mathematical convenience: the arrival process is a time-shifted demand process, i.e. $A(t) = N(t-D)$, and thus is renewal. The FCLT limit of renewal process is Brownian motion, so we can easily derived the weak convergence of queue length process for $\mathcal{Q}^r_1, \mathcal{Q}^r_2, \ldots$.

To make our model more realistic, we now allow our pre-arrival delays to be discretely distributed on a finite sample space.

Assumption 3. *There are c demand locations ($c < \infty$) and the customer pre-arrival delay related to Server i is $D_i = d_j^i$ with probability q_j^i, where $d_j^i > 0$ is a constant. Notice here that $\sum_j q_j^i = 1$ for both $i = 1, 2$.*

Remark 5.4.1. *In practice, customer demand occurs in finite many locations. The pre-arrival delay is usually proportional to the distance between the demand location and the server. Let c denote the number of demand locations, d_{ij} be the corresponding pre-arrival delay from the demand location j to the server i. Therefore, it is meaningful to assume pre-arrival delay follows a discrete distributions. Its probability, denoted by q_{ij}, can be estimated by the ratio of the population in the location j among to the gross population.*

Now we give our results for remote queues with random pre-arrival delays in Proposition 5.4.1.

Proposition 5.4.1. *Under Assumptions 2 and 3, if the sequence of remote queues adopts JSQ-REB* (5.1)*, then we have*

$$\left(\hat{Q}_{n,1}, \hat{Q}_{n,2}\right) \Rightarrow \left(\frac{\mu_1}{\mu}, \frac{\mu_2}{\mu}\right) \hat{Q} \quad in \quad \mathbb{D}^2[0, T],$$

where $T > 0$ is finite and $\hat{Q}$ is a reflected Brownian motion with drift $-\beta\lambda$ and variance $\sqrt{1 + \sum_{i=1}^{2} \frac{\mu_i^2}{\mu^2} \left(c_{s,i}^2 + \sum_{j=1}^{c} q_{ij}^2 (c_\lambda^2 - 1)\right)}$.

Remark 5.4.2. *We cannot not let the pre-arrival delay D be continuously distributed, because weak convergence of the corresponding arrival process*

$$A_{n,i}(t) = \int_0^t \int_0^\infty \mathbb{1}\left\{x + s\sqrt{n} \le t\right\} dK_n^{(i)} \left(\frac{N_{n,i}(s)}{n}, x\right) \tag{5.14}$$

is unknown, due to the appearance of scaler n in the indicator function $\mathbb{1}\{x + s\sqrt{n} \le t\}$ in (5.14)*.*

Proposition 5.4.2 in the next section generalizes Proposition 5.4.1, so we only prove Proposition 5.4.2 in § 5.6.3.

5.4.2 Remote Queues with a Finite Number of Servers

In this section, we extend JSQ-REB policy for remote queues with more than two single servers. Assume there are m single server in $\mathcal{Q}_n^r$ where $n = 1, 2, \ldots$. Service times of server i are i.i.d. distributed with rate μ_i and SCV $c_{s,i}^2$, where $i = 1, 2, \ldots, m$. We again assume $\mu = \sum_{i=1}^m \mu_i$ without loss of generality.

Assumption 4. *The number of dedicated serves is finite, i.e. $m < \infty$.*

JSQ-REB. *In $\mathcal{Q}_n^r$, if customers are routed the shortest queue with probability*

$$p_{n,s} = \frac{\mu_s}{\mu} + \theta\sqrt{1 - \rho_n},$$

or routed to the jth queue ($j \neq s$) with probability

$$p_{n,j} = \frac{\mu_j}{\mu} + \frac{\theta}{m-1}\sqrt{1-\rho_n}.$$

then the routing policy is called JSQ-REB. Here s is the label of the shortest queue and the information of s is available when customers make decisions.

Proposition 5.4.2. *Under Assumptions 2, 3 and 4, if the sequence of remote queues adopts JSQ-REB, then we have*

$$\left(\hat{Q}_{n,1}, \hat{Q}_{n,2}, \ldots, \hat{Q}_{n,m}\right) \Rightarrow \frac{1}{\mu}\left(\mu_1, \mu_2, \ldots, \mu_m\right)\hat{Q} \quad in \quad \mathbb{D}^m[0,T],$$

where $\hat{Q}$ is a reflected Brownian motion with drift $-\beta\lambda$ and variance $\sqrt{1+\sum_{i=1}^m \frac{\mu_i^2}{\mu^2}\left(c_{s,i}^2 + \sum_{j=1}^c q_{ij}^2(c_\lambda^2-1)\right)}$.

Since the proof of Proposition 5.4.2 is separated to two steps, we relegate it to § 5.6.3.

5.5 Numerical Experiments

In this section, we test the effectiveness of JSQ-REB policy through numerical simulations. We start with the base $(M+D)/(M/1)^2$/JSQ-REB with constant pre-arrival delays having different system parameters (ρ, D)'s in § 5.5.1. Then in § 5.5.2, we test if the effectiveness of JSQ-REB preserves in $(GI+D)/(GI/1)^2$, where the exponential assumption on interarrival and service times is relaxed. In § 5.5.3, we test JSQ-REB in a artificial case exhibiting more realistic features, including random pre-arrival delays and three dedicated servers.

In our numerical experiments, for each model, we run $R = 1,000$ replications. The system is initially empty. To collect steady-state queue length, we run each replication for a long time T. The time horizon increases $T = 10^3, 10^4, 10^5$ and 2×10^5 for models with traffic intensity $\rho = 0.90, 0.95, 0.99$ and 0.995, since it becomes slower for heavier loaded queue to approach steady state.

In each run, at discrete time points $\Delta t, 2\Delta t, \ldots, N\Delta t$, we collect queue length: $Q_r(1), Q_r(2), \ldots, Q_r(N)$. The mean steady-state queue length at each time k is then computed as $\bar{Q}(k) = \frac{1}{R}\sum_{r=1}^R Q_r(k)$.

5.5.1 The Base $(M+D)(M/1)^2$/JSQ-REB

According to Theorem 5.3.1, the base $(M+D)(M/1)^2$/JSQ-REB approaches to the optimal $M/M/2$ as $\rho_n \to 1$ (Assumption 2). The Markovian $M/M/2$ model is tractable and the expected

steady-state queue length is

$$\mathrm{E}\left[Q_p(\infty)\right] = \frac{2\rho}{1-\rho}.$$

Therefore, we start to test the effectiveness of JSQ-REB in $(M+D)(M/1)^2/\Pi$, by comparing to JSQ and the optimal $M/M/2$.

We first set he traffic intensity is $\rho = 0.99$ and the pre-arrival delay $D = 100$ and display the results in Figure 5.6. There are four queue lengths in Figure 5.6. The solid line is for $M/M/2$ queue; the dashed line is for $(M+D)/(M/1)^2$/JSQ-REB; the dashed-dotted line is for $(M+D)/(M/1)^2$/JSQ; and the dotted line is for $(M+D)/(M/1)^2/(p=0.5)$. It can been seen that the advantage of JSQ-REB in reducing the system congestion. Moreover, JSQ-REB brings the system near its optimal status.

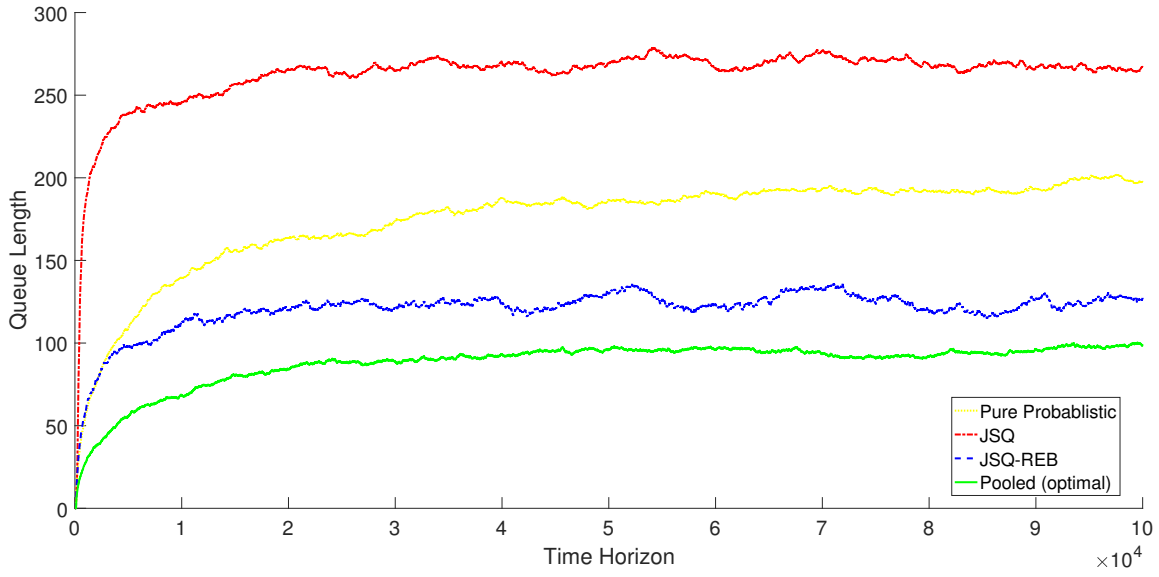


Figure 5.6: Comparisons of remote queue performance under different policies

We next want to test the robustness of JSQ-REB as ρ changes. Therefore, we compare JSQ-REB to JSQ and pure probabilistic policy (p=0.5) in four different settings: systemp parameters (ρ, D) changes from $(0.90, 10)$, $(0.95, 20)$, $(0.99, 100)$ and $(0.995, 200)$. The results are recorded in Table 5.1. The data is the mean queue length sampled at time when system is near steady state. Table 5.1 shows that JSQ-REB still significantly improves the efficiency of remote queues, though advantage becomes less obviously as the system become less congested.

Table 5.1: Steady-state queue length of $(M+D)/(M/1)^2/\Pi$ with constant pre-arrival delay

		Probalistic	JSQ	JSQ-REB	Pooling
$\rho = 0.90$	E[Q]	16.2	18.5	14.2	9.47
	CI	$\pm$0.6	$\pm$0.4	$\pm$0.4	-
$\rho = 0.95$	E[Q]	36.3	44.7	28.8	19.5
	CI	$\pm$1.1	$\pm$0.6	$\pm$0.5	-
$\rho = 0.99$	E[Q]	194	269	124	99.4
	CI	$\pm$4	$\pm$4	$\pm$4	-
$\rho = 0.995$	E[Q]	431	569	259	199.5
	CI	$\pm$17	$\pm$13	$\pm$11	-

5.5.2 General $(GI+D)(GI/1)^2/\Pi$ Model

We now investigate JSQ-REB policy for $(GI+D)/(GI/2)^2$/JSQ-REB, in which interarrival times and service times are no longer exponential distributed. Here we assume they are hyperexponentially distributed. The rate and SCV of interarrival times are $\lambda = 0.99$ and $c_\lambda^2 = 2$, and the rates and SCV's of service times in the two servers are different: $\mu_1 = \mu_2 = 1$ and $c_{s,1}^2 = 1.7$, $c_s^2 = 1.3$. The time horizon is set at $T = 3.5 \times 10^5$ to allow the system approach to steady state.

Simulation results are shown Figure 5.7. we can see that 1) JSQ-REB is superior to other routing policies such as JSQ, and 2) system adopting JSQ-REB behaves very close to the optimal case.

5.5.3 JSQ-REB Policy on a 2-D Map

In this section, we evaluate JSQ-REB policy on a more realistic two-dimensional map. On the rectangular zone on the map, we randomly pick the location of three servers specified in Figure 5.8. Customers are uniformly generated in the zone. Their pre-arrival delay, D, is proportional to the distance from the selected server. For convenience, we assume that three servers are identical and interarrival times and service times are exponentially distributed. However, general distribution should not be a problem according to our theoretical and numerical results in § 5.3 and § 5.5.2.

