ARAS, AHMET KORHAN. Heavy-Traffic Gaussian Limits for Queues with Nonstationary Arrivals, and Nonexponential Service and Patience Times. (Under the direction of Yunan Liu.)

In this thesis, we develop heavy-traffic approximations for many-server queueing systems. The queueing models considered here have servers working independently in parallel, and these models are most convenient to model moderate-to-large-scaled service systems such as call centers and healthcare systems, and to model production systems, and communication systems. Similar models are used in other fields such as finance, computing, and biology. Heavy-traffic approximations provide simpler analytical formulas to approximate the key performance measures associated with complicated systems. Such approximations are key to the analysis of the queueing models considered in this thesis because the models here are not amenable to exact analysis.

Key realistic features common in many service systems are incorporated into the models. These features include time-varying arrivals and staffing levels, random service times having a general probability distribution, and customer abandonment—customers are willing to wait in queue only for random amount of time having a general probability distribution. These features make the model non-Markovian, and hence, much less tractable. Therefore, we prove a functional weak law of large numbers (FWLLN) and functional central limit theorem (FCLT) for key performance processes such as the number in system, number of abandonments, head-of-line waiting time, and virtual waiting time. This enables us to derive analytical formulas approximating the real-time performance of the system, to compute time-dependent mean and variance formula associated with the performance of the key processes.

More specifically, we obtain heavy-traffic conditions by appropriately increasing the arrival rate and the number of servers. When scaled appropriately, the performance processes converge in distribution to “nice” processes. In particular, we use two different scalings to obtain a deterministic and stochastic limiting process for each performance process. The deterministic limiting processes are obtained by proving an FWLLN, and the stochastic limiting processes are obtained by proving an FCLT. We obtain approximation formulas for the processes by combining the associated limiting processes. These approximations can be used to improve capacity planning in and operational control of real-world systems.

We establish an FCLT for the performance processes associated with the $G_t/GI/s_t + GI$ queueing system having time-varying arrivals and staffing levels (i.e., the number of servers), and general (i.e., nonexponential) service- and patience-time distribution. To do so, we first establish an FCLT for the so-called the initial content process (ICP) in the context of infinite-server queues, but the results can be applied to finite-server queues as has been done in this thesis.
Then we establish an FCLT for the stationary finite-server $G/GI/n + GI$ queueing system. We also characterize the steady state of the associated limiting processes. We also consider an open network of parallel $G_t/GI/s_t + GI$ queueing systems, and establish an FWLLN for the key performance processes.

This thesis consists of four pieces of work presented in four separate chapters. In chapter 2, we establish a many-server heavy-traffic FWLLN for the $(G_t/GI/s_t + GI)^m/M_t$ open queueing network, with a finite number of queues (the superscript $m$), nonstationary non-Poisson external arrival processes (the $G_t$), nonexponential service times (the first $GI$), time-varying staffing levels (the $s_t$), and customer abandonment following nonexponential patience times (the $+GI$). Upon service completion, customers are either routed to one of the queues in the network or out of the system according to time-dependent probabilities (the $M_t$). The limit provides support for a previously proposed deterministic fluid approximation and extends a previously established limit for the $G_t/GI/s_t + GI$ single queue model.

In chapter 3, we focus on the evolution of the content that is initially in service at some time. For infinite-server queues, it is often fruitful to partition the content into the initial content and the new input, because these can be analyzed separately. For finite-server queues one cannot simply analyze initial content separately, however, the results turn out to be very valuable for the analysis of finite-server queueing systems. With i.i.d service times having a nonexponential distribution, the state of the initial content in infinite-server queues can be described by specifying the elapsed service times of the remaining initial customers. That initial content process is then a Markov process. We establish a heavy-traffic FCLT for the initial content process for two-parameter stochastic processes as state descriptors, assuming a FCLT for the initial age process, with the number of customers initially in service growing in the limit. The proof addresses a technical challenge: For each time, including time 0, the conditional remaining service times, given the ages, are mutually independent but in general not identically distributed.

In chapter 4, we establish a many-server heavy-traffic FCLT for the $G/GI/n + GI$ queueing system with stationary arrivals, nonexponential service times, $n$ identical servers, and non-exponential patience times. Under the assumption that the queueing system operates in the efficiency-driven (ED) regime, we establish process-level convergence for the key performance processes such as the head-of-line waiting time, virtual waiting time, queue length, abandonment and service-completion process to non-Markovian Gaussian limits as the arrival rate and the number of servers goes to infinity. We develop analytic formulas to characterize the distributions of these Gaussian limits. An important engineering value of these Gaussian limits is that they can be used to effectively approximate the steady-state performance of the $G/GI/n + GI$ queueing system.

In chapter 5, we establish a many-server heavy-traffic FCLT for the $G_t/GI/s_t + GI$ queueing
system alternating between the ED and QD regime. The system is assumed to be critically loaded only at finitely many isolated regime switching points. The treatment of switching points is crucial as the ending conditions of the previous interval becomes the initial conditions for the subsequent interval. We prove an FCLT for the performance processes associated with both the ED and QD regime provided that the queuing system is not critically loaded anywhere in the interval of interest. We conjecture that the FCLT for the number in system can be extended to the intervals that are critically loaded only at the endpoints of the interval of interest, whereas the FCLT for the other processes does not immediately carry over.
Heavy-Traffic Gaussian Limits for Queues with Nonstationary Arrivals, and Nonexponential Service and Patience Times

by
Ahmet Korhan Aras

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

Operations Research

Raleigh, North Carolina
2016

APPROVED BY:

James Wilson
Min Kang

Nilay Tanik Argon
Ward Whitt

Yunan Liu
Chair of Advisory Committee
DEDICATION

To my mother, and my lovely wife...
Ahmet Korhan ARAS was born in Ankara, Turkey. He received his B.S. degree in Industrial Engineering in 2008, and M.S. degree in Industrial Engineering in 2011 from Bilkent University in Ankara, Turkey. In 2010, he joined the Ph.D program in Operations Research at North Carolina State University. Since February 2016, he has been working at SAS Institute. His research interests include applied probability, stochastic modeling, queueing theory and its applications.
ACKNOWLEDGEMENTS

First, I would like to thank my PhD advisor Dr. Yunan Liu for his invaluable support and guidance throughout my Ph.D. study. I am very fortunate to have worked with an enthusiastic advisor who has had a keen interest in all the problems I have worked on during my Ph.D. study. I am grateful to him for his help every step of the way, but especially, for his help with proofs. He has always had time to discuss about research and, as a great teacher, he has always been very patient bearing with my endless questions. Doing research with him has been an invaluable experience.

I have had a chance to work with great collaborators. I would like to thank Dr. Xinyun Chen for her contribution to this thesis. She has provided great help with the mathematics we needed to get a very crucial step in the thesis done. It has been an honor to have worked with a pioneer of our field - Professor Ward Whitt. I am greatly thankful for his contributions to this thesis in various aspects.

I would like to thank my Ph.D. committee members James Wilson, Min Kang, and Nilay Tanik Argon for their valuable comments that have helped making this thesis better. I would like to thank Dr. Min Kang for bearing with my questions about the course material of various courses, about more general topics in probability theory, and most importantly, about some of the proofs in this thesis. I am very fortunate to have taken courses from such a great teacher.

I would like to thank Professor Thom Hodgson for being exceptionally helpful supporting me financially by having me as his TA for several semesters.

I would like to thank my wife for her patience and constant support during the course of my Ph.D. study. I have been very busy working on my thesis most of the time, but she has always found ways to create valuable time for us, and to make my life enjoyable. I feel very lucky to have met her in the freshman year at Bilkent University. My life has changed for the better since then.

I would like to thank my mother who has sacrificed a lot, as a single parent, for me so that I get the best education I could possibly get in Turkey. I would also like to thank her for convincing me to pursue Ph.D. studies abroad. I do appreciate the vision she had for me more and more as years pass. I am eternally grateful to her.

It was great having a “Turkish gang” whom I had a chance to hangout and have quality time with, and learn a lot from. I would like to thank my fellow Ph.D. students Müge Çapan, Carl Pankok, İrem Şengül Örgüt, “Res” Örgüt, Erinç Albey, Hakan Karagül, Ali Kefeli, and Çağatay Karan for their company. I would also like to thank my new friends at SAS Institute for their moral support.
# TABLE OF CONTENTS

LIST OF TABLES .......... viii

LIST OF FIGURES .......... ix

Chapter 1 Introduction .......... 1
  1.1 Time-varying arrivals and staffing levels ................. 1
  1.2 Customer Abandonment and Non-Markovian Models .............. 2
  1.3 Network Structure ................................ 3
  1.4 Many-server Heavy-traffic Approximations ........................ 3
    1.4.1 Heavy-traffic Regimes ............................. 4
    1.4.2 Scaling and Approximation Formulas ..................... 6
  1.5 Related Literature .................................. 8
    1.5.1 Finite-server queues .................................. 8
    1.5.2 Infinite-server queues ............................... 10
    1.5.3 Network of queues .................................. 11
  1.6 Organization of the thesis ................................ 11
  1.7 Notations ........................................... 12
  1.8 Preliminaries ...................................... 13
    1.8.1 Functional spaces and convergence ......................... 13
    1.8.2 Some useful results .................................. 16
  1.9 Acronyms ........................................... 21

Chapter 2 A FWLLN for the Time-Varying $(G_t/GI/s_t + GI)^m/M_t$ Queueing Network .......... 23
  2.1 Introduction ...................................... 23
  2.2 A Sequence of $(G_t/GI/s_t + GI)^m/M_t$ Queueing Networks ........ 24
  2.3 The $(G_t/GI/s_t + GI)^m/M_t$ Fluid Network ........................ 27
    2.3.1 The FQNet and Its Parameters ........................ 27
    2.3.2 Performance Functions of the FQNet ....................... 28
  2.4 FWLLN for the $(G_t/GI/s_t + GI)^m/M_t$ SQNet .................. 31
  2.5 Proof of the Main Result ................................ 33
    2.5.1 FWLLN for service related processes at OL queues ................ 34
    2.5.2 FWLLN for the Total Arrival Process .................... 40
    2.5.3 FWLLNs for Other Processes ............................. 42
  2.6 An $(M_t/H_2/s_t + E_2)^2/M_t$ Example ........................ 45
  2.7 Conclusion ....................................... 46

Chapter 3 Heavy-traffic Limit for the Initial Content Process .......... 49
  3.1 Introduction ...................................... 49
  3.2 The Model After Time 0 ................................ 52
  3.3 Main Results ...................................... 57
  3.4 Proof of Theorem 3.3.2. ................................ 62
    3.4.1 Decomposition of $\hat{X}_n^{e,o}$ ......................... 63
### Chapter 3

- 3.4.2 Convergence for the First Term in (3.40) ........................................ 64
- 3.4.3 Convergence for the Second Term in (3.40) .................................... 72
- 3.4.4 Proof of Convergence for the Other Processes ................................ 74
- 3.4.5 Proof of Alternative Representations in (3.22) and (3.27) ................. 74

### Chapter 4

- 3.5 The $G_t/GI_0, GI_\nu/\infty$ Model Starting in the Past ......................... 75
- 3.6 Steady-State Approximations ............................................................ 78
- 3.7 A Challenging Test Case ................................................................. 84
- 3.8 Conclusion ......................................................................................... 85

### Chapter 5

- 5.1 Introduction ....................................................................................... 124
LIST OF TABLES

Table 1.1 Summary of frequently used acronyms (in alphabetical order) . . . . . . . . . . 22

Table 4.1 $H_2(\lambda^{-1}, c_\lambda^2)/GI/100 + H_2(\theta^{-1}, c_\theta^2)$ with $(\lambda, \rho, \theta, c_\lambda^2, c_\theta^2) = (120, 1.2, 0.5, 4, 4)$ . . 120

Table 4.2 $H_2(\lambda^{-1}, c_\lambda^2)/GI/100 + E_2(\theta^{-1})$ with $(\lambda, \rho, \theta, c_\lambda^2, c_\theta^2) = (120, 1.2, 0.5, 4, 0.5)$ . . . 121
LIST OF FIGURES

Figure 1.1 Flowchart of the chapters................................................................. 12
Figure 2.1 Proof strategy for the FWLLN in Theorem 1................................. 33
Figure 2.2 Performance functions of the \( (M_t/H_2/s_t + E_2)^2/M_t \) FQNet, including (i) TAR \( \lambda \), (ii) queue content \( Q \), (iii) VWT \( w \), (iv) service content \( B \), (v) total fluid \( X \) and (vi) rate into service \( b(t, 0) \)...................................................... 45
Figure 2.3 A comparison of performance functions in the \( (M_t/H_2/s_t + E_2)^2/M_t \) FQNet with simulation of the corresponding \( (M_t/H_2/s_t + E_2)^2/M_t \) SQNet with (a) single sample paths and scale \( n = 3000 \), and (b) average of 100 paths and scale \( n = 100 \)............................................... 47
Figure 3.1 Example 1: Simulation comparisons of the mean and variance for the number of customers in service in an \( M_t/LN(1, 4)/\infty \) model starting empty at a finite negative time \(-t_0 = -5\), with the sinusoidal arrival rate (3.72) having parameters \( a = c = 1, b = 0.6, \phi = 0 \) and \( n = 100 \)....................................... 77
Figure 3.2 Example 2: Simulation comparisons of the mean and variance for the number of new and old customers in service of an \( M/H_2(1, 4)/\infty \) model in steady state, with a constant arrival rate \( n\lambda = 100 \)................................................. 83
Figure 3.3 Two choices of the initial-condition density \( b^*(u) \) in (3.91)............... 85
Figure 3.4 Example 3 with \( b^*_1 \) in (3.92): Simulation comparisons of the mean and variance for the number of customers in service of an \( H_2^0(1, 4)/LN(1, 4)/\infty \) model, with the sinusoidal arrival rate (3.72) having parameters \( a = c = 1, b = 0.6, \phi = 0 \) and \( n = 100 \) and general initial conditions........................................ 86
Figure 4.1 Graphic demonstration of \( E_n^\nu(t) \)...................................................... 99
Figure 4.2 Reflection of \( \hat{E} \) on \((-\infty, 0] \)....................................................... 108
Figure B.1 Example 4 with \( b^*_2 \) in (3.92): Simulation comparisons of the mean and variance for the number of customers in service of an \( H_2^0(1, 4)/LN(1, 4)/\infty \) model, with the sinusoidal arrival rate (3.72) having parameters \( a = c = 1, b = 0.6, \phi = 0 \) and \( n = 100 \) and general initial conditions........................................ 179
Chapter 1

Introduction

Many service systems have the following features (i): time-varying demand, e.g., arrivals of jobs, customers, patients, calls; (ii) time-varying staffing levels, e.g., number of agents/nurses in call centers/hospitals are not constant in a given day; (iii) abandonment from queue, e.g., impatient customers may choose to leave the system before getting service or jobs/orders may be canceled before being processed; (iv) non-Markovian probability structure, e.g., interarrival, service- and patience-time distributions are not near exponential; (v) network structure, e.g, manufacturing systems/hospitals with parts/patients moving from one workstation/department to other. Queueing systems having these realistic features are extremely complicated and not amenable to exact analysis. In this thesis, we propose an approximation for both single queues and queueing networks with (i)-(iv), and a certain type queueing network composed of finite number of queues with (i)-(iv).

In the rest of this chapter, we elaborate on the above realistic features and many-server heavy-traffic approximations. Then we review the related literature, and introduce the notation. Finally we review some existing results that we have frequently referred to.

1.1 Time-varying arrivals and staffing levels

Real-world systems typically experience arrivals that vary over time. Recent empirical studies have also revealed the time-varying nature of arrivals. A study by Brown et al. [13] has shown that the arrival rate of incoming calls in a call center can be highly time-varying in a day. Armony et al. [6] confirm the time-varying behavior of patient arrivals in an emergency department in a day and in different days of the week. Because of time-varying arrivals, methods developed for stationary queues and steady-state analysis perform poorly.

In the presence of time-varying arrivals, adjustment of the staffing level is crucial to cope with the variability, and to achieve stability in performance measures of interest such as waiting
time of customers before they start service. For example, in call centers, an important objective is to answer a certain percentage of incoming calls within a relatively short amount of time [25] (e.g., 80% of calls within 20 seconds). To achieve such an objective, the staffing level must be high enough when the number of arrivals is high. When the number of arrivals is low, on the other hand, the objective can again be achieved with high staffing levels, however, this is not desirable because of operational costs (see [31]). Therefore, in real-world systems it is crucial to be responsive to time-varying arrivals to balance the trade-off between high quality of service (i.e., high percentage of customers served within short delay) and cost effectiveness (i.e., low staffing costs).

In this thesis, motivated by these real-world systems, we develop limit theorems, and based on those limit theorems we derive useful engineering approximation formulas for queueing systems with time-varying arrivals and staffing levels. We focus on the performance analysis of a system rather than stabilization of some performance measures by setting staffing levels. In other words, we treat staffing level, which is a function of time, as an input of the model, not a decision variable.

1.2 Customer Abandonment and Non-Markovian Models

Customer abandonment is common in many service systems such as call centers and emergency rooms in hospitals. In call centers, customers may choose to hang up if they have waited long enough before being answered by an agent. Similarly, a patient may leave without being seen by a doctor after excessive delay. As shown in [26], abandonments have a big impact on the performance of the system. Garnett et al. [26] have reported that even a small number of customer abandonments can cause a drastic change in average waiting time and queue length. Therefore, it is important to incorporate customer abandonment into existing models to better understand and to accurately model real-world systems.

Customer abandonments have typically been incorporated into models through random patience times. As is common in call centers, a customer may retry later or never try again and be considered as a lost customer. In this thesis, we model customer abandonment assuming that customers have random patience times drawn from a general distribution, once abandoned they are lost customers, i.e., no retrials.

The same empirical study by Brown et al. [13] also shows that durations of calls and patience times of customers are not exponentially distributed. According to their findings, durations of calls are close to the lognormal distribution, and the hazard-rate function associated with the patience-time distribution is not constant meaning, that the patience times are not exponential. Liu and Whitt [61] show that the service-time distribution has a significant impact on the transient behavior of queues with time-varying arrivals. These findings motivate us to go beyond
the classical Erlang models such as the Erlang-A model $M/M/s/\infty + M$ with Poisson arrivals (the first $M$), service times drawn from the exponential distribution (the second $M$), $s$ servers, patience times drawn from the exponential distribution (the $+M$), and unlimited waiting space (the $/\infty$). Erlang models are desirable because they are highly tractable: the number-in-system process is a continuous-time Markov process, in particular, a birth-and-death process. In this thesis, we consider more general, hence more difficult, models than Erlang models.

1.3 Network Structure

Most real-world systems are more complicated than queue systems with a single queue. In fact, they tend to be networks of parallel and/or tandem (in series) queues in which the customers/jobs are directed from one queue to another queue multiple times before leaving the system. Some important examples are large-scale service systems arising in customer contact centers, communication networks, healthcare systems and production systems.

In this thesis, we particularly work with a network of parallel multiserver queues with Markovian routing, i.e., each customer/job is routed upon service completion either to one of other queues in the network, or out of the system with respect to time-dependent probabilities independent of the past behavior/dynamics of the system. This network consists of queues with the external arrival process at each queue being a nonstationary general (non-Poisson) process, with service and patience times being nonexponential random variables, and with time-varying staffing levels (i.e., number of servers). Such networks are denoted by $(G_t/GI/s_t + GI)^m/M_t$ where $M_t$ stands for the Markovian (probabilistic) routing policy, and $m$ denotes the (finite) number of parallel queues in the network.

In Chapter 2, we consider the $(G_t/GI/s_t + GI)^m/M_t$ the queueing network, and we prove a functional weak law of large numbers (FWLLN) for the key performance processes such as number in system, number of abandonments, and virtual waiting time. This result serves as a mathematical justification for the fluid approximation to same queueing network in [64]. The results can be used to develop stochastic refinements to the $(G_t/GI/s_t + GI)^m/M_t$ queueing network. Although the main theme of this thesis is developing stochastic refinements, we leave the study of developing stochastic refinements the queueing network for future work.

1.4 Many-server Heavy-traffic Approximations

Exact analyses are available for Markovian models but any generalization beyond rises a significant challenge. To analyze queueing systems with nonexponential interarrival, service, and patience times, we develop useful heavy-traffic approximations because analytic performance formulas are not available.
Heavy-traffic approximations for queueing systems is hardly a new topic. The general idea is to increase the system load (so that traffic intensity reaches the critical value 1) by increasing the arrival rate, and appropriately adjusting the service capacity such that the system stays stable. The number in system and waiting times tend to grow indefinitely under these conditions, however, under appropriate scaling, one may find nondegenerate limits which serve as approximations for the original queueing system. We next review some key elements of heavy-traffic analysis.

1.4.1 Heavy-traffic Regimes

We first elaborate on three extensively used ways of obtaining heavy-traffic conditions. These heavy-traffic conditions are referred to as heavy-traffic regimes that are characterized by certain limiting conditions as the scale increases indefinitely. The conventional heavy-traffic regime [52] is obtained by keeping the number of servers fixed and increasing the traffic intensity (often times denoted by $\rho$) to 1. This is accomplished by considering a sequence of queues indexed by, say, $n$, with appropriately scaled arrival rate and service rate such that the traffic intensity approaches 1 as $n \to \infty$. The Halfin-Whitt heavy-traffic regime is obtained by keeping the service rate fixed and increasing arrival rate and number of servers. More specifically, a queueing system with single waiting line and finitely-many servers is said to operate in the Halfin-Whitt regime if the following condition holds: For some constant $\beta$,

$$\sqrt{n}(1 - \rho_n) \equiv \sqrt{n}(1 - \frac{\lambda_n}{n\mu}) \to \beta \quad \text{as} \quad n \to \infty. \quad (1.1)$$

This regime is introduced in the seminal paper by Halfin and Whitt [34] and initiated a new stream of research. Other two well-known regimes are quality-driven (QD) and efficiency-driven (ED) regimes introduced by Garnett et al. [26]. The QD regime is when the limit in (1.1) is $+\infty$ whereas the ED regime is when the limit is $-\infty$. The terms “quality-driven” and “efficiency-driven” have been named after two different settings queueing systems operate in. In particular, in revenue-generating systems quality of service is important whereas in service-oriented systems such as call centers, capacity utilization is emphasized over quality service ([103]). The Halfin-Whitt regime is also known as quality-and-efficiency driven (QED) regime in that it combines high level of efficiency (high server utilization) and of service quality (short waiting times, small number of abandonments). Mandelbaum and Zeltyn [73] introduce the ED+QED regime which is a QED refinement of the ED regime in the sense that it provides an additional condition on the speed of convergence. More specifically, there exists $\bar{\beta} \in \mathbb{R}$ and a sequence of constants
\[ \{w^n\}_{n \geq 1} \text{ converging to } w \in (0, \infty) \]

\[
\lim_{n \to \infty} \frac{\lambda_n}{s_n \mu} = \rho > 1 \quad \text{and} \quad \frac{\lambda_n F_c^n(w^n) - s_n \mu}{\sqrt{\lambda_n}} \to \bar{\beta} \quad \text{as} \quad n \to \infty. \tag{1.2}
\]

In the recent paper by Atar [7], a new heavy-traffic regime called the nondegenerate slowdown (NDS) regime is proposed to serve as a midpoint between the two extreme regimes that are the conventional and Halfin-Whitt regimes. The term slowdown is defined as the ratio between the sojourn time and the service time experienced by a typical customer. This regime can be obtained by setting the number of servers to \(\sqrt{n}\) and scaling service rates by \(\sqrt{n}\). We refer interested readers to the survey paper [99] for more about the regimes (Also see [40]).

The above regimes are mostly applicable in the context of stationary queues, i.e., when the arrival rate is constant, or, applicable for queueing systems with time-inhomogeneous arrival process provided the regime does not change over the time period of interest. Many real-world systems, on the other hand, may operate in different regimes over a time period. More specifically, in the presence of time-varying arrivals, real-world systems may experience consecutive periods of overloading and underloading and this is because the staffing level may not be adjusted timely and sensitively enough due to constraints such as shortage of agents, staffing costs, limited budget, etc. Motivated by this aspect of real-world systems, Liu and Whitt [61] developed a fluid model with time-varying arrival rate. The fluid model alternates between underloaded and overloaded time intervals dictated by time-varying fluid arrival. In a sequel paper [65], the authors define regimes for the (stochastic) queueing system based on the fluid model in [61]. In particular, the system is said to be overloaded (resp. underloaded) if the associated fluid queue is overloaded (resp. underloaded). The way these two regimes are defined for systems with time-varying arrivals parallels the way the above regimes for stationary queues are defined. More specifically, (1.1)-(1.2) and other conditions in the other papers are asymptotic conditions that are assumed to hold as \(n \to \infty\). These conditions imply that there may be some systems which violate these conditions for some \(n < N_0\). However, by definition of limit, all systems that are large enough, i.e., \(n \geq N_0 \) for sufficiently large \(N_0\), must satisfy the associated conditions. Analogously, the system may be underloaded (resp. overloaded), i.e., there are some idle servers (resp. all servers are busy and possibly queue is not empty), during a time interval where the associated fluid queue is overloaded (resp. underloaded) due to randomness. However, for sufficiently-large-scaled systems, the regime of the (stochastic) system and the associated fluid queue must agree. More specifically, they agree with probability 1 as \(n \to \infty\).

Finally, the underloaded (resp. overloaded) regime in the context of nonstationary queues can be considered as the analogue of the QD (resp. ED) regime described above. In this thesis, we prove limit theorems for some key performance processes of both stationary and nonstationary queues operating in both the underloaded (QD) and the overloaded (ED) regimes.
1.4.2 Scaling and Approximation Formulas

The queueing models considered in this thesis are not amenable to exact analysis. Therefore, we resort to asymptotic analysis which requires appropriate scaling of the key performance processes to obtain nondegenerate limits as the scaling factor \( n \to \infty \). In each of the following chapters, we explicitly define the scalings we apply to the key processes such as number in system, number waiting in line, head-of-line waiting time, potential waiting etc. To give a broad sense of our asymptotic analysis, we next review our framework and two scalings that we use throughout this thesis. We discuss more details in §1.8.

For asymptotic analysis, we consider a sequence of queues indexed by \( n \geq 1 \) where the \( n \)th queue has constant or time-varying arrival rate \( \lambda_n \) scaled by \( n \) such that \( \lambda_n \to \infty \) as \( n \to \infty \), and also has \( s_n \) (constant or time-varying) many servers where \( s_n \) is scaled by \( n \) such that \( s_n \to \infty \) as \( n \to \infty \). Depending on the context, some conditions are specified to relate \( \lambda_n \) and \( s_n \) in order to characterize a heavy-traffic regime. The discussion in this section is very broad hence we do not give any details on regimes. Let \( X_n \) be a real-valued process, which may have one parameter or two parameters, associated with the \( n \)th system. We want to establish convergence of the forms

\[
\frac{X_n}{n} \Rightarrow X \quad \text{in} \quad \mathcal{X} \quad \text{and} \quad \frac{X_n - nX}{\sqrt{n}} \Rightarrow \hat{X} \quad \text{in} \quad \mathcal{X} \quad \text{as} \quad n \to \infty, \tag{1.3}
\]

where \( \Rightarrow \) denotes convergence in distribution, \( \mathcal{X} \) is an appropriate function space, and \( X_n \) and \( X \) have sample paths in \( \mathcal{X} \). We apply only spatial scaling as given in (1.3). We do not apply time scaling where the time parameter(s) is (are) also scaled by a scaling factor.

For simplicity, we base our discussion on stochastic processes with one parameter until the end of this section. All arguments below for one-parameter processes are also valid for two-parameter processes.

**Fluid approximation.** Often times the limit \( X \) in (1.3) is continuous and deterministic, and called the fluid limit. The intuition is that as the scale increases (i.e., as \( n \to \infty \)) the customers behaves like the atoms of fluid. When the scale is large, the behavior of customers moving from the queue to service resembles the flow of fluid from a “storage reservoir” into “service reservoir” where the fluid is processed and leaves the system at a certain rate.

The first limit in (1.3) suggests the following approximation formula: For each \( n \geq 1 \),

\[
X_n(t) = nX(t) + o(n), \quad t \in I, \tag{1.4}
\]

where \( I \) is a subset of \( \mathbb{R} \) (e.g., \([0, T]\), \([0, \infty)\)), and \( o(1) \) is asymptotically negligible as \( n \to \infty \) which implies that the accuracy of the approximation in (1.4) increases as \( n \) increases. The equality in (1.4) holds for almost all sample paths of \( X_n \), i.e., there exists a subset \( \Omega_0 \) of the set
of all realizations such that (1.4) holds for all paths in the subset $\Omega_0$, and $P(\Omega_0) = 1$. That is, for practical purposes, if $X$ is known, then $nX$ provides a good approximation for all realizations of $X_n$. Note that even if the convergence in (1.3) is in distribution sense, (1.4) can be used to approximate sample paths of $X_n$. This is due to Skorohod representation theorem (see, e.g., [102]). Skorohod representation theorem states that there exists $\tilde{X}_n \overset{d}{=} X_n$ for all $n$, and $\tilde{X} \overset{d}{=} X$ in a common probability space such that $P(\lim_{n \to \infty} \tilde{X}_n = \tilde{X}) = 1$ where $\overset{d}{=}$ means equal in distribution (see §1.8).

Fluid limits provide effective estimates of the “average” of the performance measures of a queueing system such as the mean queue length, the mean waiting time etc. (see e.g. [65]). Although fluid limits do not account for the fluctuations around the mean, they still provide accurate approximations for large systems ($n$ sufficiently large) such as large-scale call centers (see e.g. [65]).

**Stochastic refinement.** As discussed above, fluid limits are deterministic approximations for stochastic systems. In order to quantify the impact of randomness and to have a better approximation to stochastic systems, it is important to develop stochastic refinements to fluid approximations. Stochastic refinements are developed by identifying a stochastic process that characterizes stochastic fluctuations around the associated fluid approximation. More specifically, the limit $\hat{X}(\cdot)$ in (1.3) is a stochastic refinement that accounts for stochastic fluctuations around the fluid limit $X(\cdot)$. The phrase “fluctuations around” comes from the simple observation that the prelimit processes $\{X_n\}_{n \geq 1}$ on the right-hand side in (1.3) are “centered” around the fluid limit by subtracting the scaled fluid limit $nX$ from $X_n$ for each $n$. Roughly speaking, the numerator $X_n - nX$ in (1.3) is the deviation from the mean curve. The second convergence in (1.3) suggests the following approximation formula: For each $n \geq 1$,

$$X_n(t) = nX(t) + \sqrt{n}\hat{X}(t) + o(\sqrt{n}), \quad t \in I,$$

where $I$ is a subset of $\mathbb{R}$ (e.g., $[0,T]$, $[0,\infty)$), and $o(1)$ is asymptotically negligible as $n \to \infty$. Similar to (1.4), the formula in (1.5) is interpreted in almost sure sense, i.e., practically all realizations of $X_n$ can be approximated by the sum of a deterministic process $X$ and a stochastic process $\hat{X}$.

We say a **functional weak law of large numbers** (FWLLN) holds for a sequence of stochastic processes $\{X_n\}_{n \geq 1}$ if the first convergence in (1.3) holds. We say a **functional central limit theorem** (FCLT) holds for a sequence of stochastic processes $\{X_n\}_{n \geq 1}$ if the second convergence in (1.3) holds. In each chapter of this thesis, we prove an FWLLN and an FCLT for the key performance processes of the queueing systems of interest. To summarize, FWLLN limits serve as deterministic fluid functions whereas FCLT limits are stochastic processes that accounts for the fluctuations around the FWLLN limits.
It is of course desirable to have analytically tractable FCLT limits such as Brownian motion, reflected Brownian motion, fractional Brownian motion, Levy processes, and more general diffusion processes with known generators. FCLT limits, i.e., stochastic refinements, are called diffusion approximations when the limit process $\hat{X}$ is a Markov process. Different types of scalings in addition to those in (1.3) have been used in the literature to obtain Markov-process limits because of the analytical tractability of Markov processes. See [102] for more on diffusion approximations.

As is common in the literature, we do not scale either of service- and patience-time distribution (see §1.5 for alternatives). Because the queues we consider in this thesis have nonexponential service and patience times, along with non-Markovian arrivals, it is not possible to obtain Markov-process limits under the spatial scaling. Instead, we obtain Gaussian processes with explicit analytical forms or with known covariance functions. Consequently, our engineering approximations using the formula (1.5) provides a Gaussian-based approximation because the limit $\hat{X}$ is Gaussian.

1.5 Related Literature

There is a large body of literature on many-server heavy-traffic (MSHT) limits. In this section, we survey the most related papers. First, we review the literature on MSHT limits for finite-server queues. Second, we review the literature on heavy-traffic limits for infinite-server queues which turn out to play an important role in the analysis of finite-server queueing systems. We have used and extended existing results on the $GI/GI/\infty$ infinite-server queueing systems and use them establish MSHT for the $G/GI/n+GI$ and $G_t/GI/s_t+GI$ queueing systems. Therefore, for completeness, we also provide the most related literature on infinite-server queueing systems. Finally, we list the most relevant papers on queueing networks.

1.5.1 Finite-server queues

Many-server heavy-traffic limits have been extensively studied following the seminal paper by Halfin and Whitt [34], and have been used to provide approximations for the key performance processes such as queue length, number-in-system, and others for various types of queueing systems. See [17, 43, 71, 48, 49, 88] for heavy-traffic limits for queueing systems having different type of service-time distributions. Here we discuss the most related literature. A FWLLN for discrete-time number-in-system process associated with the $G_t(n)/GI/s+GI$ queueing system with time-dependent and state-dependent (the $n$) arrival process, and having general service- and patience-time distribution has been proved by Whitt [104]. Among the work that has been done in continuous-time framework are [89, 16, 72, 87, 48]. In [89], a many-server heavy-
The traffic diffusion limit has been established for the number of customers in the system in the $G/GI/N$ queueing system having stationary a general arrival process, and general service-time distribution. Dai and He [16] consider the $G/G/n + GI$ queueing system in the QED regime, and they prove a deterministic relationship that asymptotically holds between appropriately scaled abandonment and queue-length processes. In another study, Mandelbaum and Momcilovic [72] provide a diffusion approximation for the virtual-waiting-time process as well as the queue-length process associated with the $G/GI/N + GI$ in the QED regime. The results in [89] have been further generalized in [87] such that their results hold without the assumption of system’s operating in the QED regime. Kaspi and Ramanan [48] prove a law of large numbers for the number in system in the $G/GI/N$ queueing system by using a different framework. In particular, they define a measure-valued process, which keeps track of the ages of customers in service, to fully describe the state of the system, and establish a MSHT law of large numbers for the measure-valued process and the number-in-system process. Kang and Ramanan [46] extend the results in [48] by incorporating customer abandonment. They represent the dynamics of the $G/GI/N + GI$ queueing system in terms of two measure-valued processes one of which we have already mentioned, and the other keeps track of the waiting times of the customers in the queue. They establish a law of large numbers for those two measure-valued processes. In an independent study, Zhang [109] takes a similar approach, and provide a fluid limit for the same $G/GI/N + GI$ queueing system. Kaspi and Ramanan [49] obtain diffusion limits for the $GI/G/n$ queueing system using measure-valued processes as the state descriptor. In another study, Huang et al. [40] establish diffusion limits for the $G/GI/N_n + GI$ queueing system in the Halfin-Whitt regime. They use a general framework that unifies no scaling and hazard-rate scaling of the patience-time distribution.

Other studies involving scaling of the patience-time distribution are [92, 90, 57, 40], where the patience-time distributions are scaled through scaling of the associated hazard-rate functions. Hazard-rate scaling is introduced by Reed and Ward [92] to a develop diffusion approximation for the number-in-system process, and the offered-waiting-time process associated with the $G/GI/1 + GI$ queueing system operating in the conventional regime. A diffusion approximation for the number-in-system process using hazard rate scaling is then obtained for the $G/M/N + GI$ queueing system in the Halfin-Whitt regime by Reed and Tezcan [90]. By scaling patience-time distributions, Lee and Weerasinghe [57] obtain a diffusion limit for the appropriately scaled offered waiting time in the conventional regime. Huang et al. [40] develop a unified approach to obtain diffusion approximations that applies to both $G/GI/1 + GI$ and $G/GI/N_n + GI$ queues. Motivated by delay announcement in some service systems, Huang et al. [39] develop a refined approximation for the number in system in the $GI/M/s + GI$ queueing system operating in the ED+QED regime. They also characterize the stationary distribution of the limiting diffusion process.
Two of the recent works by Liu and Whitt [62, 65] provide many-server heavy-traffic limits for the queues with time-varying arrivals and staffing levels. In [62], the authors establish a mathematical basis for the $G_t/GI/s_t + GI$ fluid queueing system developed in [61] by proving a functional weak law of large numbers. In [65], they establish Gaussian approximations for the $G_t/M/s_t + GI$ queueing system alternating between underloaded and overloaded regimes. In the underloaded regime, the finite-server queueing system is asymptotically equivalent to an infinite-server queueing system having same arrival process and service-time distribution, so that the existing functional central limit theorem (FCLT) for infinite-server queues are valid. In the overloaded regime, on the other hand, the dynamics of the system is more complicated. In particular, they show that the limiting head-of-line waiting time process satisfies a stochastic differential equation driven by three independent Brownian motions. The limits for the rest of the processes are zero-mean Gaussian processes.

1.5.2 Infinite-server queues

Iglehart [41] shows that appropriately scaled number in system in the $M/M/\infty$ queueing system converges in distribution to an Ornstein-Uhlenbeck (OU) process as the scale increases. Borovkov [11] considers a more general infinite-server queueing system with general arrivals and service times, and shows that the limiting number-in-system process is Gaussian but not Markovian. Whitt [100] later shows that the non-Markovian structure of the limit can be attributed to the nonexponential service times, because appropriately scaled number-in-system processes in the $G/M/\infty$ queueing system converge in distribution to an OU process. Glynn [28] later analyzes the limiting process in [11], and provides a necessary and sufficient condition that ensures that the limiting process is Markovian under the assumption that appropriately scaled arrival processes converge to a Wiener process. Glynn and Whitt [29] further analyze the structure of the limiting process in [11] by considering service times following variants of exponential distributions. They also study open networks of infinite-server queues.

Krichagina and Puhalskii [54] develop heavy-traffic limit theorems for a closed queueing network consisting of an infinite server queue and a single-server queue with general service times and state-dependent arrival rates, i.e., the $G/GI/\infty$ queue. They further characterize the Gaussian limit in [11]; in particular, they show that the limit for the queue length processes can be represented as a two-dimensional stochastic integral with respect to a Kiefer process [68]. Extending [29, 54], Pang and Whitt [78] use two-parameter processes to describe the state of the $G/GI/\infty$ queueing system. Specifically, they consider two types of processes: (i) the number in system at time $t$ with elapsed service times at most $y$, and (ii) the number in system at time $t$ with residual service times greater than $y$. They develop heavy-traffic limits for these two-parameter processes by extending the results for one-parameter processes in [54]. Pang and
Whitt [80] later extend the results in [78] by incorporating weak dependence among consecutive service times.

Reed and Talreja [91] make use of distribution-valued processes to describe the state of the $G/GI/\infty$ queueing system, and obtain a diffusion limit that is a tempered-distribution-valued OU process, extending the diffusion approximation in [18] for the $M/GI/\infty$. For fluid approximations based on measure-valued processes, see [46, 48, 109]. Also see the survey paper [81] for a review on martingale-based approaches for infinite-server systems.

### 1.5.3 Network of queues

We now review the most relevant studies on non-Markovian queueing system with time-varying arrivals and staffing levels. The MSHT fluid and diffusion limits are developed by Mandelbaum et al. [69] Markovian queueing networks having possibly time-dependent Poisson arrivals, and exponential service- and patience-time distributions (i.e., $M_t/M_t/s_t + M_t$). In a recent study, Puhalskii [86] considers large-time asymptotics of the $M_t/M_t/K_t + M_t$ queueing system, and obtains fluid and diffusion limits. Liu and Whitt [61] propose a fluid approximation for the $G_t/GI/s_t + GI$ queue with time-varying arrivals and nonexponential distributions; they later extend to the framework of networks [60, 64]. An FWLLN [62] has been established to substantiate the fluid approximation in [61] and an FCLT [65] has been developed for the $G_t/M/s_t + GI$ model with exponential service times. Paralleling [64], He and Liu [35] develop a fluid approximation for the multi-class queueing network with deterministic routing paths. Extending [60] and [65], Huang and Liu [38] develop a MSHT FCLT for network of queues with exponential service times.

### 1.6 Organization of the thesis

This thesis consists of four pieces of work that we have done throughout my Ph.D. study. In Chapter 2, we consider the $(G_t/GI/s_t + GI)^m/M_t$ queueing network, and establish an heavy-traffic FWLLN for the key performance processes. We show that the fluid model developed in [64] is indeed the limit for the LLN-scaled key performance processes, and hence, we provide a mathematical justification for the fluid model in [63]. In Chapter 3, we focus on the evolution of the content that is initially in service, and establish a many-server heavy-traffic FCLT for the initial content process assuming an FCLT for the initial age (elapsed time in service) process. The results in Chapter 3 are presented in the context of infinite-server queues, however, they can be applied to finite-server queues as has been done in Chapter 5. In Chapter 4, we establish an FCLT for the key performance processes associated with the stationary $G/GI/n + GI$ queueing system. We also characterize the steady state of the FCLT limits. In Chapter 5, we consider
the $G_t/GI/s_t + GI$ queueing system with the arrival process having a time-varying rate, and with time-varying staffing levels. This queueing system experiences periods of underloading and overloading due to time-varying arrivals. We establish an FCLT for the key performance processes for both underloaded and overloaded intervals where the intervals are assumed to be critically loaded only at the left endpoints. An FCLT for the intervals with critical loading at both left and right endpoints are left for future work. Additional material associated with each chapter is provided in the appendix. More details about the organization of the chapters are provided in the chapters.

We conclude the introduction chapter by introducing the notations, reviewing some preliminary results, and providing a table of acronyms.

### 1.7 Notations

We now introduce the notations used throughout this thesis. The sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, $\mathbb{R}_+$ denote the natural numbers excluding 0, i.e., $\{1, 2, 3, \ldots\}$, the set of integers, the real numbers and the nonnegative real numbers, respectively. The cartesian products of these sets are denoted by $\mathbb{N}^k$, $\mathbb{Z}^k$, $\mathbb{R}^k$, $\mathbb{R}_+^k$, and are endowed with product topologies. For $a, b \in \mathbb{R}$, $a \vee b \equiv \max\{a, b\}$, $a \wedge b \equiv \min\{a, b\}$, $a^+ \equiv a \vee 0$, $a^- \equiv a \wedge 0$, $[a] \equiv \max\{j \in \mathbb{Z} : j \leq a\}$ is the largest integer less than or equal to $a$, and $[a] \equiv \min\{j \in \mathbb{Z} : j \geq a\}$ is the largest integer greater
then or equal to \( a \). We let \( e \) be the identity function, i.e., \( e(t) = t, \ t \geq 0; \ \chi \) be the constant one function \( \chi(t) = 1, \ t \geq 0 \). For two functions \( f \) and \( g \), \( (f \circ g)(t) \equiv f(g(t)) \) denotes the composition of \( f \) and \( g \). We say \( f(t) = O(g(t)) \) if \( \limsup_{t \to \infty} f(t)/g(t) \leq K \) for some \( K > 0 \) and \( f(t) = o(g(t)) \) if \( \limsup_{t \to \infty} f(t)/g(t) = 0 \). Let \( \| \cdot \|_{[a,b]} \) be the supremum norm over the interval \([a,b] \in \mathbb{R} \), i.e., \( \|f\|_{[a,b]} = \sup_{u \in [a,b]} |f(u)| \). For convenience, we define \( \| \cdot \|_b = \| \cdot \|_{[0,b]} \). For two-parameter functions, the supremum norm is defined as \( \|f\|_{[a,b] \times [c,d]} = \sup_{(u,v) \in [a,b] \times [c,d]} |f(u,v)| = \sup_{u \in [a,b]} \sup_{v \in [c,d]} |f(u,v)| \). Similarly, \( \| \cdot \|_{b,d} = \| \cdot \|_{[0,b] \times [0,d]} \).

We denote by \( \mathbb{D} \equiv \mathbb{D}([0,T]; \mathbb{R}) \) the space of real-valued right-continuous functions with left limits on the interval \([0,T] \), and by \( \mathbb{C} \equiv \mathbb{C}([0,T]; \mathbb{R}) \) the subset of \( \mathbb{D} \) consists of continuous functions. We define \( \mathbb{D}_\mathbb{D} \equiv \mathbb{D}([0,T_1]; \mathbb{D}([0,T_2]; \mathbb{R})) \) as the set of right-continuous functions with left limits whose images are functions that are elements of \( \mathbb{D}([0,T_2]; \mathbb{R}) \). In other words, \( \mathbb{D}_\mathbb{D} \) is the space of \( \mathbb{D}([0,T_2]; \mathbb{R}) \)-valued functions which are right-continuous on the interval \([0,T_1] \) with left limits on the interval \((0,T_1] \). The definition of the spaces \( \mathbb{C}((0,\infty), \mathbb{R}), \mathbb{D} \equiv \mathbb{D}([0,\infty); \mathbb{R}), \) and \( \mathbb{D}((0,\infty); \mathbb{D}([0,\infty); \mathbb{R})) \) is analogous. We denote the cartesian products of these spaces \( \mathbb{C}^k, \mathbb{D}^k, \mathbb{D}_\mathbb{D}^k \), respectively, with the topologies on these product spaces assumed to be the product topology.

All random variables and stochastic processes in a following chapter are assumed to be defined on a common probability space \((\Omega, \mathcal{F}, \mathbb{P}) \). We denote by \( \mathbb{E}[\cdot] \) the expectation, by \( \text{Var}(\cdot) \) the variance, and by \( \text{Cov}(\cdot, \cdot) \) the covariance function. We use \( \overset{a.s.}{=} \) to denote equality in almost sure sense. In particular, two random variables \( X \) and \( Y \) are said to be almost-surely equal, or equal in almost sure sense, if \( \mathbb{P}(\{ \omega \in \Omega : X(\omega) = Y(\omega) \}) = 1 \); two one-parameter stochastic processes \( (X(t) : t \geq 0) \) and \( (Y(t) : t \geq 0) \) are said to be almost-surely equal, or equal in almost sure sense, if \( \mathbb{P}(\{ \omega \in \Omega : X(\omega; t) = Y(\omega; t), \ \forall t \geq 0 \}) = 1 \). If two random variables \( X \) and \( Y \) have a common distribution then we write \( X \overset{d}{=} Y \). We reserve \( \overset{d}{=} \) to denote distributional equality of single-parameter stochastic processes with fixed \( t \) (see §3). We use \( 1_A \) or \( 1(A) \) to denote the indicator random variable of the event \( A \). We denote convergence in distribution by \( \Rightarrow \). We use the acronym “a.s.” for “almost surely” and “w.p.1” for “with probability 1”.

1.8 Preliminaries

In this section, we provide some mathematical background on the framework in this thesis.

1.8.1 Functional spaces and convergence

**Functional spaces.** In each chapter of this thesis, we establish convergence of certain pre-limit stochastic processes to limiting stochastic processes. This requires the notion of convergence in function spaces. In particular, for a constant \( 0 < T < \infty \), we work with the space
$\mathbb{C}([0, T]; \mathbb{R})$ of real-valued continuous functions on $[0, T]$, the space $\mathbb{D}([0, T]; \mathbb{R})$ of real-valued right-continuous functions with left limits in $(0, T)$ (a.k.a. càdlàg or RCLL functions), and the space $\mathbb{D}((0, T]; \mathbb{R})$ of $\mathbb{D}([0, \infty); \mathbb{R})$-valued functions that are right continuous on the interval $[0, T)$ and have left limits in $(0, T]$. We work with the space $\mathbb{D}((0, T]; \mathbb{D}([0, \infty); \mathbb{R}))$ to prove convergence for two-parameter stochastic processes $X \equiv \{X(t, y) : 0 \leq t \leq T, y \geq 0\}$. Note that we do not treat two-parameter stochastic processes as elements of $\mathbb{X}$ prove convergence for two-parameter stochastic processes $D$ as elements of $\mathbb{R}^{\mathbb{C}([0, T]; \mathbb{R})}$, i.e., real-valued RCLL functions defined on the domain $[0, T]$ and have left limits in $(0, T]$. This treatment has been proven to be useful and has made possible the use of the sequential empirical process to find new limits for two-parameter stochastic processes (see, e.g., [54, 78, 65]). Also, the limits in the former space are required to exist at each point in the domain (here $[0, T] \times [0, \infty)$) which is not necessarily true for the two-parameter stochastic processes in this thesis (see [78] for more discussion).

We endow the space $\mathbb{C}([0, T]; \mathbb{R})$ with the uniform topology induced by the uniform norm, that is,

$$\|x - y\|_T \equiv \sup_{s \in [0, T]} \{|x(s) - y(s)|\} \text{ for } x, y \in \mathbb{C}([0, T]; \mathbb{R}). \quad (1.6)$$

The space $\mathbb{C}([0, T]; \mathbb{R})$ is complete and separable under the uniform norm, hence, it is a Banach space. The uniform metric, however, does not work well on $\mathbb{D}((0, T]; \mathbb{R})$ because the uniform metric requires the prelimit processes and the limiting process to have discontinuities at the same time. In terms of convergence of functions with discontinuities, this is a strong requirement. Instead, we endow the space $\mathbb{D}((0, T]; \mathbb{R})$ with the Skorohod $J_1$-topology induced by the $J_1$-metric, that is,

$$d_{J_1}(x, y) = \inf_{\lambda \in \Lambda} \{\|x \circ \lambda - y\|_T \vee \|\lambda - e\|_T\} \text{ for } x, y \in \mathbb{C}([0, T]; \mathbb{R}), \quad (1.7)$$

where $e$ is the identity element in $\mathbb{D}((0, T]; \mathbb{R})$, i.e., $e(t) = t$ for $0 \leq t \leq T$. The set $\Lambda$ is the increasing homeomorphisms of the interval $[0, T]$, i.e., the set of increasing continuous functions mapping $[0, T]$ onto itself with inverse also being a continuous function (see [20, 42, 96, 102] for more details). The space $\mathbb{D}((0, T]; \mathbb{R})$ is not complete under the metric $d_{J_1}$. Fortunately, there is an equivalent metric $d_{J_1}^0$, which gives the $J_1$-topology, under which $\mathbb{D}((0, T]; \mathbb{R})$ is complete (see [10]). Therefore we say that $\mathbb{D}((0, T]; \mathbb{R})$ under the metric $d_{J_1}$ is topologically complete. The space $\mathbb{D}((0, T]; \mathbb{R})$ is also separable because $\mathbb{R}$ is separable. Products of that space are equipped with the product topology.

The domain in the definition of $\mathbb{D}((0, T]; \mathbb{R})$ can be extended from the finite interval $[0, T]$ to the positive half line $[0, \infty)$. Consider a convergent sequence $\{f_n\}$ of functions to the limit $f$ in $\mathbb{D}([0, T]; \mathbb{R})$. If the restrictions of the sequence $\{f_n\}$ to $[0, t]$ converges to the restrictions of the
limit \( f \) to \( [0, t] \) in \( \mathbb{D}([0, t]; \mathbb{R}) \) for all \( t > 0 \) that are continuity point of \( f \), then the convergence is said to be in \( \mathbb{D}([0, \infty); \mathbb{R}) \). The metric on \( \mathbb{D}([0, \infty); \mathbb{R}) \) is given as

\[
d_{\infty}(x, y) = \int_0^\infty e^{-t}(d_t(x, y) \wedge 1) \, dt
\]

(1.8)

for \( x, y \in \mathbb{D}([0, \infty); \mathbb{R}) \) where \( d_t \) is \( d_{J_1} \) in (1.7) over the interval \([0, t]\). Just as the case for the metric space \( (\mathbb{D}([0, T]; \mathbb{R}), d_{J_1}) \), there exists an equivalent metric to \( d_{\infty} \) that gives the same topology as \( d_{\infty} \) does, and under which \( \mathbb{D}([0, \infty); \mathbb{R}) \) is complete (see [10, 20]). Hence we say the metric space \( (\mathbb{D}([0, \infty); \mathbb{R}), d_{\infty}) \) is topologically complete. The space \( \mathbb{D}([0, \infty); \mathbb{R}) \) is separable because \( \mathbb{R} \) is separable.

Let \( (\mathcal{S}, r) \) be a metric space. Then \( \mathbb{D}([0, T]; \mathcal{S}) \) is separable if \( \mathcal{S} \) is separable, and complete if \( (\mathcal{S}, r) \) is complete. From the above discussion, we know that \( \mathbb{D}([0, \infty); \mathbb{R}) \) is complete (resp. topologically complete) and separable under the metric \( d_{\infty} \) (resp. \( d_{J_1} \)), and hence, the space \( \mathbb{D}([0, T]; \mathbb{D}([0, \infty); \mathbb{R})) \) falls in the Skorohod framework. The metric on \( \mathbb{D}([0, T]; \mathbb{D}([0, \infty); \mathbb{R})) \) is obtained by replacing the first norm in (1.7) by

\[
\|x(\cdot, \cdot) - y(\lambda(\cdot), \cdot)\|_{T, \infty} \equiv \sup_{(t,u) \in [0,T] \times [0,\infty)} |x(t, u) - y(\lambda(t), u)|.
\]

(1.9)

The two-parameter stochastic processes in the following chapters have sample paths in the space \( \mathbb{D}([0, T]; \mathbb{D}([0, \infty); \mathbb{R})) \).