Now we explain how to apply JSQ-REB to the system on a 2-D map. The idea of JSQ-REB is to send a little more customers to the shortest queue. As Eq. 5.1 implies, each server i serve at least $\left(\frac{\mu_i}{\mu} - \frac{\theta}{m-1}\sqrt{1-\rho_n}\right)$ percent of all customers are routed to Server i. When the server i becomes the shortest queue, labeled as s, it only serve $\frac{m-2}{m-1}\theta\sqrt{1-\rho_n}$ percent more customers. Hence, we pick a small band on the 2-D map so that the population percentage in the band is about $\sqrt{1-\rho_n}$. We send customer whose demands appears in that band to the shortest queue and let others outside of the band go to the nearest server. The band we picked is shown in Figure 5.8. The rectangle zone is divided into four sub-regions, and only customers appeared in the band are routed to the

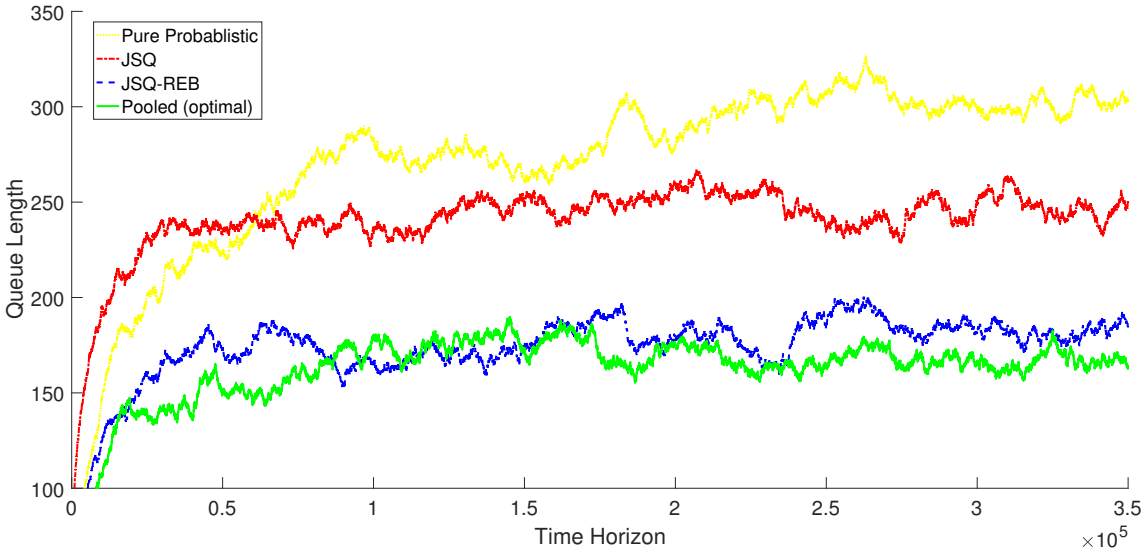


Figure 5.7: Steady-state queue length comparison in $(GI + D)/(GI/1)^2/\Pi$ under different policy Π

shortest server.

In simulation, we set $\rho = 0.99$ and $n = 100$. The distance is proportional to distance from demand location and the selected server. Total service rate is $\mu = 1$, and the rate in each server is standardized by the corresponding population size. In particular, $\mu_1 = 0.27(0.63, 0.83)$, $\mu_2 = 0.34(0.81, 0.26)$, and $\mu_3 = 0.49(-0.74, -0.80)$.

We run simulations under different pairs of system pareameters (ρ, D): $(0.90, 10)$, $(0.95, 20)$, $(0.99, 100)$ and $(0.995, 200)$. Figure 5.9 displayed the steady-state number of waiting customers in the rectangular zone. We can find that the JSQ-REB (the dashed line) is superior to JSQ (the dashed-dotted line) in all cases and JSQ-REB approaches to optimality as $\rho_n \to 1$, coinciding with our theoretical results.

5.6 Other Proofs

In this section, we give proofs for Lemma 5.3.1 and Proposition 5.4.2, as we promised.

5.6.1 Proof of Lemma 5.3.1

We write

$$\Delta_n(t, a, N) = \frac{N_n(t + an^{-1/4}) - \mu(t + an^{-1/4})}{\sqrt{n}} - \frac{N_n(t) - \mu t}{\sqrt{n}} \Rightarrow \mathcal{B}(t + an^{-1/4}) - \mathcal{B}(t) \to 0,$$

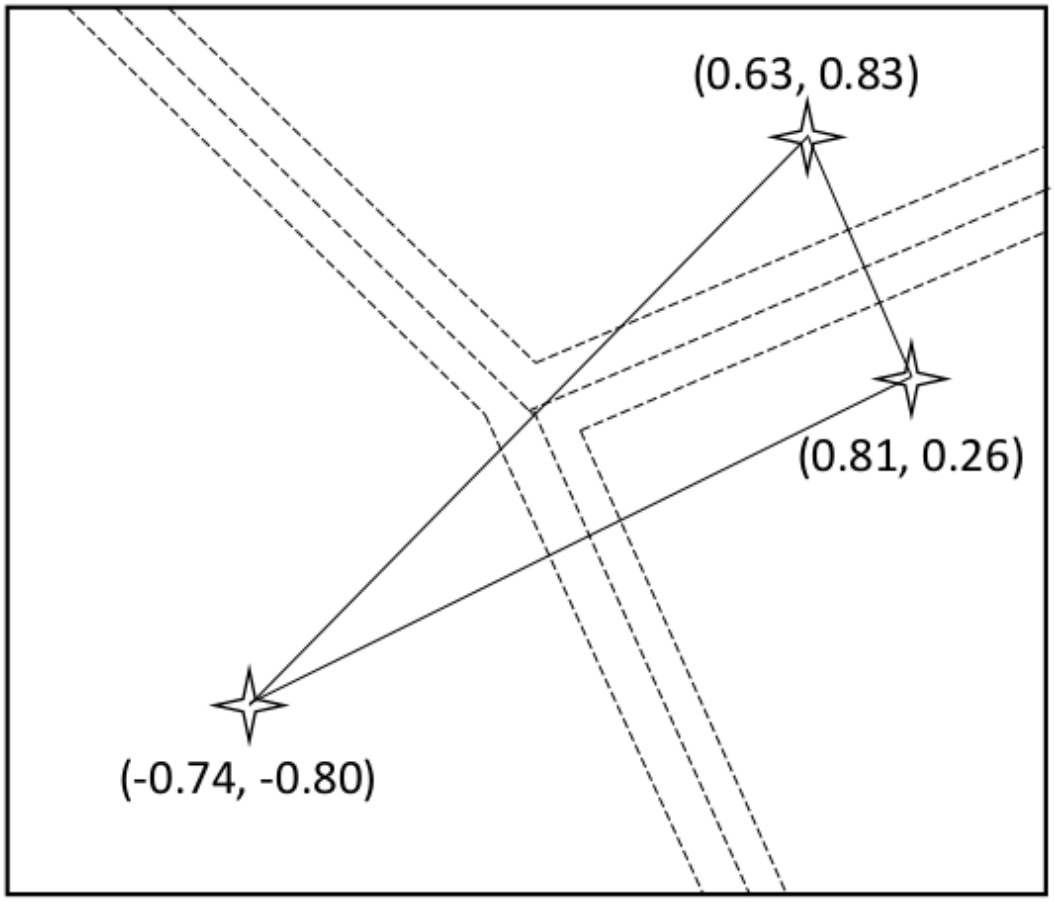


Figure 5.8: Rectangular zone on 2-D map with two servers

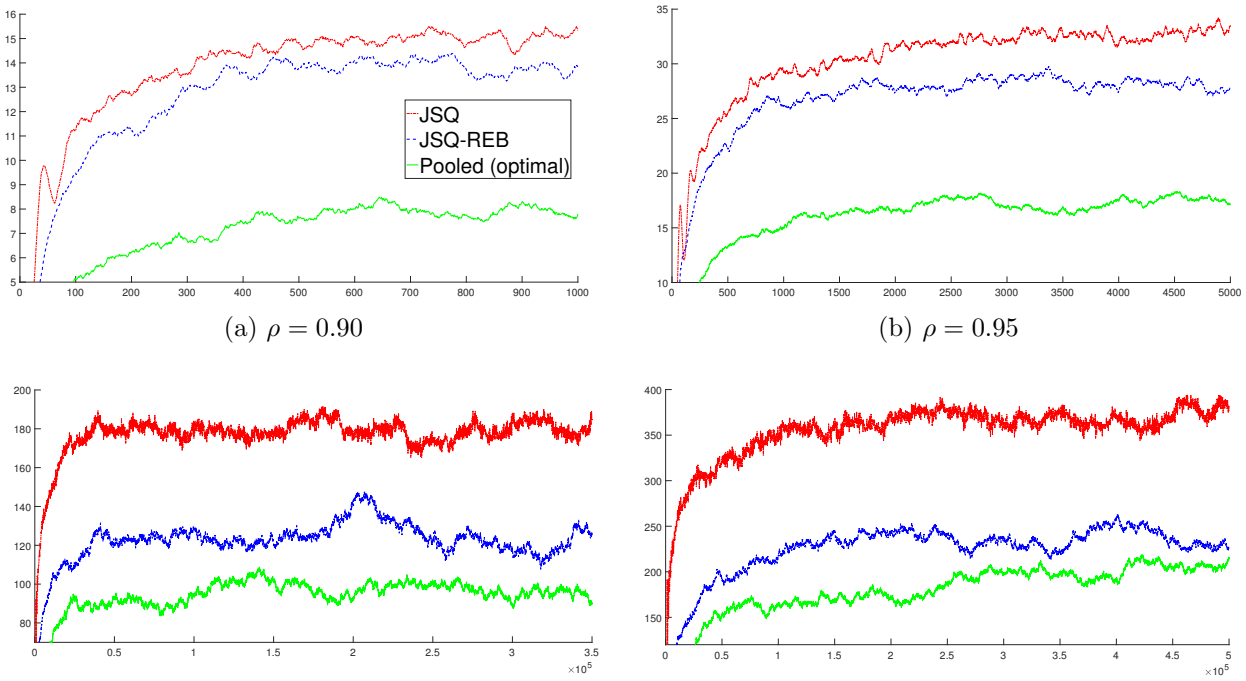
where the second last weak convergence ($\Rightarrow$) is by FCLT (see Theorem 7.2.3 in [106]) and the last convergence ($\rightarrow$) is by the continuity of Brownian motion.

5.6.2 Proof of Lemma 5.3.3

Instead of studying $N_{n,i}(t)$, we consider two modified stochastic process: $\overline{N}_{n,i}(t)$ and $\underline{N}_{n,i}(t)$, which are sampled from $N_n(t)$ with probability $\overline{p}_{n,i} = \frac{\mu_i}{\mu} + \theta\sqrt{1-\rho_n}$ and $\underline{p}_{n,i} = \frac{\mu_i}{-}\theta\sqrt{1-\rho_n}$. Namely,

$$\overline{N}_{n,i}(t) = \sum_{k=1}^{N_n(t)} \mathbb{1}_k^{(i)}, \tag{5.15}$$

where $\mathbb{1}_k^{(i)}$ is 0-1 distributed random variables with parameter $\overline{p}_{n,i}$, indicating the number customers routed to Server i at the kth epoch in the renewal process $N_n(t)$. It is easy to know that $\mathrm{E}[\mathbb{1}_k^{(i)}] = \frac{\mu_i}{\mu}$



(a) $\rho = 0.90$ (b) $\rho = 0.95$

(c) $\rho = 0.99$ (d) $\rho = 0.995$

Figure 5.9: Comparison of queue lengths (customers' waiting time) under different policies

and $\text{Var}\left(\mathbb{1}_k^{(i)}\right) = \frac{\mu_i}{\mu}(1 - \frac{\mu_i}{\mu})$. Then we define

$$S_{n,i}(t) \equiv \left(\sum_{k=1}^{\lfloor nt \rfloor} \mathbb{1}_k^{(i)} - \overline{p}_{n,i} nt \right),$$

and obtain its FCLT limit by FCLT (see Theorem 7.2.3 in [106])

$$\frac{1}{\sqrt{n}} S_{n,i}(t) \Rightarrow \tilde{S}_i(t) \equiv \sqrt{\frac{\mu_i}{\mu}(1 - \frac{\mu_i}{\mu})}\mathcal{B}(t). \tag{5.16}$$

According to FCLT for split processes (Theorem 9.5.1 in [106]), we have

$$\frac{\overline{N}_{n,i}(t) - \overline{p}_{n,i}\lambda_n nt}{\sqrt{n}} \Rightarrow \frac{\mu_i}{\mu}\tilde{N}(t) + \tilde{S}_i(t) = \sqrt{\frac{\mu_i^2}{\mu^2}(c_\lambda^2 - 1) + \mu_i}\mathcal{B}(t),$$

where $j \neq i$, $\mathcal{B}_\lambda(t)$ and $\mathcal{B}_{\mu_i}(t)$ are two independent standard Brownian motions. Similarly, we have $n^{-1/2}\underline{N}_{n,i}(t) \Rightarrow \sqrt{\frac{\mu_i}{\mu}(\frac{\mu_j}{\mu} + \frac{\mu_i}{\mu}c_\lambda^2)}\mathcal{B}(t)$. By coupling, the process $N_{n,i}(t)$ by $\underline{N}_{n,i}(t) \leq N_{n,i}(t) \leq \overline{N}_{n,i}(t)$, hence, we have $n^{-1/2}N_{n,i}(t) \Rightarrow \sqrt{\frac{\mu_i}{\mu}(\frac{\mu_j}{\mu} + \frac{\mu_i}{\mu}c_\lambda^2)}\mathcal{B}(t)$.