**Convergence.** Our main results are all about proving *convergence in distribution* in function spaces of some sequences of stochastic processes (i.e., prelimit processes) to limits that are also stochastic processes. Convergence in the sense of distribution is so dominant in this thesis that we sometimes use the terms “convergence” and “convergence in distribution” interchangeably. We also use the term “weak convergence” instead of “convergence in distribution”. Throughout, we clearly state the mode of convergence if it is different than convergence in distribution. A general definition of convergence in distribution involves a sequence of probability measures on a separable metric space \( (\mathcal{S}, r) \) with a metric \( r \). In particular, the sequence of probability measures \( \mu_n \) is said to converge weakly to the probability measure \( \mu \), i.e., \( \mu_n \Rightarrow \mu \) in \( \mathcal{S} \) if

\[
\lim_{n \to \infty} \int_{\mathcal{S}} f \, d\mu_n = \int_{\mathcal{S}} f \, d\mu
\]

(1.10)

for all continuous bounded real-valued functions \( f \) on \( \mathcal{S} \) (see, e.g., [10, 20, 102] also for alternative characterizations). The survey paper [105] provides an extensive review of convergence in distribution in function spaces.

We relate convergence of probability measures and stochastic processes by observing that
stochastic processes can be considered as random elements of the functional spaces discussed above. Observe, for instance, that for one-parameter stochastic processes on $\mathbb{R}_+$ with RCLL sample paths, $(X(\omega; t) : t \geq 0)$, the sample paths $t \mapsto X(\omega; t)$ are the elements of $\mathbb{D}([0, \infty); \mathbb{R})$ for each fixed $\omega \in \Omega$. Because each $\omega \in \Omega$ is interpreted as a realization of some random phenomenon, the sample paths $t \mapsto X(\omega; t)$ correspond to different elements of $\mathbb{D}([0, \infty); \mathbb{R})$ for each $\omega$. Hence $X(\omega; t)$ is said to be a random element of $\mathbb{D}([0, \infty); \mathbb{R})$. With this viewpoint, $X$ is a measurable mapping from $(\Omega, \mathcal{F}, \mathbb{P})$ to the measurable space $(\mathbb{D}([0, \infty); \mathbb{R}), \mathcal{B}_\mathbb{D})$ where $\mathcal{B}_\mathbb{D}$ is the Borel $\sigma$-algebra defined on $\mathbb{D}([0, \infty); \mathbb{R})$. In other words, $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{B}_\mathbb{D}$. Then the probability measure induced by $X$, $\mathbb{P}_X$, is given by $\mathbb{P}_X(A) \equiv \mathbb{P}(X^{-1}(A))$. Now consider a sequence of stochastic processes $\{X_n\}_{n \geq 1}$. The sequence $\{X_n\}_{n \geq 1}$ is said to converge in distribution or converge weakly to $X$ if the induced probability measures $\{\mathbb{P}_X\}_{n \geq 1}$ weakly converge to $\mathbb{P}_X$ in the sense of (1.10), i.e., $\mathbb{P}_{X_n} \Rightarrow \mathbb{P}_X$ as $n \to \infty$. An equivalent representation of (1.10) for random elements on complete separable metric spaces $\mathcal{S}$ is given in terms of expectations:

$$\lim_{n \to \infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$$

for all continuous bounded real-valued functions $f$ on $\mathcal{S}$. See, e.g., Theorem 11.3.4. of [102] for alternative characterizations.

Since all limits in the rest of the thesis will almost surely have continuous sample paths, convergence in $J_1$ topology is equivalent to uniform convergence over compact sets.

1.8.2 Some useful results

The material in this section plays a key role in each of the following chapters. We review some material on the sequential empirical processes, the standard Kiefer process, and on a certain construction of stochastic integrals for two-parameter martingales. The material in this section is mostly taken from the papers [54] and [78].

**Kiefer and sequential empirical process.** A stochastic process $(W(t) : t \in \mathbb{R}_+^2)$ is the two-parameter Wiener process (the Brownian sheet) if

1. $W(R)$ with $R = [x_1, x_2] \times [y_1, y_2]$ has a normal distribution with mean 0 and variance $(x_2 - x_1)(y_2 - y_1)$, where an increment is defined as

$$W(R) \equiv W(x_2, y_2) - W(x_1, y_2) - W(x_2, y_1) + W(x_1, y_1)$$

with $W(x, y) \equiv W([0, x] \times [0, y])$ for all $0 \leq x, y < \infty$.

2. $W(0, y) = W(x, 0) = 0$ for all $0 \leq x, y < \infty$. 

16
3. $W$ has independent increments, i.e., $W(R_1), W(R_2), W(R_3), \ldots, W(R_n)$, $n \geq 2$ are independent random variables if $R_1, R_2, \ldots, R_n$ are disjoint regions in $\mathbb{R}^2$.

4. The sample path function $W(\omega; t)$ is continuous in $t$ with probability 1.

Analogues to the one-parameter Brownian motion, the covariance of the Brownian sheet is

$$\text{Cov}(W(x_1, y_1), W(x_2, y_2)) = (x_1 \wedge x_2)(y_1 \wedge y_2).$$

For any fixed $0 < x < \infty$, the process $(x^{-1/2}W(x, y) : y \geq 0)$ is the one-dimensional Brownian motion, as is $(y^{-1/2}W(x, y) : x \geq 0)$. See, e.g., [50] for more properties of the Brownian sheet.

The **Kiefer process** $(K(x, y) : x \geq 0, 0 \leq y \leq 1)$ is defined as

$$K(x, y) = W(x, y) - yW(x, 1)$$

where the $W$ is the standard Brownian sheet. More specifically, $K(\cdot, y)$ is a Brownian motion for each fixed $y$, while $K(x, \cdot)$ is a Brownian bridge for each fixed $x$ where Brownian bridge $B^0$ is given by $B^0(t) = W(t) - tW(1)$ with $W$ being the standard Wiener process (Brownian motion). The Kiefer process can be understood to be the Brownian sheet conditioned to be 0 when $y = 1$. The Kiefer process has mean 0 and the covariance function

$$\text{Cov}(K(x_1, y_1), K(x_2, y_2)) = (y_1 \wedge y_2 - y_1 y_2)(x_1 \wedge x_2).$$

Moreover, the sample path functions of the Kiefer process is continuous with probability 1. See, e.g., [78] and [98] for more properties of the Kiefer process. Also see [93] for a semimartingale decomposition of the Kiefer process.

Consider the two-parameter processes

$$\hat{U}_n(t, x) \equiv \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} 1(\xi_i \leq x),$$

$$\hat{U}_n(t, x) \equiv \sqrt{n} \left( \bar{U}_n(t, x) - \mathbb{E} \left[ \bar{U}_n(t, x) \right] \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} (1(\xi_i \leq x) - x), \quad t \geq 0, 0 \leq x \leq 1, \quad (1.11)$$

where $\{\xi_i\}$ are independent and uniformly distributed on $[0, 1]$. The process $\hat{U}_n(t, x)$ is called the **sequential empirical process**. Bickel and Wichura [9] have proved that, if restricted to the region $[0, 1]^2$, the $(\hat{U}_n(t, x) : t \geq 0, 0 \leq x \leq 1)$, as a random element of $D([0, 1]^2; \mathbb{R})$, converges in distribution to the Kiefer process as $n \to \infty$. This result has further been extended by Krichagina and Puhalskii [54]. They have treated $\hat{U}_n(t, x)$ as a random element of the space
\( \mathbb{D}([0, \infty), \mathbb{D}([0, 1]; \mathbb{R})) \) and have proved that

\[
\hat{U}_n \Rightarrow \hat{U} \quad \text{in} \quad \mathbb{D}([0, \infty); \mathbb{D}([0, 1]; \mathbb{R})), \quad \text{as} \quad n \to \infty,
\]

where \( \hat{U} \) is again the standard Kiefer process.

Observe that if \( F \) is cumulative distribution function, then from continuous mapping theorem, we immediately obtain

\[
\hat{U}_n(t, F(x)) \Rightarrow \hat{U}(t, F(x)) \quad \text{in} \quad \mathbb{D}([0, \infty); \mathbb{D}([0, \infty); \mathbb{R}))) \quad \text{as} \quad n \to \infty.
\]

Note that the convergence is with respect to a stronger topology on \( \mathbb{D}([0, \infty); \mathbb{D}([0, \infty); \mathbb{R})). \)

**Stochastic integrals for two-parameter processes.** We next define stochastic integrals for the sequential empirical process and the Kiefer process. A key observation in [54] is that the total number in system at time \( t \) in an infinite-server queueing system that is initially empty

\[
X_n^\infty(t) = \sum_{i=1}^{N_n(t)} 1(\tau^n_i + \nu_i > t), \quad t \geq 0,
\]

can be represented as a double integral with respect to the sequential empirical process, i.e.,

\[
X_n^\infty(t) = n \int_0^t \int_0^\infty 1(s + x > t) d\bar{U}_n \left( \frac{N_n(s)}{n}, F(x) \right), \quad t \geq 0, \tag{1.12}
\]

where \( N_n \) is the arrival process to the \( n \)th queue, \( \tau^n_i \) is the arrival time of the \( i \)th customer in the \( n \)th system, \( \nu_i \) is the service time of the \( i \)th customer, \( F \) is the service distribution, and \( \bar{U}_n \) is as in (1.11). The integrator \( d\bar{U}_n(\cdot, \cdot) \) in (1.12) is understood as a measure that is the sum of Dirac measures. In particular, \( d\bar{U}_n(\cdot, \cdot) \) is the measure

\[
m_n(ds, dx) = n^{-1} \sum_{i=1}^{\infty} \delta_{(\tau^n_i, \nu_i)}(ds, dx)
\]

where \( \delta_{(\tau^n_i, \nu_i)} \) is the Dirac measure at \( (\tau^n_i, \nu_i) \), i.e., \( \delta_{(\tau^n_i, \nu_i)} \) is 1 at \( (\tau^n_i, \nu_i) \) for all \( i \geq 1 \), and 0 elsewhere.

Given the rigorous definition of \( d\bar{U}_n(\cdot, \cdot) \), we understand the double integral in (1.12) as a counting process counting the points \( (\tau^n_i, \nu_i) \) in \( \mathbb{R}^2_+ \) restricted to the region determined by the condition \( s + x > t \) for \( t > 0 \), which is the region above the 45° line through the points \( (t, 0) \) and \( (0, t) \) in \( \mathbb{R}^2_+ \). Because the Dirac measures have value 1 at each point \( (\tau^n_i, \nu_i) \), the double integral in (1.12) indeed gives the number of points in the region \( s + x > t \) in the first quadrant.
As in [54], (1.12) can be represented as the sum of three terms given as
\[ X_n^\infty(t) \equiv X_{n,1}^\infty(t) + X_{n,2}^\infty(t) + X_{n,3}^\infty(t), \quad t \geq 0, \]
where
\[ X_{n,1}^\infty(t) \equiv \sqrt{n} \int_0^t F^c(t - s) d\tilde{N}_n(s), \]
\[ X_{n,2}^\infty(t) \equiv \sqrt{n} \int_0^t \int_0^\infty 1(s + x > t) d\hat{U}_n(\tilde{N}_n(s), F(x)), \]
\[ X_{n,3}^\infty(t) \equiv n \int_0^t F^c(t - s) d\Lambda(s), \] (1.13)
where \( \tilde{N}_n \) is the CLT-scaled arrival process, \( \hat{U} \) is as in (1.11), and \( \Lambda(s) \) is the limit for the LLN-scaled arrival processes \( \bar{N}_n \), i.e., the fluid arrival process. As has been showed in [54], proving convergence for the first and the third term in (1.15) is straightforward. In particular, the first term, which is \( O(\sqrt{n}) \), converges in distribution to 0 under LLN-scaling whereas the third term, which is \( O(n) \), converges to a nondegenerate limit under LLN-scaling. Moreover, the first term converges to a nondegenerate limit by continuous mapping theorem whereas the third term disappears under FCLT scaling. The second term, which is \( O(\sqrt{n}) \), converges in distribution to 0 under LLN scaling. Proving FCLT for \( X_{n,2}^\infty(t) \) in (1.15), on the other hand, is complicated because it is not possible to directly apply continuous mapping theorem to prove that
\[ \frac{1}{\sqrt{n}} X_{n,2}^\infty(t) = \int_0^t \int_0^\infty 1(s + x > t) d\hat{U}_n(\tilde{N}_n(s), F(x)) \]
\[ \Rightarrow \int_0^t \int_0^\infty 1(s + x > t) d\hat{U}(\Lambda(s), F(x)) \text{ in } D([0, \infty); D([0, \infty); \mathbb{R})) \] (1.14)
\[ n \to \infty, \] where \( \hat{U} \) is the Kiefer process and \( \Lambda \) is the limit for the LLN-scaled arrival processes \( \tilde{N}_n \). The mapping
\[ (\hat{U}_n, \tilde{N}_n) \mapsto \int_0^t \int_0^\infty 1(s + x > t) d\hat{U}_n(\tilde{N}_n(s), F(x)) \]
turns out to be more complicated. The proof in [54] exploits involved martingale arguments, therefore, we omit the details of the proof of convergence in (1.14).

Pang and Whitt [78] later extend the results in [54]. The establish FCLT limits for the two-parameter processes \( X_n^{\infty,c}(t, y) \) and \( X_n^{\infty,r}(t, y) \) respectively defined as the number customers in the system at time \( t \) that have elapsed service times less than or equal to \( y \), and the number customers in the system at time \( t \) that have residual service times strictly greater than \( y \). We
next give the decomposition for $X_n^{\infty,e}(t, y)$. For $t \geq 0$, $0 \leq y \leq t$,

$$X_n^{\infty,e}(t, y) \equiv X_{n,1}^{\infty,e}(t, y) + X_{n,2}^{\infty,e}(t, y) + X_{n,3}^{\infty,e}(t, y),$$

where

$$X_{n,1}^{\infty,e}(t) \equiv \sqrt{n} \int_{t-y}^{t} F^c(t - s) \, d\hat{N}_n(s),$$

$$X_{n,2}^{\infty,e}(t) \equiv \sqrt{n} \int_{t-y}^{t} \int_{0}^{\infty} 1(s + x > t) \, d\hat{U}_n(\hat{N}_n(s), F(x)),$$

$$X_{n,3}^{\infty,e}(t) \equiv n \int_{t-y}^{t} F^c(t - s) \, d\Lambda(s).$$

The major difference between the frameworks of [78] and [54] is the underlying spaces. In particular, the process $X_n^{\infty,e}(t, y)$ in [78] has sample paths in $\mathbb{D}([0, \infty); \mathbb{D}([0, \infty); \mathbb{R}))$, therefore, the authors extend the tightness proof in [54] by proving tightness in $\mathbb{D}([0, \infty); \mathbb{D}([0, \infty); \mathbb{R}))$ instead of $\mathbb{D}([0, \infty); \mathbb{R})$. They also modify the second step of the proof in [54], that is, convergence of finite-dimensional distributions. To summarize, Pang and Whitt [78] prove that

$$\hat{X}_{n,1}^{\infty,e} \Rightarrow (0e), \quad \hat{X}_{n,2}^{\infty,e} \Rightarrow (0e), \quad \hat{X}_{n,3}^{\infty,e}(t) \Rightarrow X_3^{\infty,e}(t) \equiv \int_{t-y}^{t} F^c(t - s) \, d\Lambda(s),$$

$$\hat{X}_{n,1}^{\infty,e}(t) \Rightarrow \hat{X}_{1}^{\infty,e}(t) \equiv \int_{t-y}^{t} F^c(t - s) \, d\hat{N}(s), \quad \hat{X}_{n,2}^{\infty,e}(t) \Rightarrow \hat{X}_{2}^{\infty,e}(t) \equiv \int_{t-y}^{t} \int_{0}^{\infty} 1(s + x > t) \, d\hat{U}(\hat{N}(s, F(x))),$$

$$\hat{X}_{n,3}^{\infty,e} \Rightarrow (0e) \quad \text{in} \quad \mathbb{D}([0, \infty); \mathbb{D}([0, \infty); \mathbb{R})) \quad \text{as} \quad n \to \infty \quad (1.15)$$

Observe that when $y = t$, (1.15) is recovered. Hence the limits here with $y = t$ are the limits for the processes in (1.15). We make use of the FCLT limits established in [78] as well as the ones in [54] in the following chapters.

The limit $\hat{X}_{1}^{\infty,e}$ is understood as the form after integration by parts, i.e., for $t \geq 0$, $0 \leq y \leq t$,

$$X_{n,1}^{\infty,e}(\omega; t, y) \equiv \int_{t-y}^{t} F^c(t - s) \, d\hat{N}(\omega; s)$$

$$\equiv \hat{N}(\omega; t)F^c(0) - F^c(y)\hat{N}(\omega; t - y) - \int_{t-y}^{t} \hat{N}(\omega; s) \, dF(t - s),$$

for almost all $\omega \in \Omega$. The limit $\hat{X}_{2}^{\infty,e}$ is understood as the generalization of Ito’s integral to two-parameter martingales. There are several types of stochastic integrals with respect to two-parameter martingales. We adopt the first type of the stochastic integrals introduced by Cairoli
and Walsh [14]. This type is a direct generalization of the Ito’s integral defined as

\[ I_M(\phi) \equiv \int_{[0,s] \times [0,t]} \phi(u,v) \, dM(u,v) \]  

(1.16)

where \((M(u,v) : (u,v) \in \mathbb{R}_+^2)\) is a two-parameter (strong or weak)-martingale, \(\phi\) is a class of previsible and square integrable processes. See the definitions of strong and weak martingales, and of previsible processes in [14]. The stochastic integral (1.16) has the following properties:

- **Linearity:** \(I_M(\alpha \phi + \beta \psi) = \alpha I_M(\phi) + \beta I_M(\psi), \quad \alpha, \beta \in \mathbb{R},\)
- **Isometry:** \(E[I_M(\phi)I_M(\psi)] = \int_{[0,s] \times [0,t]} \phi(u,v) \psi(u,v) \, dudv,\)
- **Martingale:** \(I_M(\phi)\) is a martingale.

Prior to providing a definition for two-parameter processes, we first review the definition of filtrations used in this thesis for two-parameter processes. Let \((\Omega, \mathcal{F}, P)\) be a probability space and define the order \((s,t) \preceq (s',t')\) if \(s \leq s'\) and \(t \leq t'\). A collection \(\mathcal{F} = \{ \mathcal{F}_{(s,t)} : (s,t) \in [0,T] \times [0,T]\}\) of sub \(\sigma\)-fields of \(\mathcal{F}\) is said to be a filtration if \((i)\) \((s,t) \preceq (s',t')\) implies \(\mathcal{F}_{(s,t)} \subseteq \mathcal{F}_{(s',t')}\); \((ii)\) \(\mathcal{F}_{(0,0)}\) contains all the null sets of \(\mathcal{F}\); \((iii)\) \(\mathcal{F}_{(s,t)} = \bigcap \mathcal{F}_{(u,v)}\) for \(u > s\) and \(v > t\), i.e., right continuous. We are now ready to define two-parameter martingales of interest. A process \((M(s,t) : (s,t) \in [0,T] \times [0,T])\) is a martingale if \(M(s,t) \in \mathcal{F}_{(s,t)}\) is integrable and \(E[M(s',t')|\mathcal{F}_{(s,t)}] = M(s,t)\) for each \((s,t) \preceq (s',t').\)

### 1.9 Acronyms

We summarize in the below table the acronyms used in the following chapters.
Table 1.1: Summary of frequently used acronyms (in alphabetical order).

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.s.</td>
<td>almost sure, or almost surely</td>
</tr>
<tr>
<td>ccdf</td>
<td>complementary cumulative distribution function</td>
</tr>
<tr>
<td>cdf</td>
<td>cumulative distribution function</td>
</tr>
<tr>
<td>CL</td>
<td>critically loaded</td>
</tr>
<tr>
<td>CLT</td>
<td>central limit theorem</td>
</tr>
<tr>
<td>CVE</td>
<td>convolution-type Volterra equation</td>
</tr>
<tr>
<td>EAP</td>
<td>external arrival process</td>
</tr>
<tr>
<td>EAR</td>
<td>external arrival rate</td>
</tr>
<tr>
<td>ESP</td>
<td>enter-service process</td>
</tr>
<tr>
<td>FCFS</td>
<td>first come first served</td>
</tr>
<tr>
<td>FCLT</td>
<td>functional central limit theorem</td>
</tr>
<tr>
<td>fdd</td>
<td>finite-dimensional distribution</td>
</tr>
<tr>
<td>ED</td>
<td>efficiency driven</td>
</tr>
<tr>
<td>FQNet</td>
<td>fluid queue network</td>
</tr>
<tr>
<td>FPE</td>
<td>fixed-point equation</td>
</tr>
<tr>
<td>FWLLN</td>
<td>functional weak law of large numbers</td>
</tr>
<tr>
<td>HOL</td>
<td>head-of-line</td>
</tr>
<tr>
<td>i.i.d.</td>
<td>independent and identically distributed</td>
</tr>
<tr>
<td>ICP</td>
<td>initial content process</td>
</tr>
<tr>
<td>IRP</td>
<td>internal routing process</td>
</tr>
<tr>
<td>IS</td>
<td>infinite-server</td>
</tr>
<tr>
<td>LLN</td>
<td>law of large numbers</td>
</tr>
<tr>
<td>MSHT</td>
<td>many-server heavy-traffic</td>
</tr>
<tr>
<td>NHPP</td>
<td>nonhomogeneous Poisson process</td>
</tr>
<tr>
<td>ODE</td>
<td>ordinary differential equation</td>
</tr>
<tr>
<td>OL</td>
<td>overloaded</td>
</tr>
<tr>
<td>pdf</td>
<td>probability density function</td>
</tr>
<tr>
<td>QD</td>
<td>quality driven</td>
</tr>
<tr>
<td>QED</td>
<td>quality-and-efficiency driven</td>
</tr>
<tr>
<td>QLIFIS</td>
<td>queue length ignoring flow into service</td>
</tr>
<tr>
<td>SCP</td>
<td>service-completion process</td>
</tr>
<tr>
<td>SDE</td>
<td>stochastic differential equation</td>
</tr>
<tr>
<td>SQNet</td>
<td>stochastic queueing network</td>
</tr>
<tr>
<td>TAP</td>
<td>total arrival process</td>
</tr>
<tr>
<td>TAR</td>
<td>total arrival rate</td>
</tr>
<tr>
<td>UL</td>
<td>underloaded</td>
</tr>
<tr>
<td>VWT</td>
<td>virtual waiting time</td>
</tr>
<tr>
<td>w.p.1</td>
<td>with probability 1</td>
</tr>
</tbody>
</table>
Chapter 2

A FWLLN for the Time-Varying
\((G_t/GI/s_t + GI)^m/M_t\) Queueing Network

2.1 Introduction

In this chapter, we study the \((G_t/GI/s_t + GI)^m/M_t\) open queueing network having a finite number of queues (the superscript \(m\)), nonstationary non-Poisson external arrival processes (the \(G_t\)), nonexponential service times (the first \(GI\)), time-varying staffing levels (the \(s_t\)), and customer abandonment following nonexponential patience times (the \(+GI\)). Upon service completion, customers are either routed to one of the queues in the network, or out of the system according to time-dependent probabilities (the \(M_t\)).

This work is a sequel to [62, 64] which are, in turn, extensions of [60, 61]. We aim at extending the FWLLN [62] which provides a mathematical basis for the fluid model developed in [61] for the \(G_t/GI/s_t + GI\) queue. The fluid model in [61] has been later generalized in [64] for the \((G_t/GI/s_t + GI)^m/M_t\) open queueing network, having a finite number of queues, nonstationary non-Poisson external arrival processes, nonexponential service times, time-varying staffing levels, customer abandonment modeled by nonexponential patience times, and probabilistic routing. Following the terminology in [64], we call these time-varying open network fluid queueing models the fluid queueing networks (FQ Nets). Although simulation experiments have confirmed the effectiveness of the FQ Nets proposed in [64], a rigorous MSHT FWLLN supporting such FQ Net approximations has remained an open problem.

We will address this problem by establishing a FWLLN for the \((G_t/GI/s_t + GI)^m/M_t\) open queueing network (see §§2.2–2.3 for the detailed model description and model parameters). In particular, we will show that, as the scale increases, the performance functions such as the num-
ber in each queue, the number of routed customers between the queues, and the head-of-line and virtual waiting time process, of a sequence of the $(G_t/GI/s_t + GI)^m/M_t$ stochastic queueing networks (SQNets) converge in distribution to the associated deterministic performance functions of the corresponding $(G_t/GI/s_t + GI)^m/M_t$ FQNets in [64]. We will establish convergence in product spaces of $\mathbb{D}$ endowed with the appropriate Skorohod topology.

The key step is to establish a FWLLN for the total arrival process (TAP) at each queue. The TAP at each queue is the sum of the external arrival process (EAP), and the total rate of routed customers from the other queues in the network, i.e., internal routing processes (IRPs). The total fluid input rate—hereafter will be called total arrival rate (TAR)—in the associated fluid network has been proved to satisfy a functional fixed-point equation (FPE), see [64] (see (2.10)). We prove that that the prelimit TAP satisfies a stochastic analogue of the FPE (see (2.37)) by establishing the asymptotic equivalence of the two equations as the scale increases. Exploiting the compactness approach [102], we (i) first prove the tightness of the TAPs and (ii) next establish the uniqueness of the limits for all convergent subsequences of TAPs. Once the FWLLN of the TAPs is established, it remains to separately treat each queue of the network by adopting results from [62], which treats the FWLLN of the TAP as an assumption.

A key assumption here is to assume all the queues in the FQNet alternate between OL and UL intervals, or equivalently, the ED and QD regimes, respectively. This is not too restrictive because managers of service systems may not be able or willing to frequently adjust the number of servers when arrivals are time-varying. When the staffing intervals are long, such as 8 hours in hospitals, these systems inevitably experience periods of overloads and underloads [61]. Effective staffing methods have been developed [21, 44, 63, 67, 66, 107] to cope with time-varying arrivals. However, the time-stable performance can be achieved only in systems with flexible staffing. We assume that each queueing system will be critically loaded only at a finite number of regime switching points.

**Organization of the rest of the chapter.** In §2.2, we construct a sequence of $(G_t/GI/s_t + GI)^m/M_t$ SQNets and define the associated performance processes. In §2.3, we review the $(G_t/GI/s_t + GI)^m/M_t$ FQNet proposed in [64]; we specify the model assumptions and describe the system dynamics. In §2.4, we present our main result. In §4.7, we provide the detailed proofs of the main theorem. In §2.6, we provide practical confirmation of the FWLLN by considering an example. Finally, we draw conclusions in §2.7. Additional supplementary materials appear in the appendix. In Appendix A.2–A.5 we provide additional proofs to support §4.7.

### 2.2 A Sequence of $(G_t/GI/s_t + GI)^m/M_t$ Queueing Networks

The $(G_t/GI/s_t + GI)^m/M_t$ SQNet has a finite number of queues in parallel (the superscript $m$). The $i^{th}$ queue, $1 \leq i \leq m$, has a general nonstationary external arrival process (EAP), i.i.d.
service times following a nonexponential cdf $G_i$, a time-dependent staffing function (i.e., number of servers) $s_i(t)$ (the $s_i$), and customer abandonment with i.i.d. nonexponential patience times following cdf $F_i$. The service times, patience times and the EAP are mutually independent.

External arrivals directly enter service if there are available servers; otherwise, they wait in the infinite-capacity waiting line, and will receive service in order of their arrivals if they choose not to abandon.

Right after service is completed at time $t$, a customer will independently be routed either to a queue $j$ ($1 \leq j \leq m$) with a probability $p_{i,j}(t)$ (because the customer needs more service at station $j$), or directly out of the network (because the customer decides to leave the system) with probability $p_{i,0}(t)$. This routing policy is called the time-dependent probabilistic (Markovian) routing (the $M_t$). The probabilistic routing can be useful to model routing uncertainties. In hospitals, for example, patients leaving the intensive care units may be transferred to operating rooms due to sudden health deteriorations or to regular wards due to satisfying recovery.

A standard case of the EAP is the nonhomogeneous Poisson process (NHPP) which is characterized by a rate function. We consider a more general framework by relaxing that NHPP assumption, because statistical analyses show that arrival processes in real-world service systems can be different than the Poisson process [51]. If the EAPs are NHPPs (i.e., the $G_t$ simplifies to $M_t$), and the service-time and abandonment-time distributions are exponential (i.e., the $GI$ and $+GI$ degenerate to $M$ and $+M$), then this SQNet simplifies to the Markovian ($M_t/M/s_t + M$)$^m/M_t$ SQNet studied in [69].

A sequence of SQNets indexed by $n$. Using the $(G_t/GI/s_t + GI)^m/M_t$ SQNet introduced above as a base model, we now construct a sequence of $(G_t/GI/s_t + GI)^m/M_t$ SQNets indexed by $n$ where the scaling factor $n$ represents the size (in terms of arrival rates and number of servers) of the $n^{th}$ SQNet.

We assume that all the service times in a given queueing system have the same service-time cdf $G_i$, and customers have i.i.d. patience times having the cdf $F_i$. The time-dependent routing probabilities at the $i^{th}$ queueing system is $p_{i,j}(t)$. Note that these parameters are independent with the scale $n$. Let $N_n^{(0,i)}$ and $s_n^{(i)}$ be the EAP and the number of servers in the $i^{th}$ queueing system in the $n^{th}$ SQNet, respectively. We assume the following FWLLN for the EAPs and staffing levels.

**Assumption 2.2.1 (FWLLN for EAPs and staffing levels)** For each $i$, $1 \leq i \leq m$, there exists a nondecreasing function $\Lambda_i(t)$ with nonnegative derivative $\lambda_i(t)$, and a piecewise differentiable
function $s_i(t)$ with derivative $\dot{s}_i(t)$, such that

$$
\dot{N}_n^{(0,i)}(t) \equiv n^{-1} N_n^{(0,i)}(t) \Rightarrow \Lambda_i^{(0)}(t) \equiv \int_0^t \lambda_i^{(0)}(u) du,
$$

$$
\dot{s}_n^{(i)}(t) \equiv n^{-1} s_n^{(i)}(t) \Rightarrow s_i(t) \equiv \int_0^t \dot{s}_i(u) du \quad \text{in } \mathbb{D}, \quad \text{as } n \to \infty.
$$

**Remark 2.2.1 (Standard case of Assumption 2.2.1)** Note we do not require the EAP to have a known arrival-rate function for each $n$, but we do require the EAP to have an asymptotic rate function $\lambda^{(0)}$ in the limit as $n \to \infty$. Considering the standard case of NHPPs, we can simply let the EAPs of the $n$th SQNet be Poisson processes with scaled arrival rates $\lambda_n^{(i)} = n\lambda_i$. Standard cases for the staffing function are (i) $s_n^{(i)}(t) = [n \, s_i(t)]$ and (ii) $s_n^{(i)}(t) = [n \, s_i(t) + \beta \sqrt{n} s_i(t)]$, where $\beta$ is a constant. Case (ii) is called the square-root staffing (SRS), see [21, 63, 65, 107] for discussions on SRS. We remark that the $\sqrt{n}$ term will not affect the FWLLN, i.e., the fluid limit, but it may make an impact to the FCLT and diffusion limit [36, 65].

We next define the performance processes. Let $B_n^{(i)}(t,y)$ (resp. $Q_n^{(i)}(t,y)$) be the number of customers in service (resp. in line) at the $i$th station at time $t$ that have been so for time at most $y$. Let $B_n^{(i)}(t) \equiv B_n^{(i)}(t,\infty)$, $Q_n^{(i)}(t) \equiv Q_n^{(i)}(t,\infty)$ and $X_n^{(i)}(t) \equiv B_n^{(i)}(t) + Q_n^{(i)}(t)$ be the number of customers in service, in line, and the total number in $i$th queuing system at time $t$. Let $A_n^{(i)}(t)$, $D_n^{(i)}(t)$ and $E_n^{(i)}(t)$ count the total number of customers that have abandoned, completed service, and entered service by time $t$. Let $R_n^{(i,j)}(t)$ ($1 \leq i, j \leq m$) count the number of customers routed to from the $i$th system to the $j$th system by time $t$, and let $R_n^{(i,0)}(t)$ count the number of customers departed (routed out of the network) from the $i$th system by time $t$. Let $N_n^{(i)}(t)$ be the TAP (i.e., EAP plus IRPs) at the $i$th system by time $t$. Finally, let $W_n^{(i)}(t)$ and $V_n^{(i)}(t)$ denote the head-of-line (HOL) waiting time, that is, the elapsed waiting time for the customer at the head of the waiting line, and the virtual waiting time (VWT), that is the waiting time of an hypothetical arrival until he enters service if he had infinite patience.

In order to establish an FWLLN for these performance processes, we define the law-of-large-numbers-scaled (LLN-scaled) processes: For $t, y \geq 0$,

$$
\begin{align*}
\bar{B}_n^{(i)}(t,y) &\equiv n^{-1} B_n^{(i)}(t,y), & \bar{Q}_n^{(i)}(t,y) &\equiv n^{-1} Q_n^{(i)}(t,y), & \bar{X}_n^{(i)}(t) &\equiv n^{-1} X_n^{(i)}(t), \\
\bar{A}_n^{(i)}(t) &\equiv n^{-1} A_n^{(i)}(t), & \bar{E}_n^{(i)}(t) &\equiv n^{-1} E_n^{(i)}(t), & \bar{D}_n^{(i)}(t) &\equiv n^{-1} D_n^{(i)}(t), \\
\bar{R}_n^{(i,j)}(t) &\equiv n^{-1} R_n^{(i,j)}(t) & \text{and} & \bar{N}_n^{(i)}(t) &\equiv n^{-1} N_n^{(i)}(t), & 1 \leq i \leq m, 0 \leq j \leq m.
\end{align*}
$$

We remark that the waiting times $W_n^{(i)}$ and $V_n^{(i)}$ are not scaled by $n$ because $F_i$ and $G_i$ are not scaled by $n$.

We assume the following FWLLN holds for the initial number of customers in service, and
waiting in line at time 0.

**Assumption 2.2.2 (FWLLN for initial numbers)** There exist nondecreasing functions \((B_i(0, x) : x \geq 0)\) and \((Q_i(0, x) : x \geq 0)\) with nonnegative densities \(b_i(0, x)\) and \(q_i(0, x)\), such that \((s_i(t) - B_i(0, \infty))Q_i(0, \infty) = 0\) for \(1 \leq i \leq m\), and

\[
\begin{align*}
\tilde{B}_n^{(i)}(0, x) &\Rightarrow B_i(0, x) \equiv \int_0^x b_i(0, y) \, dy, \\
\tilde{Q}_n^{(i)}(0, x) &\Rightarrow Q_i(0, x) \equiv \int_0^x q_i(0, y) \, dy \quad \text{in } D \quad \text{as } n \to \infty.
\end{align*}
\]

2.3 The \((G_t/GI/s_t + GI)^m/M_t\) Fluid Network

In this section, we review the \((G_t/GI/s_t + GI)^m/M_t\) FQNet [64]. First, in §2.3.1, we introduce the FQNet system and its parameters. In §2.3.2, we describe the performance of the FQNet in two steps. First, we characterize the performance of each fluid queueing system in the FQNet in §2.3.2 assuming the TAR is a given parameter. Second, we discuss how to compute the \(m\)-dimensional vector of the TAR in §2.3.2.

2.3.1 The FQNet and Its Parameters

The deterministic FQNet is a legitimate dynamical system. There are \(m\) parallel fluid queueing systems in the FQNet. At the \(i\th\) system, \(1 \leq i \leq m\), external fluid arrives with rate \(\lambda_i(0)(t)\). Upon arrival, fluid immediately enters the service with finite service capacity \(s_i(t)\), if there is space available. Otherwise, fluid flows into a waiting space with infinite capacity. The fluid in the waiting space abandons the system; in particular, the total amount of fluid that has abandoned by time \(x\) is determined by the (nondecreasing) function \(F_i(x)\). The remaining amount enters service based on the FCFS discipline. Similarly, the total amount of fluid that has completed service by time \(x\) is determined by the (nondecreasing) function \(G_i(x)\). The proportion of processed (i.e., completed service) fluid routed to the \(j\th\) system at time \(t\) is determined by the function \(P_{i,j}(t)\). The proportion of processed fluid routed out of system at time \(t\) is \(P_{i,0}(t) \equiv 1 - \sum_{j=1}^m P_{i,j}(t)\) provided that \(P_{i,j}(t) > 0\).

Let \(R_{i,j}(t)\) be the total amount of fluid routed from \(i\) to \(j\) by time \(t\) with the associated rate function \(r_{i,j}(t)\). Let \(\Lambda_i(0)(t)\) and \(\Lambda_i(t)\) be the external cumulative fluid input and the total fluid input at the \(i\th\) fluid system, respectively, with external arrival rate (EAR) \(\lambda_i(0)(t)\) and TAR \(\lambda_i(t)\). We have the following traffic-flow equations

\[
\begin{align*}
\Lambda_i(t) &\equiv \Lambda_i(0)(t) + \sum_{k=1}^m R_{k,i}(t) \quad \text{and} \quad \lambda_i(t) \equiv \lambda_i(0)(t) + \sum_{k=1}^m r_{k,i}(t), \quad t \geq 0,
\end{align*}
\]

(2.2)
where

\[
\Lambda_i(t) \equiv \int_0^t \lambda_i(u) du, \quad \Lambda_i^{(0)}(t) \equiv \int_0^t \lambda_i^{(0)}(u) du,
\]
\[
R_{i,j}(t) = \int_0^t r_{i,j}(u) du, \quad r_{i,j}(t) = P_{i,j}(t) \sigma_i(t), \quad t \geq 0,
\]

and \(\sigma_i\) is the service-completion rate at the \(i^{th}\) fluid system (see (2.9)).

Let the two-parameter function \(B(t,y)\) (resp. \(Q(t,y)\)) be the quantity of fluid in service (resp. waiting space) at time \(t\) that has been so for at most \(y\) time units. We assume \(B(t,y)\) and \(Q(t,y)\) have density functions \(b(t,y)\) and \(q(t,y)\), namely,

\[
B(t,y) = \int_0^y b(t,x) dx \quad \text{and} \quad Q(t,y) = \int_0^y q(t,x) dx, \quad t,y \geq 0.
\] (2.3)

Let \(Q(t) \equiv Q(t,\infty)\), \(B(t) \equiv B(t,\infty)\) and \(X(t) \equiv Q(t) + B(t), t \geq 0\). We impose two constraints: (i) \(B(t) \leq s(t)\) (capacity constraint) and (ii) \(Q(t)(B(t) - s(t)) = 0\) (nonidling constraint).

In order to fully characterize the dynamics of the FQNet, we specify the model input \((P,I)\) with

\[
P \equiv \left( m, \lambda_i^{(0)}, s_i, F_i, G_i, P_{i,j}, 1 \leq i,j \leq m \right), \quad I \equiv (b_i(0,\cdot), q_i(0,\cdot), 1 \leq i \leq m),
\] (2.4)

where the six-tuple \(P\) has all model parameters of the FQNet, and the pair \(I\) provides complete information on the initial state of the FQNet. We point out that the TAR \(\lambda_i\) is not part of the model input because it includes the internal routing rates \(r_{i,j}\), which is to be determined. We assume that, for each \(i\), the cdfs \(F_i\) and \(G_i\) have pdfs \(f_i\) and \(g_i\), and hazard-rate functions \(h_{G_i}(x) \equiv g_i(x)/G_i^c(x)\) and \(h_{F_i}(x) \equiv f_i(x)/F_i^c(x)\), where \(G_i^c \equiv 1 - G_i\) and \(F_i^c \equiv 1 - F_i\) are the ccdfs of \(G_i\) and \(F_i\). We assume the service capacity function \(s_i(t)\) is piecewise continuously differentiable, and is feasible such that no fluid is forced out of service when \(s_i(t)\) decreases. See [60, 61] for more discussions and sufficient conditions on the feasibility of the service capacity function.

### 2.3.2 Performance Functions of the FQNet

In this subsection, we provide the performance formulas for the \((G_i/G_s + GI)^m/M_t\) FQNet. Algorithms based on these formulas can be used to compute effective approximations for the corresponding FQNet [60, 64]. In §2.3.2, we describe the performance of the \(i^{th}\) fluid queue as a function of the TAR. In §2.3.2, we characterize the TAR using a multi-dimensional functional FPE.
Performance of the $i$th fluid queue given the TAR $\lambda_i$.

We now provide the performance functions for the $i$th fluid queue with its TAR $\lambda_i$ regarded as a given parameter. For the sake of ease, we drop the subscript $i$ in this subsection. We next describe the OL and UL intervals, and a switching criterion for these intervals.

**OL and UL periods.** A fluid queue is said to be OL at time $t$ if (i) $Q(t) > 0$ or (ii) $Q(t) = 0$, $B(t) = s(t)$ and $\lambda(t) > \dot{s}(t) + \sigma(t)$, where $\dot{s}(t)$ is the derivative of $s(t)$. An OL period ends at time $T_1 \equiv \inf\{u \geq t : Q(u) = 0, \lambda(u) \leq \dot{s}(u) + \sigma(u)\}$. On the other hand, a fluid queue is said to be UL at time $t$ if (i) $B(t) < s(t)$ or (ii) $B(t) = s(t)$, $Q(t) = 0$ and $\lambda(t) \leq \dot{s}(t) + \sigma(t)$. A UL period ends at time $T_2 \equiv \inf\{u \geq t : B(u) = s(u), \lambda(u) > \dot{s}(u) + \sigma(u)\}$. We say the queue is critically loaded (CL) if $Q(t) = 0$, $B(t) = s(t)$ and $\lambda(u) = \dot{s}(u) + \sigma(u)$. Following [60, 61, 62, 64], we make the following assumption.

**Assumption 2.3.1 (Finite number of switches between UL and OL)** In any finite interval $[0, T]$, all the queues in the $(G_t/GI/s_t + GI)^m/M_1$ FQNet switches between OL and UL status for a finite number of times.

See [60, 61] for sufficient conditions. We next characterize the density functions $b(t, x)$ and $q(t, x)$ in (2.3) for UL and OL intervals.

**Performance in a UL interval.** In a UL interval, there is no fluid in the waiting space, and hence, no abandonment from the system. Therefore, we have $q = Q = w = v = 0$. As a result, the $G_t/GI/s_t + GI$ fluid queue is equivalent to the $G_t/GI/\infty$ fluid model with an infinite service capacity. According to Proposition 2 of [61], the service density function is given by

$$b(t, x) = G^c(x)\lambda(t - x)1_{\{x \leq t\}} + \frac{G^c(x)}{G^c(x - t)}b(0, x - t)1_{\{x > t\}}, \quad t, x \geq 0. \tag{2.5}$$

**Performance in an OL interval.** In an OL interval, the service density function

$$b(t, x) = b(t - x, 0)G^c(x)1_{\{x \leq t\}} + \frac{G^c(x)}{G^c(x - t)}b(0, x - t)1_{\{x > t\}}, \quad t, x \geq 0, \tag{2.6}$$

where the initial service density function $b(0, y)$ is part of the initial state descriptor $\mathcal{I}$ in (2.4), and the rate fluid enters service $b(t, 0)$ uniquely solves the FPE

$$b(t, 0) = \dot{a}(t) + \int_0^t b(t - x, 0)g(x)dx \quad \text{with} \quad \dot{a}(t) \equiv \dot{s}(t) + \int_0^\infty \frac{b(0, y)g(t + y)}{G^c(y)}dy, \tag{2.7}$$

for $t \geq 0$. See Theorem 2 in [61] for more details on the FPE (2.7).

We next determine the density function for queue content $q(t, x)$ during an OL interval. Let $w(t)$ and $v(t)$ be the head-of-line waiting time and potential waiting time at time $t$. According
to Corollary 3 of [61], the density function associated with the queue content is given by

\[ q(t, x) \equiv \lambda(t-x, 0)F^c(x)1_{\{x\leq w(t)\}} + q(0, x-t)\frac{F^c(x)}{F^c(x-t)}1_{\{x> w(t)\}} \quad (2.8) \]

where the initial quantity \( q(0, x) \) is part of the initial state descriptor \( I \) in (2.4), and \( w(t) \) uniquely solves the following ordinary differential equations (ODE)

\[ \dot{w}(t) = 1 - \frac{b(t, 0)}{q(t, w(t))}, \quad t \geq 0, \]

where \( b(t, 0) \) satisfies (2.7) and \( q(t, x) \) is given in (2.8). The fluid virtual waiting time \( v(t) \) uniquely satisfies

\[ v(t - w(t)) = w(t), \quad \text{or equivalently,} \quad v(t) = w(t + v(t)), \quad t \geq 0. \]

See Theorems 3 and 5 in [61] for details.

**Fluid flows.** Let \( A(t) \), \( D(t) \) and \( E(t) \) be the total amount of fluid that has abandoned, completed service, and entered service by time \( t \), respectively, with rates \( \alpha(t) \), \( \sigma(t) \) and \( b(t, 0) \).

Then

\[ A(t) \equiv \int_0^t \alpha(u)du, \quad \alpha(t) \equiv \int_0^\infty q(t, x)h_F(x)dx, \]
\[ D(t) \equiv \int_0^t \sigma(u)du, \quad \sigma(t) \equiv \int_0^\infty b(t, x)h_G(x)dx, \]
\[ E(t) \equiv \int_0^t b(u, 0)du, \quad t \geq 0, \quad (2.9) \]

where \( q(t, x) \) and \( b(t, x) \) satisfy (2.5),(2.6) and (2.8), and \( b(t, 0) \) solves (2.7).

**Characterizing the TAR for the FQNet.**

We now characterize the vector of TAR functions associated with each fluid queue in the FQNet.

Consider an interval \([0, \tau] \) during which no fluid queue changes regime. Let \( U(t) \equiv \{1 \leq i \leq m : B_i(t) \leq s_i(t), \ Q_i(t) = 0\} \) and \( O(t) \equiv \{1 \leq i \leq m : B_i(t) = s_i(t), \ Q_i(t) > 0\} \) be the sets of the indices of UL and OL queues in the FQNet. Note that the indices do not change with time, i.e., the sets \( U \equiv U(t) \) and \( O \equiv O(t) \), in the interval \([0, \tau] \).

The TAR vector \( \lambda \equiv (\lambda_1, \ldots, \lambda_m) \) satisfies the multi-dimensional FPE

\[ \lambda = \Psi(\lambda), \quad (2.10) \]
where for \( u \equiv (u_1, \ldots, u_m) \in \mathbb{D}^m \), the operator \( \Psi : \mathbb{D}^m \to \mathbb{D}^m \) is defined as

\[
\Psi(u)_i(t) \equiv \gamma_i(t) + \sum_{i \in \mathcal{U}} P_{i,j}(t) \int_0^t g_i(x) u_i(t-x) dx, \quad t \geq 0
\]

where \( \gamma_i(t) \equiv \lambda_i(0) + \sum_{k \in \mathcal{O}} P_{k,i}(t) \sigma_k(t) + \sum_{j \in \mathcal{U}} P_{j,i}(t) \int_0^\infty b_j(0,x) g_j(t+x) \frac{G_j(x)}{G_c(x)} dx. \)

for each \( 1 \leq i \leq m \). By Theorem 1 of [64], \( \Psi \) is a contraction operator in \( \mathbb{D}^m \), so that (2.10) has a unique solution \( \lambda \) in the interval \( [0, \tau] \).

### 2.4 FWLLN for the \((G_t/GI/s_t + GI)^m/M_t\) SQNet

In this section, we present an FWLLN of \((G_t/GI/s_t + GI)^m/M_t\) SQNet. We show that the performance processes associated with the sequence of SQNet defined in §2.2 converge in distribution to the deterministic performance functions associated with the \((G_t/GI/s_t + GI)^m/M_t\) FQNet reviewed in §2.3, as the scale increases. The limits for the one-parameter stochastic processes \( \bar{N}_n, \bar{D}_n, \bar{E}_n, \bar{A}_n, \bar{R}_n, \bar{X}_n, W_n, \) and \( V_n \) are established in the product space of \( \mathbb{D} \) whereas the limits for the two-parameter processes \( \bar{Q}_n \) and \( \bar{B}_n \) are established in the product space of \( \mathbb{D} \).

Using bold face symbols to denote vectors, we define the vectors of the prelimit LLN-scaled processes and the associated fluid functions as

\[
\bar{N}_n \equiv \left( \bar{N}_n^{(0,1)}, \ldots, \bar{N}_n^{(0,m)} \right), \quad \Lambda \equiv (\Lambda_1, \ldots, \Lambda_m),
\]

\[
\bar{R}_n \equiv \left( \bar{R}_n^{(i,j)}, 1 \leq i \leq m, 0 \leq j \leq m \right), \quad R \equiv (R_{i,j}, 1 \leq i \leq m, 0 \leq j \leq m),
\]

\[
\bar{Q}_n \equiv \left( \bar{Q}_n^{(1)}, \ldots, \bar{Q}_n^{(m)} \right), \quad Q \equiv (Q_1, \ldots, Q_m),
\]

and the other prelimit processes

\[
\left( \bar{N}_n^{(0)}, \bar{s}_n, \bar{Q}_n(0, \cdot), \bar{D}_n(0, \cdot), \bar{E}_n, \bar{A}_n, \bar{X}_n, W_n, V_n, \bar{Q}_n \right)
\]

and fluid functions

\[
\left( \Lambda^{(0)}, s, Q(0, \cdot), B(0, \cdot), D, E, A, X, W, V, Q \right)
\]

defined as analogs of (2.12). We are now ready to state our main result.

**Theorem 2.4.1** (FWLLN for the \((G_t/GI/s_t + GI)^m/M_t\) queueing network)

Let \([0, \tau]\) be an interval with switching points only at the endpoints, i.e., none of the queueing systems switches regime from one to the other in the interval \((0, T)\). If Assumptions 2.2.1–2.3.1
hold, then the following FWLLN holds for the \((G_t/GI/s_t + GI)^m/M_t\) queueing network,

\[
\begin{align*}
\left( \bar{N}_n(0), \bar{s}_n, \bar{Q}_n(0, \cdot), \bar{B}_n(0, \cdot), \bar{N}_n, \bar{D}_n, \bar{E}_n, \bar{X}_n, \bar{W}_n, \bar{V}_n, \bar{R}_n, \bar{Q}_n, \bar{B}_n \right) \\
\Rightarrow \left( \Lambda^{(0)}, s, Q(0, \cdot), B(0, \cdot), \Lambda, D, E, A, X, W, V, R, Q, B \right)
\end{align*}
\]

(2.13)

in \(D^{m^2+12m}([0, \tau]; \mathbb{R}) \times D^m([0, \tau]; \mathbb{R})\) as \(n \to \infty\) where the vectors of the prelimit processes are defined in (2.12) and §2.2, and the vector of deterministic fluid limit is defined in (2.12) and §2.3.

**Remark 2.4.1 (Useful engineering approximations from the FWLLN)**

Theorem 2.4.1 provides mathematical justification for the fluid approximations in [60, 64]. Simulation experiments [61, 64] show that the sample paths of the LLN-scaled performance processes agree closely with their deterministic fluid counterparts when \(n\) is large (e.g., \(n = 1000\)). When \(n\) is not large, stochastic fluctuations become significant, but the mean values of the LLN-scaled performance functions remain well approximated by the fluid functions with smaller \(n\) (e.g., \(n = 10\)).

**Remark 2.4.2 (Special case without abandonment)**

The \((G_t/GI/s_t)^m/M_t\) SQNet (FQNet) without customer abandonment can be viewed a special case of the \((G_t/GI/s_t + GI)^m/M_t\) SQNet (FQNet). Queueing models without customer abandonment are important because many service systems indeed have no abandonment or very low abandonment (e.g., health care systems and airport security lines). Moreover, the addition (removal) of the element of customer abandonment can significantly alter the system performance [108]. We remark that the proof of the FWLLN of the \((G_t/GI/s_t)^m/M_t\) SQNet is similar. In order to obtain the performance functions of the \((G_t/GI/s_t)^m/M_t\) FQNet, it suffices to let \(f_{F_i}(x) = F_i(x) = 0\) and \(F_i^*(x) = 1\) for \(x \geq 0\) in all performance formulas in §2.3.

**Remark 2.4.3 (Joint convergence in (2.13) and an arbitrary interval)**

According to Theorem 11.4.5 of [102], the joint convergence in (2.13) is equivalent to the marginal convergence of each component, because the FWLLN limits are all deterministic functions. Hence, in §4.7, we will prove Theorem 2.4.1 by establishing the weak convergence of each component of the performance processes in space \(D\) or \(D_D\). The weak convergence in Theorem 2.4.1 is equivalent to uniform convergence over the finite interval \([0, T]\), because the deterministic limits are continuous functions. Our proof strategy in §4.7 is to partition the interval \([0, T]\) into a sequence of disjoint intervals separated by a finite number of time points \(0 = t_0 < t_1 < t_2 < \cdots < t_N = T\), such that no fluid queue in the FQNet switches regime in each interval \([t_{i-1}, t_i]\). Therefore, it suffices to prove the FWLLN in Theorem 2.4.1 by focusing
on an interval \([0, \tau]\), where all initially OL (resp. UL) queues remain OL (resp. UL) throughout the interval.

### 2.5 Proof of the Main Result

**Outline of the proof.** We prove the weak convergence in Theorem 2.4.1 following the compactness approach \([10, 102, 78]\). In particular, we first show that the prelimit processes (indexed by \(n\)) are tight (see \([102]\) for definition and conditions for tightness), which implies that every subsequence has a further convergent subsequence. We next establish the full convergence by showing all convergent subsequences converge to the same limit (i.e. having a common probability law).

According to Remark 2.4.3, we will consider an interval \([0, T]\) where no queue changes its OL (UL) status (i.e., the queues that are OL (UL) at time 0 stays OL (UL) throughout the interval \([0, T]\)). We establish the FWLLN for each component of (2.13) in the following order: First, we show the FWLLN for all service-related processes of OL queues in §2.5.1, including the service-completion process (SCP) \(\overline{D}_{n}^{(i)}\), enter-service process (ESP) \(\overline{E}_{n}^{(i)}\), internal routing process (IRP) \(\overline{R}_{n}^{(i,j)}\) and the two-parameter service content \(\overline{B}_{n}^{(i)}\), for \(i \in O\) and \(1 \leq j \leq m\).