5.6.3 Proof of Proposition 5.4.2

According to Lemma 5.3.3, we know that the number of accumulated demands for both servers, $N_{n,j}(t)$ $(j = 1, 2, \ldots, m)$ have corresponding Brownian motions as their FCLT limits. Then based on Assumption 3,

Customers from each sampled demand process, say $N_{n,i}(t)$, with probability q_j^i, experience pre-arrival delay $\sqrt{n}d_{ij}$, so the demand from the jth demand location to the ith server, denoted by $N_{n,i,j}(t)$, is again a sampled process, sampling from $N_{n,i}(t)$. Apply Lemma 5.3.3 again, we have, for any $j = 1, \ldots, m$,

$$\frac{N_{n,i,j}(t) - nq_j^i\mu_i\lambda_n t}{\sqrt{n}} \Rightarrow \sqrt{\frac{\mu_i^2}{\mu^2}q_{ij}^2(c_\lambda^2 - 1) + \frac{\mu_i}{\mu}q_{ij}}\mathcal{B}(t).$$

Under Assumption 3, the arrival processes from the jth demand location to the ith server, denoted by $A_{n,i,j}(t)$, is simply a constant time-shifted demand process i.e.

$$A_{n,i,j}(t) = N_{n,i,j}(t - \sqrt{n}d_{ij}),$$

where $i = 1, 2, \ldots, m$. So the FCLT limit of arrival processes is again Brownian motions:

$$\hat{A}_n(t) = \sum_i \sum_j \frac{A_{n,i,j}(t) - nq_j^i\frac{\mu_i}{\mu}\lambda_n t}{\sqrt{n}} \Rightarrow \sqrt{1 + \sum_{i=1}^m \sum_{j=1}^c \frac{\mu_i^2}{\mu^2}q_{ij}^2(c_\lambda^2 - 1)}\mathcal{B}(t). \tag{5.17}$$

Now we are ready to explain our steps to prove Proposition 5.4.2. We follows the used techniques: firstly to show the weak proportional relation between queue lengths under CLT scaling, and then to couple a pool system so that it has the same CLT-scaled queue lengths with that of the remote queue. At last we prove that the FCLT limit of the coupled system is as general as any pooled system having the same queueing parameters.

The weak proportional relation between queue lengths is presented in the follow proposition.

Proposition 5.6.1. *Given the conditions of Theorem 5.4.2 hold, we have*

$$\lim_{n\uparrow 0} \sup_{0\le t\le T} \left|\mu_j^{-1}\hat{Q}_n^j(t) - \mu_s^{-1}\hat{Q}_n^s(t)\right| \stackrel{.d}{=} 0, \quad \textit{for any } j = 1, 2, \ldots, m.$$

Next we couple a pooled system. The idea of coupling is still to keep the total arrival process queue and all scaled service time according to the order of appearance in the remote queue. The arrival process is denoted as $A_n(t) = \sum_{i=1}^m A_{n,i}(t)$, and each $A_{n,i}(t)$ is defined in (5.14). Let $\{V_i^{(l)}, i = 1, 2, \ldots\}$ be the consecutive service time on the server l $(l = 1, 2, \ldots, m)$, and V_k^* be the k^{th} entering-service customer's processing time for the pooled system. Using $E_k^{(l)}$ to represent the event that

the k^{th} entering-service customer is in the server l, we define V_k^*.

$$V_k^* = \sum_{l=1}^{m} \mathbf{1}\left\{E_k^{(l)}\right\} \cdot \frac{\mu_l}{\mu} V_i^{(l)},$$

where $\mathbf{1}\left\{E_k^{(l)}\right\}$ in the indicator function with respect to event E. Because the each customer can be served by only one out of m single-server station, we have $\sum_{l=1}^{m} \mathbf{1}\left\{E_k^{(l)}\right\} = 1$.

Noticing that neither inter-arrival nor service time are i.i.d. random variables, we denote such a coupled pooling single-server queue as $\cdot/G^*/1$. We use "G^*" to represent the service time, as it is mixed of different i.i.d. random variables, and use "$\cdot$' to denote the arrival process, as it is more complicated than mixture of renewal processes.

Now we want to show that such a coupled $\cdot/G^*/1$ queue has the weakly equivalent CLT-scaled queue length with the remote queue. Let $Q_n^*(t)$ be the total number of customers in the coupled $\cdot/G^*/1$ queue, and $W_{n,i}$ and W_n^* be the workload of the server i in the remote queue and the pooled system, respectively.

Proposition 5.6.2. *Assuming that the conditions of Propositions 5.4.2 hold, we have*

$$\sup_{0\le t\le T} \left| \sum_{l=1}^{m} \hat{Q}_{n,l}(t) - \hat{Q}_n^*(t) \right| \to 0, \quad \text{as } n \Rightarrow \infty. \tag{5.18}$$

The proof is again divided to two steps: the first step is that we show that the scaled workload of the remote queue no difference from that of the pooled system, and the second step is to show a contradiction exists by assuming Proposition 5.6.2 does not hold. So we can prove it by reductio ad absurdum. The proof is of little creativity comparing to the proof of Proposition 5.3.2.

The last target is to show that the coupled $\cdot/G^*/1$ has the same FCLT limits with that of $\cdot/G/1$ queue, in which the service time has a unified distribution with mean μ and SCV $\sum_{i=1}^{m} c_{s,i}$. For any single-server queue $\cdot/G^*/1$, we can represent the queue length through a reflected map shown as (5.13), where the centered process $Y_n(t)$ has more complicated FCLT limit

$$\begin{aligned} n^{-1/2} Y_n(t) &= \frac{(\lambda_n - \mu)t}{\sqrt{n}} + \sum_{i=1}^{m} \left(\hat{A}_{n,i}(t) - \frac{1}{\sqrt{n}} \left(S^*(B_{n,i}(t)) - \mu B_{n,i}(t) \right) \right) \\ &\Rightarrow -\beta t + \sqrt{1 + \sum_{i=1}^{m} \frac{\mu_i^2}{\mu^2} \left(c_{s,i}^2 + \sum_{j=1}^{c} q_{ij}^2 (c_\lambda^2 - 1) \right)} \mathcal{B}(t), \end{aligned}$$

where the weak convergence holds by Eq. (5.17).

5.7 Conclusion

In this Chapter, we observe phenomenon in real service time: the pre-arrival delays. To study the phenomenon, we construct a new queueing model called remote queue. We find the commonly used JSQ become ineffective when system have nontrival pre-arrival delays. Therefore, we propose a new routing policy by adding probabilistic features to JSQ. We name the new policy as JSQ-REB, and proved its asymptotic optimality for remote queue in CHT. We then conduct numerical examples to show the effectiveness of JSQ-REB, and all examples justify our theorems.

Chapter 6

Conclusions and Future Research

This dissertation develops convenient tools to study queueing models performance and provides effective control policies to improve queueing network efficiency. In § 6.1, we highlighted the contribution of each chapter, and in §6.2, we discuss the future research directions.

6.1 Summary and Contributions

This dissertation consists of six chapters, of which Chapters 2-5 are the main contents. We discuss four different queueing models in the four chapters.

Chapter 2 is aimed to show the tightness of the number of idle servers in the $GI/GI/n/0$ loss model. This chapter is motivated by the need to establish heavy-traffic results to approximate steady-state loss models, which are too complicated to analyzed exactly. But the heavy-traffic limits for loss models is still an open problem. To show the heavy-traffic limits, we need to justify the existence and uniqueness of the weak convergence. Our contribution is that we proved the existence of the weak convergence by showing the tightness of the number of idle servers in the loss model. Another contribution is the auxiliary system that we constructed in the proof. Since the auxiliary system reveals the relationship between loss models and single-server queues, we believe it can be used in proving the tightness of more general queues.

Chapters 3 and 4 together are aimed to develop engineering approximations to stationary $GI/GI/n+GI$ and time-varying $G_t/GI/s_t+GI$ queueing models with impatient customers. Though their corresponding MSHT have been established, the approximations directly obtained from MSHT are sometimes ineffective. So we are motivated to refine approximations. Based on the observation, we proposed truncation techniques to refine old approximations and obtain new approximations (i.e. TGA). We then conduct extensive numerical experiments to test the practicability of newly developed approximations for $GI/GI/n+GI$ and $G_t/GI/s_t+GI$ in a wide range of system parameters. In particular, our contribution is that we developed refined approximations for $GI/GI/n+GI$

and $G_t/GI/s_t + GI$, which significantly improves the approximation accuracy, and tested their robustness in terms of various system parameters.

Chapter 5 is aimed to provide efficient routing policies for the dedicated system, that is a queueing model consisting of parallel servers and each server is associated with a unlimited waiting buffer. Our contribution is that we incorporate a more realistic feature to the dedicated system – the pre-arrival delays. In practice, customers, once decided which server to enter, usually experience an amount of pre-arrival delay before they actually arrive at a specific server. Simulation results show that the pre-arrival delay significantly affects system performance, and the previously widely used routing policy becomes ineffective. Therefore, we proposed a new routing policy called JSQ-REB and proved the efficiency of JSQ-REB.

This dissertation studied three single queue models and one queue network, ranging from the $GI/GI/n/0$ loss model without waiting buffer, the $GI/GI/n + GI$ model with waiting buffer and impatient customers, the time-varying $G_t/GI/s_t + GI$ model and the $(GI + D)/(GI/1)^m$ remote queue having pre-arrival delays. This dissertation provides heavy-traffic limits and effective approximations for various performance measures in different queueing models. Taking an example of queue length Q, the this dissertation is summarized in Figure 6.1. In Chapter 2, we discussed the heavy-traffic limits for the loss model because it is still open; in Chapters 3 and 4, based on the well established MSHT, we propose engineering refinements; in Chapter 5, we again establish heavy-traffic limits for new queueing networks.

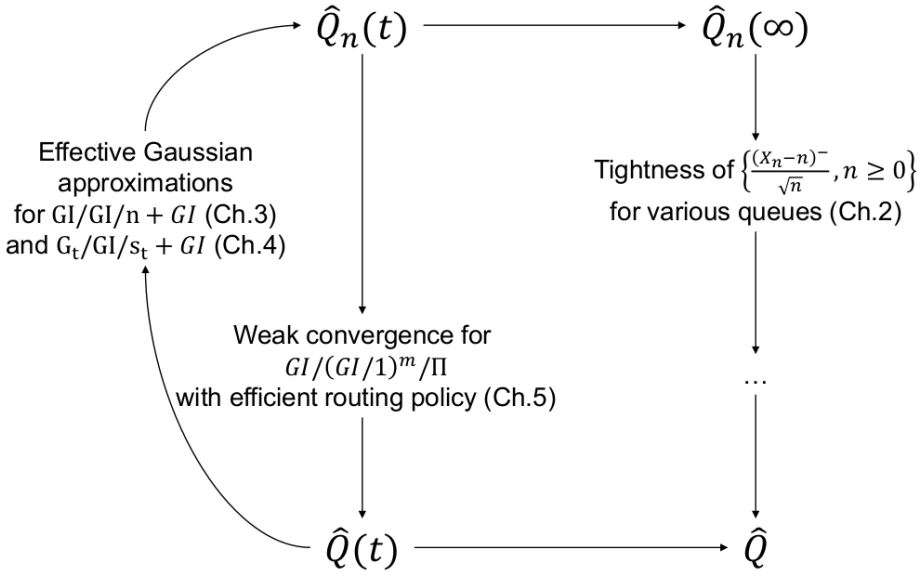


Figure 6.1: The content of the dissertation

6.2 Future Research

The dissertation enables us to further develop effective approximations for various queueing models, and we explain some directions for future research below.

An extension of our work is based on Chapter 2. Tightness of loss model the existence of convergent subsequences, so if we can establish convergent subsequences with clear limits, then we can verify the uniqueness of the heavy-traffic limits for loss models.