Using the FWLLN of the IRPs from OL queues, in §2.5.2 we next establish the FWLLNs of the TAP \(\overline{N}_{n}^{(j)}\) and IRP \(\overline{R}_{n}^{(i,j)}\), for \(i \in U\) and \(1 \leq j \leq m\). Finally, in §2.5.3 we apply the FWLLN of TAP to develop the FWLLNs of all other processes, including the service-related processes \(\overline{B}_{n}^{(i)}\) and \(\overline{D}_{n}^{(i)}\) for \(i \in U\), and the queue-related processes \(W_{n}^{(j)}\), \(V_{n}^{(j)}\), \(\overline{Q}_{n}^{(j)}\) and \(\overline{A}_{n}^{(j)}\) \(j \in O\). See Figure 2.1 for an illustration.

**Asymptotically UL and OL intervals.** Suppose queue \(i\) of the FQNet is OL in \([0, T]\) (the argument for the UL case is similar), with the net input rate \(\lambda_{i}(0) - \sigma_{i}(0) > 0\) and \(\Lambda_{i}(t) > D_{i}(t)\) for \(0 < t < T\). Because \(\overline{Q}_{n}^{(i)}(0) \Rightarrow Q_{i}(0) > 0\), queue \(i\) of the SQNet will become asymptotically OL (that is, all servers will asymptotically become busy and remain so throughout an interval \([t_{1,n}, t_{2}]\) with \(0 < t_{1,n} = o(1/n) < t_{2}\), even though some servers could be idle in the neighborhood of 0. We next construct the performance functions for the asymptotically OL queues (with index

![Figure 2.1: Proof strategy for the FWLLN in Theorem 1.](image-url)
i \in \mathcal{O}) \text{ and UL queues (with } i \in \mathcal{U}). \text{ Let } |\mathcal{O}| \text{ and } |\mathcal{U}| \text{ be the numbers of OL and UL queues (i.e., the numbers of indices in sets } \mathcal{O} \text{ and } \mathcal{U}).

### 2.5.1 FWLLN for service related processes at OL queues

In this subsection, we establish the FWLLN for the service related processes, including the SCP, ESP, IRP and service-content processes. In particular, we now show, for \( i \in \mathcal{O}, 1 \leq j \leq m, \)

\[
(\bar{D}_n^{(i)}, \bar{E}_n^{(i)}, \bar{R}_n^{(i,j)}, \bar{B}_n^{(i)}) \Rightarrow (D_i, E_i, R_{i,j}, B_i) \quad \text{in } \mathbb{D}^3 \times \mathbb{D}, \quad \text{as } n \to \infty. \tag{2.14}
\]

We are able to prove (2.14) before establishing the FWLLN of the TAP because the service-related processes of the OL queues do not directly depend on the TAP. Instead, they depend on the number of existing customers in service (i.e., old service content), the total number of servers, and how fast the servers become available to serve customers at the head of the waiting line.

**Service-completion process \( \bar{D}_n^{(i)} \) and enter-service process \( \bar{E}_n^{(i)} \).**

Our proof in this subsection draws heavily on the results in [62]. We provide the major steps here so the chapter is self contained. But we refer to [62] for some detailed proofs to avoid repetition. We also provide the omitted details in the appendix.

Flow conservation of the service content (i.e., number of customers in service) at an OL queue \( i \) implies that

\[
\bar{E}_n^{(i)}(t) = (s_n^{(i)}(t) - s_n^{(i)}(0)) + \bar{D}_n^{(i)}(t) \quad \text{for all } i \in \mathcal{O}. \tag{2.15}
\]

Hence, by Assumption 2.2.1, the tightness and weak convergence of \( \bar{D}_n^{(i)} \) will easily imply the tightness and weak convergence of \( \bar{E}_n^{(i)} \). We give the tightness results in the next lemma.

**Lemma 2.5.1** The sequence of processes \( (\bar{D}_n^{(i)}, \bar{E}_n^{(i)}, i \in \mathcal{O}) \) is C-tight in \( \mathbb{D}^{|\mathcal{O}|} \).

**Proof 2.5.1** By Theorem 11.6.7 of [102], the tightness of the big vector in Lemma 2.5.1 is equivalent to the tightness of the components \( \bar{D}_n^{(i)} \) and \( \bar{E}_n^{(i)} \), for all \( i \in \mathcal{O} \). The proof of the C-tightness of the \( \bar{D}_n^{(i)} \) closely follows from the proof in §3 of [62], so we give the proof in Appendix A.2. Given the C-tightness of \( \bar{D}_n^{(i)} \), the C-tightness of \( \bar{E}_n^{(i)} \) follows from Assumption 2.2.1, the smoothness of the limiting staffing function \( s_i(t) \) and the continuous mapping theorem with addition.

We next characterize the limit of a convergent subsequence of \( \bar{D}_n^{(i)} \). We first split the SCP and service content into two terms, corresponding to old customers (initially in service at time...
0) and \textit{new customers} (arriving after time 0). In particular, we write
\[
\bar{D}^{(i)}_n(t) = \bar{D}^{(i,o)}_n(t) + \bar{D}^{(i,\nu)}_n(t) \quad \text{and} \quad \bar{B}^{(i)}_n(t, x) = \bar{B}^{(i,o)}_n(t, x) + \bar{B}^{(i,\nu)}_n(t, x)
\] (2.16)
where \(\bar{D}^{(i,o)}_n(t) (\bar{B}^{(i,o)}_n(t, x))\) denotes the LLN-scaled number of service completions by time \(t\) (customers in service at \(t\) with ages no more than \(x\)) from those already in service at time 0, and \(\bar{D}^{(i,\nu)}_n(t) (\bar{B}^{(i,\nu)}_n(t, x))\) denotes the LLN-scaled number of service completions by \(t\) (customers in service at \(t\) with ages no more than \(x\)) from the new arrivals in the interval \([0, t]\).

To treat the first term of the SCP in (2.16), we follow [62] by writing
\[
\bar{D}^{(i,o)}_n(t) = \bar{B}^{(i)}_n(0) - \bar{B}^{(i,o)}_n(t) \quad \text{and} \quad \bar{B}^{(i,o)}_n(t) = \frac{1}{n} \sum_{k=1}^{\infty} 1 \left( \eta_{k,i}(\tau^{(k)}_{n,i}) > t \right),
\] (2.17)
where \(\left\{ 0 < \tau^{(1)}_{n,i} \leq \tau^{(2)}_{n,i} \leq \ldots \right\}\) is an ordered sequence of the ages (i.e., elapsed service times) of the old customers (i.e., those in service at time 0), and \(\{\eta_{1,i}(x), \eta_{2,i}(x), \ldots\}\) is an i.i.d. sequence of random variables following ccdf
\[
P(\eta_{1,i}(x) > t) = \frac{G^c_i(t + x)}{G^c_i(x)}, \quad x \geq 0.
\] (2.18)
By Theorem 2 of [62] and Assumption 2.2.2, we have, as \(n \to \infty\),
\[
\bar{B}^{(i,o)}_n(t) \Rightarrow B^{(i,o)}(t) \equiv \int_0^\infty \frac{b_i(0, x)}{G^c_i(x)} \, dx \quad \text{in } \mathcal{D},
\] (2.19)
which, together with (2.17) and Assumption 2.2.2, concludes the FWLLN of \(\bar{D}^{(i,o)}_n\). Namely, as \(n \to \infty\), for \(i \in \mathcal{O}\),
\[
\bar{D}^{(i,o)}_n(t) \Rightarrow D^{(i,o)}(t) \equiv B^{(i,o)}(0) - B^{(i,o)}(t) = \int_0^\infty b_i(0, x) \left( 1 - \frac{G^c_i(t + x)}{G^c_i(x)} \right) dx.
\] (2.20)
To treat the second term of the SCP in (2.16), we write
\[
\bar{D}^{(i,\nu)}_n(t) = n^{-1} \sum_{k=1}^{E^{(i)}_n(t)} 1 \left( S^{(i,n)}_k + S^{(i)}_k \right),
\] (2.21)
where \(S^{(i,n)}_k\) denotes the time the \(k\)th customer enters service, and \(\{S^{(i)}_1, S^{(i)}_2, \ldots\}\) are the i.i.d.
service times following the cdf $G_i$. By (2.15), (2.16), (2.17) and (2.21), we have

$$
\overline{E}_n^{(i)}(t) = (\overline{s}_n^{(i)}(t) - \overline{s}_n^{(i)}(0)) + \frac{1}{n} \sum_{k=1}^{n B_n^{(i)}(0)} \left( \eta_{k,i}(\tau_{n,i}^{(k)}) \leq t \right)
+ \frac{1}{n} \sum_{k=1}^{n \overline{E}_n^{(i)}(t)} \left( \mathcal{E}_k^{(i,n)} + \mathcal{S}_k^{(i)} \leq t \right).
$$

(2.22)

Following the proofs in §6 of [62] (the details omitted here and provided in Appendix A.3), we have the convergence

$$
\overline{D}_n^{(i,\nu)}(t) \Rightarrow D^{(i,\nu)}(t) \equiv \int_0^t G_i(t-s)b_i(s,0)ds \quad \text{and}
$$

(2.23)

$$
\overline{E}_n^{(i)}(t) \Rightarrow E^{(i)}(t) \equiv \int_0^t b_i(s,0)ds \quad \text{in} \quad \mathbb{D},
$$

(2.24)

where $b(0, \cdot)$ solves the FPE (2.7). The full convergence of $\{\overline{D}_n^{(i)}\}$ and $\{\overline{E}_n^{(i)}\}$ immediately follows from Lemma 2.5.1, (2.16), (2.20), (2.23), (2.24) and the continuous mapping theorem with addition.

**Two-parameter service content $\overline{B}_n^{(i)}$.**

Extending the sums in (2.17) and (2.21), we give the two-parameter representations for the LLN-scaled new service content $\overline{B}_n^{(i,\nu)}$ and old service content $\overline{B}_n^{(i,o)}$:

$$
\overline{B}_n^{(i,\nu)}(t,y) = \frac{1}{n} \sum_{k=\mathcal{E}_n^{(i)}((t-y)^+)+1}^{\mathcal{E}_n^{(i)}(t)} \left( \mathcal{E}_k^{(i,n)} + \mathcal{S}_k^{(i)} > t \right),
$$

(2.25)

$$
\overline{B}_n^{(i,o)}(t,y) = \frac{1}{n} \sum_{k=1}^{B_n^{(i)}(0,(y-t)^+)} \left( \eta_{k,i}(\tau_{n,i}^{(k)}) > t \right).
$$

(2.26)

The FWLLN of (2.25) follows from (2.24) and Theorem 3.1 of [78]. In particular,

$$
\overline{B}_n^{(i,\nu)}(t,y) \Rightarrow B^{(i,\nu)}(t,y) \equiv \int_{(t-y)^+}^t G_i^c(t-s)dE_i(s) \quad \text{in} \quad \mathbb{D}_{\mathbb{D}}.
$$

(2.27)

The FWLLN of (2.26) follows from Assumption 2.2.2 and Theorem 2 of [62]. We have

$$
\overline{B}_n^{(i,o)}(t,y) \Rightarrow B^{(i,o)}(t,y) \equiv \int_0^{(y-t)^+} b_i(0,x) \frac{G_i^c(t+x)}{G_i^c(x)} dx \quad \text{in} \quad \mathbb{D}_{\mathbb{D}}.
$$

(2.28)
Together with (2.16), (2.27) and (2.28), we apply the continuous mapping theorem with addition to obtain the FWLLN of the two-parameter service content $\bar{B}^{(i)}_n$ with the limit $B_i$ given in (2.3) and (2.6).

**Remark 2.5.1 (FWLLN for processes related to old content in service)**

Although the FWLLN for the processes related to the old service content $\bar{B}^{(i,o)}_n$ and $\bar{D}^{(i,o)}_n$ are developed here for an OL queue $i$ (i.e., $i \in \mathcal{O}$), the same arguments (of the FWLLN and fluid limit) hold for an UL queue $i$ (i.e., $i \in \mathcal{U}$). Because we assume no customer is forced out of service before completing service when the staffing level decreases (if ever), the dynamics of the old customers in service (those already in service at time 0) does not depend on if the queue is OL or UL; namely, their behavior is not affected by the ESP $E^{(i)}_n$ or the number of servers $s^{(i)}_n$, because they will continue to occupy the servers until their services are completed (in some sense they have higher priorities comparing with new customers).

However, at an OL queue, the performance of processes related to new content (e.g., $\bar{B}^{(i,\nu)}_n$ and $\bar{D}^{(i,\nu)}_n$) are precisely controlled by the amount of available service resources (here represented by $s^{(i)}_n(t) - B^{(i,o)}_n(t)$), which determines how often new customers should enter service (reflected by $E^{(i)}_n(t)$). On the contrary, the dynamics is very different at a UL queue where there is almost no constraint on the service capacity (because a UL queue is equivalent to an infinite-server queue). Therefore, in §2.5.3 we will only have to establish the FWLLNs for $\bar{D}^{(i,o)}_n$ and $\bar{B}^{(i,o)}_n$ of a UL queue $i$, because the proofs of the FWLLNs for $\bar{D}^{(i,o)}_n$ and $\bar{B}^{(i,o)}_n$ are identical to those for an OL queue with limits in the same forms as in (2.20) and (2.28).

**Internal routing flows $\bar{R}^{(i,j)}_n$ from OL queues.**

We next establish the FWLLN for the IRP $R^{(i,j)}_n$, from an OL queue $i$ (i.e., $i \in \mathcal{O}$) to another queue $j$ ($0 \leq j \leq m$), with $j = 0$ denoting the outside world (i.e., leaving the network). First we provide the representation for the routing process using independent indicators splitting the SCP. For $s \geq 0$, let $\left\{ \delta^{(1)}_{i,j}(s), \delta^{(2)}_{i,j}(s), \ldots \right\}$ and $\left\{ \tilde{\delta}^{(1)}_{i,j}(s), \tilde{\delta}^{(2)}_{i,j}(s), \ldots \right\}$ be two independent i.i.d. sequences of indicator random variables with $P(\delta^{(1)}_{i,j}(s) = 1) = P(\tilde{\delta}^{(1)}_{i,j}(s) = 1) = 1 - P(\delta^{(1)}_{i,j}(s) = 0) = 1 - P(\tilde{\delta}^{(1)}_{i,j}(s) = 0) = P_{i,j}(s)$. We write

$$\bar{R}^{(i,j)}_n(t) = \bar{R}^{(i,j,o)}_n(t) + \bar{R}^{(i,j,\nu)}_n(t), \quad (2.29)$$

where

$$\bar{R}^{(i,j,o)}_n(t) = \frac{1}{n} \sum_{k=1}^{D^{(i,o)}_n(t)} \delta^{(k)}_{i,j} \left( \eta_{k,i}(\tau^{(k)}_{n,i}) \right) \quad \text{and} \quad \bar{R}^{(i,j,\nu)}_n(t) = \frac{1}{n} \sum_{i=1}^{D^{(i,\nu)}_n(t)} \tilde{\delta}^{(i)}_{i,j} \left( \zeta^{(i)}_{n,i} \right) \quad (2.30)$$

37
denote the routing flows from old customers (those already in service at time 0) and new customers in service (from new arrivals in \([0, t]\)), \(s^{(l)}_{n, i} = \mathcal{E}^{(i, n)}_l + \mathcal{S}^{(i)}_l\) is the service-completion time of the \(l\)th new customer, with \(\mathcal{E}^{(i, n)}_l\) and \(\mathcal{S}^{(i)}_l\) defined in (2.21), and \(\eta_{k,i}(\tau^{(k)}_{n,i})\) is the service-completion time of an old customer having the \(k\)th smallest elapsed service time (age), defined in (2.17).

We obtain the FWLLN of the IRP using the continuous mapping theorem; in particular we express \(\bar{R}^{(i, j, o)}_n\) as a function of the SCP \(\bar{D}^{(i, o)}_n\). Adding and subtracting \(P_{i,j}(\eta^{(k)}_{n,i})\) in the first equation and \(P_{i,j}(\zeta^{(l)}_{n,i})\) in the second equation of (2.30) yields

\[
\bar{R}^{(i, j, o)}_n(t) = \frac{1}{n} \sum_{k=1}^{D^{(i, o)}_n(t)} \left[ \delta^{(k)}_{i,j} \left( \eta_{k,i}(\tau^{(k)}_{n,i}) \right) - P_{i,j} \left( \eta_{k,i}(\tau^{(k)}_{n,i}) \right) \right] + \frac{1}{n} \sum_{k=1}^{D^{(i, o)}_n(t)} P_{i,j} \left( \eta_{k,i}(\tau^{(k)}_{n,i}) \right)
\]

\[
= \frac{1}{n} \sum_{k=1}^{D^{(i, o)}_n(t)} \left[ \delta^{(k)}_{i,j} \left( \eta_{k,i}(\tau^{(k)}_{n,i}) \right) - P_{i,j} \left( \eta_{k,i}(\tau^{(k)}_{n,i}) \right) \right] + \int_0^t P_{i,j}(u)d\bar{D}^{(i, o)}_n(u), \tag{2.31}
\]

and

\[
\bar{R}^{(i, j, \nu)}_n(t) = \frac{1}{n} \sum_{l=1}^{D^{(i, \nu)}_n(t)} \left[ \delta^{(l)}_{i,j} \left( \zeta^{(l)}_{n,i} \right) - P_{i,j} \left( \zeta^{(l)}_{n,i} \right) \right] + \frac{1}{n} \sum_{l=1}^{D^{(i, \nu)}_n(t)} P_{i,j} \left( \zeta^{(l)}_{n,i} \right)
\]

\[
= \frac{1}{n} \sum_{l=1}^{D^{(i, \nu)}_n(t)} \left[ \delta^{(l)}_{i,j} \left( \zeta^{(l)}_{n,i} \right) - P_{i,j} \left( \zeta^{(l)}_{n,i} \right) \right] + \int_0^t P_{i,j}(u)d\bar{D}^{(i, \nu)}_n(u). \tag{2.32}
\]

**Convergence of the second terms in (2.31) and (2.32).** We next show that

\[
\int_0^t P_{i,j}(u)d\bar{D}^{(i, o)}_n(u) \Rightarrow \int_0^t P_{i,j}(u)d\bar{D}^{(i, o)}(u), \tag{2.33}
\]

\[
\int_0^t P_{i,j}(u)d\bar{D}^{(i, \nu)}_n(u) \Rightarrow \int_0^t P_{i,j}(u)d\bar{D}^{(i, \nu)}(u) \quad \text{in} \quad \mathbb{D}, \quad \text{as} \quad n \to \infty. \tag{2.34}
\]

We only show (2.33) because (2.34) is similar. We apply the continuous mapping theorem based on the next lemma, with its proof given in Appendix A.5.

**Lemma 2.5.2** For \(x \in \mathbb{D}\), the function \(\phi : \mathbb{D} \to \mathbb{D}\) defined as

\[
(\phi(x))(t) \equiv P_{i,j}(t)x(t) - \int_0^t x(s)dP_{i,j}(s)
\]

is continuous, if the \(P_{i,j}(t)\) is piecewisely differentiable.
Since $\tilde{D}_n^{(i,\nu)}(\omega,t)$ is nondecreasing in $t$ for almost all $\omega \in \Omega$, with $\tilde{D}_n^{(i,\nu)}(0) = 0$ satisfying $E[\tilde{D}_n^{(i,\nu)}(t)] < \infty$ for all $t \in [0,\infty)$, $\tilde{D}_n^{(i,\nu)}(\omega,t)$ is of bounded variation for each $n \geq 1$. Therefore, combined with the fact that $\tilde{D}_n^{(i,\nu)}(t)$ is right continuous with left limits for almost all $\omega \in \Omega$, the second term in (2.32) is a Stieltjes integral for fixed $\omega$. Then, by integration by parts,

$$\int_0^t P_{i,j}(s)d\tilde{D}_n^{(i,\nu)}(\omega,s) = P_{i,j}(t)\tilde{D}_n^{(i,\nu)}(\omega,t) - \int_0^t \tilde{D}_n^{(i,\nu)}(\omega,s)dP_{i,j}(s)$$

for all $n \geq 1$. Hence, by Lemma 2.5.2 and the FWLLN of $\tilde{D}_n^{(i,\nu)}$ in (2.23), we conclude the convergence in (2.33).

Asymptotic negligibility of the first terms in (2.31) and (2.32). We now complete the proof of the FWLLN of $\tilde{R}_n^{(i,j)}$ by showing the first terms of (2.31) and (2.32) are asymptotically negligible. Because the proofs are similar, we only show the latter.

We first condition on a realization of the sequence $\{\zeta^{(l)}_{n,i}, l \geq 1\}$. For a fixed $t$, the first term of (2.32) is the scaled random sum of independent zero-mean random variables, each taking values in the interval $[-1,1]$. Because the random variables are not identically distributed, we apply the law of large numbers for non-identically-distributed triangular arrays, see Theorem 1 on p.307 of [22], also see Appendix A.4. As a result, for fixed $t \geq 0$, $i \in \mathcal{O}$ and $1 \leq j \leq m$, we have

$$\hat{R}_n^{(i,j,\nu)}(t) = \frac{1}{n} \sum_{l=1}^n \left( \tilde{D}_n^{(i,\nu)}(\zeta^{(l)}_{n,i}) - P_{i,j}(\zeta^{(l)}_{n,i}) \right)$$

in $\mathbb{R}$, where the sum in the second equation converges in distribution to 0 by (A.7) and $\tilde{D}_n^{(i,\nu)}$ converges to $D^{(i,\nu)}$. The convergence in (2.35) for a fixed $t$ can then easily extend to uniform convergence over compact sets according to Theorem 3.2.1 in the internet supplement of [102].

Repeating the same argument for $\tilde{R}_n^{(i,j,\alpha)}$ and apply the continuous mapping theorem using addition, we complete the proof of the FWLLN of (2.29), namely,

$$\tilde{R}_n^{(i,j)}(t) \Rightarrow \int_0^t P_{i,j}(u)dD^{(i,\nu)}(u) + \int_0^t P_{i,j}(u)dD^{(i,\alpha)}(u) \quad i \in \mathcal{O}, \quad 1 \leq j \leq m. \quad (2.36)$$

It is easy to see that the right-hand side of (2.36) agrees with $\tilde{R}_{i,j}(t)$ in (2.2), by combining (2.20) and (2.23).
2.5.2 FWLLN for the Total Arrival Process

We now prove the FWLLN of the TAP $\bar{N}_n$. First, we construct equations describing the preliminaries of the TAP. We next prove the full convergence of $\bar{N}_n$ following the compactness approach \[102\], by establishing (i) the tightness of the TAP and (ii) showing the limit for all convergent subsequences uniquely solves the multi-dimensional FPE in (2.10).

Because the TAP is the sum of the EAP and IRPs, we have, for $1 \leq j \leq m$,

$$\tilde{N}_n^{(j)}(t) = \tilde{N}_n^{(0,j)}(t) + \sum_{i \in \mathcal{O}} \tilde{R}_n^{(i,j)}(t) + \sum_{i \in \mathcal{U}} \tilde{R}_n^{(i,j)}(t), \quad (2.37)$$

where $\tilde{N}_n^{(0,j)}$ is the EAP of the $j$th queue and $\tilde{R}_n^{(i,j)}$ is the IRP from queue $i$ to queue $j$. Because the FWLLN is obtained in §2.5.1 for $\tilde{R}_n^{(i,j)}$ with $i \in \mathcal{O}$ and the FWLLN for $\tilde{N}_n^{(0,j)}$ is given in Assumption 2.2.1, it remains to treat the third term in (2.37). Although the IRPs from an UL queue has the same representation as that in (2.29) and (2.30) for $i \in \mathcal{O}$, the SCP of new customers at a UL queue is different because the ESP is now the TAP, i.e., $E_n^{(i)} = N_n^{(i)}$.

Modifying (2.21), we have

$$\tilde{D}_n^{(i,\nu)}(t) = \frac{1}{n} \sum_{k=1}^{N_n^{(i)}(t)} 1(\mathcal{E}_k^{(i,n)} + S_k^{(i)} \leq t) \quad \text{for all } i \in \mathcal{U} \quad (2.38)$$

Following the compactness approach, we first establish the tightness of the TAP in the next lemma.

**Lemma 2.5.3** The TAP $\left(\tilde{N}_n^{(1)}, \ldots, \tilde{N}_n^{(m)}\right)$ is C-tight in $\mathbb{D}^m$.

**Proof 2.5.2** By Theorem 11.6.7 of \[102\], it suffices to show the C-tightness of $\tilde{N}_n^{(j)}$ in $\mathbb{D}$ for all $1 \leq j \leq m$. Based on (2.38), we first bound the routing processes $\tilde{R}_n^{(i,j)}$ with the service completions $\tilde{D}_n^{(i)}$, in particular, we have

$$\tilde{N}_n^{(i)}(t) \leq \tilde{N}_n^{(0,i)}(t) + \sum_{i \in \mathcal{O}} \tilde{D}_n^{(i)}(t) + \sum_{k \in \mathcal{U}} \tilde{D}_n^{(k)}(t) = \tilde{N}_n^{(0,i)}(t) + \sum_{i \in \mathcal{O}} \tilde{D}_n^{(i)}(t) + \sum_{k \in \mathcal{U}} \tilde{D}_n^{(k,o)}(t) + \sum_{k \in \mathcal{U}} \tilde{D}_n^{(k,\nu)}(t). \quad (2.39)$$

The convergence to continuous limits for (i) $\tilde{N}_n^{(0,i)}$ for $1 \leq i \leq m$ (Assumption 2.2.1), (ii) $\tilde{D}_n^{(i)}$ for $i \in \mathcal{O}$ (§2.5.1) and (iii) $\tilde{D}_n^{(i,o)}$ for $i \in \mathcal{U}$ (Remark 2.5.1), implies the C-tightness of the first three terms of (2.39). To complete the proof of Lemma 2.5.3, it remains to show the C-tightness of the last term in (2.39), because the C-tightness is preserved under addition (Chapter VI, Corollary 3.33 of \[42\]).
For a UL queue $k$ (i.e., $k \in \mathcal{U}$), let $s_{n}^{k, \uparrow} \equiv \sup\{s_{n}^{(k)}(t) : 0 \leq t \leq T\}$ and let $Z_{1}(t), Z_{2}(t), \ldots$ be an i.i.d. sequence of renewal processes with inter-renewal times following the cdf $G_{k}$. We can then bound the SCP $D_{n}^{(k, \nu)}(t)$ by the sum of $s_{n}^{k, \uparrow}$ independent renewal processes, in particular,

$$
\bar{D}_{n}^{(k)}(t) \leq \frac{1}{n} \sum_{r=1}^{s_{n}^{k, \uparrow}} Z_{r}(t).
$$

By the proof of Lemma 2.5.1 (see Appendix A.2), the right-hand side of (2.40) is $C$-tight. Therefore, by Chapter VI, Proposition 3.35 of [42], $\bar{D}_{n}^{(k)}$ has to be $C$-tight for each $k \in \mathcal{U}$. We thus conclude the proof.

Since Lemma 2.5.3 implies that every subsequence of $\bar{N}_{n}^{(i)}$ has a further convergent subsequence $\bar{N}_{n_{k}}^{(i)}$, we complete the proof of the FWLLN of the TAP by showing that every convergent subsequence of $\bar{N}_{n}^{(i)}$ converges to $\Lambda_{i}(t)$ as $n \to \infty$. By Remark 2.5.1, we obtain the FWLLN of $\bar{D}_{n}^{(i, \nu)}$ of an OL queue in (2.23), we easily obtain, from (2.38), that

$$
\bar{D}_{n}^{(i, \nu)}(t) \Rightarrow D_{i}^{(\nu, *)}(t) \equiv \int_{0}^{t} G_{i}(t-s) dN_{i}^{*}(s) \quad \text{in} \quad \mathbb{D}, \quad \text{as} \quad n \to \infty, \quad i \in \mathcal{U}.
$$

Paralleling the proof of the FWLLN for $\bar{R}_{n}^{(i, j, \nu)}$ with $i \in \mathcal{O}$ in §2.5.1, we have

$$
\bar{R}_{n}^{(i, j, \nu)}(t) \Rightarrow R_{i,j}^{(\nu, *)}(t) \equiv \int_{0}^{t} P_{i,j}(s) dD_{i}^{(\nu, *)}(s) \quad \text{in} \quad \mathbb{D}, \quad \text{as} \quad n \to \infty,
$$

for all $i \in \mathcal{U}, \ 0 \leq j \leq m$. By Remark 2.5.1, we obtain the FWLLN of $\bar{R}_{n}^{(i, j, o)}$ for free, namely,

$$
\bar{R}_{n}^{(i, j, o)}(t) \Rightarrow \int_{0}^{t} P_{i,j}(s) dD^{(i, o)}(s) \quad \text{in} \quad \mathbb{D}, \quad \text{as} \quad n \to \infty, \quad i \in \mathcal{U}, \quad 0 \leq j \leq m,
$$

where $D^{(i, o)}$ is defined in (2.20). Finally, combining (2.37), (2.42), (2.43), Assumption 2.2.1 and (2.36), we have

$$
N_{j}^{*}(t) \equiv \Lambda_{j}^{(0)}(t) + \sum_{i \in \mathcal{O}} \int_{0}^{t} P_{i,j}(u) dD_{i}(u) + \sum_{i \in \mathcal{U}} \int_{0}^{t} P_{i,j}(u) dD^{(i, o)}(u)
$$

$$
+ \sum_{i \in \mathcal{U}} \int_{0}^{t} P_{i,j}(u) dD^{(\nu, *)}(u),
$$

(2.44)
where $D_i \equiv D^{(i,o)} + D^{(i,\nu)}$, $D^{(i,o)}$ is given in (2.20), $D^{(i,\nu)}$ is given in (2.23) and $D^{(\nu,\ast)}$ is defined in (2.41). It is not hard to see that (2.44) agrees with the integral version of the FPE (2.10). In particular, because the proof in Appendix A.2 also implies the limit $\bar{N}^*_j$ is Lipschitz continuous, taking the derivative of (2.44) with respect to $t$ gives

$$\dot{\bar{N}}^*_j(t) = \chi_j^{(0)}(t) + \sum_{i \in O} P_{i,j}(t)\sigma_i(t) + \sum_{i \in U} P_{i,j}(t) \int_0^\infty \frac{b_i(0,y)g_i(t + y)}{G_i^c(y)} dy$$

$$+ \sum_{i \in U} P_{i,j}(t) \int_0^t g_i(t - x) dN^*_i(x),$$

(2.45)

which coincides with the FPE (2.10). Since this FPE has a unique solution (see Theorem 1 of [64]) and the choice of the subsequence is arbitrary, all convergent subsequences must have the same limit, we have $P(N^* = \Lambda) = 1$ for $\Lambda$ in (2.12) and (2.2). Hence we have completed the proof of the FWLLN of the TAP.

### 2.5.3 FWLLNs for Other Processes

We now complete the proof of Theorem 2.4.1 by establishing the FWLLNs of all other processes, including the service-related processes of UL queues (e.g., $\bar{D}_n^{(i)}$ and $\bar{B}_n^{(i)}$ for $i \in U$) and queue-related processes of OL queues (e.g., $W_n^{(j)}$, $V_n^{(j)}$, $\bar{Q}_n^{(j)}$ and $\bar{A}_n^{(j)}$ for $i \in O$). Because the FWLLN of the TAP is established, we now independently treat each queue $i$, $1 \leq i \leq m$, with a given FWLLN of its TAP $\bar{N}_n^{(i)}$. We draw heavily on the proofs in [62, 65].

**FWLLNs for service-related processes at UL queues.**

Mimicking (2.16) and the arguments in §2.5.1, we split $\bar{D}_n^{(i)}$ ($\bar{B}_n^{(i)}$) of a UL queue $i$ into the SCP (service content) of new customers $\bar{D}_n^{(i,o)}$ ($\bar{B}_n^{(i,o)}$) and the SCP (service content) of old customers $\bar{D}_n^{(i,\nu)}$ ($\bar{B}_n^{(i,\nu)}$). As discussed in Remark 2.5.1, the FWLLNs of $\bar{D}_n^{(i,o)}$ and $\bar{B}_n^{(i,o)}$ have been developed in §2.5.1 with limits in (2.20) and (2.28). It remains to prove the FWLLNs for $\bar{D}_n^{(i,\nu)}$ and $\bar{B}_n^{(i,\nu)}$. Modifying (2.21) and (2.25), we have for $i \in U$,

$$\bar{D}_n^{(i,\nu)}(t) = \frac{1}{n} \sum_{k=1}^{N_n^{(i)}(t)} 1(\mathcal{E}_k^{(i,n)} + \mathcal{S}_k^{(i)} \leq t),$$

$$\bar{B}_n^{(i,\nu)}(t, y) = \frac{1}{n} \sum_{k=N_n^{(i)}((t-y)^+)+1}^{N_n^{(i)}(t)} 1(\mathcal{E}_k^{(i,n)} + \mathcal{S}_k^{(i)} > t).$$
By the FWLLN of the TAP in §2.5.2, (2.41) and Theorem 3.1 of [78], we quickly obtain the FWLLNs for \( \bar{D}^{(j,\nu)}_n \) and \( \bar{B}^{(j,\nu)}_n \), in particular,

\[
\bar{D}^{(j,\nu)}_n(t) \Rightarrow D^{(\nu)}_j(t) \equiv \int_0^t G_i(t-s)d\Lambda_i(s) \quad \text{in} \quad D, \quad (2.46)
\]

\[
\bar{B}^{(j,\nu)}_n(t,y) \Rightarrow B^{(j,\nu)}(t,y) \equiv \int_{t-y}^t G^c_i(t-s)d\Lambda_i(s) \quad \text{in} \quad D_D, \quad \text{as} \ n \to \infty. \quad (2.47)
\]

where \( \Lambda_i \) satisfies the traffic-flow equation in (2.2).

**FWLLN for queue-related processes at OL queues.**

Since all service-related processes have already been treated in §2.5.1, it remains to prove the FWLLNs for the queue-related processes, including the two-parameter queue-content \( \bar{Q}^{(i)}_n \), HOL waiting time \( W^{(i)}_n \), HOL waiting time \( V^{(i)}_n \) and the abandonment process \( \bar{A}^{(i)}_n \), for \( i \in O \).

Following §§6.2–6.3 in [65], we let \( Q^{(i,\ast)}_n(t,x) \) be the two-parameter queue-length process ignoring flows into service (QLIFIS). Namely, \( Q^{(i,\ast)}_n(t,x) \) denotes the number of customers in queue at \( t \) with elapsed waiting times no more than \( x \), assuming no customer has been allowed to enter service since time 0. To obtain a representation of the queue-length process \( Q^{(i)}_n \) which allows the usual flow into service, we now bound the second argument \( x \) by the HOL waiting time \( W^{(i)}_n \). Namely, we have

\[
Q^{(i)}_n(t,x) = Q^{(i,\ast)}_n(t,x \wedge W^{(i)}_n) \quad (2.48)
\]

where \( Q^{(i,\ast)}_n(t,x) \) denotes the LLN-scaled number of customers in queue at \( t \) with elapsed waiting times no more than \( x \) from those customers that are in queue at time 0, and \( Q^{(i,\nu,\ast)}_n(t,x) \) denotes the LLN-scaled number of customers in queue at \( t \) with elapsed waiting times no more than \( x \) from the new arrivals in the interval \( [0,t] \). Paralleling the treatments for \( \bar{B}^{(i,o)}_n(t,x) \) and
Combining the FWLLN of the fluid limit composition and addition. It is not hard to see that the right-hand side of (2.53) coincides with the FWLLNs of \( \bar{\nu}_{n,i} \) and \( \bar{\nu}_{n,o} \) in \( \bar{\nu}_{n} \). Because (2.49) and (2.50) are analogs of (2.25) and (2.26), we parallel the proofs for the FWLLNs of \( \nu_{n,i} \) and \( \nu_{n,o} \) in §2.5.1. Because \( \bar{\nu}_{n,i} = \bar{\nu}_{n,i} \) and \( \bar{\nu}_{n,o} = \bar{\nu}_{n,o} \), we have for \( i \in \mathcal{O} \),

\[
\bar{Q}_{n}^{(i,\nu,\ast)}(t, x) = \frac{1}{n} \sum_{k=1}^{N_{n,i}(t)} 1 \left( \xi_{k}^{(i,n)} + A_{k}^{(i)} > t \right),
\]

(2.49)

\[
\bar{Q}_{n}^{(i,o,\ast)}(t, x) = \frac{1}{n} \sum_{k=1}^{N_{n,o}(t)} 1 \left( \xi_{k}^{(i,n)}(\chi_{n,i}) > t \right),
\]

(2.50)

where \( \xi_{k}^{(i,n)} \) and \( A_{k}^{(i)} \) are the arrival and patience times of the \( k \)-th new customer (i.e., arrivals after time 0) so that \( \xi_{k}^{(i,n)} + A_{k}^{(i)} \) is the time the \( k \)-th customer abandons from the queue if this customer does not enter service by then, \( \{0 < \chi_{n,i}^{(1)} \leq \chi_{n,i}^{(2)} \leq \ldots\} \) is the ordered sequence of elapsed waiting times of customers in queue at time 0, and \( \{\xi_{i}^{(1)}(x), \xi_{i}^{(2)}(x), \ldots\} \) is an i.i.d. sequence of random variables with cdf

\[
P \left( \xi_{i}^{(1)}(x) > t \right) = 1 - H_{i}^{(t)}(t) \equiv \frac{F_{i}^{(t)}(x + t)}{F_{i}^{(t)}(x)} \quad \text{for} \quad x > 0, \quad t \geq 0.
\]

Because (2.49) and (2.50) are analogs of (2.25) and (2.26), we parallel the proofs for the FWLLNs of \( \bar{Q}_{n}^{(i,\nu,\ast)} \) and \( \bar{Q}_{n}^{(i,o,\ast)} \) in §2.5.1. Because \( \bar{Q}_{n,i}(0, \cdot) \Rightarrow Q_{i}(0, \cdot) \) in \( \mathbb{D} \) by Assumption 2.2.2, we have for \( i \in \mathcal{O} \),

\[
\bar{Q}_{n,i}^{(i,\nu,\ast)}(t, x) \Rightarrow Q_{i}^{(i,\nu,\ast)}(t, x) \equiv \int_{(t-x)^{+}}^{t} F_{i}^{(t-s)}(s)ds,
\]

(2.51)

\[
\bar{Q}_{n,i}^{(i,o,\ast)}(t, x) \Rightarrow Q_{i}^{(i,o,\ast)}(t, x) \equiv \int_{0}^{(x-t)^{+}} q_{i}(0, y) \left( 1 - H_{y}^{(i)}(t) \right) dy \quad \text{in} \quad \mathbb{D}_{\mathbb{D}}.
\]

(2.52)

Combining the FWLLN of \( W_{n}^{(i)} \) and (2.51)–(2.52), we have

\[
\bar{Q}_{n}^{(i)}(t, x) = \bar{Q}_{n}^{(i,\ast)} \left( t, x \wedge W_{n}^{(i)}(t) \right)
\]

\[
\Rightarrow \bar{Q}_{n}^{(i,\ast)} \left( t, x \wedge w_{i}(t) \right) \equiv \bar{Q}_{n}^{(i,\nu,\ast)} \left( t, x \wedge w_{i}(t) \right) + \bar{Q}_{n}^{(i,o,\ast)} \left( t, x \wedge w_{i}(t) \right)
\]

\[
= \int_{(t-x \wedge w_{i}(t))^{+}}^{t} F_{i}^{(t-s)}(s)ds + \int_{0}^{(x \wedge w_{i}(t)-t)^{+}} q_{i}(0, y) \left( 1 - H_{y}^{(i)}(t) \right) dy
\]

(2.53)

in \( \mathbb{D}_{\mathbb{D}} \), as \( n \to \infty \). Here the convergence follows from the continuous mapping theorem with composition and addition. It is not hard to see that the right-hand side of (2.53) coincides with the fluid limit \( Q_{i}(t, x) \) defined in (2.3) and (2.8).

Given the FWLLNs of (i) the TAP \( \nu_{n}^{(i)(0)} \), (ii) queue length \( \bar{Q}_{n}^{(i)}(t, y) \) and (iii) ESP \( \bar{E}_{n}^{(i)} \), we can easily obtain the FWLLN of the abandonment process for an OL queue \( i \in \mathcal{O} \), defined as
Figure 2.2: Performance functions of the $(M_t/H_2/s_t + E_2)^2/M_t$ FQNet, including (i) TAR $\lambda$, (ii) queue content $Q$, (iii) VWT $w$, (iv) service content $B$, (v) total fluid $X$ and (vi) rate into service $b(t,0)$.

$$\bar{A}_n^{(i)}(t) = \bar{Q}_n^{(i)}(0) + \bar{N}_n^{(i)}(t) - \bar{E}_n^{(i)}(t) - \bar{Q}_n^{(i)}(t),$$ by the continuous mapping theorem with addition.

### 2.6 An $(M_t/H_2/s_t + E_2)^2/M_t$ Example

To provide engineering verification of Theorem 2.4.1, we now report the results of a simulation experiment. We consider a two-queue $(M_t/H_2/s_t + E_2)^2/M_t$ SQNet, with (i) NHPP arrival processes having sinusoidal arrival-rate functions $\lambda_n^{(0,j)}(t) = n \lambda^{(0)}(t)$, $\lambda_i^{(0)}(t) = a_i + b_i \sin(c_i t + \phi_i)$, (ii) two-phase hyper-exponential $(H_2)$ service times with pdf $g_i(x) = p_i \cdot \mu_i^{(i)} e^{-\mu_i^{(i)} x} + (1 - p_i) \cdot \mu_2^{(i)} e^{-\mu_2^{(i)} x}$, (iii) constant staffing levels $s_n^{(i)}(t) = \lceil n s_i \rceil$, and (iv) two-phase Erlang $(E_2)$ patience times with pdf $f_i(x) = 4 \theta_i^2 x e^{-2\theta_i x}$, for $i = 1, 2$. 

45
We let \( a_1 = 0.8, b_1 = 0.4, a_2 = 0.7, b_2 = 0.5, \phi_1 = 1.5, \phi_2 = 1, c_1 = 2, c_2 = 1, \theta_1 = 0.5, \theta_2 = 0.3, s_1 = 1, s_2 = 2, \mu_1 = 1, \mu_2 = 0.5, p_1 = p_2 = 0.5(1 - \sqrt{0.6}), \mu_1(i) = 2p_1\mu_i, \mu_2(i) = 2(1 - p_1)\mu_i, \) for \( i = 1, 2 \). We have the service-time squared coefficient of variation SCV \( c_s^2 = 4 \) and abandonment-time SCV \( c_a^2 = 1/2 \). Let the routing probabilities \( p_{1,1} = 0.15, p_{2,1} = 0.12, p_{1,2} = p_{2,2} = 0.2 \).

Figure 2.2 shows plots of the key performance functions for \( 0 \leq t \leq T \equiv 20 \), starting out empty, together with (i) EAPs \( \lambda^{(0)} \) and TARs \( \lambda \) (Subplot 1), (ii) queue contents \( Q \) (Subplot 2), (iii) HOL waiting time \( w \) (Subplot 3), (iv) service contents \( B \) (Subplot 4), (v) total fluid \( X \) (Subplot 5) and (vi) rate fluid enters service \( b(t, 0) \) (last subplot). All performance functions are continuous except for \( b(t, 0) \): in UL intervals, \( b(t, 0) = \lambda(t) \); in OL intervals \( b(t, 0) \) is the unique solution of the FPE (2.7).

To verify the accuracy of the FQNet approximation, we conduct simulation comparisons in Figure 2.3 for the LLN-scaled key performance processes of the SQNet, starting out empty (dashed lines): (i) HOL waiting time, (ii) number in queue, (iii) number in service and (iv) total number of customers. In Figure 2.3(a) we compare the fluid functions of the FQNet (the dashed lines) with the single sample paths of their corresponding LLN-scaled performance functions of the SQNet (the solid lines) with a large scale \( n = 3000 \). In Figure 2.3(b) we compare the fluid functions (the dashed lines) with the means of the LLN-scaled performance functions of the SQNet (the solid lines, estimated by averaging 100 independent samples) with a smaller scale \( n = 100 \). Figure 2.3 verifies the remarkable performance of the FQNet approximation and provides practical confirmation of the FWLLN in Theorem 2.4.1. See [64] for additional experiments supporting the FQNet approximation.

### 2.7 Conclusion

We have established a many-server heavy-traffic limit theorem for a recently proposed deterministic fluid approximation for the \((G_t/GI/s_t + GI)^m/M_t\) queuing network [64], with a non-stationary non-Poisson arrival process, nonexponential service and patience times, time-varying staffing levels and Markovian (probabilistic) routing policy. Numerical analysis and simulation experiments have been developed in [61, 64] confirming the effectiveness of this deterministic fluid approximations. However, prior to this work the functional law of large numbers of this \((G_t/GI/s_t + GI)^m/M_t\) fluid limit remained an open problem.

Here we solve this open problem by showing that all scaled performance processes, including the queue lengths (both in queue and in service), flows (of routing, abandonment and service completion), and waiting times, jointly converge in distribution to their corresponding deterministic fluid functions conjectured in [61, 64], in appropriate functional spaces. We draw heavily on the proofs in [62] which focused on the \(G_t/GI/s_t + GI\) single queue model. A key step here
Figure 2.3: A comparison of performance functions in the $(M_t/H_2/s_t + E_2)^2/M_t$ FQNet with simulation of the corresponding $(M_t/H_2/s_t + E_2)^2/M_t$ SQNet with (a) single sample paths and scale $n = 3000$, and (b) average of 100 paths and scale $n = 100$. 
is to show the convergence of the total arrival process for all queues in the network. Our proof follows the compactness approach by (i) establishing the tightness in the appropriate functional space and (ii) showing that all convergent subsequences of the performance functions converge to the same desired limits.
Chapter 3

Heavy-traffic Limit for the Initial Content Process

3.1 Introduction

The initial content process of a queueing system specifies the number of customers that were initially in service at time 0 and are still in service later at time $t$ and the elapsed service times since their arrival times before time 0. Assuming that the service times come from a sequence of i.i.d. random variables, independent of the arrival process and system history, the initial content process is a Markov process, even if the service-time distribution is nonexponential, and thus provides a useful description of the system state at each time. In this chapter, we establish a heavy-traffic FCLT for the initial content process of a large-scale queueing system. The key assumption is an FCLT for the initial age process, which requires that the number of customers initially in service grows. The technical challenge is treating non-identically-distributed remaining service times.

We establish the FCLT above in a more general context. Specifically, we consider a sequence of infinite-server (IS) queues indexed by $n$, which we refer to as $G_t/GI^o, GI^o/\infty$ models. There are infinitely many servers, so that each customer enters service immediately upon arrival. In system $n$, there is a general arrival process with a time-varying arrival rate function $\lambda_n(t) = n\lambda(t)$ (the $G_t$), so that the arrival rate is scaled by $n$, the usual many-server heavy-traffic scaling. We will specify the arrival process only by the requirement that it satisfy an FCLT; see Assumption 3.2.1 below.

We assume that the system operated in the past (prior to time 0) as a conventional $G_t/GI/\infty$, model with i.i.d. service times that are independent of the arrival process, with service times distributed according to a cdf $G$. We assume that system operates after time 0 according to a $G_t/GI/\infty$ IS model with i.i.d. service times that are independent of the arrival process.
process, with service times distributed according to the cdf $G_\nu$. As in the usual many-server heavy-traffic scaling, the two service-time cdf’s $G$ and $G_\nu$ are not scaled by $n$. Our approach is designed especially to treat the case in which these cdf’s are not exponential.

We will be interested in the system performance after time 0, which can be characterized by the pair of two-parameter stochastic processes $(X^{e,o}_n(t,y), X^{e,\nu}_n(t,y))$ with $t \geq 0$ and $y \geq 0$. The variable $X^{e,o}_n(t,y)$ counts the number of customers that were already in service at time 0 and are still in service at time $t$ and have elapsed service times that are less than or equal to $y$ (here $y > t$ since they started service prior to time 0). The variable $X^{e,\nu}_n(t,y)$ counts the number of customers that arrived after time 0 and are still in service at time $t$ and have elapsed service times that are less than or equal to $y$ (here $0 \leq y \leq t$ since they started service after to time 0). (The superscripts are chosen to help, with $e$ denoting elapsed, $o$ old and $\nu$ new.) Given the assumptions on the service times, the stochastic process $(X^{e,o}_n, X^{e,\nu}_n) \equiv \{(X^{e,o}_n(t,\cdot), X^{e,\nu}_n(t,\cdot)) : t \geq 0\}$ is a Markov process with time domain $[0, \infty)$ and state space $\mathbb{D}^2$ (see §1.8.1).

Our main result, Theorem 3.3.2, is an FCLT for $(X^{e,o}_n, X^{e,\nu}_n)$ jointly with other processes in the space $\mathbb{D}_{D^2}$ of $\mathbb{D}^2$-valued functions in $\mathbb{D}$. The use of $\mathbb{D}_{D^2}$ follows [78, 80, 97]. It is an alternative to measure-valued approaches in [18, 46, 48, 109] and distribution-valued approach in [91]. The alternative approaches are appealing for simplifying arguments and revealing structure; e.g., [91] shows that the the heavy-traffic limit for the $G/G/\infty$ model can be regarded as a tempered-distribution-valued Ornstein-Uhlenbeck diffusion process, generalizing the diffusion process limit for the $M/GI/\infty$ model in [18]. On the other hand, the $\mathbb{D}_{D^2}$ framework here evidently admits more continuous functions, and so has more immediate applications via the continuous mapping theorem. Explicit connections between the two approaches for the fluid limits are made in [45].

Theorem 3.3.2 here extends Theorems 3.2 and 5.1 of Pang and Whitt [78] by treating more general initial conditions. In particular, in §5 of [78] the remaining service times of customers initially in the system at time 0 are assumed to come from a sequence of i.i.d. random variables; similar restrictive conditions on the initial conditions are made in [48, 87, 89]. Our main contribution here is treating the initial content process (ICP) $X^{e,o}_n$; the limit for $X^{e,\nu}_n$ comes from [78]. As in [78], and in Louchard [68] and Krichagina and Puhalskii [54] before, we work with the empirical process of the service times. The key assumption for Theorem 3.3.2 here is that the initial age process $\{X^{e,o}_n(0,y) : y \geq 0\}$ (which is just the ICP at time 0) satisfies an FCLT jointly with the arrival process after time 0; see Assumption 3.2.1. To address the technical challenge of non-identically-distributed remaining service times, we draw on Chapter 25 of Shorack and Wellner [95], which in turn uses a symmetrization argument from Marcus and Zinn [74]. Substantial new arguments are required as can be seen from the tightness proof in §3.4.2. Evidently, this is the first use of symmetrization technique to analyze a queueing model with non-identically-distributed service times.
We point out that Reed and Talreja [91] also treat the ICP as part of their new approach to obtaining heavy-traffic limits for the $G/GI/\infty$ model, but the specific conditions assumed, as well as the framework, are different. We emphasize engineering relevance, e.g., by providing an explicit characterization of the limiting process, exposing key structure (see Remark 3.3.3) and providing explicit formulas for time-varying means, variances and covariances that lead to an effective algorithm for computing relevant performance measures (as confirmed by the simulation experiments at the end).

**Motivation.** Very broadly, IS models are important to represent the time-varying offered load, i.e. the time-varying demand, in a service system, by going beyond the time-varying arrival rate to include the service times in a useful way; see [31, 63, 106]. Directly including the service times is important when the service times are relatively long. Then the load through time may not be well represented by the arrival rate function alone, because the customers are in the system for a substantial time after their arrival epochs. The expected number of customers in an infinite-server queue is referred to as the offered load because it represents the expected number of servers that would actually be used as a function of time if there were no limit on their availability. The analysis here may usefully supplement previous work on the offered load, because to understand and predict the demand in a service system, it may be useful to partition the demand into the demand resulting from new input after some time and the demand resulting from customers already in service at that time. The FCLT here may be useful to predict the future demand based on observation of the initial content process.

More narrowly, within the class of IS models, the present work evidently is the first to establish a heavy-traffic limit for the performance of the $G_t/GI^o,GI^v/\infty$ model with general nonstationary arrival process after time 0 when the system started some time in the past and there is a system change at time 0, possibly leading to a new arrival process and a new service-time distribution after time 0. Corollary 3.5.1 in §3.5 shows that Theorem 3.3.2 can be applied to establish an FCLT for the $G_t/GI^o,GI^v/\infty$ model, even if the general initial conditions here hold before time 0.

Even more narrowly, for the $G_t/GI/\infty$ model, it may be important to consider IS models with initial conditions more general than assumed in §5 of [78], where the remaining service times conditioned on the initial number in service was assumed to be i.i.d.

For finite-server queueing systems, to understand the performance, it may be useful to focus on the evolution of the content that is initially in service at some time, ignoring the new input after that time. That necessarily will be the case in service systems that provide service over normal working hours each day, if the system shuts down at some time, except that all customers already in service at the termination time are allowed to complete their service. For example, terminating systems with nonhomogeneous Poisson arrival processes were considered in [30].
As discussed in [19], such performance descriptions can be useful to develop effective measures to recover after system overloads.

Finally, it is important to note that IS models are useful for studying finite-server systems. First, results for IS models can be used to develop engineering approximations for finite-server models; e.g., see the peakedness approximations in [58, 79] and references therein. Second, heavy-traffic limits for IS models can be fruitfully applied to establish associated many-server heavy-traffic limits for finite-server systems; see [2, 62, 65, 72, 87, 89].

In fact, we intend to apply the results here in [5](Chapter 5) to establish a FCLT for the $G_t/GI/s_t + GI$ model with alternating overloaded (OL) and underloaded (UL) intervals, time-varying staffing as well as arrival rate and customer abandonment (the $+GI$), extending the FCLT for the $G_t/M/s_t + GI$ model in [65]. We intend to apply the present results in three ways. First, the theory here applies directly to UL intervals, which can directly be regarded as IS models, starting off with customers in service with elapsed service times, determined from the previous OL interval. Second, the theory also directly applies to the initial content in service during an OL interval, determined from the previous UL interval (because the dynamics of the ICP is not affected by the finite service capacity). Third, the theory will once again apply to treat the number of waiting customers in an OL interval, because then we can regard the patience times as service times; see [65].