Equipped with the heavy-traffic limits for loss models, we can further extend works in Chapters 3 and 4. The new approximation TGA degrades when the approximated queueing system approaches to a loss model. Once we can obtain the heavy-traffic limits for the loss model, the restriction of abandonment when using TGA can be relaxed.

Another extension is again based on Chapter 2. In the proof, we constructed an auxiliary system which enables us to transform a problem of loss model to that of single-server queue. Next, we plan to use the auxiliary system to discuss the tightness of $GI/GI/n+GI$ model and even the heavy-traffic limits of $GI/GI/n+GI$ in Halfin-Whitt regime. The heavy-traffic limits of $GI/GI/n+GI$ can automatically extend Chapter 3, as the Halfin-Whitt limts is naturally more accurate than the used MSHT.

Though the asymptotic efficiency of the new routing policy (JSQ-REB) for the remote queue has been proved in Chapter 5, the result is theoretically restrictive, because the pre-arrival delay is assumed to be discretely distributed. When the pre-arrival delay is continuously distributed, though we can still show the resource pooling, but the heavy-traffic limits are yet to be established. Our next plan is to discuss the heavy-traffic limits when the pre-arrival delay is a continuous random vector. To this end, it is suffice to study the heavy-traffic limits of a infinity-server queue in which service times increases in a proportional manner to the square root of the system scaler.

In addition, in remote queues, service rate can be regarded as control parameters: different service rates reveal different staffing level. By discussing possible control policy on service rates, we are able to formulate an optimization problem and provide managerial insights on how to allocate resources among servers based on the solutions. Last but not least, though the remote queue approaches to a pooled system in CHT, the convergence rate is yet to be known. It is theoretically interesting and practically meaingful to know the convergence rate and parameters that affect the rate, because in practice we can select the rate parameter so that the remote queue becomes efficient at earliest convenience.

REFERENCES

[1] Zeynep Aksin, Mor Armony, and Vijay Mehrotra. The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management*, 16(6):665–688, 2007.

[2] Mor Armony, Shlomo Israelit, Avishai Mandelbaum, Yariv N Marmor, Yulia Tseytlin, Galit B Yom-Tov, et al. On patient flow in hospitals: A data-based queueing-science perspective. *Stochastic Systems*, 5(1):146–194, 2015.

[3] Md Asaduzzaman, Thierry J Chaussalet, and Nicola J Robertson. A loss network model with overflow for capacity planning of a neonatal unit. *Annals of Operations Research*, 178(1):67–76, 2010.

[4] Soren Asmussen. *Applied Probability and Queues*. Springer, 2003.

[5] Rami Atar. A diffusion regime with nondegenerate slowdown. *Operations research*, 60(2):490–500, 2012.

[6] A. N. Avramidis, A. Deslauriers, and P. L'Ecuyer. Modeling daily arrivals to a telephone call center. *Management Sci.*, 50:896–908, 2004.

[7] F. Baccelli and G. Hebuterne. On queues with impatient custoemrs. In *Performance 1981*, pages 159–179. North-Holland, Amsterdam, 1981.

[8] P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 2nd edition, 1999.

[9] Patrick Billingsley. *Convergence of Probability Measures*. Wiley-Interscience, second edition, 1999.

[10] AA Borovkov. Some limit theorems in the theory of mass service, ii multiple channels systems. *Theory of Probability & Its Applications*, 10(3):375–400, 1965.

[11] AA Borovkov. On limit laws for service processes in multi-channel systems. *Siberian Mathematical Journal*, 8(5):746–763, 1967.

[12] AA Borovkov. Stochastic processes in queueing theory, 1972.

[13] A. Brandt and M. Brandt. On the $M(n)/M(n)/s$ queue with impatient calls. *Performance Evaluation*, 35:1–18, 1999.

[14] A. Brandt and M. Brandt. Asymptotic results and a Markovian approximation for the $M(n)/M(n)/s + GI$ system. *Queueing Systems*, 41:73–94, 2002.

[15] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100:36–50, 2005.

[16] L. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100(469):36–50, 2005.

[17] H. Chen and D. D. Yao. *Fundamentals of Queueing Networks.* Springer–Verlag, New York, 2001.

[18] JW Cohen. The full availability group of trunks with an arbitrary distribution of the inter-arrival times and a negative exponential holding time distribution. *Simon Stevin*, 31(4), 1957.

[19] J. G. Dai, S. He, and T. Tezcan. Many-server diffusion limits for $G/Ph/n + GI$ queues. *Annals of Applied Probability*, 20(5):1854–1890, 2010.

[20] J. G. Dai and Shuangchi He. Customer abandonment in many-server queues. *Mathematics of Operations Research*, 35(2):347–362, 2010.

[21] Arnoud M de Bruin, René Bekker, Lillian Van Zanten, and GM Koole. Dimensioning hospital wards using the erlang loss model. *Annals of Operations Research*, 178(1):23–43, 2010.

[22] Nico M Dijk and Erik Sluis. To pool or not to pool in call centers. *Production and Operations Management*, 17(3):296–305, 2008.

[23] Bradley Efron. *The jackknife, the bootstrap and other resampling plans.* SIAM, 1982.

[24] A. K. Erlang. The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B*, 20(33-39):16, 1909.

[25] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Elektrotkeknikeren*, 13:5–13, 1917.

[26] AK Erlang. On the rational determination of the number of circuits. *The life and works of AK Erlang*, pages 216–221, 1948.

[27] Nasim Farrohknia, Maaret Castrén, Anna Ehrenberg, Lars Lind, Sven Oredsson, Håkan Jonsson, Kjell Asplund, and Katarina E Göransson. Emergency department triage scales and their components: a systematic review of the scientific evidence. *Scand J Trauma Resusc Emerg Med*, 19(42):1–13, 2011.

[28] A. Feldman, A. Mandelbaum, W. Massey, and W. Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54:324–338, 2008.

[29] W. Fischer and K. Meier-Hellstern. The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18:149171, November-December 1992.

[30] G Foschini and JACK Salz. A basic dynamic routing problem and diffusion. *IEEE Transactions on Communications*, 26(3):320–327, 1978.

[31] David Gamarnik and David Goldberg. Steady-state $GI/GI/n$ queue in the halfin-whitt regime. *Annals of Applied Probability*, 23(6):2382–2419, 2013.

[32] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002.

[33] David A Goldberg. On the steady-state probability of delay and large negative deviations for the $GI/GI/n$ queue in the halfin-whitt regime. *arXiv preprint arXiv:1307.0241*, 2013.

[34] L. V. Green, P. J. Kolesar, and W. Whitt. Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management*, 16:13–39, 2007.

[35] Linda Green and Peter Kolesar. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37:84–97, 1991.

[36] Varun Gupta, Mor Harchol Balter, Karl Sigman, and Ward Whitt. Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation*, 64(9):1062–1081, 2007.

[37] Frank A Haight. Two queues in parallel. *Biometrika*, 45(3-4):401–410, 1958.

[38] Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, May-Jun. 1981.

[39] B. He, Y. Liu, and W. Whitt. Staffing a service system with non-poisson nonstationary arrivals. *Probability in the Engineering and Informational Sciences*, forthcoming.

[40] David J Houck. Comparison of policies for routing customers to parallel queueing systems. *Operations Research*, 35(2):306–310, 1987.

[41] J. Huang, H. Zhang, and J. Zhang. A unified approach to diffusion analysis of queues with general patience-time distributions. Preprint, 2015.

[42] Junfei Huang, Boaz Carmeli, and Avishai Mandelbaum. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4):892–908, 2015.

[43] Junfei Huang, Avishai Mandelbaum, Hanqin Zhang, and Jiheng Zhang. Refined models for efficiency-driven queues with applications to delay announcements and staffing. working paper, 2015.

[44] R. Ibrahim, P. L'Ecuyer, N. Regnard, and H. Shen. On the modeling and forecasting of call center arrivals. *Proceedings of the 2012 Winter Simulation Conference*, 2012:256–267, 2012.

[45] Rouba Ibrahim and Ward Whitt. Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science*, 55(10):1729–1742, 2009.

[46] Rouba Ibrahim and Ward Whitt. Real-time delay estimation based on delay history in many-server service systems with time-varying arrivals. *Production and Operations Management*, 20(5):654–667, 2011.

[47] Rouba Ibrahim and Ward Whitt. Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations research*, 59(5):1106–1118, 2011.

[48] Donald L Iglehart. Limiting diffusion approximations for the many server queue and the repairman problem. *Journal of Applied Probability*, 2(2):429–441, 1965.

[49] Donald L Iglehart. Weak convergence of compound stochastic process, I. *Stochastic Processes and their Applications*, 1(1):11–31, 1973.

[50] Donald L Iglehart and Ward Whitt. Multiple channel queues in heavy traffic. I. *Advances in Applied Probability*, 2(1):150–177, 1970.

[51] Donald L Iglehart and Ward Whitt. Multiple channel queues in heavy traffic. II: Sequences, networks, and batches. *Advances in Applied Probability*, 2(2):355–369, 1970.

[52] David L Jagerman. Some properties of the erlang loss function. *Bell System Technical Journal*, 53(3):525–551, 1974.

[53] Predrag Jelenković, Avishai Mandelbaum, and Petar Momčilović. Heavy traffic limits for queues with many deterministic servers. *Queueing Systems*, 47(1-2):53–69, 2004.

[54] G. Jongbloed and G. Koole. Managing uncertainty in call centers using Poisson mixtures. *Applied Stochastic Models in Business and Industry*, 17:307–318, 2001.

[55] Weining Kang, Kavita Ramanan, et al. Fluid limits of many-server queues with reneging. *The Annals of Applied Probability*, 20(6):2204–2260, 2010.

[56] Haya Kaspi, Kavita Ramanan, et al. Law of large numbers limits for many-server queues. *The Annals of Applied Probability*, 21(1):33–114, 2011.

[57] Haya Kaspi, Kavita Ramanan, et al. Spde limits of many-server queues. *The Annals of Applied Probability*, 23(1):145–229, 2013.

[58] FP Kelly and CN Laws. Dynamic routing in open queueing networks: Brownian models, cut constraints and resource pooling. *Queueing systems*, 13(1-3):47–86, 1993.

[59] S.-H. Kim and W. Whitt. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? *Manufacturing and Service Oper. Management*, 16(3):464–480, 2014.

[60] Song-Hee Kim, Ponni Vel, Ward Whitt, and Won Chul Cha. Poisson and non-poisson properties in appointment-generated arrival processes: The case of an endocrinology clinic. *Operations Research Letters*, 43(3):247–253, 2015.

[61] J. F. C. Kingman. The single server queue in heavy traffic. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 57, pages 902–904. Cambridge Univ Press, 1961.

[62] J. F. C. Kingman. Two similar queues in parallel. *The Annals of Mathematical Statistics*, 32(4):1314–1323, 1961.

[63] J. F. C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 383–392, 1962.

[64] J. F. C. Kingman. The heavy traffic approximation in the theory of queues. In *Proc. Symp. on Congestion Theory*, pages 137–16. University of North Carolina Press, 1965.

[65] EV Krichagina and Anatolii A. Puhalskii. A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems*, 25(1-4):235–280, 1997.

[66] Retsef Levi and Ana Radovanovic. Provably near-optimal lp-based policies for revenue management in systems with reusable resources. *Operations Research*, 58(2):503–507, 2010.

[67] Peter A Lewis and Gerald S Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval research logistics quarterly*, 26(3):403–413, 1979.

[68] Peter AW Lewis and Gerald S Shedler. Statistical analysis of non-stationary series of events in a data base system. *IBM Journal of Research and Development*, 20(5):465–482, 1976.

[69] Andrew Li and Ward Whitt. Approximate blocking probabilities in loss models with independence and distribution assumptions relaxed. *Performance Evaluation*, 80:82–101, 2014.

[70] K Lindberger. Simple approximations of overflow system quantities for additional demands in the optimisation. In *Proceedings of 10th International Teletraffic Congress, Montreal, Canada, paper*, volume 5, 1983.

[71] Nelly Litvak, Marleen Van Rijsbergen, Richard J Boucherie, and Mark van Houdenhoven. Managing the overflow of intensive care patients. *European journal of operational research*, 185(3):998–1010, 2008.

[72] Ran Liu, Michael E Kuhl, Yunan Liu, and James R Wilson. Modeling and simulation of nonstationary non-poisson arrival processes. Submitted to INFORMS Journal on Computing, North Carolina State University, Raleigh, NC, 2015.

[73] Yunan Liu and Ward Whitt. Large-time asymptotics for the $G_t/M_t/s_t + GI$ many-server fluid queue with abandonmens. *Queueing Systems*, 67:145–182, 2011.