**Organization of the chapter.** In §3.2, we specify the model operating after time 0 in more detail. We define its key performance functions, specify the many-server heavy-traffic scaling, and our detailed assumptions. At this point, we represent the behavior before time 0 by the assumed behavior of the ICP at time 0 in Assumption 3.2.1. In §4.5 we state our main results and in §3.4 we prove them. In §3.5 we show that our results apply to the $G_t/GI^o,GI^o/\infty$ model starting some time before time 0, if we assume that the limit for the arrival process has independent increments, because then the ICP at time 0 has the properties assumed in Assumption 3.2.1. From this case, we can also see that the results are consistent with the previous results in [78]. In §3.6 we provide the first characterization of the steady-state distribution of the new and old content in the stationary $G/GI/\infty$ model. In §3.7 we report simulation results for a challenging test case having non-Markov arrival process, nonexponential service-time distribution, and general initial conditions. Finally, we provide concluding remarks in §3.8. Additional material appears in an appendix.

### 3.2 The Model After Time 0

We start by considering the model after time 0; we show that the results can be applied to the $G_t/GI^o,GI^o/\infty$ model starting with the initial conditions here at some time before time 0 in
§3.5. Even though we consider time with $t \geq 0$, we are especially interested in those customers who arrived before time 0. Their history will be captured by the initial age process, which coincides with the ICP at $t = 0$.

We are primarily interested in the ICP $X_n^{e,o}(t, y)$, but we also consider the associated process for the new input $X_n^{e,\nu}(t, y)$. In addition to the pair of two-parameter stochastic processes $(X_n^{e,o}(t, y), X_n^{e,\nu}(t, y))$, counting the old and new customers in the system at time $t$ with elapsed service times at most $y$, we also define the closely related pair of two-parameter stochastic processes $(X_n^{r,o}(t, y), X_n^{r,\nu}(t, y))$, counting the old and new customers in the system at time $t$ with remaining service times at least $y$. Of course, these remaining-time processes are usually not directly observable, but they do usefully represent the future demand. However, they are tightly linked with the other processes. In particular, they are linked via the simple relations

$$X_n^{r,o}(t, y) = X_n(0) - X_n^{e,o}(t+y), \quad t \geq 0, \; y \geq 0$$

where $\tau_i^n$ is the arrival time of the $i$th customer and $v_i^n$ is the associated service time in system $n$. The service times $v_i$ has not been scaled by $n$, hence no subscript.

Now we turn to the processes associated with initial customers already in the system at time 0. Let $v_{-i}$ denote the length of time the $i$th customer has been in service (age in service) at time 0 in the $n$th system (negative index corresponds to initial customers in service at time 0). Without loss of generality we assume the ages are ordered $0 \leq v_{-1} \leq v_{-2} \leq \ldots$. Also let $\bar{v}_{-i}$ be the remaining service time at time 0 of the $i$th customer initially in service. The remaining times depend on the ages $\{v_{-i} : i \geq 1\}$ of customers in service. Because the service-time distribution is not the exponential distribution, the remaining service times $\{\bar{v}_{-i} : i \geq 1\}$ are not i.i.d.. Then it follows that

$$X_n^{e,o}(t, y) = X_n(0(y-t)^+) = \sum_{i=1}^{X_n(0(y-t)^+)} 1(\bar{v}_{-i} > t), \quad t \geq 0, \; y \geq 0$$

$$X_n^{r,o}(t, y) = \sum_{i=1}^{X_n(0)} 1(\bar{v}_{-i} > t+y), \quad t \geq 0, \; y \geq 0$$
where \( X_0^n(0, (y - t) +) \) is the total number of customers at time 0 that have been in service for time \((y - t) +\).

The key property we will exploit is the conditional independence property: Conditional on the sequence of service age random variables \( \{v_{n,i} : i \geq 1\} \), the sequence of random variables \( \{\tilde{v}_{n,i} : i \geq 1\} \) are mutually independent with conditional tail probabilities

\[
\mathbb{P}(\tilde{v}_{n,i} > t | v_{n,i} = x) \equiv \tilde{G}_x^c(t) \equiv 1 - \tilde{G}_x(t) \equiv \frac{G_o^c(t + x)}{G_o^c(x)}, \quad t, x \geq 0, \tag{3.5}
\]

where \( G_o^c(x) \equiv 1 - G_o(x) \) is the ccdf for the service-time distribution associated with initial customers in system. The primary difficulty in the proof stems from the fact that, conditional on the sequence of service age random variables \( \{v_{n,i} : i \geq 1\} \), the random variables \( \{\tilde{v}_{n,i} : i \geq 1\} \) are not identically distributed.

Given the processes (3.1)-(3.4) and the equalities \( X_n(t) = X_n^e(t, \infty) = X_n^r(t, 0) \), we can define the service-completion process associated with initial and new customers from the \( n^{th} \) queue. Let \( D_n^o(t) \) (resp. \( D_n^r(t) \)) be the total number of initial (resp. new) customers who have completed service by time \( t \). Then necessarily \( D_n^o(t) = X_n(0) - X_n^o(t) \) and \( D_n^r(t) = N_n(t) - X_n^r(t) \). Hence \( D_n(t) \equiv D_n^o(t) + D_n^r(t) = X_n(0) + N_n(t) - X_n(t) \) represents the total number of service completions by time \( t \).

**Associated scaled processes.** Let the associated LLN-scaled processes be

\[
\tilde{N}_n(t) \equiv N_n(t)/n, \quad \tilde{X}_n^e(t,y) \equiv X_n^e(t,y)/n, \quad \tilde{D}_n(t) \equiv D_n(t)/n, \quad \tilde{X}_n^r(t,y) \equiv X_n^r(t,y)/n. \tag{3.6}
\]

Let the associated CLT-scaled processes be

\[
\hat{N}_n(t) \equiv \frac{N_n(t) - n\Lambda(t)}{\sqrt{n}}, \quad \hat{X}_n^e(t,y) \equiv \frac{X_n^e(t,y) - nX^e(t,y)}{\sqrt{n}}, \\
\hat{D}_n(t) \equiv \frac{D_n(t) - nD(t)}{\sqrt{n}}, \quad \hat{X}_n^r(t,y) \equiv \frac{X_n^r(t,y) - nX^r(t,y)}{\sqrt{n}}, \quad t, y \geq 0, \tag{3.7}
\]

where the centering terms \( \Lambda(t), X^e(t,y), X^r(t,y), D(t) \) are deterministic functions (fluid limits) to be specified below in Assumption 3.2.1 and Theorem 3.3.1.

**The spaces \( \mathbb{D} \) and \( \mathbb{D}_E \).** The limits are established in the function space \( \mathbb{D} \equiv \mathbb{D}([0, \infty), \mathbb{R}) \) and for the two-parameter processes, the processes are random elements of the space \( \mathbb{D}_{\mathbb{D}} \equiv \mathbb{D}((0, \infty), \mathbb{D}((0, \infty), \mathbb{R})) \) of \( \mathbb{D} \)-valued functions (see §1.7).

We prove convergence in these spaces by using the compactness approach, i.e., by proving convergence of the finite dimensional distributions (fdds) and the tightness of the processes; see [10, 20, 42, 102] for tightness criteria in \( \mathbb{D} \), Theorem 6.2 of [78] for tightness criteria in \( \mathbb{D}_{\mathbb{D}} \).

**Assumptions.** Our key assumption is a joint FCLT for the arrival process of new customers
after time 0 and for the initial ages. We discuss the appropriateness of this assumption in Remarks 3.2.3 and 3.2.4 below and in §3.5.

**Assumption 3.2.1** (Joint FCLT for the arrival process and initial ages) The CLT-scaled ICP and external arrival processes defined in (3.7) jointly satisfy the FCLT

\[
\left( \hat{X}^e_n(0, \cdot), \hat{N}_n \right) \Rightarrow \left( \hat{X}^e(0, \cdot), \hat{N} \right) \quad \text{in } D^2 \quad \text{as } n \to \infty \tag{3.8}
\]

where \(\hat{X}^e(0, \cdot)\) and \(\hat{N}\) are two independent zero-mean continuous Gaussian processes. We assume that the deterministic centering terms in (3.7), which come from the associated functional weak law of large numbers (FWLLN) stated below in (3.10), can be represented as

\[
X^e(0, x) = \int_0^x a(u) \, du, \quad x \geq 0, \quad \text{and} \quad \Lambda(t) = \int_0^t \lambda(u) \, du, \quad t \geq 0, \tag{3.9}
\]

where the fluid initial age density \(a(u)\) and arrival rate function \(\lambda(t)\) are nonnegative real-valued functions that are integrable over bounded intervals.

**Remark 3.2.1** (FWLLN for the arrival process and initial content in service) As an immediate consequence of Assumption 3.2.1, we have a FWLLN for \(\hat{N}_n\) and \(\hat{X}^e_n(0, \cdot)\), i.e.,

\[
(\hat{X}^e_n(0, \cdot), \hat{N}_n, \hat{X}_n(0)) \Rightarrow (X^e(0, \cdot), \Lambda, X(0)) \quad \text{in } D^2 \times \mathbb{R} \quad \text{as } n \to \infty, \tag{3.10}
\]

**Remark 3.2.2** (The zero-mean Gaussian assumption) The zero-mean Gaussian requirement of Assumption 3.2.1 is not required for the convergence, but it is required for drawing the useful conclusion that the limiting process also has this structure, as in (3.15) below. Extensions are possible, as illustrated by §10 of [65].

**Remark 3.2.3** (Independence of the limiting initial content and new input) If the arrival process is a nonhomogeneous Poisson process, then we have an \(M_t/GI^\nu, GI^\nu/\infty\) IS model, for which the new input after time 0 is independent of the initial content, so that the independence of the two limiting processes follows directly from the two separate limits in Assumption 3.2.1. But, more generally, the number of customers in service at time 0 and the ages of the service times of those customers typically will not be independent of the arrivals after time 0. Nevertheless, Assumption 3.2.1 is very reasonable. For example, consider a \(G_t/GI/\infty\) system starting empty in the finite past. Even though the arrival process may not have independent increments, from [78] we know that it is common for the limit for the arrival process to be a time-transformed Brownian motion (BM), which has independent increments. In particular, that occurs if we assume that the arrival process is a deterministic time transformation of any arrival process that satisfies an FCLT with a BM limit. For such limits, it is natural to start with a stationary
process, such as an equilibrium renewal process, but it suffices to have the FCLT with a BM limit, as discussed in §7 of [77]. With either an ordinary or equilibrium renewal process, the limiting process will be $\hat{N}(t) = c_\lambda B_a(\Lambda(t))$, where $B_a$ is a standard BM, $\Lambda(t)$ is the deterministic time transformation, corresponding to the limiting cumulative arrival rate function and $c_\lambda^2$ is the squared coefficient of variation (SCV, variance divided by the square of the mean) of an interarrival time in the ordinary renewal process. For all these representations, the arrival FCLT and the independence of the limit is satisfied, as assumed in Assumption 3.2.1. In addition to [77], see [27, 59] and §5.4 of [67] for uses of this representation of nonstationary non-Poisson arrival processes.

**Remark 3.2.4** (Performance forecasting using limits in Assumption 3.2.1) For engineering purposes, the limits in Assumption 3.2.1 can be understood as estimators (approximations) for future demand posed by new input and initial content. The goal here is to develop performance forecasting formulas as functions of the limits in Assumption 3.2.1. It will be clear from the formulas and examples that the general initial conditions (represented by the initial fluid age function $X^e(0, \cdot)$ and the associated stochastic limiting process $\hat{X}^e(0, \cdot)$) can be a significant part of the performance functions.

We also impose additional regularity assumptions, which evidently are not too restrictive for engineering applications. We first impose conditions on the two service-time cdf’s. Even though not restrictive, both assumptions are used critically in the analysis; see Remark 3.3.2 and Lemma 3.4.4.

**Assumption 3.2.2** (Regularity conditions for service-time cdf’s) The two service-time cdf’s $G$ and $G^\nu$ are assumed to be continuous. In addition, the cdf $G$ has a probability density function (pdf) $g$ satisfying $0 < g(x) \leq g^\uparrow \equiv \sup_{x \geq 0} g(x) < \infty$ for all $x \geq 0$.

We also impose some regularity conditions on the initial content. These are used in the proof of tightness in $\mathbb{D}_D$ in §3.4.2.

**Assumption 3.2.3** (Regularity conditions for the initial content) We assume that the following two properties hold:

(i) The LLN-scaled initial content is bounded, i.e.,

$$X_n(0) \leq X^\uparrow < \infty$$

for all $n \geq 1$, w.p.1. \hfill (3.11)

(ii) There exists $y^\uparrow > 0$ such that $X_n(0) - X_n(0, y^\uparrow) = 0$ for all $n \geq 1$ w.p.1.
3.3 Main Results

In this section, we present the new FWLLN and FCLT for the $G_t/GI^o, GI^v/\infty$ model. They extend the corresponding results for the $G_t/GI/\infty$ model in §3 and §5 of [78] by treating more general initial conditions. In particular, the results for the new arrivals come from [78], but unlike §5 of [78], Assumption 3.2.1 here makes the remaining service times at time 0 be conditionally independent, given the ages, but not identically distributed random variables. We state the FWLLN first, but give no separate proof, because it is a consequence of the FCLT.

**Theorem 3.3.1 (FWLLN)** Consider the sequence of $G_t/GI^o, GI^v/\infty$ queues satisfying all assumptions in §3.2. Then

$$\left(\tilde{N}_n, \tilde{X}_n^c(0, \cdot), \tilde{X}_n^r(0, \cdot), \check{X}_n^c, \check{X}_n^r, \check{D}_n\right) \Rightarrow \left(\bar{N}, \bar{X}^c(0, \cdot), \bar{X}^r(0, \cdot), \bar{X}^c, \bar{X}^r, \bar{D}\right)$$  \hspace{1cm} (3.12)

in $\mathbb{D}^3 \times \mathbb{D}_D^2 \times \mathbb{D}^2$ as $n \to \infty$ where the limit is continuous and deterministic with $X(t) = X^c(t, \infty) = X^r(t, 0)$, and for $t, y \geq 0$,

$$X^c(t, y) = X^{c,o}(t, y) + X^{c,v}(t, y), \quad X^r(t, y) = X^{r,o}(t, y) + X^{r,v}(t, y),$$

$$X^{c,o}(t, y) = \int_{0}^{(y-t)^+} a(x)H^c_x(t)dx, \quad X^{c,v}(t, y) = \int_{(t-y)^+}^{t} G^c_v(t-s)\lambda(s) ds,$$

$$X^{r,o}(t, y) = \int_{0}^{\infty} a(x)H^c_x(t+y)dx, \quad X^{r,v}(t, y) = \int_{0}^{t} G^c_v(t+y-s)\lambda(s) ds,$$

$$D(t) = \Lambda(t) - X(t) = \int_{0}^{\infty} a(x)H^c_x(t)dx + \int_{0}^{t} G^c_v(t-s)\lambda(s) ds$$  \hspace{1cm} (3.13)

and $a(x)$ being the initial fluid limit age density and $\lambda(s)$ being the arrival rate function specified in Assumption 3.2.1.

For real numbers $a$ and $b$, let $a \vee b \equiv \max\{a, b\}$ and $a \wedge b \equiv \min\{a, b\}$ as defined in §1.7.

**Theorem 3.3.2 (FCLT)** Consider the sequence of $G_t/GI^o, GI^v/\infty$ IS models satisfying all assumptions in §3.2. Then,

$$\left(\check{N}, \check{X}^c(0, \cdot), \check{X}^r(0, \cdot), \tilde{X}^c, \tilde{X}^r, \tilde{D}\right) \Rightarrow \left(\check{N}, \check{X}^c(0, \cdot), \check{X}^r(0, \cdot), \tilde{X}^c, \tilde{X}^r, \tilde{D}\right)$$  \hspace{1cm} (3.14)

in $\mathbb{D}^3 \times \mathbb{D}_D^2 \times \mathbb{D}^2$ as $n \to \infty$ where the limiting stochastic process for the two-parameter ICP, the scaled number of customers in service at $t$ with age at most $y$, is

$$\hat{X}^c(t, y) = \hat{X}^{c,v}_1(t, y) + \hat{X}^{c,v}_2(t, y) + \hat{X}^{c,o}_1(t, y) + \hat{X}^{c,o}_2(t, y), \quad t, y \geq 0,$$  \hspace{1cm} (3.15)
where \( \hat{X}_e, \hat{X}_e^c, \hat{X}_e, \hat{X}_e^c \) and \( \hat{X}_e^o \) are independent zero-mean Gaussian processes with continuous sample paths,

\[
\hat{X}_e^c(t, y) = \int_{(t-y)}^t G^c_{\nu}(t-s) d\hat{N}(s), \quad (3.16)
\]

\[
\hat{X}_e(t, y) = \int_{(t-y)}^t \int_0^\infty 1(x > t-s) d\hat{U}_\nu(\Lambda(s), G^c_{\nu}(x)), \quad t, y \geq 0, \quad (3.17)
\]

where \( \hat{N} \) is the limiting process in the assumed FCLT for the arrival process specified in Assumption 3.2.1, and \( \hat{U}_\nu \) is the standard Kiefer process as in §1.8.2, capturing the variability of the new service times, and independent of \( \hat{N} \); \( \hat{X}_e^c \) is a zero-mean Gaussian process with the covariance function

\[
C_e^c((t_1, y_1), (t_2, y_2)) \equiv \text{Cov}\left( \hat{X}_e^c(t_1, y_1), \hat{X}_e^c(t_2, y_2) \right) = \int_0^{(y_1-t_1)^+\land(y_2-t_2)^+} H^c_{u}(t_1 \land t_2) H^c_{u}(t_1 \lor t_2) dX^c(0, u), \quad (3.18)
\]

and \( \hat{X}_e^o \) has the representation

\[
\hat{X}_e^o(t, y) = \int_0^{(y-t)^+} H^c_{x}(t) d\hat{X}_e^o(0, x)
= H^c_{(y-t)^+}(t)\hat{X}_e^o(0, (y-t)^+) - \int_0^{(y-t)^+} \hat{X}_e^o(0, u-)dH^c_{u}(t), \quad t, y \geq 0, \quad (3.19)
\]

where \( \hat{X}_e^o(0, \cdot) \) is the limit for the initial age process in Assumption 3.2.1 and (3.10). The joint convergence (3.14) follows from the displayed limit. The other limiting processes \( \hat{X} \), \( \hat{D} \) and \( \hat{X} \) are specified in the corollaries below.

**Remark 3.3.1** (Correction in [78]) The limits for the new input follow from [78], so the formulas in (3.16) and (3.17) should be consistent with [78]. However, here we make a correction, noting that the upper limit of the inner integrals in (2.10), (2.15) and for \( X_{e,2}^c(t, y) \) in (3.16) of [78] all should be \( \infty \) instead of \( t \). Similarly the upper limit of the second integral in the expression for \( \sigma^2_{q,e}(t, y) \) in Theorem 4.2 of [78] also should be \( \infty \) instead of \( t \). After this correction, the formulas in (3.17) and elsewhere are consistent with [78].

We next characterize the other limiting processes using the limiting process in (3.15). Recall that \( d_t \) denotes equal in distribution for each \( t \). Let \( B_s(\cdot) \) be an independent BM (associated with service times of new customers).
Corollary 3.3.1 (Limits for the one-parameter queue length process) Under the assumptions of Theorem 3.3.2, the limit for the total number in service at $t$ is

$$
\hat{X}(t) \equiv \hat{X}^e(t, \infty) \equiv \hat{X}^v_1(t) + \hat{X}^v_2(t) + \hat{X}^o_1(t) + \hat{X}^o_2(t), \quad t \geq 0,
$$

(3.20)

where $\hat{X}^v_1$, $\hat{X}^v_2$, $\hat{X}^o_1$ and $\hat{X}^o_2$ are independent zero-mean Gaussian processes with continuous sample paths and, for $t \geq 0$,

$$
\hat{X}^v_1(t) \equiv X^v_1(t, \infty) \equiv \int^t_0 G^e_\nu(t - s) d\hat{N}(s),
$$

(3.21)

$$
\hat{X}^v_2(t) \equiv X^v_2(t, \infty) \equiv \int^t_0 \int^\infty_0 1(x > t - s) d\hat{U}_\nu(\Lambda(s), G(x))
\quad \overset{d}{=} t - \int^t_0 \sqrt{G_\nu(t - s)G^{\nu}_\nu(t - s)} d\mathcal{B}_s(\Lambda(s)),
$$

(3.22)

$\hat{X}^o_1(t) \equiv \hat{X}^o_1(t, \infty)$ is a zero-mean Gaussian process with the covariance function

$$
C^o_1(t, t') \equiv \text{Cov} \left( \hat{X}^o_1(t), \hat{X}^o_1(t') \right) = C^e_1((t_1, \infty), (t_2, \infty)) = \int^{\infty}_0 H_u(t \wedge t') H^e_u(t \vee t') dX^e(0, u),
$$

(3.23)

and

$$
\hat{X}^o_2(t) \equiv \hat{X}^o_2(t, \infty) \equiv \int^{\infty}_0 H^e_x(t) d\hat{X}^e(0, x) \quad t \geq 0.
$$

(3.24)

Corollary 3.3.2 (Limits for the one-parameter service-completion process) Under the assumptions of Theorem 3.3.2, The limit for the number of service completions by time $t$ is

$$
\hat{D}(t) = \hat{D}_1^e(t) + \hat{D}_2^e(t) + \hat{D}_1^o(t) + \hat{D}_2^o(t), \quad t \geq 0,
$$

(3.25)

where $\hat{D}_1^e$, $\hat{D}_2^e$, $\hat{D}_1^o$ and $\hat{D}_2^o$ are independent zero-mean Gaussian processes, with

$$
\hat{D}_1^e(t) \equiv \int^t_0 G_\nu(t - s) d\hat{N}(s),
$$

(3.26)

$$
\hat{D}_2^e(t) \equiv \int^t_0 \int^\infty_0 1(x \leq t - s) d\hat{U}_\nu(\Lambda(s), G(x))
\quad \overset{d}{=} \int^t_0 \sqrt{G_\nu(t - s)G^{\nu}_\nu(t - s)} d\mathcal{B}_s(\Lambda(s)), \quad t \geq 0,
$$

(3.27)

$\hat{D}_1^o(t) = -\hat{X}^o_1(t)$ being a zero-mean Gaussian process with covariance function

$$
\text{Cov} \left( \hat{D}_1^o(t), \hat{D}_1^o(t') \right) = C^o_1(t, t') \quad \text{and} \quad \hat{D}_2^o(t) \equiv \hat{X}(0) - \hat{X}_2^o(t) = \int^{\infty}_0 H_x(t) d\hat{X}^e(0, x).
$$

(3.28)
Corollary 3.3.3 (Limits for the residual-service-time process) Under the assumptions of Theorem 3.3.2, the limit $\hat{X}(t,x) = \hat{X}^r(t,x) = \hat{X}_1^r(t,x) + \hat{X}_2^r(t,x)$ for all $x \geq 0$,

$$\hat{X}(t,x) = \hat{X}_1^r(t,x) + \hat{X}_1^o(t,x) + \hat{X}_2^o(t,x),$$

(3.29)

with $\hat{X}_1^r$, $\hat{X}_2^r$, $\hat{X}_1^o$ and $\hat{X}_2^o$ being independent zero-mean Gaussian processes given by

$$\hat{X}_1^r(t,x) \equiv \int_0^t G_{\nu}(t + x - s) d\hat{N}(s),$$

(3.30)

$$\hat{X}_2^r(t,x) \equiv \int_0^t \int_0^{\infty} 1(u > t + x) d\hat{U}(u)$$

(3.31)

$$\hat{X}_1^o(t,x) \equiv \hat{X}_o^1(t + x) = \hat{X}_1^e(t + x, \infty), \quad \text{and}$$

$$\hat{X}_2^o(t,x) \equiv \hat{X}_2^o(t + x) = \hat{X}_2^e(t + x, \infty), \quad t, y \geq 0.$$ 

(3.32) (3.33)

Remark 3.3.2 (The stochastic integrals) The integrals in Theorem 3.3.2 should be interpreted just as in [78], as explained in Remark 3.2 there. In particular, the deterministic integrals in (3.18) and (3.23) are Stieltjes integrals, while the integrals in (3.17), (3.22), (3.27) and (3.31) are all stochastic integrals, just as in [78]. As in Theorem 3.2 and Remark 3.3 of [78], the continuity assumption on the cdf $G_\nu$ in Assumption 3.2.2 is used to get the representation in terms of the Kiefer process.

Of special note are the stochastic integrals with respect to $\hat{N}$ in (3.16), (3.21), (3.26) and (3.30), and with respect to $\hat{X}_e(0, \cdot)$ in (3.19), (3.24) and (3.28). As explained in Remark 3.2 of [78], these all should be interpreted as the form after the representation of integration by parts, as given on p. 336 of [12]. That is justified because the prelimit processes of the integrator process have sample paths of bounded variation. For example, the alternative representation for (3.19) is given there; see §3.4.3.

Remark 3.3.3 (Four independent stochastic effects) The expression for the limiting process $\hat{X}_e$ in (3.15) as the sum of the four independent processes $\hat{X}_1^e, \hat{X}_2^e, \hat{X}_1^o$ and $\hat{X}_2^o$ shows that the four sources of variability in the model contribute to the total variability independently. The process $\hat{X}_1^e$ captures the variability in the arrival process after time 0; the process $\hat{X}_2^e$ captures the variability in the service times after time 0; the process $\hat{X}_1^o$ captures the variability in the remaining service times at time 0 given that the initial age process is around $nX_e(0, \cdot)$; and the process $\hat{X}_2^o$ captures the variability of the ages of initial customers at time 0.

In the next corollary, we make further assumptions on the limiting processes $\hat{N}$ and $\hat{X}(0, \cdot)$ to obtain simpler covariance and variance formulas.
Corollary 3.3.4 (Variance and covariance formulas for the ICP) If \( \hat{N}(t) = c_{\lambda}B_{\lambda}(\Lambda(t)) \) and \( \hat{X}^e(0, t) \equiv c_{\nu}B_{\nu}(\Theta(t)) \), \( t \geq 0 \), then the covariance function for \( \hat{X}^e \) is given by

\[
C^e((t_1, y_1), (t_2, y_2)) = \text{Cov}(\hat{X}^e(t_1, y_1), \hat{X}^e(t_2, y_2)) = C^{e, \nu}((t_1, y_1), (t_2, y_2)) + C^{e, o}((t_1, y_1), (t_2, y_2)),
\]

where

\[
C^{e, \nu}((t_1, y_1), (t_2, y_2)) \equiv \int_{(t_1-y_1)^+\wedge(y_2-t_2)^+}^{(t_1-t_2)^+\wedge(y_2-t_2)^+} \left[ (c_\nu^2 - 1)G^e_\nu(t_1 - s)G^e_\nu(t_2 - s) + G^e_\nu((t_1 \lor t_2) - s) \right] \lambda(s) ds,
\]

\[
C^{e, o}((t_1, y_1), (t_2, y_2)) \equiv \int_0^{(y_1-t_1)^+\wedge(y_2-t_2)^+} H_u(t_1 \land t_2)H^e_u(t_1 \lor t_2) dX^e(0, u) + c_o^2 \int_0^{(y-t)^+} H^e_u(t_1 \land t_2)H^e_u(t_1 \lor t_2) d\Theta(u).
\]

so that the variances are

\[
\sigma^2_e(t, y) \equiv \text{Var}(\hat{X}^e(t, y)) = \sigma^2_{e, \nu}(t, y) + \sigma^2_{e, o}(t, y),
\]

where

\[
\sigma^2_{e, \nu}(t, y) = \sigma^2_{\nu}((t - y)^+, t),
\]

\[
\sigma^2_{e, o}(u, v) \equiv \int_u^v \left[ (c_\nu^2 - 1)G^e_\nu(v - s)^2 + G^e_\nu(v - s) \right] \lambda(s) ds
\]

and

\[
\sigma^2_{e, o}(t, y) = \int_0^{(y-t)^+} H_u(t)H^e_u(t) dX^e_0(u) + c_o^2 \int_0^{(y-t)^+} H^e_u(t)^2 d\Theta(u).
\]

Corollary 3.3.5 (Variance for \( \hat{X}(t) \) and \( \hat{D}(t) \)) Under the assumptions of Corollary 3.3.4, the variances of the one-parameter processes \( \hat{X}(t) \) and \( \hat{D}(t) \) are

\[
\sigma^2_X(t) \equiv \text{Var}(\hat{X}(t)) = \sigma^2_{\hat{X}, \nu}(t) + \sigma^2_{\hat{X}, o}(t), \quad (3.34)
\]

where

\[
\sigma^2_{\hat{X}, \nu}(t) \equiv \sigma^2_{\hat{X}, \nu}(t, \infty)
\]

\[
= \sigma^2_{\nu}(0, t) = \int_0^t \left[ (c_\nu^2 - 1)G^e_\nu(t - s)^2 + G^e_\nu(t - s) \right] \lambda(s) ds \quad (3.35)
\]

and

\[
\sigma^2_{\hat{X}, o}(t) = \sigma^2_{\hat{X}, o}(t, \infty) = \int_0^{\infty} H_u(t)H^e_u(t) dX^e_0(u) + c_\nu^2 \int_0^{\infty} H^e_u(t)^2 d\Theta(u), \quad (3.36)
\]

\[
\sigma^2_D(t) \equiv \text{Var}(\hat{D}(t)) = \sigma^2_{\hat{D}, \nu}(t) + \sigma^2_{\hat{D}, o}(t), \quad (3.37)
\]

where

\[
\sigma^2_{\hat{D}, \nu}(t) = \int_0^t \left[ (c_\nu^2 - 1)G^2_\nu(t - s) + G_\nu(t - s) \right] \lambda(s) ds
\]

and

\[
\sigma^2_{\hat{D}, o}(t) = \int_0^{\infty} H_u(t)H^e_u(t) dX^e_0(u) + c_o^2 \int_0^{\infty} H^e_u(t)^2 d\Theta(u),
\]

61
Remark 3.3.4 (Additivity of the variance formulas) The first term of the variance formula of \( \hat{X}(t) (\hat{D}(t)) \sigma^2_{\hat{X},\nu}(t) (\sigma^2_{\hat{D},\nu}(t)) \) provides the variance when the system is initially empty (which coincides with the variance formula in [78]). The second term \( \sigma^2_{\hat{X},o}(t) (\sigma^2_{\hat{D},o}(t)) \) represents the variance of the content that has been in the system since time 0.

3.4 Proof of Theorem 3.3.2.

Let \( \bar{K}(\cdot,\cdot) \equiv \bar{U}(\cdot, G(\cdot)), \hat{K}(\cdot,\cdot) \equiv \hat{U}(\cdot, G(\cdot)), \) and \( \hat{R}(\cdot,\cdot) \equiv \hat{U}(\Lambda(\cdot), G(\cdot)) \) with \( \hat{U} \) being the standard Kiefer process. By Assumption 3.2.1, it remains to show the convergence

\[
(\hat{N}_n, \bar{N}_n, \hat{K}_n, \bar{K}_n, \hat{R}_n, \hat{X}_n^{e,\nu}, \hat{X}_n^{r,\nu}, \hat{D}_n^\nu) \Rightarrow (\hat{N}, \Lambda, \hat{K}, \hat{R}, \hat{X}^{e,\nu}, \hat{X}^{r,\nu}, \hat{D}^\nu) \quad (3.38)
\]

in \( D^3 \times D^5 \) as \( n \to \infty \) where \( \Lambda \) is as in (3.9), \( K(t,x) = tG(x), \) \( \hat{K}(t,x) = \hat{U}(t,G(x)) \), and \( \hat{R}(t,x) = \hat{U}(\Lambda(t),G(x)) \) with \( \hat{U} \) being the standard Kiefer process. By Assumption 3.2.1, it remains to show the convergence

\[
(\hat{X}_n(0,\cdot), \hat{X}_n(0,\cdot), \hat{X}_n^{e,o}, \hat{X}_n^{r,o}, \hat{D}^o_n) \Rightarrow (\hat{X}(0,\cdot), X(0,\cdot), \hat{X}^{e,o}, \hat{X}^{r,o}, \hat{D}^o) \quad (3.39)
\]

in \( D^3 \times D^2 \) as \( n \to \infty \). We will then have the joint convergence of (3.38) and (3.39) in \( D^7 \times D^8 \). (The joint convergence of \( \hat{X}_n \) and \( \hat{D}_n \) follows from continuous mapping theorem for addition at continuous limits.) In §3.4.1 we show that the main two-parameter process \( \hat{X}_n^{e,o} \) can be decomposed into two other two-parameter processes \( \hat{X}_n^{e,o,1}(t,y) \) and \( \hat{X}_n^{e,o,2}(t,y) \), that can be treated separately by conditioning on the ages at time 0. We establish convergence for those two processes in §§3.4.2–3.4.3; In §3.4.4 we prove the convergence of other processes.
3.4.1 Decomposition of $\hat{X}_{n}^{e,o}$.

To prove (3.39), we use a convenient representation of $\hat{X}_{n}^{e,o}(t,y)$. Let $\hat{\nu}_{n}^i(t) \equiv 1\{\tilde{\nu}_{n}^i > t\} - H^{c}_{\nu_{n}^i}(t)$. From (3.3), (3.7) and (3.13), we can write

$$
\hat{X}_{n}^{e,o}(t,y) = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{\infty} 1\{\tilde{\nu}_{n}^i > t\} - \int_{0}^{(y-t)^{+}} a(x)\tilde{G}^{c}_{x}(t)dx \right)
$$

$$
= \frac{1}{\sqrt{n}} \sum_{i=1}^{\infty} \bigg( 1\{\tilde{\nu}_{n}^i > t\} - \tilde{G}^{c}_{\nu_{n}^i}(t) \bigg)
$$

$$
+ \sqrt{n} \left( \int_{0}^{(y-t)^{+}} \tilde{G}^{c}_{x}(t)d\hat{X}_{n}^{e}(0,x) - \int_{0}^{(y-t)^{+}} a(x)\tilde{G}^{c}_{x}(t)dx \right)
$$

$$
= \frac{1}{\sqrt{n}} \sum_{i=1}^{\infty} \tilde{\nu}_{n}^i(t) + \int_{0}^{(y-t)^{+}} \tilde{G}^{c}_{x}(t)d\hat{X}_{n}^{e}(0,x)
$$

$$
\equiv \hat{X}_{n,1}^{e,o}(t,y) + \hat{X}_{n,2}^{e,o}(t,y), \quad t,y \geq 0,
$$

(3.40)

where the second equality holds by adding and subtracting $H^{c}_{\nu_{n}^i}(t)$ in the summation. Even though $\hat{X}_{n,1}^{e,o}$ and $\hat{X}_{n,2}^{e,o}$ in (3.40) are not independent, because they both involve the age sequence $\{
u_{n}^j : j \geq 1\}$, or equivalently, the counting process $X^{e}(0,\cdot)$, they are conditionally independent given $\hat{X}_{n}^{e}(0,\cdot)$. Hence, in order to treat the two terms separately we condition upon the age sequence and then uncondition. In doing so, we apply the assumed convergence in Assumption 3.2.1 together with the following lemma, which expresses the argument used in the proof of Theorem 7.6 of [81]. The spaces are different here, but the argument is the same.

**Lemma 3.4.1** Let $\{Y_{n} : n \geq 1\}$ and $Y$ be processes with sample paths in $\mathbb{D}$, and let $\{Z_{n} : n \geq 1\}$ and $Z$ be processes with sample paths in $\mathbb{D}$. Let $Y_{n}^{Z_{n}}$ $(Y^{Z})$ denote $Y_{n}$ $(Y)$ conditioned on $Z_{n}$ $(Z)$. If $Z_{n} \Rightarrow Z$ in $\mathbb{D}$ and

$$
Y_{n}^{Z_{n}} \Rightarrow Y^{Z} \quad \text{in} \quad \mathbb{D} \quad \text{whenever} \quad Z_{n} \rightarrow Z \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \rightarrow \infty \quad \text{w.p.1},
$$

(3.41)

then $Y_{n} \Rightarrow Y$ in $\mathbb{D}$ as $n \rightarrow \infty$.

We apply Lemma 3.4.1 with the initial age process $\hat{X}_{n}^{e}(0,\cdot)$ playing the role of $Z_{n}$. The required convergence in distribution holds by Assumption 3.2.1. We will then condition on the ages and assume that

$$
\hat{X}_{n}^{e}(0,\cdot) \rightarrow \hat{X}^{e}(0,\cdot) \quad \text{in} \quad \mathbb{D} \quad \text{as} \quad n \rightarrow \infty \quad \text{w.p.1}.
$$

(3.42)

It remains to establish the limit (3.41) assuming (3.42).
Given that we condition with respect to the ages and then unconditional, in order to establish the joint convergence
\[
\left( \bar{X}_{n,1}^{e,o}, \bar{X}_{n,2}^{e,o}, \bar{X}_{n}^{e}(0, \cdot), \bar{X}_{n}^{e}(0, \cdot) \right) \Rightarrow \left( \bar{X}_{1}^{e,o}, \bar{X}_{2}^{e,o}, \bar{X}_{0}^{e}, X_{0}^{e} \right) \quad \text{in } \mathbb{D}_{\mathbb{D}}^{2} \times \mathbb{D}^{2} \quad \text{as } n \to \infty,
\]

it suffices to prove \( \left( \bar{X}_{n,1}^{e,o}, \bar{X}_{n}^{e}(0, \cdot) \right) \Rightarrow \left( \bar{X}_{1}^{e,o}, X_{0}^{e} \right) \) in \( \mathbb{D}_{\mathbb{D}} \times \mathbb{D} \), and \( \left( \bar{X}_{n,2}^{e,o}, \bar{X}_{n}^{e}(0, \cdot), \bar{X}_{n}^{e}(0, \cdot) \right) \Rightarrow \left( X_{2}^{e,o}, X_{D}^{e}, X_{0}^{e} \right) \) in \( \mathbb{D} \times \mathbb{D} \) as \( n \to \infty \), i.e., it suffices to treat the two terms separately. Aside from the conditioning, we would be using Theorems 11.4.4 and 11.4.5 in [102], which justify joint convergence. We next separately prove the convergence of two terms in (3.40).

### 3.4.2 Convergence for the First Term in (3.40).

In addition to the conditioning discussed above, we use compactness approach (see, e.g., [102]) to prove (3.41) in order to establish convergence for the first term in (3.40); i.e., we prove in two steps convergence for the fdds in \( \mathbb{D} \) to prove (3.41) in order to establish convergence for the first term in (3.40); i.e., we prove in two steps convergence for the fdds in \( \mathbb{D} \), and then, we prove tightness in the third step. In Step 1 (§3.4.2), we establish convergence for the four-parameter covariance functions of \( \bar{X}_{n,1}^{e,o} \), referred to as \( K_{n}(t, y, t', y') \), to those of \( \bar{X}_{1}^{e,o} \), defined as \( K(t, y, t', y') \) in (3.18) in Theorem 3.3.2. In Step 2 (§3.4.2), using the convergence of the covariance functions, we establish convergence for the fdds of \( \bar{X}_{n,1}^{e,o} \) in \( \mathbb{D} \), which is equivalent to the joint convergence of \( \left( \bar{X}_{n,1}^{e,o}(t_{1}, \cdot), \ldots, \bar{X}_{n,1}^{e,o}(t_{k}, \cdot) \right) \) in \( \mathbb{D}^{k} \), for all \( k \geq 1 \) and \( 0 < t_{1} < \cdots < t_{k} \). We do this in two sub-steps: First, we show the convergence of the fdds of the vector \( \left( \bar{X}_{n,1}^{e,o}(t_{1}, \cdot), \ldots, \bar{X}_{n,1}^{e,o}(t_{k}, \cdot) \right) \) in the second argument, namely, joint convergence of the bigger vector \( \left( \bar{X}_{n,1}^{e,o}(t_{i}, y_{j}), 1 \leq i \leq k, 1 \leq j \leq m \right) \) in \( \mathbb{R}^{k \times m} \), for all \( m \geq 1 \) and \( 0 < y_{1} < \cdots < y_{m} \). Second, we establish the tightness of \( \left( \bar{X}_{n,1}^{e,o}(t_{1}, \cdot), \ldots, \bar{X}_{n,1}^{e,o}(t_{k}, \cdot) \right) \) in \( \mathbb{D}^{k} \).

In Step 3 (§3.4.2) we prove that the sequence \( \{ \bar{X}_{n,1} \}_{n \geq 1} \) is tight in \( \mathbb{D} \).

**Step 1: Convergence for the covariance functions.**

As indicated above, we start by conditioning on ages in service. Let \( \mathbb{E}^{v} \) denote the conditional expectation operator, conditional on the ages, or equivalently, on the process \( X_{n}^{e}(0, \cdot) \). When conditioned on ages, the first term in (3.40) is the nonrandom sum of the independent zero-mean random variables \( \tilde{v}_{n,i}^{a}(t) \) defined at the beginning of §3.4.1. Hence,

\[
\mathbb{E}^{v} \left[ \bar{X}_{n,1}^{e,o}(t, y) \bar{X}_{n,1}^{e,o}(t', y') \right] = \frac{1}{n} \sum_{i=1}^{X_{n}^{e}(0, (y-t)^+ \wedge (y'-t')^+)} \mathbb{E}^{v} \left[ \tilde{v}_{n,i}^{a}(t) \tilde{v}_{n,i}^{a}(t') \right]
\]

\[
= \frac{1}{n} \sum_{i=1}^{X_{n}^{e}(0, (y-t)^+ \wedge (y'-t')^+)} H_{\bar{v}_{n,i}^{a}(t)}(t) H_{\bar{v}_{n,i}^{a}(t')} H_{u}(t) d\bar{X}_{n}^{e}(0, u) = \int_{0}^{(y-t)^+ \wedge (y'-t')^+} H_{u}(t) H_{u}(t') d\bar{X}_{n}^{e}(0, u). \tag{3.43}
\]
Assuming (3.42), which can be considered as convergence of finite measures, we have

\[ E^v \left[ \hat{X}^{e,o}_{n,1}(t, y) \hat{X}^{e,o}_{n,1}(t', y') \right] \to \int_0^{(y-t)^+ \wedge (y'-t')^+} H_u(t)H_u^c(t') \, dX(0, u) \equiv K(t, y, t', y'), \quad (3.44) \]

as \( n \to \infty \), from (3.43), and by the continuity of the integral (see (2.1) of §3.2 in [102]).

We now continue to prove convergence of the covariance functions. Since the random variables in (3.44) are bounded as in Assumption 3.2.3, i.e, \( \bar{X}_n(0) \leq X \uparrow < \infty \) for all \( n \), we have convergence of the means associated with the convergence in (3.44). This implies convergence of the covariance functions after unconditioning, i.e.,

\[ K_n(t, y, t', y') \equiv E^v \left[ \hat{X}^{e,o}_{n,1}(t, y) \hat{X}^{e,o}_{n,1}(t', y') \right] = E \left[ \int_0^{(y-t)^+ \wedge (y'-t')^+} H_u(t)H_u^c(t') \, dX_n(0, u) \right] \to \int_0^{(y-t)^+ \wedge (y'-t')^+} H_u(t)H_u^c(t') \, dX(0, u) \equiv K(t, y, t', y') \quad \text{as} \quad n \to \infty. \quad (3.45) \]

As an immediate consequence of (3.45), we have a formula for the sequence of variance functions, and have the limiting variance

\[ \sigma^2_n(t, y) \equiv K_n(t, y, t, y) = E \left[ \int_0^{(y-t)^+} H_u(t)H_u^c(t) \, d\bar{X}_n(0, u) \right] \to \sigma^2(t, y) = \int_0^{(y-t)^+} H_u(t)H_u^c(t) \, dX(0, u) \quad \text{as} \quad n \to \infty. \quad (3.46) \]

**Step 2: Convergence of the fdds in \( \mathbb{D}_B \).**

We again apply Lemma 3.4.1 by assuming (3.42). Hence, for each \( n \), we condition on ages in service.

**Step 2a: Joint convergence in \( \mathbb{R}^{k \times m} \).** Fix \( m \geq 1 \) and \( 0 < y_1 < \ldots < y_m \). The convergence for the fdds of the vector \( \left( \hat{X}^{e,o}_{n,1}(t_1, \cdot), \ldots, \hat{X}^{e,o}_{n,1}(t_k, \cdot) \right) \) in the second argument is equivalent to joint convergence of the bigger vector \( \left( \hat{X}^{e,o}_{n,1}(t_i, y_j), 1 \leq i \leq k, 1 \leq j \leq m \right) \) in \( \mathbb{R}^{k \times m} \). By the Cramér-Wold device (see, e.g., Theorem 4.3.3 of [102]), this is equivalent to showing that, for all \( \{a_{i,j}\} \in \mathbb{R}, \ i = 1, \ldots, k \) and \( j = 1, \ldots, m \),

\[ \sum_{i=1}^k \sum_{j=1}^m a_{i,j} \hat{X}^{e,o}_{n,1}(t_i, y_j) \Rightarrow \sum_{i=1}^k \sum_{j=1}^m a_{i,j} \hat{X}^{e,o}_1(t_i, y_j) \overset{d}{=} \mathcal{N}(0, \Sigma) \quad \text{in} \quad \mathbb{R} \quad (3.47) \]
as \( n \to \infty \) where the variance of the limit is

\[
\Sigma \equiv \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{i'=1}^{k} \sum_{j'=1}^{m} a_{i,j} a_{i',j'} K(t_i, y_j, t_{i'}, y_{j'}).
\] (3.48)

To establish (3.47), we define the random variables

\[
\tilde{X}_{n,i} \equiv \frac{1}{\sqrt{n}} \hat{\nu}_n(t_i) \quad \text{and} \quad \tilde{Y}_{n,i} \equiv \sum_{i=1}^{k} \sum_{j=1}^{m} a_{i,j} \tilde{X}_{n,i} \mathbf{1}(l \leq X_n(0, (y_j - t_i)^+)).
\]

Since \( \tilde{Y}_{n,j}, j \geq 1 \), are independent random variables, conditioned on \( \{X_n^e(0, \cdot) : n \geq 1\} \), we can rewrite the left-hand side of (3.47) as

\[
\tilde{S}_n \equiv \sum_{i=1}^{k} \sum_{j=1}^{m} a_{i,j} \tilde{X}_{n,i}^e(t_i, y_j) = \sum_{i=1}^{k} \sum_{j=1}^{m} a_{i,j} \sum_{l=1}^{\tilde{X}_n(0, (y_j - t_i)^+)} \tilde{X}_{n,i} = \sum_{l=1}^{\tilde{X}_n(0,M)} \tilde{Y}_{n,l},
\]

where \( M \equiv \max\{((y_j - t_i)^+ : 1 \leq i \leq k, 1 \leq j \leq m\} \). By the final expression above, \( \tilde{S}_n \) is the sum of independent random variables. Of course, the summands \( \tilde{Y}_{n,l} \) and the index \( X_n^e(0,M) \) both depend on \( n \), but they do so in a regular way because, to apply Lemma 3.4.1, we are assuming that (3.42) holds. For example, this means that \( n^{-1} X_n^e(0,M) \to X^e(0,M) \leq X^+ < \infty \).

Hence, we can now apply CLT for double sequences (triangular array) of non-identically-distributed independent random variables, see e.g., Theorem 7.2.4 of [33]. The variance of \( \tilde{S}_n \) is

\[
\tilde{s}_n^2 \equiv \text{Var}(\tilde{S}_n)) = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{i'=1}^{k} \sum_{j'=1}^{m} a_{i,j} a_{i',j'} K_n(t_i, y_j, t_{i'}, y_{j'})
\]

\[
\to \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{i'=1}^{k} \sum_{j'=1}^{m} a_{i,j} a_{i',j'} K(t_i, y_j, t_{i'}, y_{j'}) \equiv \Sigma,
\] (3.49)

as \( n \to \infty \) where \( \Sigma \) is defined in (3.48), and the convergence follows from (3.45). It remains to verify the Lindeberg conditions (see (2.1)-(2.2) on p.330 of [33]). But, since \( \{\tilde{X}_{n,i}\} \) take values in the interval \([-1/\sqrt{n}, 1/\sqrt{m}] \) and the variance \( \tilde{s}_n \) converges to \( \Sigma \) in (3.49) as \( n \to \infty \), the Lindeberg condition is satisfied. Therefore, by Theorem 7.2.4 of [33], if (3.42) holds, then

\[
\tilde{S}_n/\tilde{s}_n \Rightarrow \mathcal{N}(0,1) \quad \text{as} \quad n \to \infty,
\] (3.50)

which, together with (3.49), imply the desired convergence in (3.47) under the condition (3.42). Lemma 3.4.1 then proves the unconditional convergence.
Step 2b: Tightness in $D^k$. We now establish the tightness of the vectors

$$(\hat{X}_{n,1}^{e,o}(t_1, \cdot), \ldots, \hat{X}_{n,1}^{e,o}(t_k, \cdot))$$ in $D^k$,

again by assuming (3.42). This is equivalent to proving the tightness of each component $\hat{X}_{n,1}^{e,o}(t_i, \cdot)$ in $D$, for all $1 \leq i \leq k$, by Theorem 11.6.7. of [102]. We prove the tightness of the components by proving a stronger result. Paralleling §7 of [62], we show, for each $t \geq 0$, that

$$\hat{X}_{n,1}^{e,o}(t, y) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{nX_n^{e}(0,y-t)} \tilde{\gamma}_{n,i}(t) \Rightarrow B(\sigma_t^2(y))$$ in $D$ as $n \to \infty$ \quad (3.51)

where $B$ is the standard Brownian motion. To prove this, we observe that, conditional on the ages, the processes $\{\hat{X}_{n,1}^{e,o}(t, y) : y \geq 0\}$ for fixed $t$ are martingales with respect to the natural filtration augmented by the age sequence $\{\tilde{\gamma}_{n,i}\}$ since the function $y \mapsto \bar{X}_{n}(0,y-t)$ is strictly increasing for each $n \geq 1$. Then we can apply the martingale FCLT in Theorem 7.1.4 of [20], by exploiting the fact that the summands are independent $[-1,1]$-valued zero-mean random variables. The first variance function in (3.46), for fixed $t$, is the quadratic variation of $\{\hat{X}_{n,1}^{e,o}(t, y) : y \geq 0\}$, and converges to the second variance formula in (3.46). Consequently, we have the function $\sigma_t^2(y)$ in (3.51) for each fixed $t$. Note that the weak convergence of the components implies the tightness of each component.

Step 3: Proof of $C$-tightness of $\{\hat{X}_{n,1}^{e,o}\}$ in $D_D$.

To complete the proof of the convergence for the first term of (3.40) under condition (3.42), we next show that the sequence $\{\hat{X}_{n,1}^{e,o}\}$ is $C$-tight in $D_D$. To do so, we verify the following two conditions: (i) stochastic boundedness, and (ii) asymptotically negligible oscillations, as in Theorem 6.2 of [78]. The specific conditions we establish are (3.52) and (3.62) below.

**Verifying condition (i): Stochastic Boundedness.** Let $P^u$ and $E^u$ be the conditional probability and expectation given the ages $\{\nu_{n}^{u}\}$. It suffices to show, under condition (3.42), that for all $\epsilon > 0$, there exists $c > 0$ such that

$$P^u(\|\hat{X}_{n,1}^{e,o}\|_{T,y} > c) \leq \epsilon \quad \text{for all} \quad n \geq 1,$$ \quad (3.52)

where $\|\hat{X}_{n,1}^{e,o}\|_{T,y} = \sup_{(t,y) \in [0,T] \times [0,y]} |\hat{X}_{n,1}^{e,o}(t, y)|$.

To bound the probability in (3.52), we apply Chernoff's inequality (e.g., see Lemma 3.1.1.}
of [33]), obtaining, for \( r > 0 \),
\[
\mathbb{P}^\nu(\|\hat{X}_{n,1}^{\nu,e,o}\|_{T,y^1} > c) \leq e^{-rc} \mathbb{E}^\nu \left[ e^{r\|\hat{X}_{n,1}^{\nu,e,o}\|_{T,y^1}} \right],
\]  
(3.53)

To bound the right side of (3.53), we follow the symmetrization argument used in the proof of inequality 3 on p.820 of [95] (which in turn follows Lemma 1.1 of [74]). Let the sequence \( \{v_{n,i}^* : i \geq 1\} \) be an independent copy of \( \{\bar{v}_{n,i}^* : i \geq 1\} \) conditional on the ages \( \{v_{n,i}^* : i \geq 1\} \), and let \( \{\xi_i : i \geq 1\} \) be i.i.d. random variables, independent of \( \{\bar{v}_{n,i}^* : i \geq 1\} \), with \( P(\xi_i = 1) = P(\xi_i = -1) = 1/2 \). Also let \( \hat{X}_{n,1}^{\nu,e,o} \) be \( \hat{X}_{n,1}^{\nu,e,o} \) with \( \{\bar{v}_{n,i}^* : i \geq 1\} \) replaced by \( \{\tilde{v}_{n,i}^* : i \geq 1\} \). Let \( \mathbb{E}^\nu_{\tilde{v}_{n,i}} \mathbb{E}_{\xi} \mathbb{E}_{\lambda} \) denote the expectation with respect to \( \{\xi_i\} \) conditioned on \( \{\bar{v}_{n,i}^*\}, \{\tilde{v}_{n,i}^*\} \) and \( \{\bar{v}_{n,i}^*\} \); let \( \mathbb{E}^\nu_{\tilde{v}_{n,i}} \mathbb{E}_{\xi} \mathbb{E}_{\lambda} \), denote the expectation with respect to \( \{\tilde{v}_{n,i}^*\} \) conditioned on \( \{\tilde{v}_{n,i}^*\} \) and \( \{\bar{v}_{n,i}^*\} \); and so on.

The next two lemmas will be used in the proof.

**Lemma 3.4.2** Let \( X \) and \( X^* \) be two i.i.d. elements in the space \( \mathbb{D}_D \). Suppose a function \( \phi \) is convex and nondecreasing with domain \( [0, \infty) \). Then
\[
\mathbb{E} \left[ \phi(\|X - \mathbb{E}[X]\|_{T,y^1}) \right] \leq \mathbb{E} \left[ \phi(\|X - X^*\|_{T,y^1}) \right].
\]  
(3.54)

We omit the proof because it is similar to that of Lemma (A.14.15) in [95], but we do give the proof in the appendix.

**Lemma 3.4.3** Let \( \zeta_i \) and \( w_i \) be real-valued numbers with \( w_i \geq 0 \) for \( 1 \leq i \leq N \) and let \( \{w_i, 1 \leq i \leq N\} \) be the order statistics of \( \{w_i, 1 \leq i \leq N\} \) with \( w_{(i)} \leq w_{(i+1)} \) for \( 1 \leq i \leq N - 1 \). For \( T > 0 \),
\[
\sum_{i=1}^{N} \zeta_i \mathbf{1}(w_{(N-i+1)} > t) \leq \max_{1 \leq j \leq N} \left| \sum_{i=1}^{j} \zeta_i \right|, \quad 0 \leq t \leq T.
\]  
(3.55)

**Proof.** We partition the interval \( [0, T] \) into disjoint intervals with end points \( 0 = w_{(0)} \leq w_{(1)} \land T \leq \cdots \leq w_{(N)} \land T \leq w_{(N+1)} = T \). We have
\[
\left| \sum_{i=1}^{N} \zeta_i \mathbf{1}(w_{(N-i+1)} > t) \right| = \left| \sum_{j=1}^{N} \left( \sum_{i=1}^{j} \zeta_i \right) \mathbf{1}(w_{(N-j)} \land T \leq t < w_{(N-j+1)} \land T) \right|
\leq \sum_{j=1}^{N} \left( \sum_{i=1}^{j} \zeta_i \right) \mathbf{1}(w_{(N-j)} \land T \leq t < w_{(N-j+1)} \land T)
\leq \max_{1 \leq j \leq N} \left| \sum_{i=1}^{j} \zeta_i \right|.
\]
We now continue to bound the right side of (3.53). For that purpose, define

\[
\Gamma_{n,i}(t, y) = \frac{1}{\sqrt{n}} \sum_{i=1}^{X_n(0,(y-t)^+)} \xi_i \cdot 1(\bar{v}_{-i}^n > t) \quad \text{and} \\
\Gamma_{n,i}^*(t, y) = \frac{1}{\sqrt{n}} \sum_{i=1}^{X_n(0,(y-t)^+)} \xi_i \cdot 1(\bar{v}_{-i}^{*,n} > t). 
\]  

Let \( N_n^*(y, t) \equiv X_n(0, (y - t)^+) \) and \( \{\bar{v}_{(i)}^n, 1 \leq i \leq N_n^*(y, t)\} \) be the order statistics of \( \{\bar{v}_{-i}^n : 1 \leq i \leq N_n^*(y, t)\} \) so that \( \bar{v}_{(N_n^*(y, t) - i)}^{n} \) is the \( i \)th largest one.

With that preparation, we can write (explanation given afterwards)

\[
\mathbb{E}^\nu \left[ \exp \left( r \| \bar{X}_{n,1}^o - \bar{X}_{n,1}^{*,o} \|_{T,y^+} \right) \right] \leq \mathbb{E}^\nu \left[ \exp \left( r \| X_{n,1} - X_{n,1}^* \|_{T,y^+} \right) \right]
\]

\[
= \mathbb{E}^\nu \left[ \exp \left( r \| n^{-1/2} \sum_{i=1}^{N_n^*(y,t)} (1(\bar{v}_{-i}^n > t) - 1(\bar{v}_{-i}^{*,n} > t)) \|_{T,y^+} \right) \right]
\]

\[
= \mathbb{E}^\nu \left[ \exp \left( r \| n^{-1/2} \sum_{i=1}^{N_n^*(y,t)} \xi_i (1(\bar{v}_{-i}^n > t) - 1(\bar{v}_{-i}^{*,n} > t)) \|_{T,y^+} \right) \right]
\]

\[
\leq \mathbb{E}^\nu \left[ \exp \left( r \| \Gamma_{n,i}(t, y) \|_{T,y^+} + r \| \Gamma_{n,i}^*(t, y) \|_{T,y^+} \right) \right]
\]

\[
= \mathbb{E}^\nu \mathbb{E}_\xi^\nu \left[ \exp \left( r \| \Gamma_{n,i}(t, y) \|_{T,y^+} + r \| \Gamma_{n,i}^*(t, y) \|_{T,y^+} \right) \right]
\]

\[
\leq \mathbb{E}^\nu \left[ \mathbb{E}_\xi^\nu \left[ \exp \left( 2r \| \Gamma_{n,i}(t, y) \|_{T,y^+} \right) \right] \right]^{1/2} \cdot \mathbb{E}^\nu \left[ \mathbb{E}_\xi^\nu \left[ \exp \left( 2r \| \Gamma_{n,i}^*(t, y) \|_{T,y^+} \right) \right] \right]^{1/2}
\]

\[
= \mathbb{E}^\nu \mathbb{E}_\xi^\nu \left[ \exp \left( 2r \| \Gamma_{n,i}(t, y) \|_{T,y^+} \right) \right]
\]

\[
= \mathbb{E}^\nu \mathbb{E}_\xi^\nu \left[ \exp \left( 2r \| n^{-1/2} \sum_{i=1}^{N_n^*(y,t)} \xi_i (1(\bar{v}_{(N_n^*(y,t) - i)}^{n} > t) \|_{T,y^+} \right) \right]
\]

\[
\leq \mathbb{E}^\nu \left[ \exp \left( 2r \sup_{(t,y) \in [0,T] \times [0,y^+]} \left\{ \max_{1 \leq j \leq N_n^*(y,t)} \left| n^{-1/2} \sum_{i=1}^{j} \xi_i \right| \right\} \right) \right]
\]

\[
\leq \mathbb{E}^\nu \left[ \exp \left( 2r \max_{1 \leq j \leq X_n(0)} \left| n^{-1/2} \sum_{i=1}^{j} \xi_i \right| \right) \right] 
\]

(3.57)

where the first inequality holds by Lemma 3.4.2; the first equality holds because the centering terms cancel out; the second equality holds because

\[
\xi_i \left( 1(\bar{v}_{-i}^n > t) - 1(\bar{v}_{-i}^{*,n} > t) \right) \overset{d}{=} 1(\bar{v}_{-i}^n > t) - 1(\bar{v}_{-i}^{*,n} > t);
\]
the second inequality follows by (3.56) and the triangle inequality; the third equality holds by conditioning on the $\tilde{\nu}$ and $\tilde{\nu}^*$; the third inequality holds by applying the Cauchy-Schwarz inequality on the expectation $E_{\xi}^\nu$; the fourth equality holds because $\{\tilde{\nu}_i^n\}$ and $\{\tilde{\nu}^*_i^n\}$ are two i.i.d. copies; the fifth equality holds because the two sequences $\{\xi_i\}$ and $\{\tilde{\nu}_i^n\}$ are independent; the fourth inequality holds by Lemma 3.4.3; and the last inequality holds because $t$ and $y$ appear only in the upper limit of the inner maximum $N^*_n(y, t)$, which itself is bounded above by $X_n(0)$.

To bound (3.57), we apply the integration by parts on p. 150 of [23] to write the moment generating function of a nonnegative random variable $Y$ as

\[
E\left[e^{\theta Y}\right] = 1 + \int_0^\infty \theta e^{\theta x} P(Y \geq x) \, dx.
\] (3.58)

Therefore we next provide an upper bound on the tail probability using Lévy’s inequality (e.g., Theorem 3.7.1 on p. 138 of [33] (also Theorem B.2.1 in the appendix); in particular

\[
P^\nu \left( \max_{1 \leq j \leq X_n(0)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^j \xi_i \right| \geq x \right) \leq 2P^\nu \left( \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{X_n(0)} \xi_i \right| \geq x \right) \leq 4e^{-\frac{x^2}{2X_n(0)}},
\] (3.59)

where the second inequality follows from Hoeffding’s inequality (e.g., Theorem 3.1.3 on p. 120 of [33]; also Theorem B.2.2 in the appendix).

Combining (3.57)–(3.59) with $\theta = 2r$ and $Y \equiv \max_{1 \leq j \leq X_n(0)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^j \xi_i \right|$ yields that

\[
E^\nu \left[ \exp \left( r \| \hat{X}_{n,1}^{e,o} \|_{T,y} \right) \right] \leq 1 + 4 \int_0^\infty 2re^{2rx} e^{-\frac{x^2}{2X_n(0)}} \, dx
\]

\[= 1 + 8r \sqrt{2\pi X_n(0)} e^{2r^2X_n(0)} \int_0^\infty (2\pi X_n(0))^{-1/2} e^{-\frac{(x-2rX_n(0))^2}{2X_n(0)}} \, dx
\]

\[= 1 + 8r \sqrt{2\pi X_n(0)} e^{2r^2X_n(0)} \Phi \left( 2r \sqrt{X_n(0)} \right) \leq 1 + 8r \sqrt{2\pi X_n(0)} e^{2r^2X_n(0)}
\] (3.60)

where $\Phi$ is the cdf of the standard normal distribution. Here we assume, without loss of generality, that $X_n(0) > 0$ because (3.57) becomes 0 if $X_n(0) = 0$. Letting $r = 1/\sqrt{X_n(0)}$ in (3.60) and applying (3.53), we have

\[
P^\nu(\| \hat{X}_{n,1}^{e,o} \|_{T,y} > c) \leq \left( 1 + 8\sqrt{2\pi e^2} \right) E^\nu \left[ \exp \left( -\frac{c}{\sqrt{X_n(0)}} \right) \right]
\]

\[\leq \left( 1 + 8\sqrt{2\pi e^2} \right) \exp \left( -\frac{c}{\sqrt{X_n(0)}} \right),
\]

which converges to 0 as $c \to \infty$.

**Verifying condition (ii): asymptotically negligible oscillations.** We show that the
oscillations are asymptotically negligible, again by assuming (3.42). For that purpose, consider
an arbitrary sequence of uniformly bounded stopping times \( \{ \kappa_n \} \) with respect to the natural
filtration \( F_n \equiv \{ F_n(t), t \in [0, \infty) \} \) \( \vee \mathcal{N} \) where

\[
F_n(t, y) \equiv \sigma\{ 1(\vec{\nu}^e_i > x) : 1 \leq i \leq X^n_0(0, y), x \geq t, 0 \leq s \leq y \} \\
\phantom{F_n(t, y) =} \vee \sigma\{ X^n_0(0,(y-x)^+), x \geq t, 0 \leq s \leq y \},
\]

\[
F_n(t) = \bigvee_{y \geq 0} F_n(t, y),
\]

(3.61)

and \( \mathcal{N} \) being all the null sets. We will show that, for any \( \delta > 0 \) and \( \epsilon > 0 \), and for any such
sequence of stopping times \( \{ \kappa_n \} \),

\[
\lim_{\delta \downarrow 0} \lim_{n \to \infty} \sup_{\kappa_n} \mathbb{E}^\nu \left[ \left( \sup_{y \geq 0} \left| \tilde{X}^{e,o}_{n,1}(\kappa_n + \delta, y) - \tilde{X}^{e,o}_{n,1}(\kappa_n, y) \right| \right)^2 \right] = 0,
\]

(3.62)

which is a sufficient condition for condition \((ii)\) in Theorem 6.2 of [78].

To establish (3.62), we condition on the sequence \( \{ \kappa_n \} \) as well as the sequence \( \{ \nu^n_{-i} \} \). As in
Step 2b, conditional on the sequences \( \{ \kappa_n \} \) and \( \{ \nu^n_{-i} \} \), the process \( \{ \tilde{X}^{e,o}_{n,1}(t, y) : y \geq 0 \} \) with \( t \)
fixed is an adapted martingale with respect to \( \tilde{F}^t_\equiv \bigvee_{y \geq 0} F^t_n(y) \vee \{ \kappa_n \} \vee \{ \nu^n_{-i} \} \), where \( F^t_n(y) \)
denotes the \( \sigma \)-algebra \( F_n(t, y) \) in (3.61) with \( t \) being fixed. Consequently, with the conditioning,
the process \( \tilde{X}^{e,o}_{n,1}(\kappa_n + \delta, y) - \tilde{X}^{e,o}_{n,1}(\kappa_n, y) : y \geq 0 \) is an \( \tilde{F}^{\kappa_n+\delta}_n \)-adapted martingale. Let \( \mathbb{E}^{\nu,\kappa} \)
denote that the expectation is computed by conditioning on \( \{ \nu^n_{-i} \} \) and \( \{ \kappa_n \} \), let \( \text{Var}^{\nu,\kappa} \) be the
conditional variance. Then, by Doob’s maximal inequality,

\[
\mathbb{E}^{\nu,\kappa} \left( \sup_{y \geq 0} \left| \tilde{X}^{e,o}_{n,1}(\kappa_n + \delta, y) - \tilde{X}^{e,o}_{n,1}(\kappa_n, y) \right| \right)^2 \leq 4 \sup_{y \geq 0} \mathbb{E}^{\nu,\kappa} \left( \tilde{X}^{e,o}_{n,1}(\kappa_n + \delta, y) - \tilde{X}^{e,o}_{n,1}(\kappa_n, y) \right)^2
\]

\[
= 4 \sup_{y \geq 0} \text{Var}^{\nu,\kappa} \left( X^n_0(0,(y - \kappa_n - \delta)^+) \sum_{i=1} X^n_0(0,(y - \kappa_n + \delta)^+) \sum_{i=1} 1(\vec{\nu}^e_i > \kappa_n + \delta) - \sum_{i=1} 1(\vec{\nu}^e_i > \kappa_n) \right)
\]

\[
\leq 4 \sup_{y \geq 0} \text{Var}^{\nu,\kappa} \left( \sum_{i=1} X^n_0(0,(y - \kappa_n)^+) 1(\kappa_n < \vec{\nu}^e_i \leq \kappa_n + \delta) \right)
\]

\[
= 4 \sup_{y \geq 0} \sum_{i=1} X^n_0(0,(y - \kappa_n)^+) \tilde{G}^\delta_{\nu^n_{-i}}(\kappa_n)(1 - \tilde{G}^\delta_{\nu^n_{-i}}(\kappa_n))
\]

\[
= 4 \int_0^\infty \tilde{G}^\delta_u(\kappa_n)(1 - \tilde{G}^\delta_u(\kappa_n)) dX^n_u(0, u)
\]

\[
\leq 4 \int_0^\infty \tilde{G}^\delta_u(M) dX^n_u(0, u) \leq \int_0^\infty \frac{4g^\delta_T}{\tilde{G}^\delta_u(T + y^u)} dX^n_u(0, u) = \frac{4g^\delta_T}{\tilde{G}^\delta_u(T + y^u)} X^n(0),
\]

(3.63)
where \( \tilde{G}_u(t) \equiv \tilde{G}_u(t + \delta) - \tilde{G}_u(t) \), \( M = \sup_{n \geq 1} |\kappa_n| \), \( g_o \) is the pdf of the service-time distribution \( G_o \), and \( g_o^\uparrow \equiv \sup_{M \leq u \leq M + \delta} |g_o(u)| \). The first equality holds since the sums are zero-mean random variables conditioned on \( \{\nu_n\} \), \( \{\kappa_n\} \) for all \( t \geq 0, y \geq 0 \), whereas the second equality holds due to (conditional) independence. Starting from the third equality, \( \{\nu_n\} \) and \( \{\kappa_n\} \) are necessarily treated as deterministic sequences.

Taking expectation on both sides of (3.63) to uncondition implies that

\[
E^v \left( \sup_{y \geq 0} \left| \hat{X}^{e,o}_{n,1}(\kappa_n + \delta, y) - \hat{X}^{e,o}_{n,1}(\kappa_n, y) \right| \right)^2 \leq 4g_o^\uparrow \delta E \left[ \hat{X}_n(0) \right] \leq 4g_o^\uparrow X^{\uparrow \delta} \to 0,
\]

as \( \delta \to 0 \). Therefore, the condition in (3.62) is satisfied. It would still be satisfied if we took a further expectation to uncondition on the ages.

### 3.4.3 Convergence for the Second Term in (3.40).

In this section, we establish convergence for the second term in (3.40), i.e., \( \hat{X}^{e,o}_{n,2} \Rightarrow \hat{X}^{e,o}_2 \) in \( D([0, \infty); \mathbb{R}) \) as \( n \to \infty \), again by conditioning on the ages, and assuming that (3.42) holds so that we can apply Lemma 3.4.1. Since \( \hat{X}^{e}_n(0,-) \) is of bounded variation, the second term in (3.40) can be expressed as a Stieltjes integral. Therefore, we can use the integration by parts formula given on p.336 of [12] to obtain an equivalent representation

\[
\hat{X}^{e,o}_{n,2}(t, y) \equiv -\int_0^{(y-t)^+} H^{e}_{u}(t)d\hat{X}^{e}_{n}(0, u)
\]

\[
= H^{e}_{(y-t)^+}(t)\hat{X}^{e}_{n}(0, (y-t)^+) - \int_0^{(y-t)^+} \hat{X}^{e}_{n}(0, u-)dH^{e}_{u}(t), \quad t, y \geq 0.
\]

Since we are conditioning on the ages, everything in (3.64) is deterministic. Hence, we will show that the convergence follows by continuity (convergence preservation of mappings). The mapping is a measurable mapping that is continuous almost surely with respect to continuous limits. Measurability in this setting holds because the Borel \( \sigma \)-field induced by the usual topology on \( D_\mathbb{D} \) coincides with the usual Kolmogorov \( \sigma \)-field generated by the coordinate projections; see §11.5.3 of [102] and references cited there. (Hence standard measurability arguments can be used.) If we uncondition, then we would be applying the continuous mapping theorem in Theorem 3.4.3 of [102].

The next two lemmas allow us to establish the desired convergence. We first show that the function \( H^e_{x}(t) \) has finite variation in \( x \) over a bounded interval, by virtue of the Assumption 3.2.2 on the service-time cdf \( G \).

**Lemma 3.4.4 (Finite total variation in \( x \) for \( H^e_{x}(t) \) in bounded intervals)**
In an interval \([0, T^*]\), \(\int_0^{T^*} |dH^c_x(t)| < \infty\), for \(t \geq 0\).

**Proof.** Taking the derivative of \(H^c_x(t)\) with respect to \(x\) yields

\[
\int_0^{T^*} |dH^c_x(t)| = \int_0^{T^*} \frac{|g(t + x)G^c(x) - g(x)G^c(t + x)|}{(G^c(x))^2} \, dx
\]

\[
< \left\{ \frac{g^+}{G^c(T^*)} + \frac{g^+}{(G^c(T^*))^2} \right\} T^* \equiv K(T^*) < \infty. \tag{3.65}
\]

We next establish continuity in the uniform metric over compact subsets of the domain. Let

\[d_u(x_1, x_2) \equiv \sup_{t \in [0, T]} |x_1(t) - x_2(t)|\]

for \(x_1, x_2 \in \mathbb{D}\) and

\[d_u(y_1, y_2) \equiv \sup_{(t, u) \in [0, T] \times [0, \infty)} |y_1(t, u) - y_2(t, u)|\]

for \(y_1, y_2 \in \mathbb{D}_u\).

**Lemma 3.4.5** The mapping \(\phi : (\mathbb{D}, d_u) \rightarrow (\mathbb{D}_u, d_u)\) defined by

\[
\phi(x)(t, y) = H_{(y-t)^+}(t)x((y - t)^+) - \int_0^{(y-t)^+} x(s-)dH_s(t)
\]

for \(0 \leq t \leq y\) is continuous in \(\mathbb{D}_u\).

**Proof.** Let \(\{x_n\}\) be a sequence such that \(d_u(x_n, x) \equiv \|x_n - x\|_{y_0} \rightarrow 0\) as \(n \rightarrow \infty\). Then

\[
|\phi(x_n)(t, y) - \phi(x)(t, y)|
\]

\[
\leq H_{(y-t)^+}(t) \left| x_n((y - t)^+) - x((y - t)^+) \right| + \int_0^{(y-t)^+} (x_n(s-) - x(s-)) \, dH_s(t)
\]

\[
\leq H_{(y-t)^+}(t) \|x_n - x\|_{y_0} + \|x_n - x\|_{y_0} \int_0^{(y-t)^+} |dH_s(t)| \leq (1 + K(y_0)) \|x_n - x\|_{y_0},
\]

where the finite constant \(K(y_0)\) is defined in (3.65). Therefore, as \(n \rightarrow \infty\),

\[
d_u(\phi(x_n), \phi(x)) \equiv \sup_{(t, y) \in [0, T] \times [0, \infty)} |\phi(x_n)(t, y) - \phi(x)(t, y)| \rightarrow 0. \tag{3.67}
\]

Finally, we observe that Lemma 3.4.5 establishes the desired result, because (i) it suffices to consider continuous limits \(x\) by virtue of the continuity assumption included in Assumption 3.2.1 and (ii) convergence in \(\mathbb{D}\) reduces to uniform convergence over bounded intervals when the limiting function is continuous.
3.4.4 Proof of Convergence for the Other Processes.

The proof of convergence for the other processes is straightforward. First, we can apply flow conservation and continuous mapping theorem to treat the service-completion process. In particular,

\[
\hat{D}_n(t) = \hat{N}_n(t) + \hat{X}_n(0) - \hat{X}_n(t) \Rightarrow \hat{N}(t) + \hat{X}(0) - \hat{X}(t) \quad \text{in } \mathbb{D} \quad \text{as } n \to \infty,
\]

which coincides with (3.25).

To treat the residual-service-time processes \(\hat{X}_r^n(0, x)\) and \(\hat{X}_r^n(t, x)\), note that \(X_r^n(0, x) = X^o_n(x)\) which is the number of initially existing customers that are still in service at time \(x\). Hence, as \(n \to \infty\), \(\hat{X}_r^n(0, x) = X^o_n(x) \Rightarrow X^o(x)\) in \(\mathbb{D}\), which is proved earlier in this section. Next, just as for the ICP \(\hat{X}_e^n(t, x)\), we split \(\hat{X}_r^n(t, x)\) into two independent terms associated with new content and old content,

\[
\hat{X}_r^n(t, x) = \hat{X}^r_{n, \nu}(t, x) + \hat{X}^r_{n, o}(t, x), \quad t, x \geq 0
\]

where the convergence of \(\hat{X}^r_{n, \nu}(t, x)\) is proved in [78], and the convergence of the second term holds because

\[
\hat{X}^r_{n, o}(t, x) = \hat{X}_r^n(0, t + x) = X^o_n(t + x) \Rightarrow X^o(t + x), \quad \text{in } \mathbb{D} \quad \text{as } n \to \infty.
\]

3.4.5 Proof of Alternative Representations in (3.22) and (3.27).

We only prove (3.22) because (3.27) is similar. We obtain the right-hand representation in (3.22) using the fact that \(\frac{\hat{U}(t, y_0)}{\sqrt{y_0(1 - y_0)}}; t \geq 0\) is a standard BM motion for a fixed \(0 < y_0 < 1\). We have

\[
\int_0^t \int_0^\infty 1(x > t - s) d\hat{U}_\nu(\Lambda(s), G_\nu(x))
\]

\[
\overset{d}{=} \int_0^t \int_0^\infty 1(x > t - s) d \left( \tilde{B}_s(\Lambda(s)) \sqrt{G_\nu(x)(1 - G_\nu(x))} \right)
\]

\[
= \int_0^t \left( \sqrt{G_\nu(\infty)G_\nu^c(\infty)} - \sqrt{G_\nu(t - s)G_\nu^c(t - s)} \right) d\tilde{B}_s(\Lambda(s)),
\]

which coincides with the expression on the right in (3.22). To show that the two expressions in (3.22) are indeed equal in distribution for each \(t\), it suffices to show that they have the same variances because both processes are zero-mean Gaussian processes. Because the Kiefer process \(\hat{U}(\Lambda(s), G_\nu(x)) = W(\Lambda(s), G_\nu(x)) - G_\nu(x)W(\Lambda(s), 1)\) where \(W\) is a standard Brownian sheet
[78], the variance of the first expression in (3.22) is

\[
E \left[ \left( \int_0^t \int_0^\infty 1(x > t - s) d \left( W(\Lambda(s), G(\nu(x))) - G(\nu(x)W(\Lambda(s), 1) \right) \right)^2 \right]
\]

\[
= E \left[ \left( \int_0^t \int_0^\infty 1(x > t - s) dW(\Lambda(s), G(\nu(x))) - \int_0^t G(\nu(t - s) dW(\Lambda(s), 1) \right)^2 \right]
\]

\[
= \int_0^t \int_0^\infty 1(x > t - s) d\Lambda(s) dG(\nu(x)) + \int_0^t G(\nu(t - s)^2 d\Lambda(s)
\]

\[- 2 \int_0^t \int_0^\infty G(\nu(t - s)1(x > t - s) d\Lambda(s) dG(\nu(x))
\]

\[= \int_0^t G(\nu(t - s)G(\nu(t - s) d\Lambda(s),
\]

which simply coincides with the variance of the second expression in (3.22).

### 3.5 The $G_t/GI^o, GI^\nu/\infty$ Model Starting in the Past

We now show that Theorem 3.3.2 applies to the $G_t/GI^o, GI^\nu/\infty$ model starting at some time in the past, provided we impose an extra condition. We assume that the system starts at time $-t_0 < 0$, satisfying the assumptions in §3.2 with service-time cdf $G$. We let the service-time cdf change to $G_\nu$ after time 0. It suffices to show that Assumption 3.2.1 holds at time 0, which requires an additional independent-increments assumption on the arrival process to obtain the assumed independence of the processes. In particular, we assume that the limiting process in the assumed FCLT for the arrival process is a time-transformed BM.

**Corollary 3.5.1 (FCLT for the $G_t/GI^o, GI^\nu/\infty$ model starting in the past)** Consider the sequence of $G_t/GI^o, GI^\nu/\infty$ models starting at time $-t_0 < 0$ with all the assumptions in §3.2 at time $-t_0$. Let the service-time cdf change from $G$ to $G_\nu$ at time 0. If in addition $\hat{N}(t) = c_0B(\Lambda(t))$, where $\Lambda(t) = \int_{-t_0}^t \lambda(s) ds$, $t > -t_0$, then Assumption 3.2.1 also holds at time 0, so that Theorem 3.3.2 holds for $t \geq 0$, with

\[
X^e(0, y) = \int_0^y G(\nu(s) d\Lambda(s) \cdot 1(0 \leq y \leq t_0) + \int_0^{y-t_0} G(\nu(s) dX(0, y) \cdot 1(y > t_0),
\]

\[
\hat{X}^e(0, y) = \hat{X}^e_0(0, y) \cdot 1(0 \leq y \leq t_0) + \left( \hat{X}^e_0(0, y) + \hat{X}^e_0(0, y) \right) \cdot 1(y > t_0),
\]

\[
\hat{X}^e_0(0, y) = \hat{X}^e_0(0, y) \cdot 1(0 \leq y \leq t_0) + \left( \hat{X}^e_0(0, y) + \hat{X}^e_0(0, y) \right) \cdot 1(y > t_0),
\]

\[
(3.68)
\]

\[
(3.69)
\]

75
where, with $B_a$, $B_s$ and $B$ denoting three independent standard BM’s, and $\hat{U}_o$ denoting a Kiefer process that is independent with $B_a$, 

$$
\hat{X}_a^e(0, y) = c_\lambda \int_0^y G^c(-s) \, dB_a(\Lambda(s)) + \int_0^y \int_0^\infty 1(x > -s) \, d\hat{U}_o(\Lambda(s), G(x))
$$

$$
d_y \, e^{\int_0^y G^c(-s) \, dB_a(\Lambda(s))} - y \int_0^y \sqrt{G(-s)G^c(-s)} \, dB_a(\Lambda(s))
$$

$$
d_y \int_0^y \sqrt{(c_\lambda^2 - 1)G^c(-s^2) + G^c(-s^2)} \, dB(s), \quad \text{for } 0 \leq y \leq t_0,
$$

(3.70)

$\hat{X}_{b,1}^e(0, y)$ is a zero-mean Gaussian process with covariance

$$
\text{Cov} \left( \hat{X}_{b,1}^e(0, y_1), \hat{X}_{b,1}^e(0, y_2) \right) = \int_0^{y_1 \wedge y_2 - t_0} \overline{G}_u(t_0) \overline{G}_c^c(t_0) \, dX^e(-t_0, u), \quad \text{for } y_1, y_2 > t_0,
$$

and

$$
\hat{X}_{b,2}^e(0, y) = \int_0^{y - t_0} \overline{G}_c^c(t_0) \, d\hat{X}^e(-t_0, u), \quad \text{for } y > t_0.
$$

If the system starts empty at time $-t_0$, then the variance formula for the FCLT limit for the number in service $\hat{X}(t)$ for $t \geq -t_0$ is

$$
\sigma^2_{\hat{X}}(t) = \int_{-t_0}^0 \left[ (c_\lambda^2 - 1)G^c(t - s)^2 + G^c(t - s) \right] \lambda(s) \, ds + \int_0^t \left[ (c_\lambda^2 - 1)G^c_\nu(t - s)^2 + G^c_\nu(t - s) \right] \lambda(s) \, ds.
$$

If in addition $G = G_\nu$, we have

$$
\sigma^2_{\hat{X}}(t) = \sigma^2_{\hat{X}}(-t_0, t) = \int_{-t_0}^t \left[ (c_\lambda^2 - 1)G^c_\nu(t - s)^2 + G^c_\nu(t - s) \right] \lambda(s) \, ds.
$$

(3.71)

**Remark 3.5.1** (Verifying consistency with [78]) Corollary 3.5.1 provides an important consistency check by allowing us to compare with the previous results in [78]. In particular, we see that we get strong verification through (3.71).

We illustrate Corollary 3.5.1 with the following example.

**Example 3.5.1** (Simulation comparison) We consider an $M_t/LN(1, 4)/\infty$ model over the time interval $[-t_0, T] = [-5, 20]$, having a nonhomogeneous Poisson arrival process ($M_t$) with the sinusoidal arrival rate function

$$
\lambda_n(t) = n \left( a + b \sin(\pi t + \phi) \right), \quad t_0 \leq t \leq T,
$$

(3.72)

with $a = c = 1$, $b = 0.6$ and $\phi = 0$. This example has a lognormal ($LN$) service distribution with mean $1/\mu = 1$ and $c_s^2 = 4$. We set $n = 100$. 

76
We let the system start empty at time $-t_0$ and use the arrivals in the negative time interval $[-t_0, 0]$ to generate the initial number of customers in service and the age process at time 0. We expect our FWLLN and FCLT limits to provide effective engineering approximations for the mean and variance of the performance functions. For instance, Theorems 3.3.1 and 3.3.2 imply that $X_n(t) = nX(t) + \sqrt{n}\hat{X}(t) + o(1)$ when $n$ is large. Therefore, we expect $\mathbb{E}[X_n(t)] = nX(t) + o(1)$ and $\text{Var}(X_n(t)) = n\text{Var}(\hat{X}(t)) + o(1)$. We next provide simulation comparison results. Each simulation experiment in this chapter is based on performing 2000 independent replications of the system.
Figure 3.1 shows close approximations of the fluid and variance formulas provided by Theorem 3.3.1 and Corollaries 3.3.5 and 3.5.1. In particular, the arrival rate after time 0 is shown in the top plot. Then the expected number in system of the old customers and the new customers are shown together with the total expected number in the second plot; while the variances of the number in system of the old customers and the new customers are shown together with the total variances in the third plot. In both cases, we see the additivity. As expected, the old content dissipates by about time $t = 6$. The bottom plot shows the variance of the initial age process at time 0, with age $0 \leq y \leq t_0 = 5$ (because no customer has an age greater than $t_0$). As expected the right endpoint in the bottom plot coincides with the left endpoint in the third plot.

3.6 Steady-State Approximations

An important application of the results in this chapter is generating useful approximations for the steady-state behavior of general stationary $G/GI/\infty$ IS models. Since we can apply Little’s law to conclude that the steady-state mean number in system is $\rho \equiv \lambda E[S]$, where $E[S]$ is a service time, we assume that $E[S] < \infty$ in this section.

Finding general conditions for the existence of such steady-state distributions is complicated, even for the special case of the number in system with renewal ($GI$) arrival processes, as can be seen from Remark 2 of [100]. However, assuming that steady state for the process $\{\hat{X}_n^{e,v}(t, \cdot) : t \geq 0\}$ in $D_D$ for system $n$ is well defined, it is natural to approximate by the steady-state stationary process associated with the limiting process $\{\hat{X}^{e,v}(t, \cdot) : t \geq 0\}$ in $D_D$, which is itself an element of $D$.

As observed by Glynn and Whitt [29], pp.193–195, the steady-state behavior of the limiting process is relatively easy to analyze. For example, they observed, for service-time distributions with finite support in an interval $[0, y^*]$, that the number in system is in steady state after time $y^*$. We establish a generalization of that result, which holds because the driving processes $B_a(s)$ and $\hat{K}(s, \cdot)$ have stationary increments in $s$.

**Corollary 3.6.1** (Stationary version of the limiting IS age process) If $\hat{N}(\cdot) = c_a \sqrt{\lambda} B_a(\cdot)$ and the system starts empty at time $-t_0 \leq 0$, then the limiting processes (as $n \to \infty$) of the new input in the $G/GI/\infty$ model can be represented as

$$
(\hat{X}_1^{e,v}(t, y), \hat{X}_2^{e,v}(t, y)) = \left( \sqrt{\lambda a_2} \int_{(t-y)+}^{t} G^c(t-s) dB_a(s), \int_{(t-y)+}^{t} \int_{t-s}^{\infty} d\hat{U}(\lambda s, G(x)) \right), \tag{3.73}
$$

$$
(\hat{X}_1^{r,v}(t, y), \hat{X}_2^{r,v}(t, y)) = \left( \sqrt{\lambda a_2} \int_{t-t_0}^{t} G^c(t + y - s) dB_a(s), \int_{t-t_0}^{t} \int_{t+y-s}^{\infty} d\hat{U}(\lambda s, G(x)) \right). \tag{3.74}
$$
(a) If \( G^c(y^*) = 0 \), then the distribution of \((\hat{X}^e_1(t, \cdot), \hat{X}^e_2(t, \cdot))\) as a process in \( \mathbb{D}^2 \) is independent of \( t \) for \( t + t_0 \geq y^* \) and thus reaches steady state at time \( y^* \) if \( t_0 = 0 \) (or is in steady state at time 0 if \( t_0 \geq y^* \));

(b) As \( -t_0 \downarrow -\infty \), corresponding to the system starting empty in the distant past, the processes in (3.73) and (3.74) converge w.p.1 to the associated stationary processes (as functions of \( t \))

\[
(\hat{X}^e_1(t, y), \hat{X}^e_2(t, y)) = \left( \sqrt{\lambda e_a} \int_{t-y}^t G^c(t-s) \, dB_a(s), \int_{t-y}^t \int_{t-s}^\infty d\hat{U}(\lambda s, G(x)) \right), \quad (3.75)
\]

\[
(\hat{X}^{r,e}_1(t, y), \hat{X}^{r,e}_2(t, y)) = \left( \sqrt{\lambda e_a} \int_{-\infty}^t G^c(t+y-s) \, dB_a(s), \int_{-\infty}^t \int_{t+y-s}^\infty d\hat{U}(\lambda s, G(x)) \right),
\]

whose marginal distribution as a function of \( y \) (as a process in \( \mathbb{D} \)) can be seen by setting \( t = 0 \), yielding the steady-state processes

\[
(\hat{X}^{e,s}_1(y), \hat{X}^{e,s}_2(y)) \equiv \left( \sqrt{\lambda e_a} \int_{-\infty}^0 G^c(-s) \, dB_a(s), \int_{-\infty}^0 \int_{-s}^\infty d\hat{U}(\lambda s, G(x)) \right), \quad y \geq 0, \quad (3.77)
\]

\[
(\hat{X}^{r,s}_1(y), \hat{X}^{r,s}_2(y)) \equiv \left( \sqrt{\lambda e_a} \int_{-\infty}^0 G^c(y-s) \, dB_a(s), \int_{-\infty}^0 \int_{y-s}^\infty d\hat{U}(\lambda s, G(u)) \right), \quad y \geq 0, \quad (3.78)
\]

with covariance formulas given by

\[
\text{Cov}(\hat{X}^{e,s}_1(y_1), \hat{X}^{e,s}_1(y_2)) = \lambda c_2 \int_0^{y_1 \wedge y_2} G^c(s)^2 \, ds, \quad \text{Cov}(\hat{X}^{e,s}_1(y_1), \hat{X}^{e,s}_2(y_2)) = \lambda \int_0^{y_1 \wedge y_2} G^c(s)G(s)ds,
\]

\[
\text{Cov}(\hat{X}^{e,s}_1(y_1), \hat{X}^{e,s}_2(y_2)) = \lambda c_2 \int_0^\infty G^c(y_1+s)G^c(y_2+s) \, ds,
\]

\[
\text{Cov}(\hat{X}^{e,s}_2(y_1), \hat{X}^{e,s}_2(y_2)) = \lambda \int_0^\infty (G^c((y_1 \vee y_2) + s) - G^c(y_1+s)G^c(y_2+s)) \, ds. \quad (3.79)
\]

Proof We have already established (3.73) and (3.74). The other representations hold because both \( B_a(s) \) and \( \hat{U}(\lambda s, G(\cdot)) \) have stationary increments in \( s \). The limit as \( -t_0 \downarrow -\infty \) is relatively easy because the processes \( B_a(s) \) and \( \hat{U}(\lambda s, G(\cdot)) \) do not change over the interval \([-t_0, t]\) if we expand the interval on the left and consider the process over \((-\infty, t]\). The new contribution over the interval \((-\infty, -t_0]\) decreases as \(-t_0 \downarrow -\infty\). This can be quantified through the variance of the zero-mean Gaussian random variable, which is asymptotically negligible. It thus remains to derive the covariance formulas in (3.79). We only prove the first two because the proofs of
the others are similar. First, by the isometry of Brownian integrals,

$$\text{Cov}(\hat{X}^{e,s}_{1}(y_1), \hat{X}^{e,s}_{1}(y_2)) = \lambda \int_{-y_1 \wedge y_2}^{0} G^e_c(-s)^2 \, ds = \lambda \int_{0}^{y_1 \wedge y_2} G^e(s)^2 \, ds. \quad (3.80)$$

Next, exploiting the representation of the Kiefer process in terms of the Brownian sheet, i.e., $\hat{U}(x,y) = W(x,y) - yW(x,1)$ (see the appendix of [78]), we have

$$\text{Cov}(\hat{X}^{e,s}_{2}(y_1), \hat{X}^{e,s}_{2}(y_2))$$

$$= E \left[ \int_{-y_1}^{0} \int_{-s}^{\infty} d(W(\lambda s, G(x)) - G(x)W(\lambda s, 1)) \right] \left. \int_{-y_2}^{0} \int_{-s}^{\infty} d(W(\lambda s, G(x)) - G(x)W(\lambda s, 1)) \right]$$

$$= \lambda \int_{-y_1 \wedge y_2}^{0} G^e(-s) ds + \lambda \int_{-y_1 \wedge y_2}^{0} G^e(s)^2 ds - 2\lambda \int_{-y_1 \wedge y_2}^{0} G^e(-s)^2 ds, \quad (3.81)$$

which coincides with the second covariance formula in (3.79).

Corollary 3.6.1 adds to the insight about the stationary model provided by Corollaries 3.1, 4.1 and 4.2 of [78]. As indicated in Corollary 4.1 of [78], we see that the approximating stationary distribution depends on the arrival process beyond its constant arrival rate $\lambda$ and the assumed FCLT only through the asymptotic variability parameter $c_a^2$. Thus this distribution is the same as for the $M/GI/\infty$ model if and only if $c_a^2 = 1$. It is thus instructive to also consider what can be established for the $M/GI/\infty$ model directly by exploiting its special structure. So now we consider the $M/GI/\infty$ model. It is well known that the steady-state number of customers $X_n(0) \equiv X_n^e(0,\infty)$ is a Poisson random variable with $E[X_n(0)] = Var(X_n(0)) = n\lambda E[S] = n\lambda \int_{0}^{\infty} G^e(u) \, du$ and, conditioned on that number, the ages (and the residual service times) are i.i.d. with the stationary-excess cdf $G^e_e(t) \equiv (1/E[S]) \int_{0}^{t} G^e_c(u) \, du$, $t \geq 0$ and ccdf $G^e_{ce}(t) \equiv 1 - G^e_e(t)$; e.g., see [19, 30]. Thus we have the following corollary.

**Corollary 3.6.2** *(FCLT for the $M/GI/\infty$ model in steady state)* Consider a sequence of $M/GI/\infty$ models in steady state at time 0, with service-time cdf $G$ and constant arrival rate $n\lambda$. Then Assumption 3.2.1 holds with the FWLLN and FCLT limits for the initial age processes (and all $t \geq 0$) given by

$$X^{e,s}(0, y) = \rho G^e_e(y) = \int_{0}^{y} a(y) dy = \int_{0}^{y} \lambda G^e_c(u) \, du \quad \text{and}$$

$$\hat{X}^{e,s}(0, y) \overset{d}{=} \hat{X}^{e,s}_{1}(y) + \hat{X}^{e,s}_{2}(y), \quad (3.82)$$
where \( \rho \equiv \lambda E[S] = X^e(0, \infty) = X(0) \), and \( \hat{X}_1^{e,ss} \) and \( \hat{X}_1^{e,ss} \) are independent processes with

\[
\begin{align*}
\hat{X}_1^{e,ss}(y) &\equiv \tilde{U}^*(\rho, G_e(y)) \overset{d}{=} \sqrt{\rho} B^*_s(G_e(y)) \quad \text{and} \\
\hat{X}_2^{e,ss}(y) &\equiv G_e(y) \hat{X}(0) \overset{d}{=} \sqrt{\rho} G_e(y) Z_0,
\end{align*}
\]

(3.83)

where \( \tilde{U}^* \) is a standard Kiefer process associated with old customers, \( B^*_s \) is a standard Brownian bridge, \( Z_0 \) is a standard Gaussian random variable, independent of \( \tilde{U}^* \). The steady-state version of the residual-service-time process

\[
X^{r,s}(0, y) = \rho G_e(y) = \int_0^\infty \lambda G^c(u) \, du \quad \text{and} \quad \hat{X}_1^{r,s}(0, y) = \hat{X}_1^{r,ss}(y) + \hat{X}_2^{r,ss}(y), \quad t \geq 0,
\]

(3.84)

where \( \hat{X}_1^{r,ss} \) and \( \hat{X}_2^{r,ss} \) are independent zero-mean Gaussian processes, with

\[
\begin{align*}
\text{Cov}\left( \hat{X}_1^{r,ss}(x_1), \hat{X}_1^{r,ss}(x_2) \right) &= \int_0^\infty G^c_u(x_1 \wedge x_2) \tilde{G}^c_u(x_1 \lor x_2) \, dX^{c,0}(0, u), \\
\text{and} \quad \hat{X}_2^{r,ss}(x) &\overset{d}{=} \sqrt{\rho} \int_0^\infty G^c_u(x) \, dB^*_o(G_e(u)) + \sqrt{\rho} G^c_e(y) Z_0.
\end{align*}
\]

The variance of \( \hat{X}^{e,ss}(y) \) is

\[
\text{Var}(\hat{X}^{e,ss}(y)) = \int_0^y \lambda G^c(u) \, du, \quad y \geq 0.
\]

(3.85)

As a consequence, the variance of the total content, as the sum of variances of the new content (that is (3.35)) and old content (that is (3.36)), is

\[
\sigma_X^2 = \sigma_{X,\nu}^2 + \sigma_{X,a}^2 = \lambda \int_0^t G^c(u) \, du + \lambda \int_0^\infty H^c_u(t) G^c(u) \, du = \lambda \int_0^\infty G^c(u) \, du = \rho.
\]

(3.87)

**Proof** Let \( \nu_1, \nu_2, \ldots \) be the ages (e.g., elapsed time in service) of the customers in service at time 0. To prove the FWLLN in (3.82), we have

\[
\hat{X}^e(0, y) = \frac{1}{n} \sum_{i=1}^{nX_n(0)} \mathbf{1}(\nu_i \leq y) \Rightarrow X(0) G_e(y) = \rho G_e(y) \quad \text{in } \mathbb{D}, \quad \text{as } n \to \infty,
\]

81
where the convergence holds because the age \( v_i \) follows the equilibrium distribution \( G_e \) (see Theorem 1 of [19]). To prove the FCLT in (3.82), we have

\[
\hat{X}_e(0, y) = \frac{1}{\sqrt{n}} \left( \sum_{i=1}^{n\bar{X}_n(0)} 1(v_i \leq y) - nX(0)G_e(y) \right)
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n\bar{X}_n(0)} (1(v_i \leq y) - G_e(y)) + \hat{X}_n(0)G_e(y)
\]

\[
= \hat{K}_n (\bar{X}_n(0), G_e(y)) + \hat{X}_n(0)G_e(y)
\]

(3.88)

\[
\Rightarrow \hat{U}^* (X(0), G_e(y)) + \hat{X}(0)G_e(y) \overset{d}{=} \sqrt{X(0)} B^*_e (G_e(y)) + \sqrt{X(0)} G_e(y) Z_0 \quad \text{in } \mathbb{D},
\]

as \( n \to \infty \), where the second equality holds by adding and subtracting \( G_e(y) \) in the sum, and the convergence holds by (i) the marginal convergence of each of the two terms in (3.88) (due to the Gaussian CLT limit for a Poisson random variable) and by (ii) Lemma 3.4.1 and the conditional independence of these two terms conditioning on \( X_n(0) \) (with arguments similar to §3.4.1). The variance formulas (3.86) and (3.87) immediately follow from (3.84) and (3.36) with \( \nu = G \). The proof of the FWLLN limit in (3.84) follows from (3.82) and Theorem 3.3.1, and the proof of (3.85) follows from (3.84), Theorem 3.3.2 and Corollary 3.3.3.

**Corollary 3.6.3** (Two equivalent decompositions) For the M/GI/\( \infty \) model (with \( c^2_a = 1 \)), the two representations (3.77) and (3.82) ((3.78) and (3.84)) are equivalent independent decompositions, i.e.,

\[
\hat{X}_{1,e,s}^{e,s} + \hat{X}_{2,e,s}^{e,s} \overset{d}{=} \hat{X}_{1,e,s}^{e,s} + \hat{X}_{2,e,s}^{e,s} \quad \text{and} \quad \hat{X}_{1,r,s}^{e,s} + \hat{X}_{2,r,s}^{e,s} \overset{d}{=} \hat{X}_{1,r,s}^{e,s} + \hat{X}_{2,r,s}^{e,s}.
\]

(3.89)

**Proof** We only prove the first equality in (3.89) since the second equality follows similarly. Because all four terms in the first equality of (3.89) are zero-mean Gaussian processes, with \( \hat{X}_{1,e,s}^{e,s} \) independent of \( \hat{X}_{2,e,s}^{e,s} \) and \( \hat{X}_{1,e,s}^{e,s} \) independent of \( \hat{X}_{2,e,s}^{e,s} \), it suffices to show that

\[
\sum_{k=1}^{2} \text{Cov}(\hat{X}_k^{e,s}(y_1), \hat{X}_k^{e,s}(y_2)) = \sum_{k=1}^{2} \text{Cov}(\hat{X}_k^{e,s}(y_1), \hat{X}_k^{e,s}(y_2)),
\]

(3.90)

By (3.83), we have

\[
\text{Cov}(\hat{X}_1^{e,s}(y_1), \hat{X}_1^{e,s}(y_2)) = \rho \left[ G_e(y_1) \wedge G_e(y_2) - G_e(y_1)G_e(y_2) \right],
\]

\[
\text{Cov}(\hat{X}_2^{e,s}(y_1), \hat{X}_2^{e,s}(y_2)) = \rho G_e(y_1)G_e(y_2).
\]

82
so that the left-hand side of (3.90) is \( \rho G_e(y_1) \wedge G_e(y_2) = \lambda \int_0^{y_1 \wedge y_2} G^c(u) \, du \), which coincides with the right-hand side of (3.90), according to (3.79) with \( c_a = 1 \).

The corollaries above show that the evolution of new and old content after some time is somewhat complicated for this basic \( M/GI/\infty \) model, even though the steady-state distribution at one time is remarkably simple. We now illustrate with an example.

**Example 3.6.1** *(Simulation comparison for an \( M/GI/\infty \) in steady state)* We consider an \( M/H_2(1, 4)/\infty \) model in \([0,T]\), having a Poisson arrival process with constant arrival rate \( \lambda_n = n\lambda \) for \( \lambda = 1 \), an \( H_2 \) service distribution with balanced means, i.e., a mixture of two exponential r.v.’s with rates \( \mu_1 \) and \( \mu_2 \) with probability \( p \). We set \( \mu_1 = 2p\mu \), \( \mu_2 = 2(1-p)\mu \), \( \mu = 1 \) and \( p = 0.5(1 - \sqrt{0.6}) \) so that the mean \( 1/\mu = 1 \) and \( c_2^2 = 4 \). We set \( n = 100 \).

Since the model is in steady state at time 0, we use a Poisson distribution with mean \( n\lambda/\mu \) to generate the number of customers at time 0. To generate the elapsed times in service (e.g., ages) and the residual service times for these customers, we use the equilibrium version of the service distribution, which again follows an \( H_2 \) distribution, but with altered parameters, specifically with \( p^* = p\mu_2/(p\mu_2 + (1-p)\mu_1) \), \( \mu_1^* = \mu_1 \) and \( \mu_2^* = \mu \). See Theorem 1 of [19]. See Figure 3.2 for the simulation comparison, which is an analog of Figure 3.1.

![Figure 3.2: Example 2: Simulation comparisons of the mean and variance for the number of new and old customers in service of an \( M/H_2(1, 4)/\infty \) model in steady state, with a constant arrival rate \( n\lambda = 100 \).](image-url)
3.7 A Challenging Test Case

In order to more fully substantiate the theoretical results, we consider a $G_t/GI/\infty$ IS model with non-Markov arrival process, nonexponential service-time distribution and unconventional initial conditions. We let the (artificial) initial conditions be generated by a time changed renewal process, and exploit

**Corollary 3.7.1** *(Initial customers generated from a renewal process with a time change)* In the $n^{th}$ $G_t/GI/\infty$ system, if the initial age process $X_n^e(0,y) \equiv N^*(nX^*(0,y))$ where $N^*$ is a rate-1 renewal process with interrenewal SCV $\bar{c}_0^2$ and the deterministic function

\[ X^*(y) \equiv \int_0^y b^*(y) dy < \infty, \quad (3.91) \]

then Assumption 2 is satisfied with $X(0,y) = X^*(y)$ and $\tilde{X}(0,y) = \bar{c}_0 B_0(X^*(y))$, for $y \geq 0$, where $B_0$ is a standard BM. In addition, the variance of the ICP (that is (3.36)), is

\[ \sigma_{X,o}^2(t) = \int_0^{\infty} \tilde{G}_u(t) \tilde{G}_u(c)(u) dX^*(u) + \bar{c}_0 \int_0^{\infty} \tilde{G}_u(t)^2 dX^*(u) \]

\[ = \int_0^{\infty} \left[ (\bar{c}_0^2 - 1) H_u(t) + 1 \right] H_u(t) b^*(u) du. \]

**Example 3.7.1** *(Simulation comparison for an example of Corollary 3.7.1)* In addition to the initial conditions specified above, we let the new input come from an $H_t^2(1,4)/LN(1,4)/\infty$ model in $[0,T]$, having a $G_t$ arrival process $N_n(t) \equiv N^e(n\Lambda(t))$, where $N^e$ is a rate-1 equilibrium renewal process having an $H_2$ interrenewal distribution with balanced means and $\bar{c}_2^2 = 4$, while $\Lambda(t) \equiv \int_0^t \lambda(u) du$, where $\lambda(t)$ is the sinusoidal arrival rate with in (3.72) having parameters $a = c = 1$, $b = 0.6$, $\phi = 0$ and $n = 100$. This example has an LN service distribution with mean $1/\mu = 1$ and $\bar{c}_s^2 = 4$. For the initial conditions, let $N^*$ be a rate-1 Poisson process (so that $\bar{c}_0^2 = 1$) in Corollary 3.7.1 and consider two density functions in (3.91)

\[ b_1^*(u) = u \mathbf{1}_{(0 \leq u \leq \Phi)} + (2d - u) \mathbf{1}_{(d \leq u \leq 2d)} \quad \text{and} \quad b_2^*(u) = \frac{1}{3} u^2 \mathbf{1}_{(0 \leq u \leq 2d)}, \quad (3.92) \]

with $d = 1.5$, as shown in Figure 3.3.

Comparisons with simulations are shown in Figure 3.4. The first three plots in Figure 3.4 are analogs of those in Figure 3.1. In the fourth and fifth plots we give simulation comparisons for the variances of the two-parameter process $X^e(t,y)$. We provide the simulations for the case of $b_2^*$ in the appendix.
3.8 Conclusion

Our main result is an FCLT for the initial content process (ICP), which specifies the elapsed service times of all customers in the system at time $t$ who were also initially in the system at time 0, having arrived sometime earlier. The ICP in system $n$ is the process $X_{n}^{e,o}(t,y)$ in $D_{D}$, defined in (3.3). The key assumptions are (i) that the service times come from a sequence of i.i.d. random variables independent of the arrival process and (ii) an FCLT holds for the initial ICP at time 0 in $D$, which includes having the total number at time 0 grow proportional to $n$.

The limiting process for the scaled ICP is $\hat{X}_{e,o}(t,y) \equiv \hat{X}_{1}^{e,o}(t,y) + \hat{X}_{2}^{e,o}(t,y)$, where $\hat{X}_{1}^{e,o}(t,y)$ and $\hat{X}_{2}^{e,o}(t,y)$ are independent zero-mean Gaussian processes specified in (3.18) and (3.19).