[74] Yunan Liu and Ward Whitt. The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems*, 71:405–444, 2012.

[75] Yunan Liu and Ward Whitt. A many-server fluid limit for the $G_t/GI/s_t+GI$ queueing model experiencing periods of overloading. *Operations Research Letters*, 40:307–312, 2012.

[76] Yunan Liu and Ward Whitt. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Operations Research*, 60:1551–1564, 2012.

[77] Yunan Liu and Ward Whitt. Many-server heavy-traffic limits for queues with time-varying parameters. *The Annals of Applied Probability*, 24:378–421, 2014.

[78] Yunan Liu, Ward Whitt, and Yao Yu. Appendix to: Approximations for heavily-loaded $G/GI/n + GI$ queues. http://yunanliu.wordpress.ncsu.edu/files/2014/02/LiuWhittYuApprox090215app.pdf, 2016.

[79] Sharon E Mace and Thom A Mayer. Triage. chapter 155 in pediatric emergency medicine, baren, jm, rothrock, sg, brennan, ja and brown, l.(eds.), philadelphia: Saunders, 2008.

[80] A. Mandelbaum, W. A. Massey, and M. Reiman. Strong approximations for Markovian service networks. *Queueing Systems*, 30:149–201, 1998.

[81] A. Mandelbaum and S. Zeltyn. Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. In D. Spath and K. Fahnrich, editors, *Advances in Services Innovations*, pages 17–48. Springer, 2007.

[82] Avishai Mandelbaum and Petar Momcilovic. Queues with many servers and impatient customers. *Mathematics of Operations Research*, 37(1):41–65, 2012.

[83] Avishai Mandelbaum and Sergey Zeltyn. The impact of customers' patience on delay and abandonment: some emempirical-driven experiemnts with the $M/M/n+G$ queue. *OR Spectrum*, 26:377–411, 2004.

[84] Avishai Mandelbaum and Sergey Zeltyn. Staffing many-server queues with impatient customers: constraint satisfaction in call centers. *Operations Research*, 57(5):1189–1205, 2009.

[85] W. A. Massey and J. Pender. Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems*, 75:243–277, 2013.

[86] William A Massey and Jamol Pender. Gaussian skewness approximation for dynamic rate multi-server queues with abandonment. *Queueing Systems*, 75(2-4):243–277, 2013.

[87] Michael Mitzenmacher. How useful is old information? *IEEE Transactions on Parallel and Distributed Systems*, 11(1):6–20, 2000.

[88] CBC News. Woman questions ahs online wait times, 2013.

[89] Guodong Pang and Ward Whitt. Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems*, 65:325–364, 2010.

[90] Yu Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214, 1956.

[91] A. A. Puhalskii and M. I. Reiman. The multiclass $GI/PH/N$ queue in the Halfin-Whitt regime. *Advances in Applied Probability*, 32(2):564–595, 2000.

[92] Anatolii A Puhalskii, Josh E Reed, et al. On many-server queues in heavy traffic. *The Annals of Applied probability*, 20(1):129–195, 2010.

[93] Josh Reed et al. The g/gi/n queue in the halfin–whitt regime. *The Annals of Applied Probability*, 19(6):2211–2269, 2009.

[94] Josh Reed and Tolga Tezcan. Hazard rate scaling of the abandonment distribution for the $GI/M/n+GI$ queue in heavy traffic. *Operations Research*, 60:981–995, 2012.

[95] M. I. Reiman. Some diffusion approximations with state space collapse. In *Modelling and Performance Evaluation Methodology*, volume 60 of *Lecture Notes in Control and Information Sciences*, pages 209–240. Springer–Verlag, Berlin, 1984.

[96] Mateo Restrepo, Shane G Henderson, and Huseyin Topaloglu. Erlang loss models for the static deployment of ambulances. *Health care management science*, 12(1):67–79, 2009.

[97] Sheldon M Ross et al. *Stochastic processes*, volume 2. John Wiley & Sons New York, 1996.

[98] Jun Shao and Dongsheng Tu. *The jackknife and bootstrap*. Springer Science & Business Media, 2012.

[99] Władysław Szczotka. Tightness of the stationary waiting time in heavy traffic. *Advances in Applied Probability*, pages 788–794, 1999.

[100] Lajos Takacs. On the generalization of erlang's formula. *Acta Mathematica Hungarica*, 7(3-4):419–433, 1956.

[101] Amy R Ward. Asymptotic analysis of queueing systems with reneging: A survey of results for fifo, single class models. *Surveys in Operations Research and Management Science*, 17(1):1–14, 2012.

[102] Amy R Ward and Peter W Glynn. A diffusion approximation for a markovian queue with reneging. *Queueing Systems*, 43(1-2):103–128, 2003.

[103] Richard R Weber. On the optimal assignment of customers to parallel servers. *Journal of Applied Probability*, pages 406–413, 1978.

[104] W Whitt. Heavy-traffic approximations for service systems with blocking. *AT&T Bell Laboratories Technical Journal*, 63(5):689–708, 1984.

[105] W. Whitt. The pointwise stationary approximation for $M_t/M_t/s$ queues is asymptotically correct as the rates increase. *Management Science*, 37(3):307–314, 1991.

[106] W. Whitt. *Stochastic-Process Limits*. Springer–Verlag, New York, 2002.

[107] W. Whitt. Engineering solution of a basic call-center model. *Management Science*, 51:221–235, 2005.

[108] Ward Whitt. Multiple channel queues in heavy traffic. III: random server selection. *Advances in Applied Probability*, 2(2):370–375, 1970.

[109] Ward Whitt. Approximating a point process by a renewal process, I: two basic methods. *Operations Research*, 30:125–147, 1982.

[110] Ward Whitt. Deciding which queue to join: Some counterexamples. *Operations research*, 34(1):55–62, 1986.

[111] Ward Whitt. Asymptotic formulas for Markov processes with applications to simulation. *Operations Research*, 40(2):279–291, 1992.

[112] Ward Whitt. Approximations for the $GI/G/m$ queue. *Production and Operations Management*, 2(2):114–161, 1993.

[113] Ward Whitt. *Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and Ther Application to Queues*. Springer, 2002.

[114] Ward Whitt. A diffusion approximation for the $G/GI/n/m$ queue. *Operations Research*, 52(6):922–941, November-December 2004.

[115] Ward Whitt. Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, 50:1449–1461, 2004.

[116] Ward Whitt. Heavy-traffic limits for the $G/H_2^*/n/m$ queue. *Mathematics of Operations Research*, 30(1):1–27, 2005.

[117] Ward Whitt. Fluid models for multiserver queues with abandonments. *Operations Research*, 54:37–54, 2006.

[118] Ward Whitt. Sensitivity of performance in the Erlang A model to changes in the model parameters. *Operations Research*, 54:247–260, 2006.

[119] Peter Whittle. Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability & Its Applications*, 5(3):302–305, 1960.

[120] James Wilson. personal communication.

[121] Wayne Winston. Optimality of the shortest line discipline. *Journal of Applied Probability*, pages 181–189, 1977.

[122] Galit B Yom-Tov and Avishai Mandelbaum. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014.

[123] S. Zeltyn and A. Mandelbaum. Call centers with impatient customers: Many-server asymptotics of the $M/M/n + G$ queue. *Queueing Systems*, 51(3–4):361–402, 2005.

[124] J. Zhang. Fluid models of many-server queues with abandonment. *Queueing Systems*, 73(2):147–193, 2013.

[125] X. Zhang, L. J. Hong, and P. W. Glynn. Timescales in modeling call center arrivals. working paper, Department of Industrial Engineering and Logistics Management, The Hong Kong University of Science and Technology, 2014.

APPENDICES

Appendix A

Appendix for Chapter 2

A.1 Overview

This appendix supplements the Chapter 3 by providing additional numerical examples evaluating the performance of the Gaussian approximations DGA, TGA and TGA-G.

First, in §A.2, we give more examples providing examples of the good performance. Then, in §A.3, we give more examples that further expose the limitations of the approximations. In §A.3.1 we give examples showing that the abandonment rates cannot be too much larger than 1; otherwise the system ceases to be OL. In §A.3.2 we give more examples showing that the performance of underloaded models is not good except for a few performance measures. Here we consider models with smaller scale. In §A.3.3 we give examples showing the degraded performance of the approximations when the traffic intensity ρ is too close to 1. We fix $n = 100$ and let ρ decrease toward 1.

A.2 More Good examples

We first give additional positive examples for Markov models in §§A.2.1–A.2.2 and non-Markov models in §A.2.3.

A.2.1 More Examples of Markov $M/M/n + M$ Models

In the main chapter, we compare the exact solution with DGA's and TGA's in heavily loaded regime ($\rho = 1.2$). Table A.1 provides results for $\rho = 1.5$. In this very highly loaded system, the TGA and DGA approximations are nearly identifical, but we see some small differences in the cases with $\theta > 1$.

Table A.1: A comparison of the TGA and DGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $\lambda = 100\rho$ and $\rho = 1.5$ for six values of θ, $0.10 \leq \theta \leq 4.00$.

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Sim.	DGA	TGA	Sim.	DGA	TGA	Sim.	DGA	TGA
E[X]	5.99E+2	6.00E+2	same	3.00E+2	3.00E+2	same	2.00E+2	2.00E+2	same
rel. err.	±9.19E-1	0%		±4.01E-1	0%		±2.01E-1	0%	
Var(X)	1.49E+3	1.50E+3	same	6.00E+2	6.00E+2	same	3.00E+2	3.00E+2	same
rel. err.	±1.10E+3	1%		±2.41E+2	0%		±8.10E+1	0%	
E[Q]	4.99E+2	5.00E+2	5.00E+2	2.00E+2	2.00E+2	2.00E+2	1.00E+2	9.99E+1	9.99E+1
rel. err.	±9.19E-1	0%	0%	±4.01E-1	0%	0%	±2.01E-1	0%	0%
Var(Q)	1.49E+3	1.50E+3	1.50E+3	6.00E+2	6.00E+2	6.00E+2	3.00E+2	3.00E+2	3.00E+2
rel. err.	±9.19E+2	1%	1%	±1.61E+2	0%	0%	±4.08E+1	0%	0%
E[W]	4.05E+0	4.06E+0	4.06E+0	1.63E+0	1.62E+0	1.62E+0	8.17E-1	8.11E-1	8.11E-1
rel. err.	±7.53E-3	0%	0%	±3.28E-3	0%	0%	±1.64E-3	1%	1%
Var(W)	9.98E-2	1.00E-1	1.00E-1	4.01E-2	4.00E-2	4.00E-2	2.00E-2	2.00E-2	2.00E-2
rel. err.	±7.53E-3	0%	0%	±3.28E-3	0%	0%	±1.64E-3	0%	0%
PoD	1.00E+0	1.00E+0	same	1.00E+0	1.00E+0	same	1.00E+0	1.00E+0	same
rel. err.	±0.00E+0	0%		±0.00E+0	0%		±0.00E+0	0%	
PoA	3.33E-1	3.33E-1	same	3.33E-1	3.33E-1	same	3.34E-1	3.33E-1	same
rel. err.	±1.34E-3	0%		±1.37E-3	0%		±1.33E-3	0%	

	$\theta = 2$			$\theta = 4$			$\theta = 10$		
Perf.	Sim.	DGA	TGA	Sim.	DGA	TGA	Sim.	DGA	TGA
E[X]	1.25E+2	1.25E+2	same	1.12E+2	1.12E+2	same	1.05E+2	1.05E+2	same
rel. err.	±5.05E-2	0%		±2.64E-2	0%		±1.36E-2	0%	
Var(X)	7.53E+1	7.49E+1	same	3.79E+1	3.74E+1	same	1.71E+1	1.49E+1	same
rel. err.	±1.27E+1	1%		±5.96E+0	1%		±2.84E+0	13%	
E[Q]	2.50E+1	2.49E+1	2.49E+1	1.25E+1	1.23E+1	1.24E+1	5.03E+0	4.90E+0	5.08E+0
rel. err.	±5.05E-2	0%	0%	±2.60E-2	2%	1%	±1.17E-2	3%	1%
Var(Q)	7.52E+1	7.49E+1	7.46E+1	3.68E+1	3.74E+1	3.59E+1	1.35E+1	1.49E+1	1.24E+1
rel. err.	±2.63E+0	0%	1%	±7.08E-1	2%	2%	±1.49E-1	10%	9%
E[W]	2.08E-1	2.03E-1	2.03E-1	1.06E-1	1.01E-1	1.01E-1	4.55E-2	4.10E-2	4.25E-2
rel. err.	±4.13E-4	2%	2%	±2.14E-4	5%	5%	±1.01E-4	10%	7%
Var(W)	5.07E-3	5.00E-3	4.98E-3	2.52E-3	2.50E-3	2.40E-3	9.87E-4	1.00E-3	8.44E-4
rel. err.	±4.13E-4	1%	2%	±2.14E-4	1%	5%	±1.01E-4	2%	14%
PoD	9.98E-1	9.98E-1	same	9.79E-1	9.78E-1	same	8.62E-1	9.02E-1	same
rel. err.	±1.27E-4	0%		±4.53E-4	0%		±1.03E-3	5%	
PoA	3.32E-1	3.34E-1	same	3.34E-1	3.32E-1	same	3.35E-1	3.36E-1	same
rel. err.	±1.35E-3	0%		±1.34E-3	0%		±1.31E-3	0%	

A.2.2 Low Abandonment Rates

We next consider the Markovian $M/M/n+M$ queueing system, with the arrival rate, number of servers and service rate fixed at $(\lambda, n, \mu) = (105, 100, 1)$, but decreasing abandonment rate. As studied in [115], the queueing system tends to heavily overloaded when abandonment rates decrease. The Theorem 4 in [115] states that in an $M/M/n/r+M$ model, a scaled process of number in system converges to a OU process as $s/\theta \to 0$. To reshow the results, we give the Markovian $M/M/n+M$ queueing system, with the arrival rate, number of servers and service rate fixed at $(\lambda, n, \mu) = (105, 100, 1)$ but decreasing abandonment rate θ from $0.05, 0.02$ to 0.01. Table A.2 shows that our TGAs continue to work effectively for smaller abandonment rates. Notice that little difference between DGAs and TGAs are presented; it is because when $\theta \to 0$, improvements brought about by truncation become less effective as the queue tends to ED regime.