The proof of the FCLT in §3.4 exploits results in [95] for empirical processes of independent non-identically-distributed random variables.

In Theorem 3.3.2, our FCLT for the ICP is expressed in the more general context of an $G_{t}/GI/\infty$ infinite-server queueing model, which includes new input as well as the ICP. In this context, the key assumption is Assumption 3.2.1, which requires a joint FCLT for the arrival process and the ICP at time 0, where the limiting processes are independent zero-mean Gaussian processes. The limit for the ICP alone is obtained by having a null arrival process. The additional results for the new input are obtained from [78]. Corollary 3.5.1 shows that Theorem 3.3.2 implies a corresponding FCLT for the key processes in the more general $G_{t}/GI^{o},GI^{o}/\infty$ model starting at time $-t_{0} < 0$, which has one service-time cdf $G$ before time 0 and then another service-time cdf $G_{\nu}$ after time 0. The limit holds if Assumption 3.2.1 holds at time $-t_{0}$, provided that the limiting process is a time-transformed Brownian motion, which is the common case for applications.

As a further consequence, in Corollary 3.6.1 we characterize the steady-state new and old content in the stationary $G/GI/\infty$ model. Using the explicit structure of the $M/GI/\infty$ model, we obtain an alternative characterization of the steady-state new and old content in Corollary
Figure 3.4: Example 3 with $b^*_1$ in (3.92): Simulation comparisons of the mean and variance for the number of customers in service of an $H_2(1,4)/LN(1,4)/\infty$ model, with the sinusoidal arrival rate (3.72) having parameters $a = c = 1$, $b = 0.6$, $\phi = 0$ and $n = 100$ and general initial conditions.

3.6.2. We then show that the two different representations are equivalent in Corollary 3.6.3. In §3.5-§3.7 we also report results of simulation experiments that verify the accuracy of the formulas and show that they can be useful engineering approximations.
Chapter 4

Many-Server Heavy-traffic Limit for Stationary Queues with Customer Abandonment and Nonexponential Service Times

4.1 Introduction

We consider a queueing system with many parallel servers operating in the efficiency-driven (ED) regime. In particular, we establish many-server heavy-traffic limits for the $G/GI/n + GI$ system with general stationary arrivals (the $G$), independent and identically distributed (i.i.d.) nonexponential service times (the $GI$), and patience times according to i.i.d. nonexponential distribution (the $+GI$). Under spatial scaling, we prove that the key performance processes such as head-of-line waiting time, virtual waiting time, number in system, and queue-length converge in distribution to Gaussian processes in the limit as the scale increases. We also analyze the long-run asymptotic behavior of the Gaussian limits which can provide effective approximations for the original system.

Literature review. There is a large body of literature on many-server heavy-traffic (MSHT) limits for queueing models. We hereby review the related work on the efficiency-driven (ED), or equivalently, overloaded (OL) regime. A fluid model for the $G/GI/n + GI$ queue is developed by Whitt [104] using two-parameter performance functions keeping track of elapsed service and waiting time of customers; in addition, a discrete-time functional weak law of large numbers (FWLLN) is established in [104]. This fluid model has subsequently been extended to incor-
porate time-varying arrivals and staffing levels in [61]. FWLLN results for the $G/GI/n + GI$ queue has been obtained in [62, 46, 109].

Functional central limit theorem (FCLT) results for queueing systems in the ED regime exploit in different ways Markovian structure to obtain Markov processes as limits. Whitt [103] has established an Ornstein-Uhlenbeck (OU) process as the limit of scaled number in system for the $M/M/s/r + M$ queue. Liu and Whitt [65] have considered the $G_t/M/s_t + GI$ alternating between the quality-driven (QD) (equivalently, underloaded) and ED regimes. For the head-of-line waiting time in the ED regime, they have obtained by exploiting Markovian service a stochastic differential equation driven by independent Brownian motions for the head-of-line waiting time. They have also established limits for the number in system, the queue length, the virtual waiting time, and the number of abandonments. In a recent paper by Huang et al. [39], the authors have proved a diffusion limit for the virtual waiting time associated with the $G/M/n + GI$ model in the ED+QED regime (see [73] for a definition) and applied their results in the context of delay announcements. It is evident that the FCLT limits for the key processes are bound to become non-Markovian for fully non-Markovian queueing systems when spacial scaling is applied, that is, one cannot obtain diffusion approximations when service distribution is also nonexponential. To obtain a diffusion FCLT limit, He [37] considers the $G/GI/n + GI$ model and applies both a spacial and time scaling to the virtual-waiting-time process, and the number-in-system process. Specifically, in the framework of [37], time is scaled by the mean patience time and asymptotic analysis is done as the mean patience time goes to infinity. As a result, a one-dimensional OU process is obtained in the limit for the scaled number in system as time scaling reduces the correlation between increments of the service-completion process, thus, yields a Brownian limit. Although simple and convenient, the diffusion approximations in [37] degrades in systems having small and medium mean patience times.

Main difficulty of GI service. In this paper, we establish process-level convergence by proving an FCLT under a special condition on the initial state of the queueing systems, and characterize the steady state of the FCLT limits. More specifically, we assume that the queueing systems indexed by $n$ are initially critically loaded, i.e., all servers are busy and no customer is waiting in line at time 0. This assumption is restrictive in terms of approximating the transient behavior of the original queueing system by using FCLT limits. Our ultimate goal is to provide effective approximations to the long-run behavior of the original system by using the steady-state characterization of the FCLT limits. We first argue that the impact of initial conditions is negligible in the long-run analysis. Therefore, the assumption of a special initial condition is not restrictive in this regard.

The main difficulty in considering general service times along with general patience times is that the scaled service-completion processes no longer converge to a Brownian motion as is
the case in [65] and [39]. In these papers, the scaled service-completion processes converge to (time-changed) Brownian motions due to Markovian service distribution. This, in turn, yields stochastic differential equations (SDEs) driven by independent Brownian motions for the scaled head-of-line waiting time in the former paper and for the virtual waiting time in the latter. In our paper, by assuming the a special initial condition, we guarantee that the scaled service-completion processes converge to a centered Gaussian process with a known covariance function. Consequently, we are able to obtain an SDE driven by both the Brownian motion and the centered Gaussian process for the sequence of scaled head-of-line waiting times and its limit. In particular, we solve for an SDE involving an Itô integral with respect Brownian motion, and a stochastic integral with respect to a centered Gaussian process where the integrand is a deterministic two-parameter function.

Our proof technique is based on the analysis of the number of customers who has entered service. We introduce a new decomposition different than those in [39, 65, 78] for the enter-service process. Then using the new decomposition, we derive an SDE for each of the prelimit head-of-line waiting processes, and prove they converge to a limiting SDE. The main steps of our proof technique involves a martingale FCLT, Gronwall’s inequality, and continuous mapping theorem. Unlike [39, 65], we do not use the compactness approach (see, e.g., [102]) to prove distributional convergence. Consequently, we avoid a possibly complicated tightness proof given the complicated proofs in [39, 65] even for the case of $M$ service. In particular, we prove stochastic boundedness of the sequence of SDEs for the scaled head-of-line processes and then prove convergence of the sequence by applying Gronwall’s inequality. We further characterize the limiting SDE by computing the covariance function of the Gaussian solution to the limiting SDE. Convergence for the other processes is established by applying continuous mapping theorem. We also remark that the martingale FCLT in this paper is different than those in [39, 16]. More is discussed in the proof of Lemma 4.7.2.

To characterize the limiting processes we construct a stochastic integral with respect to centered Gaussian processes with almost-surely Hölder-continuous sample paths where the integrand is a two-parameter deterministic function. We show that such stochastic integrals can be defined pathwise, and satisfy an integration-by-parts formula. We remark that we do not attempt to develop an equivalent of Ito calculus for general Gaussian processes. Equivalents of Ito calculus have been constructed for different types of Gaussian processes – that are not Brownian motions– inspired by real-world applications. Examples of such Gaussian processes include fractional Brownian motions (see [8] and the references therein), multifractional Brownian motions, and Volterra processes (see [1, 56] and the references therein).

**Our Contributions.** Here we summarize our contributions:

1. We prove an FCLT for the key performance measures, and identify their FCLT limits.
We solve an open problem by considering a queueing system with nonexponential service- and patience-time distributions (generalizing the work of [65, 39]). To the best of our knowledge, neither an FCLT limit for the $G/GI/n + GI$ queue nor its steady state is known.

2. We provide a new proof technique which involves (i) a new decomposition for the enter- service process; (ii) a new martingale FCLT; (iii) a new approach to prove convergence other than the compactness approach. Along the way, we solve a stochastic differential equation involving more general Gaussian processes than Brownian motions, and hence, requires characterization of a certain type of stochastic integrals with respect to certain Gaussian processes more general than Brownian motions. We provide an explicit expression for its solution , and further characterize by deriving its covariance function . We remark that our proof in §4.7 allows for more general time-varying non-Poisson ($G_t$) arrivals.

3. We characterize the steady state of the FCLT limits for the $G/GI/n + GI$ queue. Our FCLT results (along with FWLLN fluid limits developed in [104]) can provide effective Gaussian approximations with mean and variance formulas for the steady-state performance of the $G/GI/n + GI$ queue.

**Organization of the Paper.** In §4.2 we describe a sequence of the $G/GI/n + GI$ queueing systems and specify the model assumptions. In §5.3, we review the associated fluid queueing system and the fluid functions for the $G/GI/n + GI$ queue. We give preliminary results that are building blocks of the main results in §5.4. In §4.5, we present our main results under a special condition on the initial state of the systems. In §4.6, we provide further characterization of the FCLT limits by deriving analytical formulas, and by characterizing the steady-state distributions of the FCLT limits. We give a proof of the main results in §4.7. In §4.8, we conduct numerical experiments to evaluate the engineering approximation formulas; we give simulation comparison results. In §4.9, we consider a more general staffing function, and provide a generalization of the main results in §4.5. Additional material including proofs appear in the appendix.

**Notations.** All random variables and processes are defined on a common probability space $(\Omega, \mathcal{F}, P)$. We denote by the sets $\mathbb{N}$ and $\mathbb{R}$ the strictly positive integers and the real numbers, respectively. For any $a, b \in \mathbb{R}$, $a \wedge b \equiv \min\{a, b\}$, $a \vee b \equiv \max\{a, b\}$ and $a^+ \equiv a \vee 0$. For $a \in \mathbb{R}$, the function $\lfloor a \rfloor$ gives the largest integer less than or equal to $a$. Let $\|\cdot\|_{[a,b]}$ be the supremum norm over the interval $[a,b]$. For convenience, we define $\|\cdot\|_b = \|\cdot\|_{[0,b]}$. We let $1(A)$ be the indicator random variable of a set $A$. Also, we let $e$ be the identity function, i.e., $e(t) = t, t \geq 0$; $\chi$ be the
constant one function \( \chi(t) = 1, t \geq 0 \). A function \( f(t) \) is \( o(g(t)) \) if \( \limsup_{t \to \infty} f(t)/g(t) \to 0 \); and is \( O(g(t)) \) if \( \limsup_{t \to \infty} f(t)/g(t) < \infty \). A measure \( \mu \) is said to be locally finite if, for a measurable space \((S, \mathcal{S}, \mu)\), \( \mu(K) < \infty \) for all compact \( K \in \mathcal{S} \).

We denote by \( \mathcal{D} \equiv \mathcal{D}([0, T]; \mathbb{R}) \) the space of real-valued right-continuous functions with left limits on the interval \([0, T]\), and by \( \mathcal{C} \equiv \mathcal{C}([0, T]; \mathbb{R}) \) the subset of \( \mathcal{D} \) consists of continuous functions. Convergence in \( \mathcal{D} \) is characterized through Skorokhod’s \( J_1 \)-topology. Therefore, \( J_1 \)-convergence to a continuous limit implies uniform convergence over all compact sets. We use the notation \( \Rightarrow \) to denote convergence in distribution in \( \mathcal{D} \). See [10, 102] for details.

### 4.2 A sequence of \( G/GI/n + GI \) queues

We consider a sequence of \( G_t/GI/s_t + GI \) queueing systems, which are indexed by the number of servers \( n \), having stationary arrivals from a general process (the \( G \)), independent and identically distributed (i.i.d.) nonexponential service times (the \( GI \)), and customer abandonment with i.i.d. nonexponential patience times (the \( +GI \)). Upon arrival, customers start service immediately if there are any idle servers. Otherwise, if all servers are busy, customers join a waiting line with an infinite capacity. Customers waiting in line may abandon the system without getting service if they wait longer than their patience time. Customers do not abandon the system once they begin service. The service discipline is first-come first-served (FCFS).

Let \( N_n(t) \) be the number of customers who have arrived in the interval \([0, t]\), \( E_n(t) \) be the total number of customers who have started service in \([0, t]\), \( A_n(t) \) be the number of customers who have abandoned in \([0, t]\). Let the two-parameter process \( B_n(t, y) \) \( (Q_n(t, y)) \) denote the number of customers in service (in queue) at time \( t \) for at most \( y \) units of time in the \( n \)th system. Also, define \( B_n(t) \equiv B_n(t, \infty) \) \( (Q_n(t) \equiv Q_n(t, \infty)) \) as the total number of customers in service (line) at time \( t \). Then, the total number in the \( n \)th system at time \( t \) is represented as \( X_n(t) = B_n(t) + Q_n(t) \). We let \( D_n(t) \) be the total number of service completions in \([0, t]\). The process \( W_n(t) \) represents head-of-line (HOL) waiting time, i.e., the elapsed waiting time of the customer at the head of line at time \( t \), i.e., the waiting time of the customer who has been waiting the longest. Finally, we let \( V_n(t) \) be the virtual waiting time at time \( t \), i.e., the waiting time of a potential arriving customer assuming the customer has infinite patience. By definition of \( W_n(t) \) and \( V_n(t) \), we have

\[
V_n(t - W_n(t)) = W_n(t) + O(1/n) \quad (4.1)
\]
\[
V_n(t) = W_n(t + V_n(t) + O(1/n)) + O(1/n), \quad t \geq 0, \quad (4.2)
\]

where (5.1) suggests that the virtual waiting time at the time of arrival of the head-of-line customer at time \( t \) is the head-of-line customer’s elapsed waiting time in line at time \( t \) plus the
additional time after time $t$ until a server becomes idle. The additional time is $O(1/n)$ because the minimum of $n$ i.i.d. remaining service times divided by $1/n$ converges in distribution to an exponential random variable with rate $h_G(0)$, that is, the hazard-rate function of the service-time distribution $G$ evaluated at 0. In particular, the remaining service times of customers in service at any time $t$ are asymptotically (as $n \to \infty$) i.i.d. because the minimum of $n$ i.i.d. remaining service times divided by $1/n$ converges in distribution to an exponential random variable with rate $h_G(0)$, that is, the hazard-rate function of the service-time distribution $G$ evaluated at 0. In particular, the remaining service times of customers in service at any time $t$ are asymptotically (as $n \to \infty$) i.i.d. because the remaining service times of initial customers in service at time 0 are assumed to be i.i.d., and the system is assumed to operate in the ED regime in the interval $[0, T]$. The equality in (5.2) is obtained by applying change of variable.

**New customers.** For the $i$th customer who arrives at the $n$th system after time 0, we let $\tau^n_i$ be the arrival time, $\nu^n_i$ be the service time, $\gamma^n_i$ be the patience time and $w^n_i$ be the offered waiting time, i.e., time the $i$th customer would have to wait until starting service if the customer were infinitely patient. We assume that $\{\nu^n_i : i \geq 1\}$ are i.i.d. with cdf $G$ and $\{\gamma^n_i : i \geq 1\}$ are i.i.d. with cdf $F$ for each $n \geq 1$. Also, it holds that $w^n_i = V_n(\tau^n_i -)$ for $i \geq 1$ (see e.g., [16]).

**Initial customers.** For customers in service at time 0, we let $\nu^n_j$, $1 \leq j \leq n$, be the remaining service time of the $j$th initial customer in the $n$th system. We assume the remaining service times are i.i.d. drawn from the equilibrium cdf $G_e$ given by

$$G_e(t) = \frac{1}{E[\nu_1]} \int_0^t [1 - G(s)] \, ds \quad \text{for} \quad t \geq 0, \ 1 \leq j \leq n.$$  

Without loss of generality, assume that $\nu^n_j$ is the remaining service time of the customer in server $j$, and let $D_j(t)$ count the number of service completions at server $j$ by time $t$, $1 \leq j \leq n$. Then, $(D_j(t) : t \geq 0)$ is an equilibrium renewal process provided that server $j$ remains busy at all times. Then the total number of service completions in $[0, t]$ is given by $D_n(t) = \sum_{j=1}^n D_j(t)$ for $t \geq 0$. Because the system operates in the ED regime, all servers will be busy at all times with probability 1 as $n \to \infty$. Hence it holds that the service-completion process is asymptotically equivalent to superposition of equilibrium renewal processes. Then, by the convergence-together theorem (see, e.g., Theorem 11.4.7. of [102]), we deduce that the CLT-scaled service-completion process and the CLT-scaled superposition process must converge to the same limit, that is, the Gaussian FCLT limit in Theorem 2 of [101].

We remark that the assumption that all $n$ remaining service times follow the equilibrium distribution is not too restrictive (in fact, this assumption is commonly used in the literature [89, 72, 37, 24, 49]), because we plan to later focus on characterizing the long-run behavior, on which initial conditions have negligible impact. We intend to approximate the steady-state performance by that of the limiting processes (as $n \to \infty$) under the conjecture that the interchange of the two limits (as $n \to \infty$ and as $t \to \infty$) holds. However, establishing the
interchange of the two limits is beyond the scope of this paper.

We further assume that the arrival process, service times and patience times are mutually independent. We impose a condition on the service and patience distributions.

**Assumption 4.2.1 (Regularity conditions on service- and patience-time distributions)**

(i) The patience-time cdf $F$ is differentiable with pdf $f$ and satisfies $F'(x) > 0$ for all $x \geq 0$. Moreover, the pdf $f$ is bounded on bounded intervals.

(ii) The service-time cdf $G$ satisfies is differentiable with pdf $g$ which is bounded on bounded intervals. Consequently,

\[
\limsup_{t \downarrow 0} \frac{G(t) - G(0)}{t} < \infty. \tag{4.3}
\]

Condition (4.3) is necessary for weak convergence results involving scaled renewal processes (see [101, 102]). This condition allows for an atom at 0 and is satisfied for any distribution function that is absolutely continuous in a neighborhood of 0. Two special classes of distributions that satisfy (4.3) are all discrete distributions with no mass at time 0 and all continuous distributions with finite density at time 0 (see [24]).

We now introduce two scalings for a $\mathbb{R}$-valued or $\mathcal{D}$-valued process $K_n$ (see [102, 10]). Then LLN-scaled and CLT-scaled versions of $K_n$ are

\[
\bar{K}_n \equiv n^{-1} K_n \quad \text{and} \quad \hat{K}_n \equiv n^{-1/2} (K_n - nK) \tag{4.4}
\]

where $K$ is the limit of the sequence $\{\bar{K}_n : n \geq 1\}$ (see §5.3 for more details). We apply the LLN and CLT scaling to the processes $E_n, D_n, A_n, B_n, Q_n$ and $X_n$. For virtual waiting time $V_n$ and HOL waiting time $W_n$, we define the CLT-scaled version as

\[
\hat{W}_n \equiv \sqrt{n} (W_n - w) \quad \text{and} \quad \hat{V}_n \equiv \sqrt{n} (V_n - v)
\]

where $w$ and $v$ are the limits of the sequences $\{W_n : n \geq 1\}$ and $\{V_n : n \geq 1\}$, respectively.

**Assumption 4.2.2 (FCLT for the arrival process)** The sequence of CLT-scaled arrival processes $\{\hat{N}_n\}$ satisfies

\[
\hat{N}_n(t) \equiv n^{-1/2} (N_n(t) - n\Lambda(t)) \Rightarrow \hat{N}(t) \quad \text{in} \quad \mathbb{D}([0,T];\mathbb{R}) \quad \text{as} \quad n \to \infty
\]

where $\Lambda(t) = \lambda t$, $\lambda$ is the asymptotic arrival rate, and the limit $(\hat{N}(t) : 0 \leq t \leq T)$ has continuous sample paths and independent increments.
Remark 4.2.1 An important special case is $\hat{N}(t) = c_\lambda B_\lambda(\Lambda(t))$ where $B_\lambda$ is a standard Brownian motion (BM). In this case, the time variability and stochastic variability are characterized by the rate $\lambda t$ and the constant $c_\lambda^2$. For instance, for Poisson arrivals, $c_\lambda = 1$.

As an immediate corollary of Assumption 4.2.2, FWLLN for the sequence $\{\hat{N}_n\}$ follows. In particular,

$$\hat{N}_n(t) \equiv n^{-1}N_n(t) \Rightarrow \Lambda(t) \text{ in } \mathbb{D} \text{ as } n \to \infty.$$ 

Assumption 4.2.3 (Initial critical loading) All queues indexed by $n$ are initially critically loaded, that is, $Q_n(0) = W_n(0) = 0$ and $B_n(0) = n$ for all $n \geq 1$.

Assumption 4.2.4 (ED regime) The system operates in the ED regime over the interval $[0, T]$ with $\rho \equiv \lambda/\mu > 1$.

4.3 The $G/GI/n + GI$ fluid model

In this section, we review the $G/GI/n + GI$ fluid queueing system based on the fluid queueing system introduced in [61]. The associate fluid functions are the limits for the LLN-scaled processes in Theorem 4.5.1.

Consider a system where fluid flows into the system from an external source with a deterministic arrival rate $\lambda$. The input fluid content moves directly to service if there is available space otherwise the fluid content flows into waiting room with unlimited capacity. The total amount of input by time $t$ is given by the deterministic function $\Lambda(t) = \lambda t$. The cdfs $G$ and $F$ in the context of stochastic queueing systems are interpreted as proportions in the context of the fluid queueing systems. That is, $G(x)$ is the proportion of the fluid that has completed service by time $x$ to the total amount of fluid which started service by time $x$, and $F(x)$ is the proportion of the total fluid input that has abandoned the system before starting service process by time $x$.

The key performance measures of the fluid queueing system are the two-parameter deterministic processes $B(t, y)$ and $Q(t, y)$ defined as the amount of fluid in service (resp. in waiting space) at time $t$ that has been so for at most $y$ units of time. These quantities can also be represented as an integral of density function:

$$B(t, y) = \int_0^y b(t, s) \, ds \quad \text{and} \quad Q(t, y) = \int_0^y q(t, s) \, ds \quad \text{for} \quad t, y \geq 0.$$ 

where the density functions $b(\cdot, \cdot)$ and $q(\cdot, \cdot)$ will be defined later in this section. Given these definitions, the total amount of fluid in the system at time $t$, $X(t)$, is given by $X(t) = B(t) + Q(t)$ with $B(t) \equiv B(t, \infty)$ and $Q(t) \equiv Q(t, \infty)$ by convention. Finally, the total amount of abandoned
Fluid over \([0, t]\) satisfies

\[
A(t) = \Lambda(t) - E(t) - X(t) + X(0), \quad t \geq 0,
\]

where \(E(t)\) denotes the amount of fluid that has entered service by time \(t\) and defined in (4.6).

**Initial content.** So far we have discussed the dynamics of fluid content that enters the system from outside after time 0. We now discuss fluid content present in service at time 0. The equilibrium distribution \(G_e\) of \(G\) is chosen to determine the remaining service times of initial customers in service in §4.2. In the context of the fluid queueing system, \(G_e(x)\) is the proportion of initial fluid content that has completed service by time \(x\).

**Overloaded regime.** The fluid queueing system is said to be in the ED regime at time \(t\) if either fluid content in waiting space is strictly positive, i.e., \(Q(t) > 0\), or there is no fluid in waiting space, fluid content in service is equal to the service capacity and arrival rate at time \(t\) is greater than service-completion rate \(\mu\), i.e., \(Q(t) = 0, B(t) = 1\) for all \(t \geq 0\), and \(\lambda > \mu\).

**Service content.** The density function associated with fluid in service at time \(t\) that has been in service for exactly \(x\) units of time is given as

\[
b(t, x) = b(t - x, 0)G^c(x)1(x < t) + b(0, x - t)G_e^c(x - t)1(x \geq t), \quad t \geq 0.
\]  

The term \(b(t - x, 0)\) on the right is the amount of the fluid content that entered service exactly \(x\) units of time in the past. Consequently, the amount of fluid at time \(t\) after \(x\) units of time has passed, \(b(t, x)\), is what’s left over from the initial amount \(b(t - x, 0)\) after the departure of processed fluid during the period of length \(x\). The condition \(\{x < t\}\) suggests that the first term is associated with fluid that enters system from outside after time 0. Hence the service-completion rate of processed fluid in service is determined by the function \(G^c\). The term \(b(0, x - t)\), on the other hand, is the amount of fluid that has been present in service for exactly \(x\) (\(\geq t\)) units of time at time 0. That is, the second term describes the behavior of initial content. Necessarily, the remaining amount of fluid content is determined by \(G_e^c\) instead of \(G^c\). In particular, the amount of fluid at time \(t\) given the fluid content has been in service \(x\) (\(\geq t\)) units of time, \(b(t, x)\), is what’s left over from the initial amount at time 0, \(b(0, x - t)\), after the departure of processed fluid during the period of length \(t\). The density \(b(0, x - t)\) is part of the initial data.

The expression in (5.5) requires \(b(t, 0)\). Because the fluid queueing system is OL, the flow rate into service at time \(t\), \(b(t, 0)\), must be equal to the rate at which fluid is processed out of
Queue content. Because the waiting space is initially empty by assumption, the density function associated with fluid content in waiting space at time \( t \) that has been waiting for exactly \( x \) units of time is given by

\[
q(t, x) = \lambda(t - x) F^c(x) \mathbf{1}(x \leq w(t) \wedge t), \quad t, x \geq 0,
\]

where \( w(t) \) is such that, for all \( t \geq 0 \), \( q(t, x) = 0 \) if \( x > w(t) \). The function \( w(t) \) can be interpreted as the head-of-line waiting time function for the fluid queueing system. Similar to the interpretation of (5.5), the term on the right-hand side is associated with the fluid content that enters the waiting space after time 0. The amount of fluid content that has been waiting for exactly \( x \) units of time at time \( t \) is what is left over from the amount \( x \) units ago, \( q(t - x, 0) \), after some proportion determined by \( F^c \) has abandoned.

According to [61], the fluid head-of-line waiting time \( w(t) \) uniquely solves the ordinary differential equation (ODE)

\[
w'(t) = 1 - \frac{b(t, 0)}{q(t, w(t))}, \quad t \geq 0,
\]

where \( w' \) is the derivative of \( w \), and \( q(\cdot, \cdot) \) is as in (5.11). In addition, by Theorem 5 of [61], \( w(t) \) and the fluid virtual waiting time \( v(t) \) uniquely satisfy

\[
v(t) = w(t + v(t)) \quad \text{and} \quad v(t - w(t)) = w(t) \quad t \geq 0.
\]
Hölder continuous of order $0 < \alpha < 1$ if there is a constant $c$ such that

$$|f(s) - f(t)| \leq c|s - t|^{\alpha} \quad \text{for all} \quad a < s < t < b.$$ 

Let $Z(\omega; t)$ be a centered Gaussian process, i.e., $\mathbb{E}[Z(\omega; t)] = 0$ for all $t \geq 0$, with Hölder-continuous sample paths for almost all $\omega \in \Omega$, and covariance function $C_Z(s, t) \equiv \text{Cov}(Z(s), Z(t))$. We next consider the stochastic integral

$$L(t) \equiv \int_0^t J(t, u) \, dZ(\omega; u), \quad t \geq 0, \quad (4.10)$$

where $J(t, u)$ is some deterministic two-parameter function and is differentiable with respect to the second component. To make sure (4.10) is well defined and to be able to characterize its distribution, we define a sequence of discrete version of (4.10), that is, $\{L^{(m)} : m \geq 1\}$, where

$$L^{(m)}(t) \equiv \sum_{i=0}^{m-1} J(t, u_i) (Z(\omega; u_{i+1}) - Z(\omega; u_i)), \quad t \geq 0, \quad (4.11)$$

for a given partition on the interval $[0, t]$, $0 = u_0 < u_1 < \ldots < u_m = t$.

**Proposition 4.4.1 (Integral with respect to Hölder-Continuous Gaussian Processes)**

Suppose $(Z(\omega; t) : t \geq 0)$ is a centered Gaussian process on the interval $[0, T]$ with almost-surely Hölder-continuous sample paths, and for each fixed $t$, the integrand $J$ is continuously differentiable with respect to the second component with $J_u(t, u) \equiv (\partial/\partial u)J(t, u)$. Then the stochastic integral in (4.10) is well-defined for each $\omega \in \Omega$ for which $(Z(\omega; t) : t \geq 0)$ has Hölder-continuous sample paths. In particular, for $0 \leq t \leq T$,

$$\int_0^t J(t, u) \, dZ(\omega; u) = J(t, t)Z(\omega; t) - J(t, 0)Z(\omega; 0) - \int_0^t Z(\omega; u) \, dJ(t, u) \quad (4.12)$$

where the equality holds almost surely, and the integral on the right-hand side is understood as the Riemann-Stieltjes integral.

Having defined a stochastic integral with respect to Hölder-continuous Gaussian processes, we next give a characterization of the Gaussian integral in (4.10).

**Theorem 4.4.1 (Characterization of the Gaussian integral)** The process $L(t)$ defined in
is a centered Gaussian process with the covariance function

\[
C_L(t_1, t_2) = J(t_1, t_1)J(t_2, t_2)C_Z(t_1, t_2) + J(t_1, 0)J(t_2, 0)C_Z(0, 0) - J(t_2, t_2)J(t_1, 0)C_Z(0, t_2) - J(t_1, t_1)J(t_2, 0)C_Z(0, t_1) - \int_0^{t_1} J(t_1, s)J_s(t_2, s)C_Z(s, t_1) ds - \int_0^{t_2} J(t_2, s)J_s(t_2, s)C_Z(s, t_2) ds - \int_0^{t_1} \int_0^{t_2} J_s(t_1, s)J_r(t_2, r)C_Z(s, r) dr ds
\]  

(4.13)

for \(0 \leq t_1 < t_2\) where, for fixed \(t \geq 0\), \(J_t(\cdot, \cdot)\) is the derivative with respect to the second component.

The proofs of Proposition 4.4.1 and Theorem 4.4.1 are given in §C.1.2 of the appendix.

4.4.2 Prelimit Processes

We next provide a representation of the prelimit enter-service process and queue-length process, that will be useful in the proving the main results.

**Enter-service process.** First, we give an expression for the \(E_n(t)\), that is, the total number of customers who have entered service in the interval \([0, t]\). Then Then

\[
E_n(t) = \sum_{i=1}^{N_n(t-W_n(t))} 1(\gamma_i^n > V_n(\tau_i^n))
\]  

(4.14)

\[
= n \int_0^{t-W_n(t)} \int_0^1 1(y > F(V_n(s))) d\bar{U}_n(N_n(s), y).
\]  

(4.15)

where \(\bar{U}_n\) is the sequential empirical processes reviewed in §1.8.2. The random variables \(0 \leq \tau_1^n \leq \tau_2^n \leq \cdots\) are the arrival times of the customers, and \(\gamma_1^n, \gamma_2^n, \ldots\) are i.i.d. patience times of those customers.

**Remark 4.4.1 (Understanding the representations (4.14)–(4.15))** We first remark that our new decomposition is more convenient than those in [65] in terms of facilitating our new proof technique. In particular, we use the virtual-waiting-time process \(V_n\) inside the indicator function and hence the nonlinear curve in Figure 4.1. We next carefully explain why Equations (4.14)–(4.15) hold. First, note that Assumption 4.2.3 implies \(W_n(t) \leq t\) for all \(t \geq 0\). To understand (4.14), note that all arrivals before time \(t - W_n(t)\) have already entered service if they don’t abandon; also the condition inside the indicator guarantees that customer \(i\) with arrival time \(\tau_i^n\) does not abandon (because its patience time \(\gamma_i^n\) is bigger than its potential
waiting time $V_n(\tau^n_i)$); see the blue shaded area in the right-hand figure of Figure 4.1. Following [54, 78] we obtain the equivalent integral representation in (4.15).

The process $E_n(t)$ in (4.15) can be decomposed into three terms, i.e,

$$E_n(t) \equiv E_{n,1}(t) + E_{n,2}(t) + E_{n,3}(t), \quad t \geq 0,$$

(4.16)

where

$$E_{n,1}(t) \equiv \sqrt{n} \int_0^{t-W_n(t)} F^c(V_n(s)) d\hat{N}_n(s),$$

(4.17)

$$E_{n,2}(t) \equiv \sqrt{n} \int_0^{t-W_n(t)} \int_0^\infty \mathbf{1}(y > F(V_n(s))) d\hat{U}_n(\hat{N}_n(s), y),$$

(4.18)

$$E_{n,3}(t) \equiv n \int_0^{t-W_n(t)} F^c(V_n(s)) d\Lambda(s), \quad t \geq 0,$$

(4.19)

and $\Lambda(\cdot)$ is as in Assumption 4.2.2.
Queue-length process. Just as in [65], the number of customer waiting in line at time $t$ can be represented as

$$Q_n(t) = \sum_{i=N_n(t-W_n(t))}^{N_n(t)} 1(\gamma_i^n + \tau_{n,i} > t)$$

$$= n \int_{t-W_n(t)}^{t} \int_{0}^{1} 1(y > F(t-s)) d\bar{U}_n(\bar{N}_n(s), y), \quad t \geq 0. \quad (4.20)$$

where $W_n(t)$ is the head-of-line waiting at time $t$. Similarly to (4.15), (5.24) can be represented as the sum of three terms, i.e,

$$Q_n(t) \equiv Q_{n,1}(t) + Q_{n,2}(t) + Q_{n,3}(t), \quad t \geq 0,$$

where

$$Q_{n,1}(t) \equiv \sqrt{n} \int_{t-W_n(t)}^{t} F^c(t-s) d\tilde{N}_n(s), \quad (4.21)$$

$$Q_{n,2}(t) \equiv \sqrt{n} \int_{t-W_n(t)}^{t} \int_{0}^{\infty} 1(x > F(t-s)) d\tilde{U}_n(\bar{N}_n(s), y), \quad (4.22)$$

$$Q_{n,3}(t) \equiv n \int_{t-W_n(t)}^{t} F^c(t-s) d\Lambda(s), \quad t \geq 0, \quad (4.23)$$

and $\Lambda(\cdot)$ is as in Assumption 4.2.2.

4.5 An FCLT for $G/GI/n + GI$ and Gaussian Limits

In this section, we present our new FCLT results for the $G/GI/n + GI$ model. We first give an FWLLL result (Theorem 4.5.1) drawing on [62]. We give an alternative proof based on our new representation in §4.4.2 in the appendix. We next give a new FCLT (Theorem 4.5.2).

**Theorem 4.5.1 (FWLLL)** If Assumptions 4.2.1-4.2.4 hold, then as $n \to \infty$,

$$(W_n, V_n, D_n, E_n, B_n, Q_n, X_n, N_n, A_n) \Rightarrow (w, v, D, E, X, Q, X, \Lambda, A) \quad \text{in} \quad D^9([0,T]; \mathbb{R}) \quad (4.24)$$

where $\chi(t) = 1, \ t \geq 0,$ and $\Lambda(t) = \lambda t, t \geq 0,$ as implied by the FCLT in Assumption 4.2.2. The deterministic limits $w$ and $v$ satisfy the fluid equations (4.9) with $w$ satisfying (5.12), or equivalently,

$$w(t) = \int_{0}^{t} \left( 1 - \frac{\mu}{\lambda F^c(w(u))} \right) du \quad \text{and} \quad v(t) = w(t + v(t)), \quad t \geq 0. \quad (4.25)$$
Moreover, for \( t \geq 0 \),
\[
D(t) = E(t) = \mu t, \quad Q(t) = \lambda \int_{t-w(t)}^{t} F_c(t-s) \, ds, \quad X(t) = Q(t) + 1, \tag{4.26}
\]
and \( A(t) = \Lambda(t) - E(t) - Q(t) \).

**Theorem 4.5.2 (FCLT)** Suppose Assumptions 4.2.1-4.2.4 hold and the limit \( \hat{N} \) in Assumption 4.2.2 is a zero-mean process. Then, as \( n \to \infty \),
\[
(W_n, \hat{V}_n, \hat{D}_n, \hat{E}_n, \hat{B}_n, \hat{Q}_n, \hat{X}_n, \hat{N}_n, \hat{A}_n) \Rightarrow (\hat{W}, \hat{V}, \hat{D}, \hat{E}, 0, \hat{Q}, \hat{N}, \hat{A}) \quad \text{in} \quad D^9([0,T];\mathbb{R}) \tag{4.27}
\]
where \( e(t) = t \), and \( \hat{A}(t) = \hat{N}(t) - \hat{Q}(t) - \hat{E}(t) \), for \( t \geq 0 \).

The FCLT limit for the enter-service process \( (\hat{E}(t) : t \geq 0) \) is a centered Gaussian process with covariance
\[
C_{E}(t,t') \equiv \text{Cov}(\hat{E}(t),\hat{E}(t')) = E[S_1(t)S_1(t')] - \mu^2 tt', \quad t, t' \geq 0, \tag{4.28}
\]
where \( S_1 \) is the equilibrium renewal process associated with service completions at each server as described in §4.2.

The FCLT limit for the HOL waiting time \( (\hat{W}(t) : t \geq 0) \) uniquely solves the SDE
\[
\hat{W}(t) = -\frac{1}{F_c(w(t))} \int_{0}^{t} f(w(s)) \hat{W}(s) \, ds + \frac{1}{\lambda F_c(w(t))} \hat{G}(t), \quad t \geq 0, \tag{4.29}
\]
where \( w \) is as in (4.25), \( f \) is the derivative of the patience-time distribution \( F \),
\[
\hat{G}(t) \equiv \int_{0}^{t} F_c(w(s)) \, d\hat{N}(s-w(s)) + B \left( \lambda \int_{0}^{t} F_c(v(u)) F(v(u)) \, du \right) - \hat{E}(t), \quad t \geq 0, \tag{4.30}
\]
with \( (B(t) : t \geq 0) \) being the standard Brownian motion independent of the processes \( \hat{N} \) and \( \hat{E} \). The first term in (5.38) is interpreted as the form after integration by parts. The deterministic function \( v \) in (5.38) is characterized by (4.25).

The FCLT limit for the virtual waiting time satisfies
\[
\hat{V}(t) = \frac{\hat{W}(t + v(t))}{1 - \hat{w}(t + v(t))}, \quad t \geq 0, \tag{4.31}
\]
where \( \hat{w} \) is the derivative of \( w \), \( w \) and \( v \) are as in (4.25).

The FCLT limit for the queue-length process is the sum of three independent terms, i.e.,
\[
\hat{Q}(t) \equiv \hat{Q}_1(t) + \hat{Q}_2(t) + \hat{Q}_3(t)
\]
with

\[ \hat{Q}_1(t) \equiv \int_{t-w(t)}^{t} F^c(t-s) \, d\hat{N}(s), \quad \hat{Q}_2(t) \equiv \int_{t-w(t)}^{t} \int_{0}^{1} 1(x > F(t-s)) \, d\hat{U}(\lambda s, x), \]

\[ \hat{Q}_3(t) \equiv \lambda F^c(w(t)) \hat{W}(t), \quad t \geq 0, \tag{4.32} \]

where \( \hat{U} \) is the standard Kiefer process reviewed in §1.8.2 of the appendix.

**Remark 4.5.1 (M service)** For the \( G/M/n + GI \) model having an exponential service distribution, we remark that the FCLT limit for the enter-server process is a Brownian motion, i.e., \( \hat{E}(t) = B_s(\mu t) \), where \( B_s \) is the standard BM independent of the other terms in (5.38), because \( S_1 \) becomes a Poisson process with rate \( \mu \). This is consistent with the results in [39, 65].

**Remark 4.5.2 (Separation of variability)** The process \( \hat{G} \) in (5.38) is characterized by three independent terms. The first term captures the randomness of arrivals (as a function of \( \hat{N} \)); the second term captures the randomness of the patience times of customers waiting in queue at time 0; the third term captures the randomness of the service times of all customers (through \( \hat{E} \)). Asymptotic independence of the limits \( \hat{N} \) and \( \hat{E} \) is implied by the assumption that the queue operates in the ED regime throughout the interval \([0, T]\) whereas independence of the time-changed Brownian motion in (5.38) follows from mutual independence of arrivals, service times, and patience times. To argue independence of three terms in (4.32), we note that applying change of variable to \( \hat{Q}_1 \) with \( s = u - w(u) \) along with the fact that \( \hat{N} \) has independent increments implies independence of the first terms in (5.38) and (4.32), respectively. Similarly, change of variable can be applied to the second term in (5.38) to show that second term is independent of \( \hat{Q}_2 \) by using the fact that Kiefer process has independent increments with respect to the first component.

### 4.6 Characterizing the Distributions of the FCLT Limits

In §4.6.1, we give an explicit solution \( \hat{W}(t) \) to the SDE (5.37). In §4.6.2, we provide covariance formulas for the limit HOL waiting time \( \hat{W} \) and queue-length process \( \hat{Q} \). Those formulas will be useful in deriving the steady-state variance formulas in §4.6.3. The results in §4.4.1 will be useful in characterizing the limit \( \hat{W} \) (see (5.49)), and in deriving variance formulas.

#### 4.6.1 Analytic Solution for \( \hat{W} \)

In this section, we solve for the SDE (5.37) and show that its solution \( \hat{W} \) has the form (5.49) involving a Gaussian integral. To prove that the Gaussian integral is a well-defined stochastic integral that falls in the framework introduced in §4.4.1, we first give a result on computing
the covariance of a stationary renewal process (Theorem 4.6.1), which is used to derive the covariance function for the limiting Gaussian process \( \hat{E} \) in (4.28). Next, we justify that, for almost all \( \omega \in \Omega \), the limiting process \( \hat{E} \) in (4.28) has Hölder continuous sample paths (Lemmas 4.6.1–4.6.2 and Proposition 4.6.1). Finally, we compute the covariance functions of the FCLT limits (Theorem 4.6.2).

We start with a theorem on the covariance function of stationary renewal processes.

**Theorem 4.6.1 (Theorem 7.2.4. of [102] with a correction)** Suppose \( A \) is a stationary renewal counting process (having stationary increments with \( A(0) = 0 \)) with interrenewal-time cdf \( F \) having pdf \( f \) and mean \( \lambda^{-1} \). Then, for \( t < u \),

\[
\text{Cov}(A(t), A(u)) = \text{Var}(A(t)) + \text{Cov}(A(t), A(u) - A(t)),
\]

where

\[
\text{Var}(A(t)) = 2\lambda \int_0^t (M(s) - \lambda s + 0.5) \, ds = 2\lambda \int_0^t M(s) \, ds - \lambda^2 t^2 + \lambda f(4.34)
\]

\[
\text{Cov}(A(t), A(u) - A(t)) = \mathbb{E}[A(t)(A(u) - A(t))] - \lambda^2 t(u - t),
\]

\[
\mathbb{E}[A(t)(A(u) - A(t))] = \lambda \int_0^t da \int_0^{u-t} db f(a+b)[1 + M(t-a)][1 + M(u-t-b)](4.36)
\]

where \( M(t) \) is the renewal function of the associated ordinary renewal process, satisfying the renewal equation (or fixed-point equation)

\[
M(t) = F(t) + \int_0^t M(t-x)f(x) \, dx.
\]

**Proof 4.6.1** The proof of (4.34) is given on p.57 of [15]. For (4.36), consider the first renewal occurs after \( t \); it falls at \( t+b \) with the stationary-excess pdf \( f_{x}(b) \equiv \lambda F^{c}(b) \). Conditional on that renewal being at \( t+b \), the last renewal by \( t \) occurs at \( t-a \) with pdf \( f(a+b)/F^{c}(b) \). Conditional on the time of these two renewals at \( t-a \) and \( t+b \), we have

\[
\mathbb{E}[A(t)(A(u) - A(t))] = \int_0^t da \int_0^{u-t} db \mathbb{E}[A(t)(A(u) - A(t))|S_{N(t)} = t-a, S_{N(t)+1} = t+b] \lambda f(a+b)
\]

\[
= \lambda \int_0^t da \int_0^{u-t} db [M(t-a) + 1][M(u-t-b) + 1]f(a+b).
\]

In the following two lemmas, we give two results on Hölder continuity of sample paths of Gaussian processes. Then using these results we show in Proposition 4.6.1 that the Gaussian process \( \hat{E} \) indeed has Hölder-continuous sample paths.
Lemma 4.6.1 (Kolmogorov’s Lemma in [94]) If \((X_t : t \in \mathbb{R}^n)\) is a stochastic process with values in a complete separable metric space \((S, \rho)\), and if there exist positive constants \(\alpha, C, \varepsilon\) such that, for all \(s, t \in \mathbb{R}^n\)

\[
\mathbb{E}[\rho(X_s, X_t)^\alpha] \leq C|s - t|^{n+\varepsilon}
\]

then there exists a continuous version of \(X\). This version is Hölder continuous of order \(\theta\) for each \(\theta < \varepsilon/\alpha\).

The following is a corollary to Kolmogorov’s lemma. See [94] for a proof.

Lemma 4.6.2 (Corollary (25.6) of [94]) Let \((X_t : t \in \mathbb{R}^n)\) be a centered Gaussian process with covariance function \(\rho(s, t) \equiv \mathbb{E}[X_s X_t]\). A sufficient condition for the existence of a continuous version is that \(\rho\) should be locally Hölder continuous: for each \(N \in \mathbb{N}\) there exists \(\theta = \theta(N) > 0\) and \(C = C(N)\) such that, for \(|s|, |t| \leq N\),

\[
|\rho(s, t) - \rho(t, t)| \leq C|s - t|^\theta.
\]

(4.38)

Proposition 4.6.1 The FCLT limit for the enter-service process \(\hat{E}(t) : t \geq 0\) is Hölder continuous for almost all \(\omega \in \Omega\) for some \(\alpha > 0\).

Proof 4.6.2 Combining (4.33)-(4.36) and (4.38), we obtain for \(u > t\)

\[
|\rho(u, t) - \rho(t, t)| = |\text{Cov}(\hat{E}(t), \hat{E}(u) - \hat{E}(t))|
\]

\[
= \left| \mu \int_0^t da \int_0^{u-t} db g(a + b)[1 + M(t - a)][1 + M(u - t - b)] - \mu^2 t(u - t) \right|
\]

where \(\mu > 0\) is the service rate and \(g\) is the service-time pdf. Then for any \(0 \leq t < u \leq N\) and \(a \in [0, t], b \in [0, u - t]\),

\[
\rho(u, t) - \rho(t, t) \leq \mu \int_0^t da \int_0^{u-t} db \|g\|_N[1 + M(N)]^2 - \mu^2 t(u - t)
\]

\[
= \mu t(u - t)\|g\|_N[1 + M(N)]^2 - \mu^2 t(u - t)
\]

\[
\leq \left( \mu N\|g\|_N[1 + M(N)]^2 + \mu^2 N \right) (u - t)
\]

which satisfies the sufficient condition (4.38) after taking the absolute value of both sides. Hence we deduce that \(\hat{E}(t)\) has a version with continuous sample paths. Moreover, by Lemma 4.6.1, that version has sample paths that are Hölder continuous of order \(\alpha\) for some \(\alpha > 0\).

Having proved that the Gaussian process \(\hat{E}\) has the desired sample-path properties, we next provide a closed-form solution \(\hat{W}\) to the SDE (5.37) in the following corollary.
Corollary 4.6.1 The stochastic differential equation (5.37) has a unique strong solution

\[
\hat{W}(t) \equiv \hat{W}_1(t) + \hat{W}_2(t) + \hat{W}_3(t) \\
= \int_0^t \frac{F_c(w(u))H(t,u)}{q(t,w(t))} d\hat{N}(u - w(u)) + \int_0^t \frac{\sqrt{\lambda F_c(v(u))F(v(u))}H(t,u)}{q(t,w(t))} dB(u) \\
- \int_0^t \frac{H(t,u)}{q(t,w(t))} d\hat{E}(u), \quad t \geq 0, \tag{4.39}
\]

where the first term on the right-hand side defined in Lemma 4.7.1, the third term is defined in §4.4.1; \(q(t,w(t)) = \lambda F_c(w(t))\), and

\[
H(t,u) \equiv e^{\int_u^t h(r) \, dr} \quad \text{with} \quad h(r) \equiv -\frac{\lambda f(w(r))}{q(t,w(t))} = -\frac{f(w(r))}{F_c(w(t))}, \quad 0 \leq r \leq t. \tag{4.40}
\]

To verify (5.49) is indeed the unique strong solution to (5.37), we argue that the rules of Itô calculus can be applied despite the fact that the Gaussian processes \(\hat{E}(t)\) and \(\hat{N}(t)\) are not necessarily Brownian motions or semimartingales. This is because the non-Brownian integrals are Riemann-Stieltjes integrals with deterministic integrands for almost all \(\omega \in \Omega\). Moreover, the integrators \(\hat{N}, \hat{E}\) and the Brownian motion in (5.49) are all independent. First, we define \(\check{q}(u,w(u)) = q(t,w(t))\) for \(0 \leq s \leq t\) where \(q(t,w(t)) = \lambda F_c(w(t))\) for \(t \geq 0\). Then we rewrite (5.37) as

\[
\hat{W}(t) = -\int_0^t \frac{\lambda f(w(u))}{\check{q}(u,w(u))} \hat{W}(u) \, du + \int_0^t \frac{\sqrt{\lambda F_c(v(u))F(v(u))}}{\check{q}(u,w(u))} dB(u) \\
- \int_0^t \frac{1}{\check{q}(u,w(u))} d\hat{E}(u) + \int_0^t \frac{F_c(w(u))}{\check{q}(u,w(u))} d\hat{N}(u - w(u)), \quad t \geq 0, \tag{4.41}
\]

where the time-changed Brownian term in (5.37) is replaced with an equivalent Itô integral. The uniqueness and existence of a solution to (4.41) immediately follows from Ito theory because the last two terms of (4.41) do not involve \(\hat{W}(t)\) and are independent of the Brownian motion in the second term. Furthermore, the last two integrals are Riemann-Stieltjes integrals for almost all \(\omega \in \Omega\). Consequently, we can use Ito’s formula to solve (4.41). In particular, using the differential form, we have

\[
d\hat{W}(t) = h(t)\hat{W}(t) \, dt + K_1(t)dB_t + K_2(t)d\hat{E}(t) + K_3(t) d\hat{N}(t - w(t))
\]

which implies

\[
d \left( e^{-\int_0^t h(r) \, dr} \hat{W}(t) \right) = K_1(t)dB_t + K_2(t)d\hat{E}(t) + K_3(t) d\hat{N}(t - w(t))
\]
Integrating both sides and multiplying through by $H(t, 0) \equiv e^{\int_0^t h(r) \, dr}$ yields (5.49).

### 4.6.2 Covariance Formulas

We now further characterize the distribution of the FCLT limits. Since all the FCLT limits are Gaussian processes, it suffices to compute their means and covariances. For simplicity, we assume that the limit for the CLT-scaled sequence of arrival processes is a time-changed Brownian motion. The formulas given in the following theorem can be modified by using (4.13) to recover the case where $\hat{N}$ is centered Gaussian process with properties given in Assumption 4.2.2.

**Theorem 4.6.2 (Further characterization of the FCLT limits)** Suppose that $\hat{N}(t) = \sqrt{\lambda c} B_\lambda(t)$, $t \geq 0$, with $B_\lambda$ being the standard Brownian motion. Then $\hat{W}$, $\hat{V}$ and $\hat{Q}$ are all centered Gaussian processes with the covariance functions

$$\text{Cov}(\hat{W}(t), \hat{W}(t')) = C_W(t, t') = C_{W_1}(t, t') + C_{W_2}(t, t') + C_{W_3}(t, t')$$

$$\text{Cov}(\hat{V}(t), \hat{V}(t')) = C_V(t, t') = \frac{C_{\hat{W}}(t, t')}{(1 - \hat{w}(t))(1 - \hat{w}(t'))},$$

$$\text{Cov}(\hat{Q}(t), \hat{Q}(t')) = C_Q(t, t') = C_{Q_1}(t, t') + C_{Q_2}(t, t') + C_{Q_3}(t, t'),$$

for $t \geq 0$, $t' \geq 0$ where

$$C_{W_1}(t, t') = \lambda c^2 \int_0^{t \wedge t'} \frac{F^c(w(u))^2 H(t, u)^2}{q(u, w(u))^2} (1 - \hat{w}(u)) \, du,$$

$$C_{W_2}(t, t') = \lambda \int_0^{t \wedge t'} \frac{F^c(v(u)) F(v(u)) H(t, u)^2}{q(u, w(u))^2} \, du,$$

$$C_{W_3}(t, t') = C_E(t, t') - \int_0^{t'} J_u(t', u) C_E(u, t) \, du - \int_0^{t} J_u(t, u) C_E(u, t') \, du$$

$$+ \int_0^{t} \int_0^{t'} J_u(t, u) J_u(t', v) C_E(u, v) \, dv \, du,$$

$$C_{Q_1}(t, t') = \lambda c^2 \int_{(t-w(t)) \vee (t'-w(t'))}^{t \wedge t'} F^c(t - s) F^c(t' - s) \, ds,$$

$$C_{Q_2}(t, t') = \lambda \int_{(t-w(t)) \vee (t'-w(t'))}^{t \wedge t'} F(t \wedge t' - s) F^c(t \vee t' - s) \, ds,$$

$$C_{Q_3}(t, t') = \lambda^2 F^c(w(t)) F^c(w(t')) C_{\hat{W}}(t, t').$$

(4.42)
with \( J(t,u) \equiv H(t,u)/q(u,w(u)) \) where \( H(t,u) \) and \( q(u,w(u)) \) are as in (5.50), and

\[
J_u(t,u) \equiv \frac{\partial}{\partial u} J(t,u) = -\frac{h(u)H(t,u)}{q(u,w(u))} - \frac{q(u,w(u))H(t,u)}{q(u,w(u))^2} \\
= \frac{(1 - \dot{w}(u))h_F(w(u))\exp\left(-\int_u^t (1 - \dot{w}(s))h_F(w(s))\,ds\right)}{\lambda F^c(w(u))} \\
+ \frac{f(w(u))\exp\left(-\int_u^t (1 - \dot{w}(s))h_F(w(s))\,ds\right)}{\lambda F^c(w(u))^2}, \quad u \geq 0, \quad t \geq 0.
\]

(4.43)

A proof is given in Appendix C.3.

### 4.6.3 Approximating the Steady-State Distributions

We now characterize the steady-state distributions of the FCLT limits as \( t \to \infty \). In particular, we show that the steady state of the FCLT limits are centered Gaussian random variables and we compute their variances to fully characterize their distributions. Consequently, we obtain approximations for the long-run performance of the \( n \)th queueing system, i.e.,

\[
Q_n(\infty) \approx nQ(\infty) + \sqrt{n} \hat{Q}(\infty), \quad W_n(\infty) \approx w(\infty) + \frac{1}{\sqrt{n}} \hat{W}(\infty), \quad \text{for all } \quad n \geq 1,
\]

where \( Q(\infty) \) and \( w(\infty) \) are the long-run limits (i.e., \( t \to \infty \)) of the fluid functions in §5.3 and they are given in (4.44).

First, Theorem 3.1 of [104] and Theorem 4.1 of [61] imply that the fluid queueing system stays in steady state once in the steady state, that is,

\[
w(t) = w = F^{-1}\left(1 - \frac{1}{\rho}\right), \quad v(t) = v = w, \quad q(t,w(t)) = \lambda F^c(w) \quad \text{and}
\]

\[
Q(t) = \lambda \int_0^w F^c(x)\,dx \quad \text{for all } \quad t \geq t^*
\]

(4.44)

where \( t^* \) is the time when fluid queueing system is assumed to be in the steady state.

Next, we characterize the steady state of the FCLT limits in Theorem 4.5.2 by letting \( t \to \infty \). In particular, we first obtain the steady-state limit for the head-of-line waiting time process \( \hat{W} \) which is also necessary for characterization of the steady-state limit for the queue length process \( \hat{Q} \). For simplicity, we continue to assume that the FCLT limit for the arrival process is a (time-change) Brownian motion, i.e., \( \hat{N}(t) \equiv \sqrt{\lambda} c_B(t), \quad t \geq 0 \). However, as discussed in Remark C.3.1, we can extend the results in this section to cover a more general case in which the limit \( \hat{N} \) has stationary increments in addition to the properties in Assumption 4.2.2. To characterize
Figure 4.2: Reflection of $\hat{E}$ on $(-\infty, 0]$.

the steady-state limit for $\hat{W}$, we define in Lemma 4.6.3 another process that can be interpreted as the reflection of the Gaussian limit $\hat{E}$ given in Theorem 4.5.2 on the interval $(-\infty, 0]$. Prior to stating the main result in this section we remark that, with the steady-state definitions in (4.44), (5.50) becomes

$$H(t, u) = e^{-h_F(w)(t-u)} + o(1) \quad (4.45)$$

for sufficiently large $t \geq t^*$ where $h_F$ is the hazard-rate function associated with the cdf $F$ and $w$ is as in (4.44).

We next derive the expression for the variance of the steady state for $\hat{W}(t)$ as $t \to \infty$. The convergence, as $t \to \infty$, of the variance function of $\hat{W}_i(t)$ in Corollary 5.6.1 is straightforward for $i = 1, 2$. It suffices to set $t = t'$ in Theorem 4.6.2 and let $t \to \infty$. Unfortunately, the treatment of $\hat{W}_3(t)$ is not straightforward because, from (4.42), we observe that $C_{\hat{W}_3} \to \infty$ since $\text{Var}(\hat{E}(t)) = C_{\hat{E}}(t, t) \to \infty$ as $t \to \infty$. Therefore, we define another Gaussian process $\tilde{E}$ in Lemma 4.6.3 that can be understood as a reflection of $\hat{E}$ on the negative half line to resolve this issue (see Figure 4.2). The proof of Lemma 4.6.3 is given in the appendix.

**Lemma 4.6.3** (Reflection of $\hat{E}$ on $(-\infty, 0]$) There exists a Gaussian process $\{\tilde{E}(t) : -\infty < t \leq 0\}$ such that (i) $\tilde{E}(0) = 0$; (ii) $\mathbb{E}[\tilde{E}(t)] = 0$ for all $-\infty < t \leq 0$; and (iii) the covariance function

$$\tilde{C}(-x, -y) \equiv C_{\hat{E}}(x \lor y, x \lor y) - C_{\hat{E}}(x \lor y, |x - y|) \quad \text{for} \quad x \geq 0, y \geq 0, \quad (4.46)$$

with $C_{\hat{E}}$ being the covariance function in (4.28). In addition, $\tilde{E}$ has the same stationary increment distribution as $\hat{E}$. In particular, for any $t > 0$,

$$(\tilde{E}(s-t) - \tilde{E}(-t) : 0 \leq s \leq t) \overset{d}{=} (\hat{E}(s) : 0 \leq s \leq t). \quad (4.47)$$

Using the reflected version $\tilde{E}$, we next obtain a more convenient expression for $\text{Var}(\hat{W}_3(t))$ in Theorem 4.6.3. The proof of Theorem 4.6.3 is in the appendix.

**Theorem 4.6.3** (Steady state of the FCLT limits) Suppose $\hat{N}(t) \equiv \sqrt{\lambda} B_{\lambda}(t), t \geq 0$, with $B_{\lambda}$ being the standard Brownian motion. Then the steady-state distributions as $t \to \infty$
for the FCLT limits for $\hat{W}(t)$, $\hat{V}(t)$ and $\hat{Q}(t)$ in Theorem 4.5.2, denoted by $\hat{W}(\infty)$, $\hat{V}(\infty)$ and $\hat{Q}(\infty)$, are Gaussian random variables with means 0 and variances

$$
\text{Var}(\hat{W}(\infty)) = \text{Var}(\hat{V}(\infty)) = \sigma_W^2 \equiv \frac{c^2}{2h_F(w)} + \frac{F(w)}{2\lambda f(w)} + 2 \frac{h_F(w)^2}{\lambda^2 F_c(w)^2} \int_0^x \int_0^y e^{-h_F(w)(x+y)} \tilde{C}(-x,-y) \, dy \, dx.
$$

(4.48)

$$
\text{Var}(\hat{Q}(\infty)) \equiv \sigma_Q^2 \equiv \lambda c^2 \int_0^w F_c(u)^2 \, du + \lambda \int_0^w F(u) F_c(u) \, du + \lambda^2 F_c(w)^2 \sigma_W^2,
$$

(4.49)

where $\tilde{C}(-x,-y)$ is as in (4.46), $h_F$ is the hazard-rate function associated with the patience-time distribution $F$, and $w$ is the steady-state fluid HOL waiting time given in (4.44).

We are able to recover special cases in the literature. In the following corollary, we show that our expressions agree with those given in Proposition 2 of Huang et al. [39] in which service distribution is the exponential distribution. In particular, we let $f_w(x) \equiv f(w)x$ since the patience-time distribution is not scaled in our framework, and $\beta = 0$ since we do not consider the ED+QED regime. The variance in (4.50) is slightly different than the variance of the stationary distribution $\pi$ in Proposition 2 of [39]. In particular, (4.50) contains $\lambda$ unlike the variance of $\pi$ in [39]. We remark that this is due to a minor difference in the way arrival processes are scaled, and that this does not cause inconsistency. We also recover Theorem 2.1 of [103] for the fully Markovian $M/M/n + M$ queueing system.

**Corollary 4.6.2 (Special cases of $M/M/n + M$ and $G/M/n + GI$)** Suppose $\tilde{N}(t) \equiv \sqrt{\lambda}c_\lambda B_\lambda(t)$. Then for the $G/M/n + GI$ queue having i.i.d. exponential service times, steady-state variance of the limit $\hat{W}$ reduces to

$$
\sigma^2_W = \frac{(c^2 - 1) + 2\rho}{2\lambda h_F(w)}.
$$

(4.50)

For the $M/M/n + M$ queue with Poisson arrivals, i.i.d. exponential service times, and i.i.d. exponential patience times, the steady-state variance of the limiting process $\hat{W}$ reduces to

$$
\sigma^2_W = \frac{1}{\mu \theta}, \quad \sigma^2_Q = \frac{\lambda}{\theta},
$$

(4.51)

where $\theta > 0$ is the abandonment rate.

**Proof.** See §C.3.3 in appendix for a proof.
4.7 Proof of Main Results

We now prove the two main results in §4.5. The proof of Theorem 4.5.1 is similar to the proof of Theorem 4.1 of [65], and it is given in §C.2. We hereby prove Theorem 4.5.2. The key step is to use Gronwall’s inequality to first prove stochastic boundedness of and then convergence for the sequence of scaled processes \( \{ \hat{W}_n : n \geq 1 \} \) in \( D([0, T]; \mathbb{R}) \). Once convergence for \( \{ \hat{W}_n : n \geq 1 \} \) is established, the limits for the other processes are characterized by continuous mapping theorem.

**Remark 4.7.1 (Nonstationary arrivals)** Although both Theorem 4.5.1 and Theorem 4.5.2 are stated under the assumption of stationary arrivals, i.e., \( \lambda(t) = \lambda \) and \( \Lambda(t) = \lambda t, 0 \leq t \leq T \), we hereby present a more general framework in which the arrival process is nonstationary in the sense that

\[
\hat{N}_n \Rightarrow \Lambda \quad \text{in} \quad D([0, T]; \mathbb{R}) \quad \text{as} \quad n \to \infty, \quad \text{with} \quad \Lambda(t) = \int_0^t \lambda(u) \, du, \quad t \in [0, T], \quad (4.52)
\]

and \( \lambda(\cdot) \) is not necessarily constant. The proofs provided in §C.2 and §C.4 cover the more general case of nonstationary arrivals. Overloaded systems with time-varying arrivals are common in healthcare applications, therefore, the FCLT with time-varying arrivals can stimulate future research on transient analysis of \( G_t/GI/n + GI \) with time-varying arrivals. The results in Theorem 4.5.1 and Theorem 4.5.2 are recovered when \( \lambda(t) = \lambda \) and \( \Lambda(t) = \lambda t, 0 \leq t \leq T \).

In §4.7.1, we establish an FCLT for the sequences \( \{ \hat{W}_n : n \geq 1 \} \) and \( \{ \hat{V}_n : n \geq 1 \} \). Then, in §4.7.2, we establish FCLT for the other processes given the FCLT in §4.7.1.

**4.7.1 FCLT for \( \hat{W}_n \)**

It might be possible to prove the FCLTs for \( \hat{W}_n \) using the compactness approach: (i) tightness and (ii) characterization of the limit for every convergent subsequences [39, 37, 65]. But that approach would involve a complicated proof of tightness. We hereby take a different approach: we first establish the convergence \( \hat{W}_n \Rightarrow \hat{W} \) using continuous mapping theorem and Gronwall’s inequality. We show that the limiting process \( \hat{W} \) uniquely solves an SDE. Our SDE here generalizes the SDE given in (6.65) of [65]. The generalization of service-time distribution from exponential to nonexponential distributions cause a change in (6.65) of [65]. In particular, the stochastic integral with respect to the Brownian motion \( B_s \) therein now have to be replaced by a stochastic integral with respect to a centered Gaussian process as in §4.4.1. Our approach has two advantages. First, it is simpler in the sense that it avoids proof of tightness in space \( D \). Second, this method may be used to treat other complicated queueing systems.
Overview of proof.

The key step of our proof is establishing convergence for the scaled HOL-waiting-time processes \( \{ \hat{W}_n : n \geq 1 \} \), and characterizing its limit. To do so, we first derive an SDE for the scaled prelimit head-of-line waiting-time process \( \hat{W}_n \) for each \( n \geq 1 \) (see (4.66)). Then by using Gronwall’s inequality (Lemma 4.7.3), we show that the sequence \( \{ \hat{W}_n : n \geq 1 \} \) converges uniformly to a limiting process \( \hat{W} \) as \( n \to \infty \) (see (4.70)). Given the convergence \( \hat{W}_n \Rightarrow \hat{W} \) in \( \mathbb{D}([0,T]; \mathbb{R}) \), we establish an FCLT for the other processes, and characterize their limits in \( \S 4.7.2 \).

The decomposition for the prelimit enter-service process \( E_n(t) \) given in (4.16)-(4.19) will be useful in deriving an SDE for \( \hat{W}_n \). From (C.6) and the discussion in \( \S C.2.3 \), we know that \( \bar{E}_{n,1}(t) \Rightarrow (0 \epsilon)(t) \), \( \bar{E}_{n,2}(t) \Rightarrow (0 \epsilon)(t) \) in \( \mathbb{D}([0,T]; \mathbb{R}) \), \( \bar{E}_{n,3}(t) \Rightarrow E_3(t) \equiv \int_0^{t-w(t)} F_c(v(s)) d\Lambda(s) \) in \( \mathbb{D}([0,T]; \mathbb{R}) \) as \( n \to \infty \). (4.53)

Recall from (4.6) that we also have \( E(t) = \mu t \), \( t \geq 0 \). The equality of \( E(t) = \mu t \) and \( E_3(t) \) in (4.53) is then guaranteed by the fact that (5.12) has a unique solution. In particular, taking the derivative of both sides of the equality \( E_3(t) = \mu t \) yields (5.12). By definition, it follows that

\[
\hat{E}_n(t) = \frac{1}{\sqrt{n}} E_{n,1}(t) + \frac{1}{\sqrt{n}} E_{n,2}(t) + \frac{1}{\sqrt{n}} (E_{n,3}(t) - nE_3(t)), \quad 0 \leq t \leq T. \quad (4.54)
\]

Convergence of the First and Second Terms in (4.54)

In Lemma 4.7.1 and Lemma 4.7.2, we respectively establish convergence for the first and the second component on the right-hand side of the equality in (4.54). The former is proved using the continuous mapping approach; the latter is proved by using the martingale convergence theorem (see Theorem 7.1.4. of [20]). Their proofs are given in \( \S C.4.1 \) and \( \S C.4.2 \) of the appendix.

Lemma 4.7.1 (Convergence of the first term in (4.54)) Suppose

\[
(\hat{N}_n, \bar{D}_n) \Rightarrow (\bar{N}, D) \quad \text{in} \quad \mathbb{D}([0,T]; \mathbb{R}) \quad \text{as} \quad n \to \infty.
\]

Then

\[
\frac{1}{\sqrt{n}} E_{n,1}(t) \Rightarrow \int_0^t F_c(w(s)) d\bar{N}(s - w(s)) \quad \text{in} \quad \mathbb{D}([0,T]; \mathbb{R}) \quad \text{as} \quad n \to \infty \quad (4.55)
\]

where the deterministic function \( w(t) \) satisfies (4.9) and the integral is understood as integration by parts for almost all \( \omega \in \Omega \).
Lemma 4.7.2 (Convergence of the second term in (4.54)) Suppose
\[(\hat{N}_n, \bar{D}_n) \Rightarrow (\hat{N}, D) \text{ in } \mathbb{D}([0, T]; \mathbb{R}) \text{ as } n \to \infty.\]

Then
\[
\frac{1}{\sqrt{n}} E_{n,2}(t) \Rightarrow \mathcal{B} \left( \int_0^t F^c(v(u))F(v(u)) \, d\Lambda(u) \right) \quad \text{in } \mathbb{D}([0, T]; \mathbb{R}) \text{ as } n \to \infty \quad (4.56)
\]
where \(\mathcal{B}\) is the standard Brownian motion, \(\Lambda(t)\) is as in (4.52) and \(v(t)\) is as in (4.9).

Treating the third term in (4.54).

Following (4.19) and (4.53), we have
\[
E_{n,3}(t) - nE_3(t)
= n \int_0^{t-W_n(t)} F^c(V_n(s)) \, d\Lambda(s) - n \int_0^{t-w(t)} F^c(V_n(s)) \, d\Lambda(s) \\
+ n \int_0^{t-w(t)} [F^c(V_n(s)) - F^c(v(s))] \, d\Lambda(s) \\
= n \int_0^{t-W_n(t)} F^c(V_n(s)) \, d\Lambda(s) + n \int_0^{t-w(t)} [F^c(V_n(s)) - F^c(v(s))] \, d\Lambda(s) \\
= -\sqrt{n}F^c(\theta_2,n(t))\lambda(t - w(t))\hat{W}_n(t) - \sqrt{n} \int_0^{t-w(t)} f(\theta_3,n(s))\hat{V}_n(s) \, d\Lambda(s), \quad 0 \leq t \leq T, \quad (4.57)
\]
where \(f\) is the derivative of the patience-time distribution \(F\), and for \(0 \leq t \leq T\),
\[
V_n(t - W_n(t)) \land V_n(t - w(t)) \leq \theta_2,n(t) \leq V_n(t - W_n(t)) \lor V_n(t - w(t)), \quad (4.58)
V_n(t) \land v(t) \leq \theta_3,n(t) \leq V_n(t) \lor v(t). \quad (4.59)
\]

Deriving an SDE for \(\hat{W}_n\). Having treated all the terms in (4.54), we next derive an SDE for the prelimit processes \((W_n(t) : 0 \leq t \leq T)\) by using two representations of the enter-service process \((E_n(t) : 0 \leq t \leq T)\). Then we will obtain a stochastic integral equation for \((\hat{W}_n(t) : 0 \leq t \leq T)\) for each \(n \geq 1\), and use Gronwall’s inequality to prove that \(\{\hat{W}_n : n \geq 1\}\) is stochastically bounded in \(\mathbb{D}([0, T]; \mathbb{R})\). Gronwall’s inequality will also be used for establishing the convergence \(\hat{W}_n \Rightarrow \hat{W}\) in \(\mathbb{D}([0, T]; \mathbb{R})\) as \(n \to \infty\).
From Lemma 4.7.1, Lemma 4.7.2 and (4.57), we have

\[
E_n(t) = E_{n,1}(t) + E_{n,2}(t) + E_{n,3}(t) \\
= \sqrt{n} \int_0^t F^c(w(s)) \, d\hat{\lambda}(s) \, d\hat{\lambda}(s) + \sqrt{n} \mathcal{B} \left( \int_0^t F^c(v(u)) F(v(u)) \, d\lambda(u) \right) \\
- \sqrt{n} F^c(\theta_2_n(t)) \lambda(t - w(t)) \hat{W}_n(t) - \sqrt{n} \int_0^{t - w(t)} f(\theta_3_n(s)) \hat{V}_n(s) \, d\lambda(s) \\
+ nE(t) + o(\sqrt{n}), \quad 0 \leq t \leq T.
\] (4.60)

We want to rewrite the last integral term in (4.60) as a function of \( \hat{W}_n \) instead of \( V_n \) to be able to apply Gronwall’s inequality to prove stochastic boundedness of the sequence \( \{\hat{W}_n : n \geq 1\} \) in \( \mathbb{D}([0, T]; \mathbb{R}) \). To rewrite the integral term as a function of \( \hat{W}_n \), we will apply change of variable. However, first, we present some results on the relation between \( \hat{W}_n \) and \( \hat{V}_n \) that will be useful as we will be applying change of variable.

Let \( \Delta V_n(t) \equiv V_n(t) - v(t) \) and \( \Delta W_n(t) \equiv W_n(t) - w(t) \), \( 0 \leq t \leq T \), where \( w(t) \) and \( v(t) \) are as in (4.9). Using (5.2) we write

\[
\Delta V_n(t) = \Delta W_n(t + V_n(t) + O(1/n)) + w(t + V_n(t)) - w(t + v(t)) + O(1/n) \\
= \Delta W_n(t + V_n(t) + O(1/n)) + w(t + v(t)) \Delta V_n(t) + o(\Delta V_n(t)) + O(1/n), \quad 0 \leq t \leq T,
\]

where \( w(t) \equiv (dw/dt)(t) \). This implies

\[
\Delta V_n(t) = \frac{\Delta W_n(t + V_n(t) + O(1/n))}{1 - \hat{w}(t + v(t))} + o(\Delta V_n(t)) + O(1/n), \quad 0 \leq t \leq T. \tag{4.61}
\]

Combining (4.61) with (C.9), we deduce that \( o(\Delta V_n(t)) = o(1/n) \) since \( \Delta V_n(t) \) is of \( O(1/n) \) for all \( 0 \leq t \leq T \). Hence we have

\[
\sup_{0 \leq t \leq T} \left| \frac{\hat{V}_n(t) - \hat{W}_n(t + v(t))}{1 - \hat{w}(t + v(t))} \right| = \sqrt{n} o(1/n) = o(1/\sqrt{n}) \tag{4.62}
\]

which concludes the discussion of the relation between \( \hat{V}_n(t) \) and \( \hat{W}_n(t) \).

We next proceed to apply change of variable to the last integral term of (4.60). By using (4.62) and the fluid equations \( w(t) = v(t - w(t)) \) and \( v(t) = w(t + v(t)) \), \( 0 \leq t \leq T \), in (4.9), we
\[
\sqrt{n} \int_0^{t-w(t)} f(\theta_3, n(s)) \hat{V}_n(s) \, d\Lambda(s) \\
= \sqrt{n} \int_0^{t-w(t)} f(\theta_3, n(s)) \left( \frac{\hat{W}_n(s + v(s))}{1 - \hat{w}(s + v(s))} + o(1/\sqrt{n}) \right) \, d\Lambda(s) \\
= \sqrt{n} \int_0^t f(\theta_3', n(s)) \left( \frac{\hat{W}_n(s)}{1 - \hat{w}(s)} + o(1/\sqrt{n}) \right) (1 - \hat{w}(s)) \lambda(s - w(s)) \, ds \\
= \sqrt{n} \int_0^t f(\theta_3', n(s)) \hat{W}_n(s) \lambda(s - w(s)) \, ds + o(1) \quad 0 \leq t \leq T,
\]

for some sequence \( \{\theta_3, n(t) : n \geq 1\} \) satisfying

\[
V_n(t - w(t)) \land v(t - w(t)) \leq \theta_3', n(t) \leq V_n(t - w(t)) \lor v(t - w(t)) 
\]

for all \( 0 \leq t \leq T \) where the last equality in (4.63) holds by dominated convergence theorem since \( \|\lambda\|_t < \infty \), \( 1 - \hat{w}(t) \leq 1 \) for any \( 0 < t < \infty \), and the pdf \( f \) is bounded on bounded intervals by Assumption 4.2.2.

On the other hand, we write

\[
E_n(t) = nE(t) + \sqrt{n} \hat{E}(t) + o(\sqrt{n}), \quad 0 \leq t \leq T,
\]

by the implied convergence \( \hat{E}_n \Rightarrow \hat{E} \) in \( \mathbb{D}([0, T]; \mathbb{R}) \) as \( n \to \infty \) due to Theorem 2 of [101]. In particular, the service-completion process at each of \( n \) servers is asymptotically identical to an equilibrium renewal process since initial customers are assumed to have i.i.d. remaining service times drawn from the corresponding equilibrium distribution, and the system is operate in the ED regime (Assumption 4.2.4). In particular, there may be idle servers eventhough the system is in the ED regime, however, we know that there will be less idle servers and for shorter amount of times as the scale of the system increases (\( n \to \infty \)). In fact, all servers will be busy at all times in the limiting system with probability 1. Therefore, the service-completion process at each server converges to an equilibrium process with i.i.d. interevent times having distribution \( G \) except the first interevent time. Mathematically speaking, \( \sup_{0 \leq t \leq T} |D_n^i(t) - Z_n(t)| \to 0 \) in \( \mathbb{R} \) as \( n \to \infty \) where \( D_n^i(t) \) is the serve-completion process at the \( i \)th server, \( 1 \leq i \leq n \), and \( Z_n \) is the equilibrium renewal process. Consequently, the total service-completion process is superposition of equilibrium process in the limiting regime. Then we know from the FCLT for superposition of equilibrium processes in Theorem 2 of [101] that \( \hat{D}_n \Rightarrow \hat{D} \) in \( \mathbb{D}([0, T]; \mathbb{R}) \) as \( n \to \infty \). Together with Assumption 4.2.4, this implies that \( \sup_{0 \leq t \leq T} |\hat{E}_n(t) - \hat{D}_n(t)| \to 0 \) in \( \mathbb{R} \) as \( n \to \infty \).
for $0 \leq t \leq T$ where $q_n(t,x) \equiv F^c(\theta_{2,n}(t))\lambda(t-x)$ for $0 \leq x \leq w(t)$ for all $0 \leq t \leq T$, and

$$
\hat{G}(t) = \int_0^t F^c(w(s))d\hat{N}(s-w(s)) + \mathcal{B} \left( \int_0^t F^c(v(u))F(v(u))d\Lambda(u) \right) - \hat{E}(t).
$$

Plugging (4.63) in (4.60) and equating (4.60) with (4.65) yields

$$\hat{W}_n(t) = -\frac{1}{q_n(t,w(t))} \int_0^t f(\theta'_{\delta,n}(s))\hat{W}_n(s)\lambda(s-w(s))ds + \frac{1}{q_n(t,w(t))}\hat{G}(t) + o(1) \quad (4.66)$$

for $0 \leq t \leq T$. Then, (4.66) implies that

$$
\left| \hat{W}_n(t) \right| \leq \frac{1}{q_n(t)} \int_0^t f(\theta'_{\delta,n}(s))|\hat{W}_n(s)|\lambda(s-w(s))ds + \frac{1}{q_n(t)}|\hat{G}(t)| \quad (4.67)
$$

for $0 \leq t \leq T$.

We next introduce a version in [69] of Gronwall’s inequality that will be used to prove stochastic boundedness of $\{\hat{W}_n : n \geq 1\}$ in $\mathbb{D}([0,T];\mathbb{R})$.

**Lemma 4.7.3 (Gronwall's Inequality)** Consider measurable functions $f, h \geq 0 : [0,T] \rightarrow [0,\infty)$ and a locally-finite nonnegative measure $\mu$ on $[0,\infty)$, i.e., $\mu(K) < \infty$ on all compact sets $K$ of $[0,\infty)$, and $\mu(A) \geq 0$ for all Borel subsets of $[0,\infty)$. If

$$f(t) \leq h(t) + \int_0^t f(u)\mu(du) \quad \text{with} \quad \int_0^T h(u)\mu(du) < \infty,$$

then

$$f(t) \leq h(t) + \int_0^t h(u)\exp \left( \int_u^T \mu(r)dr \right) \mu(du).$$

**Stochastic boundedness of $\{\hat{W}_n\}$.** For fixed $n$, we apply Gronwall’s inequality to (4.67) with

$$h_n(t) = \frac{1}{q_n(t,w(t))}|\hat{G}(t)| + o(1), \quad \mu_n(u) = \frac{1}{q_n(t,w(t))}f(\theta'_{\delta,n}(u))\lambda(u-w(u)), \quad 0 \leq u \leq t \leq T,$$

which yields that

$$
\left| \hat{W}_n(t) \right| \leq h_n(t) + \int_0^t h_n(u)\exp \left( \|f\|_{[0,T]}/q_n(t,w(t)) \right) \mu_n(u) du \leq h_n(t) + \exp \left( \|f\|_{[0,T]}/q_n(t,w(t)) \right) \int_0^t h_n(u)\mu_n(u) du, \quad 0 \leq t \leq T. \quad (4.68)
$$
The exponential term in (4.68) is bounded because $\lambda$ is bounded over any closed finite interval with $\lambda(t) > 0$ for all $t \geq 0$. Furthermore, since $F^c(t) > 0$ for all $t \geq 0$, $g_n(t) > 0$ for all $t \geq 0$. This implies that the norm in (4.68) is bounded by a constant different than 0 over all bounded intervals. Hence the second term in (4.68) is bounded since the integral on the right-hand side is bounded.

The fact that the sequence $\{\hat{G}_n : n \geq 1\}$ is stochastically bounded in $D([0, T]; \mathbb{R})$, and that the second term in (4.68) is finite for all $n \geq 1$, we deduce that the sequence $\{\hat{W}_n : n \geq 1\}$ is stochastically bounded in $D([0, T]; \mathbb{R})$.

Stochastic boundedness of $\{\hat{W}_n : n \geq 1\}$ plays a key role because we want the error caused by replacing $\hat{\theta}_{2,n}(t)$ by $w(t)$, $\hat{\theta}_{3,n}(t)$ by $v(t)$, and $\hat{\theta}_{3,n}(t)$ by $w(t)$ to be asymptotically negligible. More specifically, from (4.58)-(4.59), (4.64), and $(W_n, V_n) \Rightarrow (w, v)$ in $D^2([0, T]; \mathbb{R})$, we deduce that $\hat{\theta}_{2,n}(t) = v(t - w(t)) + o(1) = w(t) + o(1)$, $\hat{\theta}_{3,n}(t) = v(t) + o(1)$, and $\hat{\theta}_{3,n}(t) = v(t - w(t)) + o(1) = w(t) + o(1)$. When $\hat{\theta}_{2,n}(t)$, $\hat{\theta}_{3,n}(t)$, and $\hat{\theta}_{3,n}(t)$ are to be replaced by respective expressions on the right-hand side of these equalities, we end up with extra terms where $o(1)$ terms are multiplied by $\hat{W}_n$. To guarantee that such terms are not too irregular, we need stochastic boundedness of $\{\hat{W}_n : n \geq 1\}$. Once stochastic boundedness is proved, such extra terms can be treated as $o(1)$.

**Finishing the proof of the FCLT for $\hat{W}_n$.** In light of the above discussion, (4.66) can be rewritten as

$$
\hat{W}_n(t) = -\frac{1}{F^c(w(t))\lambda(t - w(t))} \int_0^t f(w(s)) \hat{W}_n(s) \lambda(s - w(s)) ds \\
+ \frac{1}{F^c(w(t))\lambda(t - w(t))} \hat{G}(t) + o(1), \quad 0 \leq t \leq T.
$$

(4.69)

In order to show that $\hat{W}_n \Rightarrow \hat{W}$ where $\hat{W}$ satisfies

$$
\hat{W}(t) = -\frac{1}{F^c(w(t))\lambda(t - w(t))} \int_0^t f(w(s)) \hat{W}(s) \lambda(s - w(s)) ds \\
+ \frac{1}{F^c(w(t))\lambda(t - w(t))} \hat{G}(t),
$$

(4.70)

for $0 \leq t \leq T$, it suffices to show that $\|\hat{W}_n - \hat{W}\|_T \to 0$ for $\hat{W}_n$ and $\hat{W}$ satisfying (4.69) and (4.70). Equations (4.69) and (4.70) imply that

$$
\left| \hat{W}_n(t) - \hat{W}(t) \right| \leq \frac{1}{F^c(w(t))\lambda(t - w(t))} \int_0^t f(w(s)) |\hat{W}_n(s) - \hat{W}(s)| \lambda(s - w(s)) ds + o(1) \\
\equiv \int_0^t |\hat{W}_n(s) - \hat{W}(s)| \tilde{\mu}(s) ds + o(1), \quad 0 \leq t \leq T.
$$

116
Once again a direct application of Gronwall’s inequality similar to (4.68) yields
\[
\sup_{0 \leq t \leq T} \left| \hat{W}_n(t) - \hat{W}(t) \right| \leq \exp \left( \|\bar{\mu}\|_1 \right) \int_0^t o(1)\mu(u) \, du + o(1).
\] (4.71)
for all \( n \geq 1 \). This concludes the convergence \( \|\hat{W}_n - \hat{W}\|_T \to 0 \) since the exponential term is finite, and the integral is continuous in the sense of Theorem 11.5.1. of [102].

### 4.7.2 FCLT for the Other Processes

Thus far, we have proved that \((\hat{N}_n, \hat{D}_n, \hat{E}_n, \hat{W}_n) \Rightarrow (\hat{N}, \hat{D}, \hat{E}, \hat{W})\) in \( \mathbb{D}^4([0,T];\mathbb{R}) \) as \( n \to \infty \) where \( \hat{W} \) is as in (4.70). In particular, we start with \((\hat{N}_n, \hat{D}_n) \Rightarrow (\bar{N}, \bar{D})\) in \( \mathbb{D}^2([0,T];\mathbb{R}) \) as \( n \to \infty \) because the system is assumed to operate in the ED regime, and the system starts with the special initial condition in Assumption 4.2.3. In particular, the joint convergence holds since the arrival process and the service-completion process are asymptotically independent. Then \((\hat{N}_n, \hat{D}_n, \hat{E}_n) \Rightarrow (\bar{N}, \bar{D}, \bar{E})\) in \( \mathbb{D}^3([0,T];\mathbb{R}) \) immediately follows from convergence-together theorem (see Theorem 11.4.7. of [102]) since \( \|\hat{E}_n - \hat{D}_n\|_{[0,T]} \to 0 \) in \( \mathbb{R} \). Finally, joint convergence of \( \{W_n\} \) follows from continuous mapping theorem. We next characterize the limits for the other processes and argue their joint convergence.

By Theorem 2 of [101], the limit \( \hat{E} \) is a centered Gaussian process with the covariance function
\[
C_{E}(t,t') \equiv Cov(\hat{E}(t), \hat{E}(t')) = E[S_1(t)S_1(t')] - \mu^2 t t', \quad t, t' \geq 0
\]
where \( S_1 \) is an equilibrium renewal process with rate equal to the service rate \( \mu > 0 \).

Having established \( \hat{W}_n \Rightarrow \hat{W} \) in \( \mathbb{D}([0,T];\mathbb{R}) \) as \( n \to \infty \), joint convergence of \( \{V_n : n \geq 1\} \) with the above processes follows from (4.62) and the continuous mapping theorem. In particular, \( \hat{V}_n \Rightarrow \hat{V} \) in \( \mathbb{D}([0,T];\mathbb{R}) \) as \( n \to \infty \) with
\[
\hat{V}(t) = \frac{\hat{W}(t + v(t))}{1 - \hat{w}(t + v(t))}, \quad 0 \leq t \leq T.
\]

We next prove the FCLT for the queue-length process using the FCLT for \( \{\hat{W}_n\} \) and continuous mapping theorem. Since waiting line is initially empty in each system, i.e., \( Q_n(0) = 0 \) for each \( n \), it suffices to establish convergence for the scaled versions of (5.25)-(5.27). Recall from (C.10) the FWLLN limits
\[
Q_1(t) = Q_2(t) = (0e)(t), \quad Q_3(t) = \int_0^{t-w(t)} F^c(t-s)\lambda(s) \, ds, \quad 0 \leq t \leq T.
\]
Therefore, we have

\[ \hat{Q}_{n,1}(t) = \frac{1}{\sqrt{n}} Q_{n,1}(t) \Rightarrow \hat{Q}_1(t) \equiv \int_{t-w(t)}^{t} F^c(t-s) \, d\hat{N}(s), \quad 0 \leq t \leq T, \]

\[ \hat{Q}_{n,2}(t) = \frac{1}{\sqrt{n}} Q_{n,2}(t) \Rightarrow \hat{Q}_2(t) \equiv \int_{t-w(t)}^{t} \int_0^1 1(x > F(t-s)) \, d\hat{U}(\Lambda(s), x), \quad 0 \leq t \leq T, \]

\[ \hat{Q}_{n,3}(t) = \frac{1}{\sqrt{n}} (Q_{n,3}(t) - n Q_3(t)) = \sqrt{n} \int_{t-W_n(t)}^{t-w(t)} F^c(t-s) \lambda(s) \, ds = \hat{W}_n(t) F^c(w(t)) \lambda(t-w(t)) + o(\sqrt{n}), \quad 0 \leq t \leq T, \]

in \( D([0,T]; \mathbb{R}) \) as \( n \to \infty \). This implies

\[ \hat{Q}_{n,3}(t) \Rightarrow \hat{Q}_3(t) \equiv \hat{W}(t) F^c(w(t)) \lambda(t-w(t)), \quad 0 \leq t \leq T, \]

in \( D([0,T]; \mathbb{R}) \) as \( n \to \infty \). Note that \( \hat{Q}_3 \) involves \( \hat{W} \) which, as seen in (5.49), involves stochastic integrals with respect to \( \hat{N} \) and \( \hat{E} \), and involves an Itô integral that can be represented as a time-changed Brownian motion

\[ \mathcal{B} \left( \int_0^{\Gamma(t)} F^c(w(u)) F(w(u)) \lambda(u-w(u))(1-\dot{w}(u)) \, du \right), \quad 0 \leq t \leq T, \quad (4.72) \]

where \( \Gamma(t) \) is the inverse of the function \( u \mapsto u-w(u) \) evaluated at \( u = t \). The limits \( \hat{Q}_3 \) and \( \hat{Q}_2 \) are independent because Kiefer process has independent time increments, i.e., independent increments with respect to the first component. In particular, from the inequality \( \Gamma(t) \leq t \), we deduce that the interval of (4.72) and the limit \( \hat{Q}_2 \) do not overlap, and hence independent. Similarly, one can apply change of variable to \( \hat{Q}_1 \) to obtain an alternative representation

\[ \int_t^{\Gamma(t)} F^c(w(u)) \, d\hat{N}(u-w(u)), \quad 0 \leq t \leq T, \]

which is independent of the first integral in (5.49) because \( \hat{N} \) has independent increments. Consequently, we establish

\[ (\hat{N}_n, \hat{D}_n, \hat{E}_n, \hat{W}_n, \hat{V}_n, \hat{Q}_n) \Rightarrow (\hat{N}, \hat{E}, \hat{E}, \hat{W}, \hat{V}, \hat{Q}) \text{ in } D^6([0,T]; \mathbb{R}) \text{ as } n \to \infty \]

by applying continuous mapping theorem with addition. Finally, given the above convergence, we immediately obtain, by continuous mapping theorem, the limit \( \hat{A} = \hat{N} - \hat{Q} - \hat{E} \) and \( \hat{X} = \hat{Q} + \hat{B} \) where \( \hat{B} \) is trivially zero. The joint convergence in Theorem 4.5.2 then follows.
4.8 Numerical Examples

In this section, we provide numerical examples to demonstrate the effectiveness of the engineering formulas based on the FCLT limits in §4.5. We provide the results of a simulation experiment for the steady-state performance of the $G/GI/n+GI$ queueing system. We consider two different queueing systems. For the first system, we let interarrival times be i.i.d. having an hyperexponential distribution, and patience times be i.i.d. again having an hyperexponential distribution but, with different parameters; for the second system, we let interarrival times be i.i.d. having an hyperexponential distribution, and patience times be i.i.d. having an Erlang distribution. We compare the performance of the two systems under three different service-time distributions, namely, the Erlang, the exponential, and the hyperexponential distribution. In particular, we compare the simulation results with the numerical results obtained by numerically computing the mean and variance formula for the limiting head-of-line process $\hat{W}$ and queue-length process $\hat{Q}$. The difference between the simulation and numerical computation is indicated in Table 4.1 and Table 4.2 by the term “rel. err.” which is the half length of the associated 95% confidence interval.

We now give the parameters used in our numerical experiment. The interarrival times for both systems have the same distribution. We assume that they are drawn from an $H_2$ service distribution with balanced means, i.e., a mixture of two exponential random variables with rates, say, $\lambda_1$ and $\lambda_2$, and with probabilities $p$ and $1 - p$, respectively. More specifically, an interarrival time can be denoted as $X = pX_1 + (1 - p)X_2$ where $X_1$ and $X_2$ are exponential random variables with rates $\lambda_1$ and $\lambda_2$, respectively. Then we set $\lambda_1 = 2p\lambda$, $\lambda_2 = 2(1 - p)\lambda$, $\lambda = 120$ and $p = 0.5(1 - \sqrt{0.6})$ so that the mean $1/\lambda = 1/120$ and the squared coefficient of variation is $c_s^2 = 4$. All three of the service distributions considered in both systems have same parameters. We assume that the service rate $\mu = 1$ is same for all three distributions. The first service distribution is Erlang-2 with rate $\mu$, that is, service times are equal in distribution to the sum of two i.i.d. exponential variables with rate $\mu$. When $\mu = 1$, the squared coefficient of variation for the Erlang-2 distribution is 0.5, i.e., $c_s^2 = 0.5$. The second service distribution is the exponential distribution with rate $\mu = 1$. The squared coefficient of variation for the exponential distribution is 1, i.e., $c_s^2 = 1$. The third service distribution is the $H_2$ distribution with rates $\mu_1$, $\mu_2$, and probabilities $p$ and $1 - p$, where $\mu_1 = 2p\mu$, $\mu_2 = 2(1 - p)\mu$, $\mu = 1$, and $p = 0.5(1 - \sqrt{0.6})$. Consequently, the mean $1/\mu = 1$, and the squared coefficient of variation $c_s^2 = 4$. The patience distribution in the first system is assumed to be $H_2$ with parameters $\theta_1, \theta_2$, and probabilities $p$ and $(1 - p)$ such that $\theta_1 = 2p\theta$, $\theta_2 = 2(1 - p)\theta$, $\theta = 0.5$ and $p = 0.5(1 - \sqrt{0.6})$ so that the mean $1/\theta = 2$ and $c_a^2 = 4$. The patience distribution in the second system, on the other hand, is Erlang-2 ($E_2$) with rate 0.5. Finally, we set the number of servers $n = 100$ so that the the traffic intensity $\rho = \lambda/(n\mu) = 1.2$ in both systems.
We observe that the steady-state expected waiting time and queue length, i.e., $E[W]$ and $E[Q]$, are insensitive to the service-time distribution provided their means are the same. The results in Table 4.1 and Table 4.2 are in line with the findings of Whitt [104], that is, the head-of-line waiting time and the queue length in fluid queueing systems depend upon the service-time distribution only through its mean. However, the variance of the two steady-state performance measures is indeed sensitive to the service-time distribution. This is best observed in the last row of Table 4.2. The illustrated impact of service-time distributions on the performance of the system emphasizes that stochastic refinements to fluid approximations can provide valuable additional insight into the dynamics of the system.

Table 4.1: $H_2(\lambda^{-1}, c_s^2)/GI/100 + H_2(\theta^{-1}, c_s^2)$ with $(\lambda, \rho, \theta, c_\lambda^2, c_s^2) = (120, 1.2, 0.5, 4, 4)$

<table>
<thead>
<tr>
<th>Perf.</th>
<th>$E_2$ service ($c_s^2 = 0.5$)</th>
<th>$M$ service ($c_s^2 = 1$)</th>
<th>$H_2$ service ($c_s^2 = 4$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sim</td>
<td>Num</td>
<td>Sim</td>
</tr>
<tr>
<td>$E[W]$</td>
<td>0.237</td>
<td>0.240</td>
<td>0.239</td>
</tr>
<tr>
<td>rel. err.</td>
<td>±1.4E-3</td>
<td>1.4%</td>
<td>±1.5E-3</td>
</tr>
<tr>
<td>Var($W$)</td>
<td>0.022</td>
<td>0.022</td>
<td>0.024</td>
</tr>
<tr>
<td>rel. err.</td>
<td>±7.5E-4</td>
<td>0.9%</td>
<td>±8.3E-4</td>
</tr>
<tr>
<td>rel. err.</td>
<td>±0.168</td>
<td>1.4%</td>
<td>±0.177</td>
</tr>
<tr>
<td>Var($Q$)</td>
<td>302.78</td>
<td>316.83</td>
<td>320.97</td>
</tr>
<tr>
<td>rel. err.</td>
<td>±10.31</td>
<td>4.6%</td>
<td>±12.12</td>
</tr>
</tbody>
</table>

4.9 Refined Staffing Levels

In this section, we consider a refined staffing function given by

$$s_n \equiv \lceil ns_1 + \sqrt{ns_2} \rceil \quad \text{where} \quad s_1, s_2 > 0. \quad (4.73)$$

The refined staffing function in (4.73) is closely related to the staffing functions introduced in [73]. The general form of (4.73) enables us to recover two of the staffing functions considered in [73] that respectively lead to the ED and ED+QED operating regimes. More specifically, the
Table 4.2: \( H_2(\lambda^{-1}, c_2^3) / GI / 100 + E_2(\theta^{-1}) \) with \((\lambda, \rho, \theta, c_2^3, c_a^2) = (120, 1.2, 0.5, 4, 0.5)\)

<table>
<thead>
<tr>
<th>Perf.</th>
<th>( E_2 ) service ((c_s^2 = 0.5))</th>
<th>( M ) service ((c_s^2 = 1))</th>
<th>( H_2 ) service ((c_s^2 = 4))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sim</td>
<td>Num</td>
<td>Sim</td>
</tr>
<tr>
<td>( E[W] )</td>
<td>0.705</td>
<td>0.732</td>
<td>0.7046</td>
</tr>
<tr>
<td>rel. err.</td>
<td>±3.4E-3</td>
<td>3.7%</td>
<td>±3.7E-3</td>
</tr>
<tr>
<td>( \text{Var}(W) )</td>
<td>0.051</td>
<td>0.047</td>
<td>0.0576</td>
</tr>
<tr>
<td>rel. err.</td>
<td>±4.5E-3</td>
<td>7.8%</td>
<td>±4.9E-3</td>
</tr>
<tr>
<td>( E[Q] )</td>
<td>79.84</td>
<td>82.32</td>
<td>79.65</td>
</tr>
<tr>
<td>rel. err.</td>
<td>±0.427</td>
<td>3.0%</td>
<td>±0.454</td>
</tr>
<tr>
<td>( \text{Var}(Q) )</td>
<td>818.86</td>
<td>785.86</td>
<td>883.42</td>
</tr>
<tr>
<td>rel. err.</td>
<td>±67.60</td>
<td>4.0%</td>
<td>±71.62</td>
</tr>
</tbody>
</table>

Two staffing functions in [73] are given by

\[
\begin{align*}
  n_{\text{ED}} &= \lceil (1 - \gamma) R_n \rceil, \\
  n_{\text{ED+QED}} &= \lceil (1 - \gamma) R_n + \delta \sqrt{R_n} \rceil
\end{align*}
\]

where \( 0 < \gamma < 1 \) and \( R_n \) is the offered load defined as \( R_n = n\lambda / \mu \). Letting \( s_1 = (1 - \gamma)\lambda / \mu \) and \( s_2 = 0 \) yields (4.74) whereas letting \( s_1 = (1 - \gamma)\lambda / \mu \) and \( s_2 = \delta \sqrt{\lambda / \mu} \) yields (4.75).

We next briefly discuss the changes resulting from considering the staffing function \( s_n \) instead of \( n \). In the previous sections, the staffing function happens to coincide with our scaling factor \( n \), i.e., \( s_n = n \). In this section, we let \( n \) and \( \sqrt{n} \) be the scaling factors for FWLLN and FCLT, respectively, and let the staffing function have a more general form \( s_n = \lfloor ns_1 + \sqrt{n} s_2 \rfloor \) where \( s_1, s_2 > 0 \). To indicate the processes associated with the new staffing function, we use superscript \( r \), whereas, to indicate the processes associated the case where \( s_n = n \), we do not use additional notation. Because the arrival process is independent of the staffing level, it holds that \( \hat{N}_n(t) = \hat{N}_n(t) \) for all \( n \geq 1 \) and \( t \geq 0 \), and hence, \( \hat{N}^r(t) = \hat{N}(t) \), for \( t \geq 0 \). The FWLLN limit and the FCLT limit for the service-completion process, on the other hand, becomes \( D^r(t) = s_1 D(t), t \geq 0 \), and \( \hat{D}^r(t) = \sqrt{s_1 \hat{D}(t) + s_2 D(t)} \), respectively, where \( \hat{D}(t) \) is a centered Gaussian process with covariance function \( C_E \) in Theorem 4.5.2, and \( D(t) = \mu t \) as in (4.26). Hence \( \hat{D}^r(t) \) is a Gaussian process with covariance function \( C^r(\cdot, \cdot) = s_1 C_E(\cdot, \cdot) \) and mean \( s_2 \mu t, t \geq 0 \). Consequently, by Assumption 4.2.4, the enter-service process satisfies \( \hat{E}^r(t) = \sqrt{s_1 \hat{D}(t) + s_2 \mu t}, t \geq 0 \).

The following theorem is an analogue of Theorem 4.5.1 and Theorem 4.5.2 given the refined staffing function \( s_n \). A proof is provided in Appendix C.5.

**Theorem 4.9.1 (FWLLN and FCLT with refined staffing)** Suppose that the staffing level
at any time is given by \( s_n \equiv ns_1 + \sqrt{ns_2} \) for constants \( s_1, s_2 > 0 \). Then

(i) Under the conditions of Theorem 4.5.1, an analogue of joint convergence in (4.24) holds as \( n \to \infty \) where \( \Lambda^r(t) = \lambda t \),

\[
D^r(t) = E^r(t) = s_1 \mu t, \quad w^r(t) = \int_0^t \left( 1 - \frac{s_1 \mu}{\lambda F^c(w^r(u))} \right) du, \quad v^r(t) = w^r(t + v^r(t)), \quad t \geq 0.
\]

(4.76)
The limits \( Q^r(t), X^r(t) \) and \( A^r(t) \) have the same mathematical form as their counterparts in Theorem 4.5.1 with modified components.

(ii) Under the conditions of Theorem 4.5.1, an analogue of joint convergence in (4.27) holds as \( n \to \infty \) where

\[
\hat{W}^r(t) = \int_0^t \frac{F^c(w^r(u))H^r(t, u)}{q(t, w^r(t))} d\hat{N}(u - w^r(u)) + \int_0^t \frac{\lambda F^c(v^r(u))F(v^r(u))H^r(t, u)}{q(t, w^r(t))} d\mathcal{B}(u) - \sqrt{s_1} \int_0^t H^r(t, u) q(t, w^r(t)) \mathcal{E}(u) - s_2 \mu \int_0^t H^r(t, u) q(t, w^r(t)) du, \quad t \geq 0,
\]

(4.77)

\( w^r(t) \) and \( v^r(t) \) are as in (4.76), \( H^r(\cdot, \cdot) \) and \( q(\cdot, w^r(\cdot)) \) as in (5.50) with \( w(t) \) replaced by \( w^r(t) \). The virtual waiting time \( \hat{V}^r(t) \) and the queue-length process \( \hat{Q}^r(t) \) have the same mathematical forms as in (5.39) and (4.32), respectively, with \( w(t), v(t) \) and \( \hat{W}(t) \) replaced by their counterparts \( w^r(t), v^r(t) \) and \( \hat{W}^r(t) \). Finally, the abandonment process is given by \( A^r(t) = \hat{N}(t) - Q^r(t) - E^r(t), t \geq 0 \).

Note that (4.77) is different than (5.49) in that the third term on the right-hand side is scaled by \( \sqrt{s_1} \) and that there is an additional deterministic term. This implies that both the variance and mean of HOL waiting times changes and so do those of the virtual waiting time and queue length. The corresponding steady-state formulas are given in the following corollary.

**Corollary 4.9.1 (Steady state of limits with refined staffing)** Under the assumptions of Theorem 4.6.3 the steady-state random variables \( \hat{W}^r(\infty) \), \( \hat{V}^r(\infty) \) and \( \hat{Q}^r(\infty) \) have Gaussian distributions with means

\[
\mu_{W^r} \equiv E[\hat{W}^r(\infty)] = E[\hat{V}^r(\infty)] = -\frac{s_2 \mu}{\lambda F^c(w^r)}, \quad E[\hat{Q}^r(\infty)] = \lambda F^c(w^r)E[\hat{W}^r(\infty)] = -\frac{s_2 \mu}{h_F(w^r)},
\]

and variances

\[
\text{Var}(\hat{W}^r(\infty)) = \text{Var}(\hat{V}^r(\infty)) \equiv \sigma_{W^r}^2 \equiv \frac{c_\lambda^2}{2h_F(w^r)^2} + \frac{F(w^r)}{2\lambda F(w^r)} + s_1 \sigma_{W^r}^2,
\]

122
where

\[ \sigma_{W^r}^2 \equiv 2 \frac{h_F(w^r)^2}{\lambda^2 F'(w^r)^2} \int_0^\infty \int_0^x e^{-h_F(w^r)(x+y)} \tilde{C}(-x,-y) \, dy \, dx, \]

the covariance function \( \tilde{C}(-x,-y) \) is as in (4.46), \( w^r = F^{-1}(1 - 1/\rho^r) \) with \( \rho^r = \lambda/(s_1 \mu) \) and \( h_F \) is the hazard-rate function associated with \( F \). The variance of the steady-state queue length \( \hat{Q}(\infty) \) has the same mathematical form with \( w \) and \( \hat{W} \) replaced by \( w^r \) and \( \hat{W}^r \).

It has been showed in [73] that refined staffing functions can be useful in the context of constraint satisfaction. In particular, optimization problems of the form (4.78) have been proposed, and asymptotically optimal staffing levels have been characterized:

\[ n^* \equiv \min n \]

subject to \( U(n, \lambda) \leq M \)

(4.78)

where \( M \) is a predetermined constant, \( U(n, \lambda) \) is the cost function whose form depends on which staffing function in (4.74)-(4.75). If (4.74) is to be minimized, then \( U(n, \lambda) \) can be chosen as a function of mean waiting time and probability to abandon with corresponding penalties. On the other hand, if (4.75) is chosen to be minimized, then \( U(n, \lambda) \) can be chosen as penalty associated with tail probability of wait. The “right” values of the parameters \( \gamma \) and \( \delta \) depend on the staffing function and the constraint function \( U(n, \lambda) \). Given the steady-state limits in Theorem 4.6.3 and the refined staffing function \( s_n \), a natural direction of research is finding optimal staffing levels by solving the optimization problem (4.78) extending the results given in [73] for the \( M/M/n + G \) queue. We will not consider optimal staffing issues in this paper. We leave it for future research.
Chapter 5

Future Work

5.1 Introduction

We consider the $G_t/GI/s_t + GI$ queueing system having a general arrival process with a time-varying arrival rate (the $G_t$), independent and identically distributed (i.i.d.) nonexponential service times with cdf $G$ (the GI), customer abandonment with i.i.d. nonexponential patience times with cdf $F$ (the $+GI$), and time-varying number of servers $s_t$. Customers are served on first-come first-served (FCFS) basis, and are not allowed to abandon service. The system is assumed to alternate between underloaded (UL) and overloaded (OL) regimes where the system’s loading is determined by the associated fluid queue. More specifically, the system is said be UL (resp. OL) if the associated fluid queue is UL (resp. OL). The fluid queue switches regimes at isolated switching points. We further assume that the fluid queue is critically loaded only at these isolated switching points. These assumptions are needed for the functional central limit theorem (FCLT) established in this chapter.