Table A.2: A comparison of the TGA and DGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $\lambda = 100\rho$ and $\rho = 1.2$ for three low abandonment rates $\theta < 0.1$.

	$\theta = 0.05$			$\theta = 0.02$			$\theta = 0.01$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	2.01E+2	2.00E+2	same	3.50E+2	3.50E+2	same	5.99E+2	6.00E+2	same
rel. err.		0%			0%			0%	
Var(X)	2.00E+3	2.10E+3	same	5.22E+3	5.25E+3	same	1.03E+4	1.05E+4	same
rel. err.		5%			1%			2%	
E[Q]	1.01E+2	1.00E+2	1.00E+2	2.50E+2	2.50E+2	2.50E+2	4.99E+2	5.00E+2	5.00E+2
rel. err.		1%	1%		0%	0%		0%	0%
Var(Q)	1.99E+3	2.10E+3	2.05E+3	5.22E+3	5.25E+3	5.25E+3	1.03E+4	1.05E+4	1.05E+4
rel. err.		5%	3%		1%	1%		2%	2%
E[W]	9.90E-1	9.76E-1	9.78E-1	2.44E+0	2.44E+0	2.44E+0	4.88E+0	4.88E+0	4.88E+0
rel. err.		1%	1%		0%	0%		0%	0%
Var(W)	1.90E-1	2.00E-1	1.95E-1	4.97E-1	5.00E-1	5.00E-1	9.85E-1	1.00E+0	1.00E+0
rel. err.		5%	3%		1%	0%		2%	2%
PoD	9.93E-1	9.85E-1	same	1.00E+0	1.00E+0	same	1.00E+0	1.00E+0	same
rel. err.		1%			0%			0%	
PoA	4.81E-2	4.75E-2	same	4.76E-2	4.75E-2	same	4.76E-2	4.76E-2	same
rel. err.		1%			0%			0%	

A.2.3 More Examples of $GI/GI/n + GI$ Models

We now consider examples with various combinations of high and low variabilities for the interarrival, service and patience times. We use simply phase-type (PH) distributions to achieve both high and low variabilities: Erlang-n (E_n) for low variabilities (with SCV $1/n$) and H_2 for high variabilities (with SCV greater than 1). Other parameters remain the same as those in Table 3.2. Table A.3 shows that TGA-G works well except when the SCV of service time is high (e.g. $c_s^2 = 4$).

Table A.3: A comparison of the TGA-G approximations to simulation estimates in the $PH/PH/n + PH$ models model with $\lambda = 100\rho$ and $\rho = 1.0.5$ for for values of the service scv c_s^2, $0.25 \leq \theta \leq 4.00$.

Perf. Meas.	SCV		c_s^2					
			0.25			0.5		
	c_λ^2	c_{ab}^2	Sim	CI	TGA-GA	Sim	CI	TGA-GA
E[Q]	0.5	0.5	3.40E+1	±2.15E-1	3.57E+1	3.34E+1	±2.50E-1	3.57E+1
		2	8.67E+0	±6.46E-2	8.42E+0	9.00E+0	±7.46E-2	8.67E+0
	2	0.5	3.37E+1	±3.14E-1	3.63E+1	3.37E+1	±3.23E-1	3.64E+1
		2	1.07E+1	±9.32E-2	1.01E+1	1.09E+1	±1.02E-1	1.03E+1
Var(Q)	0.5	0.5	1.84E+2	±1.44E+1	2.14E+2	2.30E+2	±1.66E+1	2.51E+2
		2	5.14E+1	±1.48E+0	5.55E+1	6.10E+1	±1.84E+0	6.36E+1
	2	0.5	4.30E+2	±2.29E+1	4.89E+2	4.55E+2	±2.38E+1	5.18E+2
		2	1.12E+2	±3.01E+0	1.12E+2	1.21E+2	±3.44E+0	1.19E+2
E[W]	0.5	0.5	3.33E-1	±2.04E-3	3.46E-1	3.28E-1	±2.40E-3	3.47E-1
		2	8.86E-2	±6.32E-4	8.30E-2	9.21E-2	±7.41E-4	8.55E-2
	2	0.5	3.26E-1	±2.90E-3	3.51E-1	3.26E-1	±3.00E-3	3.52E-1
		2	1.06E-1	±8.77E-4	9.84E-2	1.07E-1	±9.71E-4	1.00E-1
Var(W)	0.5	0.5	1.66E-2	±2.04E-3	1.97E-2	2.13E-2	±2.40E-3	2.34E-2
		2	4.96E-3	±6.32E-4	5.32E-3	6.01E-3	±7.41E-4	6.13E-3
	2	0.5	3.68E-2	±2.90E-3	4.32E-2	3.94E-2	±3.00E-3	4.62E-2
		2	1.00E-2	±8.77E-4	1.05E-2	1.10E-2	±9.71E-4	1.12E-2
PoD	0.5	0.5	9.90E-1	±8.51E-4	9.93E-1	9.79E-1	±1.37E-3	9.87E-1
		2	8.18E-1	±2.62E-3	7.97E-1	7.96E-1	±2.89E-3	7.76E-1
	2	0.5	9.30E-1	±2.49E-3	9.43E-1	9.22E-1	±2.66E-3	9.36E-1
		2	7.28E-1	±3.18E-3	7.04E-1	7.19E-1	±3.31E-3	6.97E-1
PoA	0.5	0.5	4.82E-2	±7.19E-4	4.78E-2	4.81E-2	±7.62E-4	4.79E-2
		2	5.54E-2	±7.12E-4	5.36E-2	5.84E-2	±7.76E-4	5.52E-2
	2	0.5	5.16E-2	±8.66E-4	4.90E-2	5.26E-2	±8.83E-4	4.93E-2
		2	6.54E-2	±8.09E-4	6.32E-2	6.66E-2	±8.67E-4	6.44E-2

Perf. Meas.	SCV		c_s^2					
			2			4		
	c_λ^2	c_{ab}^2	Sim	CI	TGA-GA	Sim	CI	TGA-GA
E[Q]	0.5	0.5	3.27E+1	±3.38E-1	3.59E+1	3.27E+1	±4.14E-1	3.62E+1
		2	9.73E+0	±1.03E-1	9.44E+0	1.03E+1	±1.36E-1	1.00E+1
	2	0.5	3.35E+1	±3.76E-1	3.68E+1	3.39E+1	±4.38E-1	3.71E+1
		2	1.13E+1	±1.29E-1	1.09E+1	1.13E+1	±1.52E-1	1.14E+1
Var(Q)	0.5	0.5	4.03E+2	±2.31E+1	3.70E+2	5.13E+2	±2.97E+1	4.60E+2
		2	9.28E+1	±3.04E+0	8.96E+1	1.14E+2	±4.38E+0	1.10E+2
	2	0.5	5.84E+2	±2.91E+1	6.15E+2	6.74E+2	±3.50E+1	6.91E+2
		2	1.47E+2	±4.69E+0	1.43E+2	1.61E+2	±5.75E+0	1.62E+2
E[W]	0.5	0.5	3.23E-1	±3.34E-3	3.49E-1	3.25E-1	±4.14E-3	3.52E-1
		2	1.01E-1	±1.06E-3	9.33E-2	1.07E-1	±1.43E-3	9.91E-2
	2	0.5	3.26E-1	±3.61E-3	3.56E-1	3.31E-1	±4.28E-3	3.60E-1
		2	1.13E-1	±1.28E-3	1.07E-1	1.14E-1	±1.54E-3	1.11E-1
Var(W)	0.5	0.5	3.91E-2	±3.34E-3	3.53E-2	5.10E-2	±4.14E-3	4.42E-2
		2	9.72E-3	±1.06E-3	8.74E-3	1.22E-2	±1.43E-3	1.08E-2
	2	0.5	5.31E-2	±3.61E-3	5.60E-2	6.30E-2	±4.28E-3	6.37E-2
		2	1.42E-2	±1.28E-3	1.36E-2	1.60E-2	±1.54E-3	1.55E-2
PoD	0.5	0.5	9.25E-1	±3.13E-3	9.63E-1	8.89E-1	±4.13E-3	9.41E-1
		2	7.32E-1	±3.74E-3	7.27E-1	7.13E-1	±4.75E-3	7.02E-1
	2	0.5	8.80E-1	±3.66E-3	9.12E-1	8.57E-1	±4.36E-3	8.94E-1

A.3 More Examples Revealing Limitations of the Approximations

A.3.1 High Abandonment Rates

In the most simple $M/M/n+M$ model, both DGA and TGA can not correctly estimate the key performances when abandonment rate $\theta = 4, 10$. Moreover, the estimate of means (and probabilities) deteriorate faster than that of variances, see Table A.4 for details.

Table A.4: A comparison of the TGA and DGA approximations to exact values for the Markovian $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $n = 100, \rho = 1.05$ and $\theta = 4, 10$

	$\theta = 4$			$\theta = 10$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	96.6	100	-	95.0	100	-
rel. err		4%			5%	
Var(X)	56.1	25.1	-	44.9	10.0	-
rel. err		55%			78%	
E[Q]	1.47	0.100	2.05	0.665	0	1.26
rel. err		83%	39%		100%	90%
Var(Q)	7.77	25.1	8.76	2.47	10.0	3.41
rel. err		223%	13%		303%	38%
E[W]	1.66E-2	2.40E-3	2.11E-2	8.20E-3	1.00E-3	1.31E-2
rel. err		86%	27%		88%	61%
Var(W)	8.80E-4	2.50E-3	9.00E-4	3.10E-4	1.00E-3	3.50E-4
rel. err		182%	2%		219%	13%
PoD	0.379	-	0.508	0.282	-	0.500
rel. err			34%			77%
PoA	5.83E-2	9.50E-3	8.12E-2	6.58E-2	9.95E-2	1.23E-1
rel. err		84%	39%		85%	87%

A.3.2 Smaller UL Systems

To suppplement the UL results shown in §3.8.3, we now show the experimental results for smaller UL systems. We consider systems with smaller arrival rates and numbers of servers $n = 20, 10, 5, 3$ and 1. For different n, we choose different values for the traffic intensity, following (3.25), with the QoS factor fixed at $\beta = 0.5$. Just as in Table 3.16, Table A.5 shows good performance for X_n and B_n for $0.5 \leq \theta \leq 2.0$, but poor performance otherwise.

Table A.5: A comparison of the TGA and DGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/n + M(2)$ with $n = 50, 20, 10, 5, 3$ and 1, $\rho = 1 - \beta/\sqrt{n}$, $\lambda = n\rho$ and $\theta = 0.5$.

	$n = 50, \rho = 0.93$			$n = 20, \rho = 0.88$			$n = 10, \rho = 0.84$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	4.75E+1	4.65E+1	same	1.84E+1	1.78E+1	same	8.84E+0	8.42E+0	same
rel. err.		2%			3%			5%	
Var(X)	5.90E+1	4.65E+1	same	2.26E+1	1.78E+1	same	1.08E+1	8.42E+0	same
rel. err.		21%			22%			22%	
E[B]	4.55E+1	4.65E+1	4.52E+1	1.72E+1	1.78E+1	1.70E+1	8.00E+0	8.42E+0	7.89E+0
rel. err.		2%	1%		4%	1%		5%	1%
Var(B)	2.48E+1	4.65E+1	2.61E+1	9.39E+0	1.78E+1	1.01E+1	4.39E+0	8.42E+0	4.79E+0
rel. err.		87%	5%		89%	7%		92%	9%
E[Q]	2.00E+0	0.00E+0	1.31E+0	1.23E+0	0.00E+0	7.95E-1	8.34E-1	0.00E+0	5.35E-1
rel. err.		100%	35%		100%	35%		100%	36%
Var(Q)	1.60E+1	0.00E+0	7.69E+0	6.27E+0	0.00E+0	2.88E+0	3.06E+0	0.00E+0	1.33E+0
rel. err.		100%	52%		100%	54%		100%	56%
E[V]	4.54E-2	0.00E+0	2.62E-2	7.51E-2	0.00E+0	3.97E-2	1.12E-1	0.00E+0	5.35E-2
rel. err.		100%	42%		100%	47%		100%	52%
Var(V)	7.48E-3	0.00E+0	3.60E-3	2.02E-2	0.00E+0	9.19E-3	4.42E-2	0.00E+0	1.87E-2
rel. err.		100%	52%		100%	54%		100%	58%
PoD	3.67E-1	0.00E+0	6.69E-1	3.73E-1	0.00E+0	6.61E-1	3.80E-1	0.00E+0	6.52E-1
rel. err.		100%	82%		100%	77%		100%	72%
PoA	2.16E-2	NaN	1.30E-2	3.45E-2	NaN	1.97E-2	4.95E-2	NaN	2.64E-2
rel. err.		NaN	40%		NaN	43%		NaN	47%

	$n = 5, \rho = 0.78$			$n = 3, \rho = 0.71$			$n = 1, \rho = 0.5$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	4.16E+0	3.88E+0	same	2.33E+0	2.13E+0	same	5.82E-1	5.00E-1	same
rel. err.		7%			9%			14%	
Var(X)	5.00E+0	3.88E+0	same	2.77E+0	2.13E+0	same	6.61E-1	5.00E-1	same
rel. err.		22%			23%			24%	
E[B]	3.61E+0	3.88E+0	3.55E+0	1.94E+0	2.13E+0	1.93E+0	4.18E-1	5.00E-1	5.00E-1
rel. err.		8%	2%		10%	0%		20%	20%
Var(B)	1.99E+0	3.88E+0	2.11E+0	1.08E+0	2.13E+0	1.03E+0	2.43E-1	5.00E-1	1.60E-1
rel. err.		95%	6%		98%	4%		106%	34%
E[Q]	5.53E-1	0.00E+0	3.50E-1	3.97E-1	0.00E+0	2.49E-1	1.64E-1	0.00E+0	9.98E-2
rel. err.		100%	37%		100%	37%		100%	39%
Var(Q)	1.47E+0	0.00E+0	5.93E-1	8.45E-1	0.00E+0	3.12E-1	2.27E-1	0.00E+0	6.00E-2
rel. err.		100%	60%		100%	63%		100%	74%
E[V]	1.70E-1	0.00E+0	7.01E-2	2.36E-1	0.00E+0	8.31E-2	5.20E-1	0.00E+0	9.98E-2
rel. err.		100%	59%		100%	65%		100%	81%
Var(V)	1.01E-1	0.00E+0	3.77E-2	1.93E-1	0.00E+0	6.24E-2	9.27E-1	0.00E+0	1.60E-1
rel. err.		100%	63%		100%	68%		100%	83%
PoD	3.88E-1	0.00E+0	6.41E-1	3.96E-1	0.00E+0	6.30E-1	4.18E-1	0.00E+0	5.99E-1
rel. err.		100%	65%		100%	59%		100%	43%
PoA	7.12E-2	NaN	3.44E-2	9.31E-2	NaN	4.07E-2	1.64E-1	NaN	4.87E-2
rel. err.		NaN	52%		NaN	56%		NaN	70%

Table A.6: A comparison of the TGA and DGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/n + M(\theta^{-1})$ model with $(n, \theta, \lambda) = (100, 0.5, 100\rho)$ and $\rho \to 1$

	$\rho = 1.01$			$\rho = 1.005$			$\rho = 1.001$		
Perf.	Exact	DGA	TGA	Exact	DGA	TGA	Exact	DGA	TGA
E[X]	1.05E+2	1.02E+2	same	1.04E+2	1.01E+2	same	1.03E+2	1.00E+2	same
rel. err.		3%			3%			3%	
Var(X)	1.55E+2	2.02E+2	same	1.52E+2	2.01E+2	same	1.49E+2	2.00E+2	same
rel. err.		30%			32%			34%	
E[Q]	7.63E+0	1.91E+0	6.67E+0	7.11E+0	9.02E-1	6.12E+0	6.70E+0	1.00E-1	5.69E+0
rel. err.		75%	13%		87%	14%		99%	15%
Var(Q)	8.68E+1	2.02E+2	7.99E+1	8.12E+1	2.01E+2	7.36E+1	7.68E+1	2.00E+2	6.88E+1
rel. err.		133%	8%		147%	9%		161%	10%
E[V]	7.92E-2	2.00E-2	6.70E-2	7.40E-2	1.00E-2	6.16E-2	7.01E-2	2.00E-3	5.74E-2
rel. err.		75%	15%		86%	17%		97%	18%
Var(V)	8.90E-3	2.00E-2	7.98E-3	8.38E-3	2.00E-2	7.39E-3	7.97E-3	2.00E-2	6.93E-3
rel. err.		125%	10%		139%	12%		151%	13%
PoD	6.42E-1	5.56E-1	same	6.20E-1	5.28E-1	same	6.01E-1	5.06E-1	same
rel. err.		13%	same		15%	same		16%	same
PoA	3.78E-2	3.20E-2	same	3.54E-2	2.94E-2	same	3.35E-2	2.75E-2	same
rel. err.		15%	same		17%	same		18%	same

A.3.3 Critically Loaded Systems with $\rho > 1$

So far, we have concentrated on $G/GI/n + GI$ models with $(\rho, n) = (1.05, 100)$, or models with smaller scale n but the same QoS factor $\beta = \sqrt{n}(1 - \rho) = -0.5$. We now consider systems with lighter loading, closer to the critical loading at $\rho = 1.000$. These include systems with $n = 100$ but $1.00 < \rho < 1.05$ and systems with smaller n but $-0.5 < \beta < 0.0$.

Table A.6 provides results for $M(\lambda^{-1})/M/n+M(\theta^{-1})$ (Erlang-A) models with parameter triples $(\lambda, \theta, n) = (n\rho, 0.5, 100)$ and traffic intensity ρ ranges among $\{1.01, 1.005, 1.001\}$. Table A.6 shows that the TGA performs well in the extreme cases ($\rho = 1.001$) and surprisingly does not degenerate even the system become (critically) underloaded, when the abandonment rate $\theta = 0.5$. However, as θ decreases in light traffic models, the performance of DGA and TGA degenerates, see results of $\theta \leq 0.25$ in Tables A.10 and A.11. More experiments for ligher loading models with different queues are presented in Tables A.7-A.11 showing performance of TGA for the $M/M/n + M$ model with abandonment rate $0.1 \leq \theta \leq 4$ and different traffic intensity 1.1, 1.03, 1.02, 1.01 and 1.001. As we have observed before, the performance of TGA is very good if θ is sufficiently close to 1, because in the case $\theta = 1$, the number in system coincides with the number of busy servers in the infinite-server $M/M/\infty$ model, which has exactly a Poisson distribution.

Table A.7: A comparison of the TGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $(\lambda, \rho) = (110, 1.10)$ and $0.1 \leq \theta \leq 4$

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Exact	TGA	rel. err.	Exact	TGA	rel. err.	Exact	TGA	rel. err.
E[X]	2.00E+2	2.00E+2	0%	1.40E+2	1.40E+2	0%	1.20E+2	1.20E+2	0%
Var(X)	1.10E+3	1.10E+3	0%	4.23E+2	4.40E+2	4%	2.08E+2	2.20E+2	6%
E[Q]	1.00E+2	1.00E+2	0%	4.04E+1	4.02E+1	1%	2.08E+1	2.06E+1	1%
Var(Q)	1.09E+3	1.10E+3	0%	4.13E+2	4.18E+2	1%	1.87E+2	1.88E+2	0%
E[V]	9.58E-1	9.53E-1	1%	3.90E-1	3.83E-1	2%	2.03E-1	1.96E-1	3%
Var(V)	9.96E-2	9.98E-2	0%	3.77E-2	3.80E-2	1%	1.72E-2	1.71E-2	1%
PoD	9.99E-1	9.99E-1	0%	9.80E-1	9.72E-1	1%	9.29E-1	9.11E-1	2%
PoA	9.09E-2	9.05E-2	1%	9.19E-2	9.03E-2	2%	9.46E-2	9.16E-2	3%
	$\theta = 1$			$\theta = 2$			$\theta = 4$		
Perf.	Exact	TGA	rel. err.	Exact	TGA	rel. err.	Exact	TGA	rel. err.
E[X]	1.10E+2	1.10E+2	0%	1.04E+2	1.05E+2	1%	1.01E+2	1.02E+2	2%
Var(X)	1.10E+2	1.10E+2	0%	6.53E+1	5.50E+1	16%	4.41E+1	2.75E+1	38%
E[Q]	1.09E+1	1.09E+1	0%	5.77E+0	6.10E+0	6%	3.06E+0	3.57E+0	17%
Var(Q)	8.26E+1	8.08E+1	2%	3.57E+1	3.45E+1	4%	1.51E+1	1.49E+1	1%
E[V]	1.08E-1	1.04E-1	4%	5.88E-2	5.82E-2	1%	3.23E-2	3.41E-2	5%
Var(V)	7.72E-3	7.35E-3	5%	3.43E-3	3.14E-3	9%	1.52E-3	1.36E-3	11%
PoD	8.42E-1	8.30E-1	1%	7.31E-1	7.50E-1	3%	6.12E-1	6.83E-1	12%
PoA	9.92E-2	9.59E-2	3%	1.05E-1	1.05E-1	0%	1.11E-1	1.18E-1	6%

Table A.8: A comparison of the TGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $(\lambda, \rho) = (103, 1.03)$ and $0.1 \leq \theta \leq 4$

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Exact	TGA	rel. err.	Exact	TGA	rel. err.	Exact	TGA	rel. err.
E[X]	1.36E+2	1.30E+2	4%	1.16E+2	1.12E+2	3%	1.08E+2	1.06E+2	2%
Var(X)	7.63E+2	1.03E+3	35%	3.04E+2	4.12E+2	36%	1.68E+2	2.06E+2	23%
E[Q]	3.63E+1	3.30E+1	9%	1.73E+1	1.55E+1	11%	9.97E+0	9.21E+0	8%
Var(Q)	7.10E+2	7.50E+2	6%	2.45E+2	2.44E+2	0%	1.10E+2	1.07E+2	3%
E[V]	3.63E-1	3.25E-1	10%	1.75E-1	1.52E-1	13%	1.02E-1	9.08E-2	11%
Var(V)	6.93E-2	7.28E-2	5%	2.41E-2	2.37E-2	2%	1.10E-2	1.04E-2	6%
PoD	9.13E-1	8.25E-1	10%	8.18E-1	7.23E-1	12%	7.28E-1	6.62E-1	9%
PoA	3.53E-2	3.17E-2	10%	4.20E-2	3.67E-2	13%	4.84E-2	4.32E-2	11%
	$\theta = 1$			$\theta = 2$			$\theta = 4$		
Perf.	Exact	TGA	rel. err.	Exact	TGA	rel. err.	Exact	TGA	rel. err.
E[X]	1.03E+2	1.03E+2	0%	9.98E+1	1.01E+2	2%	9.77E+1	1.01E+2	3%
Var(X)	1.03E+2	1.03E+2	0%	7.02E+1	5.15E+1	27%	5.29E+1	2.57E+1	51%
E[Q]	5.70E+0	5.72E+0	0%	3.22E+0	3.67E+0	14%	1.78E+0	2.42E+0	36%
Var(Q)	4.94E+1	4.78E+1	3%	2.17E+1	2.20E+1	1%	9.33E+0	1.03E+1	11%
E[V]	5.94E-2	5.64E-2	5%	3.44E-2	3.62E-2	5%	1.98E-2	2.39E-2	20%
Var(V)	5.04E-3	4.65E-3	8%	2.29E-3	2.14E-3	7%	1.03E-3	1.00E-3	3%
PoD	6.29E-1	6.16E-1	2%	5.29E-1	5.83E-1	10%	4.34E-1	5.59E-1	29%
PoA	5.54E-2	5.27E-2	5%	6.25E-2	6.61E-2	6%	6.92E-2	8.41E-2	22%

Table A.9: A comparison of the TGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $(\lambda, \rho) = (102, 1.02)$ and $0.1 \leq \theta \leq 4$

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Exact	TGA	rel. err.	Exact	TGA	rel. err.	Exact	TGA	rel. err.
E[X]	1.29E+2	1.20E+2	7%	1.13E+2	1.08E+2	5%	1.06E+2	1.04E+2	2%
Var(X)	6.70E+2	1.02E+3	52%	2.82E+2	4.08E+2	45%	1.62E+2	2.04E+2	26%
E[Q]	2.97E+1	2.52E+1	15%	1.48E+1	1.27E+1	15%	8.75E+0	7.91E+0	10%
Var(Q)	6.03E+2	6.19E+2	3%	2.14E+2	2.07E+2	3%	9.84E+1	9.34E+1	5%
E[V]	2.98E-1	2.49E-1	16%	1.51E-1	1.26E-1	17%	9.01E-2	7.84E-2	13%
Var(V)	5.95E-2	6.07E-2	2%	2.14E-2	2.03E-2	5%	9.96E-3	9.16E-3	8%
PoD	8.72E-1	7.34E-1	16%	7.74E-1	6.54E-1	16%	6.86E-1	6.10E-1	11%
PoA	2.91E-2	2.43E-2	16%	3.64E-2	3.03E-2	17%	4.29E-2	3.74E-2	13%

	$\theta = 1$			$\theta = 2$			$\theta = 4$		
Perf.	Exact	TGA	rel. err.	Exact	TGA	rel. err.	Exact	TGA	rel. err.
E[X]	1.02E+2	1.02E+2	0%	9.91E+1	1.01E+2	2%	9.71E+1	1.00E+2	3%
Var(X)	1.02E+2	1.02E+2	0%	7.11E+1	5.10E+1	28%	5.45E+1	2.55E+1	53%
E[Q]	5.09E+0	5.10E+0	0%	2.91E+0	3.37E+0	16%	1.62E+0	2.27E+0	40%
Var(Q)	4.46E+1	4.31E+1	3%	1.98E+1	2.03E+1	3%	8.54E+0	9.71E+0	14%
E[V]	5.34E-2	5.06E-2	5%	3.13E-2	3.34E-2	7%	1.82E-2	2.25E-2	24%
Var(V)	4.61E-3	4.23E-3	8%	2.11E-3	1.99E-3	6%	9.58E-4	9.54E-4	0%
PoD	5.92E-1	5.78E-1	2%	4.96E-1	5.56E-1	12%	4.06E-1	5.40E-1	33%
PoA	4.99E-2	4.74E-2	5%	5.70E-2	6.11E-2	7%	6.37E-2	7.97E-2	25%

Table A.10: A comparison of the TGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $(\lambda, \rho) = (101, 1.01)$ and $0.1 \leq \theta \leq 4$

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Exact	TGA	rel. err.	Exact	TGA	rel. err.	Exact	TGA	rel. err.
E[X]	1.23E+2	1.10E+2	10%	1.10E+2	1.04E+2	6%	1.05E+2	1.02E+2	3%
Var(X)	5.78E+2	1.01E+3	75%	2.60E+2	4.04E+2	55%	1.55E+2	2.02E+2	30%
E[Q]	2.40E+1	1.83E+1	24%	1.26E+1	1.02E+1	19%	7.63E+0	6.72E+0	12%
Var(Q)	4.96E+2	4.78E+2	4%	1.85E+2	1.71E+2	8%	8.68E+1	8.05E+1	7%
E[V]	2.43E-1	1.82E-1	25%	1.29E-1	1.01E-1	22%	7.92E-2	6.69E-2	15%
Var(V)	4.96E-2	4.73E-2	5%	1.87E-2	1.69E-2	9%	8.90E-3	7.97E-3	10%
PoD	8.23E-1	6.23E-1	24%	7.27E-1	5.79E-1	20%	6.42E-1	5.56E-1	13%
PoA	2.37E-2	1.78E-2	25%	3.12E-2	2.45E-2	22%	3.78E-2	3.20E-2	15%
	$\theta = 1$			$\theta = 2$			$\theta = 4$		
Perf.	Exact	TGA	rel. err.	Exact	TGA	rel. err.	Exact	TGA	rel. err.
E[X]	1.01E+2	1.01E+2	0%	9.84E+1	1.00E+2	2%	9.66E+1	1.00E+2	4%
Var(X)	1.01E+2	1.01E+2	0%	7.21E+1	5.05E+1	30%	5.61E+1	2.52E+1	55%
E[Q]	4.52E+0	4.53E+0	0%	2.61E+0	3.09E+0	18%	1.47E+0	2.13E+0	45%
Var(Q)	3.99E+1	3.85E+1	4%	1.79E+1	1.86E+1	4%	7.77E+0	9.09E+0	17%
E[V]	4.78E-2	4.51E-2	6%	2.84E-2	3.08E-2	8%	1.66E-2	2.12E-2	27%
Var(V)	4.18E-3	3.82E-3	9%	1.94E-3	1.85E-3	5%	8.85E-4	9.03E-4	2%
PoD	5.53E-1	5.40E-1	2%	4.63E-1	5.28E-1	14%	3.79E-1	5.20E-1	37%
PoA	4.47E-2	4.23E-2	5%	5.17E-2	5.64E-2	9%	5.83E-2	7.51E-2	29%

Table A.11: A comparison of the TGA approximations to exact numerical values in the $M(\lambda^{-1})/M(1)/100 + M(\theta^{-1})$ model with $(\lambda, \rho) = (100.1, 1.001)$ and $0.1 \leq \theta \leq 4$

	$\theta = 0.1$			$\theta = 0.25$			$\theta = 0.5$		
Perf.	Exact	TGA	rel. err.	Exact	TGA	rel. err.	Exact	TGA	rel. err.
E[X]	1.18E+2	1.01E+2	14%	1.08E+2	1.00E+2	7%	1.03E+2	1.00E+2	3%
Var(X)	5.00E+2	1.00E+3	100%	2.41E+2	4.00E+2	66%	1.49E+2	2.00E+2	34%
E[Q]	1.96E+1	1.31E+1	33%	1.08E+1	8.18E+0	24%	6.70E+0	5.74E+0	14%
Var(Q)	4.08E+2	3.54E+2	13%	1.60E+2	1.40E+2	13%	7.68E+1	6.93E+1	10%
E[V]	2.00E-1	1.31E-1	34%	1.12E-1	8.18E-2	27%	7.01E-2	5.74E-2	18%
Var(V)	4.12E-2	3.54E-2	14%	1.63E-2	1.40E-2	15%	7.97E-3	6.93E-3	13%
PoD	7.72E-1	5.13E-1	34%	6.81E-1	5.08E-1	25%	6.01E-1	5.06E-1	16%
PoA	1.96E-2	1.29E-2	34%	2.70E-2	1.98E-2	27%	3.35E-2	2.75E-2	18%
	$\theta = 1$			$\theta = 2$			$\theta = 4$		
Perf.	Exact	TGA	rel. err.	Exact	TGA	rel. err.	Exact	TGA	rel. err.
E[X]	1.00E+2	1.00E+2	0%	9.77E+1	1.00E+2	2%	9.61E+1	1.00E+2	4%
Var(X)	1.00E+2	1.00E+2	0%	7.30E+1	5.00E+1	31%	5.76E+1	2.50E+1	57%
E[Q]	4.04E+0	4.04E+0	0%	2.36E+0	2.84E+0	20%	1.34E+0	2.00E+0	49%
Var(Q)	3.59E+1	3.45E+1	4%	1.62E+1	1.72E+1	6%	7.10E+0	8.55E+0	20%
E[V]	4.30E-2	4.04E-2	6%	2.59E-2	2.85E-2	10%	1.53E-2	2.00E-2	31%
Var(V)	3.81E-3	3.45E-3	9%	1.78E-3	1.72E-3	4%	8.19E-4	8.56E-4	5%
PoD	5.17E-1	5.04E-1	3%	4.33E-1	5.03E-1	16%	3.54E-1	5.02E-1	42%
PoA	4.03E-2	3.80E-2	6%	4.72E-2	5.22E-2	11%	5.36E-2	7.11E-2	33%

Appendix B

Computing Confidence Interval for Variances by Jackknife Method

In Chapter 3, we approximate the confidence interval for variances by the confidence interval of second moment information. However, suggested by Wilson [120], there is a more accurate method to approximate the confidence intervals for variance based on jackknifing (see details in [23, 98]), and we explain the method below.

In each simulation replication of r, where $r = 1, 2, \ldots, R$ and $R = 2,000$, we collect $N_r = 551$ queue lengths (and other system performance measures) time-averagely during interval $[0.4T, 0.95T]$: $Q_{r,1}, Q_{r,2}, \ldots, Q_{r,N_r}$. Here T is the constant time horizon and is long enough for system to approach to steady state.

Then we calculate the sample mean $\tilde{Q}_r$ and second moment $\tilde{Q}_r^{(2)}$ by

$$\tilde{Q}_r = \frac{1}{N_r}\sum_{l=1}^{N_r} Q_{r,l}, \quad \text{and} \quad \tilde{Q}_r^{(2)} = \frac{1}{N_r}\sum_{l=1}^{N_r} Q_{r,l}^2,$$

and estimate the steady-state variance of queue length based on all R replications

$$\hat{V} = \frac{1}{R}\sum_{r=1}^{R} \tilde{Q}_r^{(2)} - \left[\frac{1}{R}\sum_{r=1}^{R} \tilde{Q}_r\right]^2.$$

Now, to build a jackknifed confidence interval estimation of steady-state variance, we averagely divide R replications into $g = 20$ blocks, so each block has $h = R/g$ replications. In the jth block, we have sampled mean and second moment $\tilde{Q}_k$ and $\tilde{Q}_k^{(2)}$ where

$$k \in B_j \equiv \{(j-1)h + 1, (j-1)h + 2, \ldots, jh\}, \quad \text{for } j = 1, 2, \ldots, g.$$

Based on the constructed block, we compute the jackknifed estimator of variance for each block, by removing the data in the block.

$$\hat{V}_j = \frac{1}{R-h}\sum_{r\notin B_j}\tilde{Q}_r^{(2)} - \left[\frac{1}{R-h}\sum_{r\notin B_j}\tilde{Q}_r\right]^2, \quad \text{for } j = 1,\ldots,g.$$

We use overall mean variance $\hat{V}$ and jackknifed mean variance for each block $\hat{V}_j$ to generate pseudo values $\tilde{V}_j$.

$$\tilde{V}_j = g\hat{V} - (g-1)\hat{V}_j, \quad \text{for } j = 1,\ldots,g.$$

These pseudo values are regarded as if they are i.i.d. so the their sample mean is approximating steady-state variance without bias.

$$\bar{\tilde{V}} = \frac{1}{g}\tilde{V}_j, \quad \text{and} \quad s^2_{\bar{\tilde{V}}} = \frac{1}{g-1}\sum_{j=1}^{g}\left(\bar{\tilde{V}} - \tilde{V}_j\right).$$

where $s^2_{\bar{\tilde{V}}}$ is the sample variance of the pseudo values.

Hence we attain the confidence interval for variance shown as

$$\bar{\tilde{V}} \pm t_{g-1,0.025}s^2_{\bar{\tilde{V}}},$$

where $t_{g-1,0.025}$ is the 97.5% quantile of t-distribution with $g-1=19$ degree of freedom.

Then we compare the jackknife method to our original method. A simple numerical comparison is given in Table B.1 below, which shows that the two method have significant gaps between each other. However, since the original method cannot be justified, we should use the jackknife method to calculate the confidence interval for variances.

Table B.1: Comparison of CI's

$\hat{V} + s^2_{\tilde{Q}^{(2)}}$	$\bar{\tilde{V}} \pm t_{g-1,0.025}s^2_{\bar{\tilde{V}}}$
161.8821 ± 0.0233	161.8821 ± 0.4016