We believe that this model is relevant from engineering viewpoint because many systems may be incapable of dynamically adjust the staffing level sensitively and timely enough to stabilize performance when arrivals are time-varying. Therefore, systems may experience periods of underloading and overloading. Even if the arrival rate slightly deviates from a constant level, and the fluid limit is close to the staffing level, systems may still experience periods of underloading and overloading in very short consecutive intervals. In particular, we refer to the circumstances where the fluid limit stays close the staffing level but crosses it upwards (switches from UL to OL) and downwards (switches from OL to UL) at countable number of time points with measure zero (called nearly critically loaded in [53]). Then our framework is still relevant, and can provide accurate approximations. If, on the other hand, the fluid limit is level with the staffing level for a time window of positive measure, then the approximations developed in this chapter do not perform as well. From theoretical point of view, the assumption of critical
loading only on sets of measure zero may be restrictive. However, this assumption is reasonable from practical point of view because the staffing level is usually piecewise constant, and the fluid limit is continuous as a function of time. Therefore, it is unlikely for the fluid limit to stay level with the staffing level for a time window of positive measure. Our assumption of isolated switching points serves that purpose, and is in line with [70, 53].

One stream of research exploits strong approximation results, and Gaussian-based approximation methods to estimate time-dependent mean and covariance. Ko and Gautam [53] develop adjusted fluid and diffusion limits using the uniform acceleration method along with strong approximation results in [55, 69], and propose a Gaussian-based approximation for the mean and covariance that performs better than that in [69]. Massey and Pender [75] improve the results in [53] by incorporating skewness of the queue-length process. Pender [82] further extends the results of [75] by incorporating kurtosis of the queue length via the Gram-Charlier expansion. In another work by Pender [84] is used the Laguerre expansions to estimate not only the mean and variance, but the first four moments the queue length process. Also see [83] for another type of expansion, and [76] for approximations for Jackson networks. In a similar vein, Pender and Ko [85] develop a method for approximating the number in system for the $G_t/G_t/s_1 + G_1$ queue (no abandonment) with time-varying service times. These methods have been shown to provide accurate estimates for mean and covariance even for small-scaled queueing systems, however, these methods do not provide any information about sample-path properties and others that heavy-traffic analysis provides. Our work extends these papers in two directions: (i) we consider a fully non-Markovian queueing system with abandonment; and (ii) we establish heavy-traffic limits that are Gaussian processes with either closed-form expressions or known covariance functions.

We provide a Gaussian approximation by establishing many-server heavy-traffic limits. In particular, we prove an FCLT for the key performance processes such as the number in system, head-of-line waiting time, virtual waiting processes for UL and OL intervals with critical loading allowed only at the left endpoints (see §5.5). The FCLTs established in the previous chapters play a key role in establishing the FCLT for UL and OL intervals in this chapter. The FCLTs in §5.5 are valid over closed intervals with critical loading only at the left endpoints. Consequently, in the framework of the alternating $G_t/G_t/s_1 + G_1$ queue, the FCLTs are valid only on the intervals that are closed at the left endpoint and open at the right endpoint because the right endpoint of every interval is a switching point, i.e., critically loaded. In other words, we establish FCLTs on the intervals $[t_i, t_{i+1})$ for all switching points $\{t_i : i \geq 0\}$. A key assumption is that the elapsed service times of customers in service satisfy an FCLT at the left switching point of every interval. This assumption holds due to the lack of memory property if the service distribution is the exponential distribution. However, it is harder to justify this assumption when the service distribution is not the exponential distribution. We conjecture that a joint
FCLT for elapsed service times, and the scaled arrival process holds at every switching point given that elapsed service times satisfy an initial FCLT at time 0. In particular, if
\[(\hat{N}_n, \hat{B}_n(0, \cdot)) \Rightarrow (\hat{N}, \hat{B}(0, \cdot)) \quad \text{in} \quad \mathcal{D}([0, T]; \mathbb{R}) \times \mathcal{D}([0, \infty); \mathbb{R}) \quad \text{as} \quad n \to \infty,\]
then
\[(\hat{N}_n, \hat{B}_n(t_i, \cdot)) \Rightarrow (\hat{N}, \hat{B}(t_i, \cdot)) \quad \text{in} \quad \mathcal{D}([0, T]; \mathbb{R}) \times \mathcal{D}([0, \infty); \mathbb{R}) \quad \text{as} \quad n \to \infty,\]
where \(\{t_i : i \geq 0\}\) are switching points, \(\hat{N}_n\) is the CLT-scaled arrival process, and \(\hat{B}_n(t, y)\) is the CLT-scaled two-parameter process tracking the number of customers in service at time \(t\) that has been so at most \(y\) units of time. Simulation experiments suggest that the age-in-service process \(B_n(\cdot, \cdot)\) is well-behaved around switching points. We leave the treatment of switching points, and hence, a proof of an FCLT over closed intervals that are also critically loaded on the right endpoint as future work. The FCLTs for intervals that are critically loaded at the left endpoints in this chapter extend the existing results in [65].

This chapter is closely related with the paper by Liu and Whitt [65] in which the authors establish an FCLT for the \(G_t/M/st + GI\) queueing system alternating between underloading and overloading. This chapter is also related with [61] in which the authors provide a fluid model for the \(G_t/GI/st + GI\) queueing system, and with [62] in which the authors prove a FWLLN to provide mathematical justification for the fluid model in [61]. The FCLT limits here provide a refined Gaussian approximation to the fluid model developed in [61], and they are the generalizations of the FCLTs in [65]. Just as in [3], the main difficulty is that the service-completion process is no longer a nonhomogeneous Poisson process, hence, does not converge to a Brownian limit after scaling. Therefore, we have to work with Gaussian integrals, and stochastic differential equations (SDEs) driven by Gaussian processes. Moreover, since the regime switches between overloading and underloading, characterizing the enter-service process is more complicated especially for OL intervals.

During OL intervals, which start with a critically loaded switching point, the enter-service process has an implicit form, i.e., it is represented as a function of itself. Also, initial customers in service have non-i.i.d. elapsed service times unlike [3, 65]. The key step is to characterize the limit for the initial number in system, and for the enter-service process. The enter-service process is proven to be a centered (zero-mean) Gaussian process that almost-surely satisfies a Volterra equation under an assumption on the initial state of the system. Then, drawing on [3], we obtain an SDE for the scaled head-of-line waiting time given the initial state. We later prove an FCLT for the number in system under a more general assumption on the initial state, and conjecture an FCLT for the other processes given the initial state. The conjectured FCLT
is proven to hold on the open intervals excluding critically-loaded interval endpoints. This is because the other processes have discontinuous sample paths around switching points unlike the number in system.

**Organization of the chapter.** In §5.2, we describe a sequence of $G_t/GI/s_t + GI$ model and specify the model assumptions. In §5.3, we review the associated fluid functions for the $G_t/GI/s_t + GI$ queue. In §5.4, we define the key processes used to establish an FCLT for UL and OL intervals. In §5.5, we present our main results under certain initial conditions. In §5.6, we provide further characterization of the FCLT limits. We give detailed analytic formulas for their distributions. In §5.7–§5.8 we prove our main results. In §5.9, we discuss how FCLTs in §5.5 can be applied to the $G_t/GI/s_t + GI$ alternating queue. Additional material including proofs are provided in the appendix.

### 5.2 A sequence of $G_t/GI/s_t + GI$ queues

We consider the $G_t/GI/s_t + GI$ queueing system having a general arrival process with a time-varying arrival rate (the $G_t$), *independent and identically distributed* (i.i.d.) nonexponential service times with cdf $G$ (the GI), and customer abandonment with i.i.d. nonexponential patience times with cdf $F$ (the +GI). Upon arrival, customers start service immediately if there are any idle servers. Otherwise, if all servers are busy, customers join a waiting line has an infinite capacity. Customers waiting in line may abandon the system without entering service if they wait longer than their patience time. Customers do not abandon the system once they begin service. Service is assumed to be on first-come first-served (FCFS) basis. We obtain a sequence of $G_t/GI/s_t + GI$ queues indexed by $n$ by letting the time-varying arrival rate to the $n$th queue be $\lambda_n(t) \equiv n\lambda(t) + o(1)$ and the staffing level (i.e., number of servers) at time $t$ be $s_n(t) \equiv \lceil ns(t) \rceil$ where $s(t)$ is continuous and deterministic.

For the $n$th model we define the following processes. Let $N_n(t)$ be the number of customers who arrived in the interval $[0, t]$, $E_n(t)$ be the total number of customers who have started service in $[0, t]$, $A_n(t)$ be the number of abandoned customers in $[0, t]$. Let the two-parameter process $B_n(t, y)$ (resp. $Q_n(t, y)$) denote the number of customers in service (resp. in the waiting line) at time $t$ for at most $y$ units of time. Also, define $B_n(t) \equiv B_n(t, \infty)$ (resp. $Q_n(t) \equiv Q_n(t, \infty)$) as the total number of customers in service (resp. in the waiting line) at time $t$. We let $D_n(t)$ be the *total* number of service completions in $[0, t]$. Then, the total number in system at time $t$ is represented as $X_n(t) = B_n(t) + Q_n(t)$. The process $W_n(t)$ represents head-of-line (HOL) waiting time, i.e., the elapsed waiting time of the customer at the head of line at time $t$, which is the waiting time of the customer who has been waiting the longest. Finally, we let $V_n(t)$ be the virtual waiting time at time $t$, i.e., the waiting time of a potential arriving customer assuming
the customer has infinite patience. By definitions of $W_n(t)$ and $V_n(t)$, we have

$$V_n(t - W_n(t)) = W_n(t) + O(1/n)$$

$$V_n(t) = W_n(t + V_n(t) + O(1/n)) + O(1/n), \quad t \geq 0,$$

where (5.1) suggests that the virtual waiting time at the arrival time of the head-of-line customer at time $t$ is the head-of-line customer's waiting time in line at time $t$ plus the additional time until a server becomes idle after time $t$. The equality in (5.2) is obtained by applying change of variable.

The queue is assumed to alternate between underloaded (UL) and overloaded (OL) regimes with critical loading only at isolated switching points. The queue is said to operate in the UL (resp. OL) regime in the interval $[t_1, t_2]$ if the associated fluid queue is UL (resp. OL). Critical loading at isolated switching points is characterized as a time instant at which the associated fluid queue has no more service capacity, and has no content in queue. The associated fluid queue, and conditions (in terms of fluid functions) that characterize the UL and OL intervals are reviewed in §5.3.

Just as their counterparts the QD and ED regimes in the context of stationary queues, the UL and OL regimes are characterized by limiting conditions. As will be proven in this chapter, the fluid functions are the limit of the LLN-scaled processes. When we say the system is underloaded or overloaded, we refer to certain conditions that hold asymptotically. These conditions are characterized through the limiting fluid functions discussed in §5.3. Therefore, it is possible that the $n$th system may have idle servers during OL intervals if $n$ is not large enough, or the $n$th system may have customers waiting in line during UL intervals if $n$ is not large enough. However, for large $n$, these circumstances are less likely to happen. In particular, the (stochastic) system will not have idle servers (resp. waiting customers) during OL (resp. UL) intervals with probability 1 as $n \to \infty$.

**New customers.** For the $i$th customer who arrives at the $n$th system after time 0, we let $\tau^n_i$ be the arrival time, $\nu^n_i$ be the service time, $\gamma^n_i$ be the patience time, and $w^n_i$ be the offered waiting time, i.e., time $i$th customer would have to wait until starting service if the customer were infinitely patient. We also let $e^n_i$ be time $i$th customer starts service given that the customer does not abandon, i.e., $e^n_i \equiv \tau^n_i + w^n_i$. Otherwise, $e^n_i \equiv \infty$ by convention. We assume that $\{\nu^n_i : i \geq 1\}$ are i.i.d. with cdf $G$, and $\{\gamma^n_i : i \geq 1\}$ are i.i.d. with cdf $F$ for each $n \geq 1$. Also, it holds that $w^n_i = V_n(\tau^n_i - )$ for $i \geq 1$ (see, e.g., [16]). If some servers are idle at the time $\tau^n_i$ of an $i$th arrival, then it necessarily holds that $w^n_i = 0$ and $e^n_i = \tau^n_i$.

**Initial customers.** We refer to initial customers as those present in system at the beginning of a UL or an OL interval, i.e., a left-end switching point. As mentioned earlier, switching points are critically loaded, and are determined by the associated fluid model. Of course, the $n$th
queueing system is random, and therefore, it is not necessarily critically loaded at a (left-end or right-end) switching point \( t^* \), i.e., \( X_n(t^*) \neq s_n(t^*) \), even if the associated fluid queue at time \( t^* \) is. For instance, if \( X_n(t^*) > s_n(t^*) \) with \( t^* \) being a left-end switching point, then all servers are busy and the waiting line is nonempty. Those customers in the waiting line (resp. in service) can be considered as initial customers in the waiting line (resp. in service) for that interval.

Consider an interval \([t_1, t_2]\) limited by switching points, and suppose there are customers waiting in line at time \( t_1 \).

(i) Initial customers in the waiting line. For the \( i \)th customer initially in the waiting line, \( i < 0 \), we let \( \gamma^n_i \) be the remaining patience time, \( \eta^n_i \) be the elapsed waiting time (i.e., age in line), and \( \nu^n_i \) be the remaining offered waiting time at time 0. In this setting, -1th customer index refers to the customer with the smallest elapsed time, -2th customer index refers to the customer with the second smallest elapsed time, and so on. In this case, the customer with customer index \(-Q_n(t_1)\) is waiting at the head of line, and the customer with customer index \(-1\) is waiting at the end of line so that there are \( Q_n(t_1) \) customers in the waiting line at time 0 in the \( n \)th system. For given elapsed times in line \( \{\eta_1^n, \eta_2^n, \ldots\} \), the remaining patience times \( \{\gamma^n_i : i < 0\} \) follow the distribution

\[
\tilde{F}_x(s) \equiv \mathbb{P}(\gamma^n_i > s \mid \eta^n_i = x) = \frac{F^c(s + x)}{F^c(x)} \quad \text{for } s, x \geq 0.
\]

where \( F^c \equiv 1 - F \). We remark that conditional on the elapsed times \( \{\eta^n_i\} \), the remaining patience times \( \{\gamma^n_i\} \) are independent.

(ii) Initial customers in service. Similarly, we index initial customers in service on negative integers \( j < 0 \). Conditioned on elapsed service times \( \{\xi_1^n, \xi_2^n, \ldots\} \), remaining service times has the distribution

\[
\tilde{G}_x(s) \equiv \mathbb{P}(\nu^n_j > s \mid \xi^n_j = x) = \frac{G^c(s + x)}{G^c(x)} \quad \text{for } s, x \geq 0.
\]

The remaining service times \( \{\nu^n_i\} \) are independent conditional on the elapsed service times \( \{\xi^n_i\} \).

We further assume that the arrival process, service times, and patience times are mutually independent. We next impose conditions on service- and patience-time distributions.

**Assumption 5.2.1 (Regularity conditions on service and patience distributions)** (i) The patience-time cdf \( F \) is differentiable with pdf \( f \), and satisfies \( F^c(x) > 0 \) for all \( x \geq 0 \); (ii) The service-time cdf \( G \) is differentiable with pdf \( g \) satisfying \( g(x) > 0 \) for all \( x \geq 0 \).

**Scaling.** We now introduce two different scalings applicable to both one- and two-parameter processes. Let \( K_n \) be either a one- or two-parameter process. Then LLN-scaled and CLT-scaled
versions of $K_n$ are
\[
\bar{K}_n \equiv n^{-1}K_n \quad \text{and} \quad \hat{K}_n \equiv n^{-1/2}(K_n - nK) \tag{5.3}
\]
where $K$ is the limit for the sequence $\{\bar{K}_n : n \geq 1\}$. We define the LLN- and CLT- scaled versions for the processes $E_n, D_n, A_n, B_n, Q_n$ and $X_n$ as in (5.3). For virtual waiting time $V_n$ and HOL waiting time $W_n$, on the other hand, we do not consider LLN-scaling, and only define the CLT scaling as
\[
\hat{W}_n \equiv \sqrt{n}(W_n - w) \quad \text{and} \quad \hat{V}_n \equiv \sqrt{n}(V_n - v)
\]
where $w$ and $v$ are the limits for the sequences $\{W_n : n \geq 1\}$ and $\{V_n : n \geq 1\}$, respectively.

We have the following assumption on the external arrival process, and initial number in service.

**Assumption 5.2.2 (Joint FCLT for the arrival process and initial elapsed times) The CLT-scaled initial number in service and external arrival process jointly satisfy the FCLT**

\[
(\hat{N}_n, \hat{B}_n(0, \cdot), \hat{X}_n(0)) \Rightarrow (\check{N}, \check{B}(0, \cdot), \check{X}(0)) \quad \text{in} \quad \mathcal{D}([0, T]; \mathbb{R}) \times \mathcal{D}([0, \infty); \mathbb{R}) \times \mathbb{R}^2,
\]
as $n \to \infty$, where $X(0)$ is deterministic, the limit $(\hat{N}(t) : t \geq 0)$ has continuous sample paths, and independent increments, and is independent of $\hat{B}(0, \cdot)$ which has continuous sample paths.

For the first UL interval $X(0) \leq s(0)$, and for all subsequent intervals $[t_i, t_{i+1}]$, $i \geq 1$, $X(t_i) \leq s(t_i)$ for all $i \geq 1$.

As a corollary, we have a FWLLN for the sequence $\{(\hat{N}_n, \hat{B}_n(0, \cdot), \hat{X}_n(0)) : n \geq 1\}$. In particular,
\[
(\hat{N}_n(t), \hat{B}(0, \cdot), \hat{X}_n(0)) \Rightarrow (\Lambda(t), B(0, \cdot), X(0)) \quad \text{in} \quad \mathcal{D}([0, T]; \mathbb{R}) \times \mathcal{D}([0, \infty); \mathbb{R}) \times \mathbb{R},
\]
as $n \to \infty$, where
\[
\Lambda(t) = \int_0^t \lambda(u) \, du \quad \text{and} \quad B(0, y) = \int_0^y b(0, u) \, du, \quad t, y \geq 0. \tag{5.4}
\]

### 5.3 Review of $G_t/GI/s_t + GI$ fluid queue

In this section, we review the $G/GI/n + GI$ fluid queueing system based on the fluid model introduced in [61]. They are the limits of the LLN-scaled processes (see Theorem 5.5.1).

Fluid flows into the system according to a deterministic arrival process with time-varying rate $(\lambda(t) : t \geq 0)$. The input content moves directly to service if there is available space...
otherwise the input content flows into waiting space with unlimited capacity. Service capacity at any time is given by the differentiable function $s(t)$ with derivative $s'(t)$. The total amount of input by time $t$ is given by $\Lambda(t) = \int_0^t \lambda(u) \, du$, $t \geq 0$. The cdfs $G$ and $F$ in the context of stochastic queue are interpreted as proportions in the context of the fluid queue. That is, $G(x)$ is the proportion of fluid that has completed service by time $x$, and $F(x)$ is the proportion of fluid that has abandoned the system by time $x$ before starting service process. We assume that $G$ and $F$ are differentiable with densities $g$ and $f$, respectively. Moreover, we assume that $\lambda$, $s'$ and $f$ are piecewise continuous with finitely many discontinuities on bounded intervals.

The key performance measures of the fluid queue are the two-parameter deterministic processes $B(t, y)$ and $Q(t, y)$ defined as the amount of fluid in service, respectively in the waiting space, that has been so for at most $y$ units of time. These quantities can also be represented as integrals of density functions:

$$B(t, y) = \int_0^y b(t, s) \, ds \quad \text{and} \quad Q(t, y) = \int_0^y q(t, s) \, ds \quad \text{for} \quad t, y \geq 0,$$

where the density functions $b(\cdot, \cdot)$ and $q(\cdot, \cdot)$ are defined as in (5.5) and (5.11), respectively. Given these definitions, the total amount of fluid in the system at time $t$, $X(t)$, is given by $X(t) = B(t) + Q(t)$ with $B(t) \equiv B(t, \infty)$ and $Q(t) \equiv Q(t, \infty)$. Finally, the total amount of abandoned fluid over $[0, t]$ satisfies

$$A(t) = \Lambda(t) - E(t) - X(t) + X(0), \quad t \geq 0,$$

where $E(t)$ denotes the amount of fluid that has entered service by time $t$ and defined in (5.10).

**Regimes.** The fluid queue is said to be overloaded (OL) at time $t$ if either fluid content in the waiting space is strictly positive, i.e., $Q(t) > 0$, or there is no fluid in the waiting space, and fluid content in service is equal to the service capacity and arrival rate at time $t$ is greater than the service-completion rate $s'(t) + \sigma(t)$, i.e., $Q(t) = 0$, $B(t) = s(t)$ and $\lambda(t) > s'(t) + \sigma(t)$, where $s'(t)$ is the rate of change in the staffing level, and $\sigma(t)$ is the service-completion rate given by

$$\sigma(t) = \int_0^\infty b(t, x) h_G(x) \, dx, \quad t \geq 0,$$

where $h_G$ is the hazard rate function associated with $G$, i.e., $h_G = g(x)/G^c(x)$, and $b(\cdot, \cdot)$ is as in (5.5). We assume that $F^c(x) \equiv 1 - F(x) > 0$ for all $x$, $\lambda_{\inf} \equiv \inf\{\lambda(t) : 0 \leq t \leq T\} > 0$ and $s_{\inf} \equiv \inf\{s(t) : 0 \leq t \leq T\} > 0$.

The fluid queue is said to be underloaded (UL) at time $t$ if either fluid content in service is strictly less than service capacity, i.e., $B(t) < s(t)$, or there is no fluid in waiting space, fluid
content in service is equal to the service capacity and arrival rate at time \( t \) is less than the service-completion rate \( s'(t) + \sigma(t) \), i.e., \( Q(t) = 0 \), \( B(t) = s(t) \) and \( \lambda(t) < s'(t) + \sigma(t) \).

The fluid queue is said to be critically loaded at time \( t \) if \( Q(t) = 0 \), \( B(t) = s(t) \), and \( \lambda(t) = s'(t) + \sigma(t) \). In this paper, we assume that the fluid queueing system starts out underloaded with some initial content \( X(0) \leq s(0) \) in service, and that it alternates between underloaded and overloaded regimes on finite time intervals. The fluid queue switches from one regime to the other at switching points. We assume that the fluid queue has only finitely many switches between intervals in finite time horizons. We further impose the stronger condition that the fluid queue is not critically loaded at any time point except at switching points. This stronger condition is needed to establish FCLT limits well defined on the intervals (both UL and OL) separated by a switching point. The fluid model in [61] incorporates critical loading at times other than switching points. Without loss of generality, we discuss both UL and OL intervals on the interval \([0, T]\) throughout this section.

**Service content.** The density function associated with fluid in service at time \( t \) that has been in service for exactly \( x \) units of time can be represented as

\[
b(t, x) = b(t - x, 0)G_c(x)1(x < t) + b(0, x - t)\frac{G_c(x)}{G_c(x - t)}1(x \geq t), \quad t, x \geq 0.
\]  
(5.5)

The function \( b(t - x, 0) \) on the right is the amount of the fluid content that entered service exactly \( x \) units of time in the past. Consequently, the amount of fluid at time \( t \) after \( x \) units of time has passed, \( b(t, x) \), is what’s left over from the initial amount \( b(t - x, 0) \) after the departure of processed fluid during the period of length \( x \). The condition \( \{x < t\} \) implies that the first term is associated with fluid that enters system from outside after time 0. Hence the rate of departure of processed fluid in service is determined by the proportion \( G_c \). The term \( b(0, x - t) \), on the other hand, is the amount of fluid that has been present in service for exactly \( x \) \((\geq t)\) units of time at time 0. That is, the second term describes the behavior of initial content. In particular, the density at time \( t \) given that fluid has been in service \( x \) \((\geq t)\) units of time, \( b(t, x) \), is what’s left over from the initial amount at time 0, \( b(0, x - t) \), after the departure of processed fluid during the period of length \( t \). The density \( b(0, x - t) \) is part of initial data.

For UL intervals, the rate fluid enters service \( b(\cdot, 0) \) in (5.5) is equal to the external arrival rate \( \lambda(\cdot) \) because of the fact that there is excess service capacity during UL intervals, and hence, arriving fluid immediately start service process. For OL intervals, on the other hand, determining \( b(\cdot, 0) \) is more complicated because arriving fluid does not immediately start service process. In particular, in OL intervals, the flow rate into service at time \( t \), \( b(t, 0) \), must be equal to the rate at which fluid is processed out of service. As discussed in §6.2 of [61], rate into
service is given by
\[ b(t, 0) = \hat{a}(t) + \int_0^t b(t - x, 0)g(x) \, dx, \quad 0 \leq t \leq T, \tag{5.6} \]
with
\[ \hat{a}(t) \equiv s'(t) + \int_0^\infty b(0, y) \frac{g(t + y)}{G^c(y)} \, dy, \quad 0 \leq t \leq T, \tag{5.7} \]
where \( s'(t) \) is the derivative of \( s(t) \), and the integral in (5.7) is assumed to be finite. It has been shown in Theorem 2 of [61] that there exists a unique \( b(t, 0) \) satisfying (5.6) whenever the tail of \( b(0, \cdot) \) is bounded in the following sense: For \( T > 0 \),
\[ \sup_{0 \leq s \leq T} \int_0^\infty b(0, y) \frac{g(s + y)}{G^c(y)} \, dy < \infty. \tag{5.8} \]
Some sufficient conditions for (5.8) has been provided in Remark 4 of [61]. We also assume that the flow rate into service is strictly positive during OL intervals, i.e., there exists a constant \( b > 0 \) such that
\[ b(t, 0) \geq b > 0 \quad \text{for all} \quad 0 \leq t \leq T. \tag{5.9} \]
Finally, for OL intervals, the amount of fluid that has entered service by time \( t \)
\[ E(t) = \int_0^t b(u, 0) \, du, \quad 0 \leq t \leq T, \tag{5.10} \]
where \( b(\cdot, 0) \) is as in (5.6).

**Queue content.** We next review performance functions associated with queue content, i.e., fluid content in the waiting space, in OL intervals assuming that no fluid content is present waiting in queue for service initially. The density of fluid content in waiting space at time \( t \) that has been waiting for exactly \( x \) units of time is given by
\[ q(t, x) = \lambda(t - x)F^c(x)1(x \leq w(t) \land t), \quad 0 \leq t \leq T, \tag{5.11} \]
where \( w(t) \) is such that, for all \( t \geq 0 \), \( q(t, x) = 0 \) if \( x > w(t) \). The function \( w(t) \) can be interpreted as the fluid head-of-line waiting time. Similar to the interpretation of (5.5), the first term on the right-hand side is associated with the fluid content that enters the waiting space after time 0. The amount of fluid content that has been waiting for exactly \( x \) \((< t)\) units of time at time \( t \) is what is left over from the fluid amount \( x \) time ago, \( q(t - x, 0) \), after some proportion determined by \( F^c \) has abandoned.
By Theorem 3 of [61], the fluid head-of-line waiting time \( w(t) \) uniquely solves the ordinary differential equation (ODE)

\[
w'(t) = 1 - \frac{b(t,0)}{q(t,w(t))} = 1 - \frac{b(t,0)}{\lambda(t-w(t))F^c(w(t))}, \quad 0 \leq t \leq T,
\]  

whenever \( 0 \leq w(0) < \infty \) and \( \inf_{0 \leq u \leq t} \lambda(t) > 0 \). Our assumptions imply that \( -\infty < \dot{w}(t) < 1 \) for all \( t \) during OL intervals.

By Theorem 5 of [61], the fluid virtual waiting time \( v(t) \) is the unique solution satisfying

\[
v(t) = w(t + v(t)) \quad \text{and} \quad v(t - w(t)) = w(t), \quad 0 \leq t \leq T.
\]  

Moreover \( v \) is continuous due to (5.9). Finally, the fluid abandonment during an OL interval is

\[
A(t) = \int_0^t a(u) \, du \quad \text{where} \quad a(u) = \int_0^\infty Q(u,x)h_F(x) \, dx
\]

with \( h_F \) is the hazard rate function associated with the cdf \( F \), i.e., \( h_F(x) = f(x)/F^c(x) \) for \( x \geq 0 \).

### 5.4 Prelimit processes

We now define the key prelimit processes for UL and OL intervals. In this section, we consider closed intervals of underloading and overloading, respectively, under the assumption of no critical loading any time. For simplicity, we refer to intervals (both UL and OL) in question as \([0,T]\). This does not cause any loss of generality because given appropriate initial conditions, treatment of arbitrary intervals \([t_1,t_2]\) with the same regime is identical. Characterization of the sequence of enter-service processes and its limit is key to establish limit theorems for UL and OL intervals.

#### 5.4.1 UL intervals

In UL intervals, the dynamics of the system is asymptotically identical to an equivalent \( G_t/GI/\infty \) queue with identical arrival process, service times, and initial conditions. Therefore, we make use of the expressions applicable to infinite-server queues in [4] to represent number of customers
in service at time \( t \) that has been so for at most \( y \). For \( 0 \leq t \leq T, \ y \geq 0 \), we have

\[
B_n(t, y) = \sum_{i=N_n((t-y)^+)+1}^{N_n(t)} 1(\tau_i^n + \nu_i^n > t) + \sum_{j=1}^{B_n(0,(y-t)^+)} 1(\nu_{-j}^n > t) + \epsilon_n
\]

\[= n \int_{(t-y)^+}^{t} \int_0^{\infty} 1(s + x > t) d\hat{U}_n(\tilde{N}_n(s), G(x)) + \sum_{j=1}^{B_n(0,(y-t)^+)} 1(\nu_{-j}^n > t) + \epsilon_n \]

\[= B_{n,1}^\nu(t, y) + B_{n,2}^\nu(t, y) + \epsilon_n \]

where the \( \epsilon_n \) is the asymptotically negligible error term that accounts for the difference due to infinite-server approximation to the finite-server queueing system. By convention \( \sum_{\emptyset} = 0 \). The first term in (5.14) is associated with new customers who arrive after time 0 whereas the second term is associated with initial customers in service at time 0 with elapsed service times. The condition \( y > t \) holds only if there are initial customers in service at time 0. For new customers who arrive after time 0, \( y \leq t \) must hold. If an initial customer has elapsed service time \( y > t \) at time \( t \), then the customer must have elapsed service time \( y - t > 0 \) at time 0. Hence the upper limit of the second term in (5.14) follows. The indicator function ensures that from the customers with age at most \( y - t \) at time 0, the ones who will still be in service at time \( t \) are counted. On the other hand, since each new customer starts service immediately in the limiting system, those who arrive in the interval \( [t - y, t] \) and have not completed service by time \( t \) have elapsed service time at most \( y \) at time \( t \). Hence the first term in (5.14).

From [4, 78], the first term in in (5.14) can be represented as

\[B_{n,1}^\nu(t, y) = B_{n,1}^\nu(t, y) + B_{n,2}^\nu(t, y) + B_{n,3}^\nu(t, y), \quad 0 \leq y \leq t \leq T,\]

with

\[B_{n,1}^\nu(t, y) \equiv \sqrt{n} \int_{t-y}^{t} G_c(t - s) d\tilde{N}_n(s), \quad 0 \leq y \leq t \leq T,\]

\[B_{n,2}^\nu(t, y) \equiv \sqrt{n} \int_{t-y}^{t} \int_0^{\infty} 1(x + s > t) d\hat{U}_n(\tilde{N}_n(s), G(y)), \quad 0 \leq y \leq t \leq T,\]

\[B_{n,3}^\nu(t, y) \equiv n \int_{t-y}^{t} G_c(t - s) d\Lambda(s), \quad 0 \leq y \leq t \leq T,\]

the integrator \( \hat{U}_n \) being the sequential empirical process, and \( \Lambda \) being as in Assumption 5.2.2. Moreover, it holds that the total number in system \( X_n(t) = B_n(t, \infty) + o(1) \) and \( Q_n(t) = o(1) \) for all \( 0 \leq t \leq T \). Finally, since the system is underloaded, an arriving customer finds an idle server upon arrival with probability 1 as \( n \to \infty \), and enters service immediately. Hence
\[ E_n(t) = N_n(t) + o(1) \text{ for all } 0 \leq t \leq T. \]

### 5.4.2 OL intervals

In OL intervals, all servers are busy throughout the entire interval \([0, T]\) with probability 1 as \(n \to \infty\). Therefore, in the limiting system, customers enter service only if some servers become idle. Unlike during UL intervals, customers do not start service at the same rate of arrivals. Instead, they enter service at the rate of service completions. Consequently, the arrival process does not have a direct impact on the enter-service process. Hence we write

\[ E_n(t) = (s_n(t) - s_n(0)) + \sum_{j=1}^{B_n(0)} 1(\nu - j \leq t) + \sum_{i=1}^{E_n(t)} 1(e^n_i + \nu_i \leq t) + \epsilon_n, \quad 0 \leq t \leq T, \quad (5.20) \]

where the first term is the change in the number of servers, the second term is the number of service completions associated with initial customers in service at time 0, and the third term is the number of service completions associated with new arrivals after time 0. The error term \(\epsilon_n\) is negligible after scaling, and accounts for possible deficiency in the number of customers in system due to randomness, i.e., negligible time periods with idle servers.

We obtain a similar mathematical form as in (5.14) for the second sum on the right in (5.20) by replacing the external arrival process with the enter-service process, and arrival times with service start times. The term associated with initial customers in (5.20) remains unchanged. In particular, for \(0 \leq t \leq T, \ y \geq 0\),

\[ B_n(t, y) = \sum_{i=E_n((t-y)^+)}^{E_n(t)} 1(\nu^n_i + e^n_i > t) + \sum_{j=1}^{B_n(0,(y-t)^+)} 1(\nu^n_{-j} > t) \]

\[ = n \int_{(t-y)^+}^t \int_0^1 1(y > G(t-s)) \, d\tilde{U}_n(\tilde{E}_n(s), y) + \sum_{j=1}^{B_n(0,(y-t)^+)} 1(\nu^n_{-j} > t) \quad (5.21) \]

\[ \equiv B^n_n(t, y) + B^n_o(t, y) \quad (5.22) \]

where \(B^n_n(t, y)\) has a similar representation to (5.17)-(5.19) with arrival times replaced with service start times \(\{e^n_i\}\), \(N_n(t)\) replaced with \(E_n(t)\), and \(\Lambda\) replaced with \(E\) - the fluid limit for enter-service process in [62].

We next give an expression for the queue-length process, i.e., the number in the waiting line, associated with new arrivals only. The reason is that the impact of initial customers waiting in the line on the FCLT at switching points are asymptotically negligible as we will later prove. Following the discussion in [65], the number of customers waiting in line at time \(t\) can be
represented as

\[ Q_n(t) = \sum_{i=N(t)}^{N_n(t)} 1(\tau_i^n + \gamma_i^n > t) \]

\[ = n \int_{t}^{t-W_n(t)} \int_{0}^{1} 1(y > F(t-s)) d\tilde{\Upsilon}_n(\tilde{N}_n(s), y), \quad 0 \leq t \leq T. \]  

(5.24)

Similarly, (5.24) can be represented as the sum of three terms, i.e.,

\[ Q_n(t) \equiv Q_{n,1}(t) + Q_{n,2}(t) + Q_{n,3}(t), \quad 0 \leq t \leq T, \]

where

\[ Q_{n,1}(t) \equiv \sqrt{n} \int_{t-W_n(t)}^{t} F^c(t-s) d\tilde{N}_n(s), \quad 0 \leq t \leq T, \]  

(5.25)

\[ Q_{n,2}(t) \equiv \sqrt{n} \int_{t-W_n(t)}^{t} \int_{0}^{\infty} 1(x > F(t-s)) d\tilde{\Upsilon}_n(\tilde{N}_n(s), y), \quad 0 \leq t \leq T, \]  

(5.26)

\[ Q_{n,3}(t) \equiv n \int_{t-W_n(t)}^{t} F^c(t-s) d\Lambda(s), \quad 0 \leq t \leq T, \]  

(5.27)

and \( \Lambda(s) \) is as in Assumption 5.2.2.

5.5 Main results

In this section, we establish an FCLT for the key performance processes associated with both underloaded and overloaded intervals. An FCLT for UL intervals is provided in §5.5.1. The FCLT in §5.5.1 is different than the one proven in [4] in that we impose a different assumption on the initial state of the system. An FCLT for OL intervals is given in §5.5.2 extending the FCLT in [3].

The key difference in this paper compared to [65] is the treatment of the initial state of the system at the beginning of UL and OL intervals. An FCLT for UL and OL intervals for the \( G_t/M/s_t + GI \) queue has been established in [65] under the assumption that customers initially in service have i.i.d. remaining service times. This assumption is appropriate for \( G_t/M/s_t + GI \) because at any time \( t \), the remaining service times of customers in service are independent and exponentially distributed due to memoryless property. When service distribution is nonexponential, on the other hand, the remaining service times of customers in service depend on those customers' elapsed service times (age in service). At the beginning of UL and OL intervals, the remaining service times of customers in service are not exponentially distributed in the
$G_t/GI/s_t+GI$ queueing system. Therefore, we utilize the FCLT for the initial content process established in [4] that accounts for non-i.i.d. ages of initial customers. The FCLT in [4] draws on Assumption 5.2.2.

5.5.1 Heavy-Traffic Limits for UL Intervals

We establish an FCLT for the scaled total number in system for a UL interval $[0,T]$ where the system is UL at all times in the interval $[0,T]$. In this case, the scaled number-in-system processes $\{\hat{\bar{X}}_n : n \geq 1\}$ associated with the finite-server system are asymptotically equivalent to the associated number-in-system processes $\{\hat{\bar{X}}_n^\infty : n \geq 1\}$ in an $G_t/GI/\infty$ queueing system with the same arrival process, service times, and initial conditions (the initial state of the system). The analysis of initial customers and new arrivals can be done separately in the context of infinite-server queues because these customers do not interact when they enter service. As a consequence, the FCLT limit for the number in system in Theorem 5.5.1 is represented as the sum of independent terms associated with initial customers and new arrivals. The following theorem is slightly different than the FCLT in [4] in terms of initial conditions. In addition to the assumption on initial conditions in [4], we further assume an FCLT for initial number in system jointly with the arrival process and initial number in service.

Theorem 5.5.1 (FWLLN and FCLT for UL intervals) Consider a UL interval $[0,T]$. Suppose Assumption 5.2.1 and Assumption 5.2.2 hold with $X(0) \leq s(0)$, i.e., the system is initially UL in fluid-scale. Then, as $n \to \infty$,

$$(N_n, B_n(0,\cdot), B_n(\cdot,\cdot), \hat{X}_n, \hat{B}_n(0,\cdot), \hat{B}_n(\cdot,\cdot), \hat{\bar{X}}_n) \Rightarrow (\Lambda, B(0,\cdot), B(\cdot,\cdot), \hat{X}, \hat{B}(0,\cdot), \hat{B}(\cdot,\cdot), \hat{\bar{X}})$$

in $D^6([0,T];\mathbb{R}) \times D^2([0,T];D((0,\infty);\mathbb{R}))$ where the two-parameter FWLLN limit $B(t,y)$ is continuous and deterministic satisfying

$$B(t,y) \equiv B^\circ(t,y) + B''(t,y), \quad 0 \leq t \leq T, y \geq 0,$$

with

$$B^\circ(t,y) = \int_0^{(y-t)^+} b(0,x)\bar{G}_x(t)dx, \quad y \geq 0, \quad 0 \leq t \leq T,$$

$$B''(t,y) \equiv \int_{(t-y)^+}^t G^c(t-s)\lambda(s)ds, \quad y \geq 0, \quad 0 \leq t \leq T,$$

with $X(t) = B(t,\infty)$, and $b(0,x)$ being the initial fluid limit age density as in §5.3 and $\lambda(s)$ being the arrival rate function specified in Assumption 5.2.2.
Theorem 5.5.2 (FWLLN for OL intervals)
Consider an OL interval $[0, T]$. Suppose Assumption 5.2.1 and Assumption 5.2.2 with $X(0) = s(0)$. Then, as $n \to \infty$,

$$ (\tilde{N}_n, \tilde{D}_n, \tilde{E}_n, \tilde{\bar{B}}_n, \tilde{\bar{B}}_n(\cdot, \cdot), \tilde{Q}_n, W_n, V_n, \tilde{A}_n, \tilde{X}_n) \Rightarrow (\Lambda, D, E, s, B(\cdot, \cdot), Q, w, v, A, X) \quad (5.28) $$

in $\mathbb{D}^9([0, T]; \mathbb{R}) \times \mathbb{D}([0, T]; \mathbb{D}([0, \infty); \mathbb{R}))$ where the prelimit processes are given in §5.2, and the deterministic limiting functions are given in §5.3 with $Q(t) \geq 0$, $X(t) = s(t) + Q(t)$, and $E(t) = s(t) - s(0) + D(t)$ for $0 \leq t \leq T$.

We establish an FCLT for the key performance processes in OL intervals by first considering special conditions on the initial state of the system. In particular, we first establish an FCLT
under the assumption that all queueing systems are critically loaded at time 0, i.e., all servers are busy, and no customer is waiting in line. More specifically, \( X_n(0) = B_n(0) = s_n(0) \) and \( Q_n(0) = W_n(0) = 0 \) for all \( n \geq 1 \). Then it necessarily holds that \( \bar{X}_n(0) \Rightarrow s(0) \) and \( \hat{X}_n(0) \Rightarrow 0 \) in \( \mathbb{R} \), which is a special case of the initial conditions in Assumption 5.2.2. We later establish an FCLT under more general conditions extending the previously established FCLT. In this section, we denote the performance processes associated with the special initial conditions with superscript \( * \) whereas we denote the processes associated with more general initial conditions without superscript.

In the next theorem, we establish an FCLT for the total number in system under general initial conditions exploiting an FCLT for the other key processes under special initial conditions.

**Theorem 5.5.3 (FCLT for OL intervals)** Consider an OL interval \([0, T]\). Suppose Assumption 5.2.1 and Assumption 5.2.2 with \( X^*(0) = s(0) = X(0), \hat{X}^*(0) = 0, X^*(t) > s(t) \) for \( t \in (0, T] \), and \( X(t) > s(t) \) for \( t \in (0, T] \), i.e., the system is strictly OL in fluid-scale in the interval \((0, T]\). Then, as \( n \to \infty \),

\[
\begin{align*}
(\hat{N}^*, \hat{D}^*, \hat{E}^*, \hat{B}^*, \hat{B}^*_n(\cdot), \hat{Q}^*, \hat{W}^*, \hat{V}^*, \hat{A}^*, \hat{X}^*, \hat{X}_n) \\
\Rightarrow (\hat{N}^*, \hat{E}^*, 0^*, \hat{B}^*(\cdot, \cdot), \hat{Q}^*, \hat{W}^*, \hat{V}^*, \hat{A}^*, \hat{Q}^*, \hat{X})
\end{align*}
\tag{5.29}
\]

in \( \mathbb{D}([0, T]; \mathbb{R}) \times \mathbb{D}([0, T]; \mathbb{D}([0, \infty); \mathbb{R})) \) where \( \hat{A}^*(t) = \hat{N}^*(t) - \hat{Q}^*(t) - \hat{E}^*(t) \), and

\[
\hat{X}(t) = \hat{Q}^*(t) + \hat{X}(0) F^c_w(t) \quad \text{with} \quad F^c_w(t) = e^{-\int_0^t h_F(w(r)) dr}, \quad 0 \leq t \leq T,
\tag{5.30}
\]

where \( h_F \) is the hazard rate function associated with the cdf \( F \), \( w \) is given in (5.13), and \( Q^* \) is given in (5.40)–(5.41).

The FCLT limit for the enter-service process \( (\hat{E}^*(t) : 0 \leq t \leq T) \) uniquely solves the convolution-type Volterra equation (CVE)

\[
\hat{E}^*(t) = \hat{Y}(t) + \int_0^t g(t-u) \hat{E}^*(u) du
\tag{5.31}
\]

where (5.31) holds almost surely,

\[
\hat{Y}(t) = \hat{E}^o_1(t) + \hat{E}^o_2(t) + \hat{E}^o_3(t), \quad 0 \leq t \leq T,
\tag{5.32}
\]

with \( \hat{E}^o_1(t) \) being a centered Gaussian process with continuous sample paths, and covariance
function
\[ C_{\mathcal{E}_1}(t, t') = \text{Cov} \left( \bar{\mathcal{E}}_1^c(t), \bar{\mathcal{E}}_1^c(t') \right) = -\int_0^\infty \tilde{G}_u(t \wedge t') \tilde{G}_u^c(t \vee t') dB(0, u), \quad (5.33) \]

and
\[ \bar{\mathcal{E}}_2^c(t) = \int_0^\infty \tilde{G}_u(t) d\bar{B}(0, u), \quad \bar{\mathcal{E}}_2^c(t) = \int_0^t \int_0^t 1(x + s \leq t) d\bar{U}(E(s), G(x)) \quad (5.34) \]

with \( \bar{\mathcal{E}}_1^c(t), \bar{\mathcal{E}}_2^c(t), \bar{\mathcal{E}}_2^c(t) \) being independent processes independent of \( \bar{B}(0, \cdot) \), and of the standard Kiefer process \( \hat{U} \). The first integral in (5.34) is interpreted as the form after integration by parts. The integrator \( B(0, \cdot) \) in (5.33) and \( E \) in (5.34) are as in [62].

The limit for the two-parameter process \( \hat{B}_n(t, y) \) can be represented as
\[ \hat{B}_n^*(t, y) = \hat{B}_1^c(t, y) + \hat{B}_2^c(t, y) + \hat{B}_2^c(t, y) + \hat{B}_2^c(t, y), \quad 0 \leq t \leq T, \ y \geq 0, \]

where \( \hat{B}_1^c(t, y), \hat{B}_2^c(t, y), \hat{B}_2^c(t, y) \) are independent processes whereas \( \hat{B}_1^c(t, y) \) is dependent with the three processes. In particular, \( \hat{B}_1^c(t, y) \) is a zero-mean Gaussian process with continuous paths, and covariance function
\[ C_{\hat{B}_1^c}(t, t', y, y') = \text{Cov}(\hat{B}(t, y), \hat{B}(t', y')) = \int_0^{(y-t)^+ \wedge (y'-t')^+} \tilde{G}_u(t \wedge t') \tilde{G}_u^c(t \vee t') dB(0, u) \quad (5.35) \]

for \( 0 \leq t' \leq T, \ y' \geq 0, \) and
\[ \hat{B}_2^c(t, y) = \int_0^{(y-t)^+} \tilde{G}_u^c(t) d\hat{B}(0, u), \quad \hat{B}_1^c(t, y) = \int_{(t-y)^+}^{t} G_c(t - s) d\hat{E}(s), \quad (5.36) \]

\[ \hat{B}_2^c(t, y) = \int_{(t-y)^+}^{t} \int_0^\infty 1(s + x > t) d\hat{U}(E(s), G(x)) \quad 0 \leq t \leq T, \ y \geq 0. \]

where \( \hat{U} \) is a standard Kiefer process, and is independent of \( \hat{B}(0, \cdot) \). Both integrals in (5.36) are interpreted as the form after integration by parts.

The FCLT limit for the HOL waiting time \( \hat{W}(t) : 0 \leq t \leq T \) uniquely solves the SDE
\[ \hat{W}^*(t) = -\frac{1}{\lambda(t - w(t))F_c(w(t))} \int_0^t \lambda(s - w(s)) f(w(s)) \hat{W}^*(s) ds + \frac{1}{\lambda(t - w(t))F_c(w(t))} \hat{Z}(t), \quad (5.37) \]

where (5.37) holds almost surely, \( w \) is as in (5.12), \( f \) is the derivative of the patience-time
distribution \( F \),

\[
\hat{Z}(t) = \int_0^t F^c(w(s)) d\hat{N}(s - w(s)) + B \left( \int_0^t F^c(v(u)) F(v(u)) d\Lambda(u) \right) - \hat{E}(t), \quad t \geq 0, \quad (5.38)
\]

with \( (B(t) : t \geq 0) \) being the standard Brownian motion independent of the processes \( \hat{N} \) and \( \hat{E} \).

The first term in (5.38) is interpreted as the form after integration by parts. The deterministic function \( v \) in (5.38) is characterized by (5.13).

The FCLT limit for the virtual waiting time \( (\hat{V}(t) : 0 \leq t \leq T) \) satisfies

\[
\hat{V}^*(t) = \frac{\hat{W}^*(t + v(t))}{1 - w'(t + v(t))}, \quad t \geq 0,
\]

where \( w' \) is the derivative of \( w \), and \( w \) and \( v \) are as in (5.12) and (5.13), respectively.

The FCLT limit for the queue-length process is the sum of three independent terms, i.e.,

\[
\hat{Q}^*(t) \equiv \hat{Q}^*_1(t) + \hat{Q}^*_2(t) + \hat{Q}^*_3(t), \quad 0 \leq t \leq T,
\]

with

\[
\hat{Q}^*_1(t) \equiv \int_{t-w(t)}^t F^c(t-s) d\hat{N}^*(s), \quad \hat{Q}^*_2(t) \equiv \int_{t-w(t)}^t \int_0^1 1(x > F(t-s)) d\hat{U}(\lambda(s), x), \quad \hat{Q}^*_3(t) \equiv \lambda F^c(w(t)) \hat{W}^*(t), \quad 0 \leq t \leq T.
\]

Finally, (5.29) is a centered Gaussian process with continuous sample paths if \( \hat{X}(0) \) in (5.30) is a Gaussian random variable.

An FCLT for the other processes under general initial conditions requires a more careful treatment due to possible discontinuities that the limits for the other processes might have depending on the sign of \( \hat{X}(0) \). We discuss the details in §5.9, and conjecture an FCLT for the other processes.

### 5.6 Characterizing the FCLT Limits

We first characterize the limit \( \hat{E} \) in (5.31) by exploiting the fact that (5.31) is defined almost surely. Theorem 5.6.1 draws on the results in [32]. We later assume, for tractability, that \( (B(0, y) : y \geq 0) \) is a centered Gaussian process with Hölder-continuous sample paths, and satisfies a regularity condition (see Assumption 5.6.1). Then \( \hat{Y} \) in (5.32) becomes a Gaussian process with continuous sample paths since the first integral in (5.34) is a centered Gaussian process when \( \hat{B}(0, \cdot) \) is Gaussian with Hölder-continuous sample paths (see Proposition 4.1 of
Moreover, the solution \( \hat{E} \) of (5.31) is Gaussian with continuous sample paths. Consequently, we are able to obtain an explicit solution \( \hat{W} \) for the SDE in (5.37) in Corollary 5.6.1.

Let \( \Omega^o \subset \Omega \) be such that \( \mathbb{P}(\Omega^o) = 1 \) and (5.31) holds for each \( \omega \in \Omega^o \).

**Theorem 5.6.1 (Characterization of the limit \( \hat{E} \))** For each \( \omega \in \Omega^o \), (5.31) has a unique solution

\[
\hat{E}(t) = \hat{Y}(t) - \int_0^t r(t-s)\hat{Y}(s) \, ds, \quad 0 \leq t \leq T,
\]

where \( \hat{Y} \) is as in (5.32), and \( r \) - the resolvent of (5.31) - is the unique solution to the (deterministic) integral equation

\[
r(t) = -g(t) + \int_0^t g(t-s)r(s) \, ds, \quad 0 \leq t \leq T,
\]

where \( g \) is the derivative of the cdf \( G \). If \( (Y(t) : 0 \leq t \leq T) \) is Gaussian with \( \mathbb{E}[\hat{Y}(t)] = 0 \), \( 0 \leq t \leq T \), then so is \( \hat{E}(t) : 0 \leq t \leq T \). Moreover, if \( (Y(\omega;t) : 0 \leq t \leq T) \) is continuous for each \( \omega \in \Omega^o \), then so is \( \hat{E}(\omega;t) : 0 \leq t \leq T \) for each \( \omega \in \Omega^o \).

**Proof of Theorem 5.6.1.** From (5.42), we immediately have that \( \hat{E} \) is zero-mean Gaussian when \( \hat{Y}(t) \) is so. The existence and uniqueness of a solution \( r(t) \) to (5.43) directly follows from Theorem 2.3.1 of [32] because the service-time pdf \( g \) is bounded on every compact interval of \([0, \infty)\). From Theorem 2.3.5 of [32] we deduce that (5.42) has a unique continuous solution for each \( \omega \in \Omega^o \) since \( Y(\omega;t) \) is continuous in \( t \) for each \( \omega \in \Omega^o \). This completes the proof.

Representation (5.42) and the fact that \( \hat{Y} \) is the sum of the three independent centered Gaussian processes allow us to derive its covariance function rather easily. In particular, for \( 0 \leq t_1, t_2 \leq T \), we obtain

\[
C_{\hat{E}}(t_1, t_2) = C_Y(t_1, t_2) - \int_0^{t_2} r(t_2 - v)C_Y(t_1, v) \, dv \\
- \int_0^{t_1} r(t_1 - u)C_Y(u, t_2) \, du + \int_0^{t_1} \int_0^{t_2} r(t_1 - u)r(t_2 - v)C_Y(u, v) \, dudv
\]

where

\[
C_Y(t_1, t_2) = \int_0^\infty \tilde{G}_x(t_1 \wedge t_2) \tilde{G}^c_x(t_1 \lor t_2) \, dB(0, u) + \int_0^\infty \int_0^\infty C_{\tilde{g}_u}(u, v) \tilde{g}_u(t_1 \wedge t_2) \tilde{g}_v(t_1 \lor t_2) \, dudv
\]

\[
+ \int_0^{t_1 \wedge t_2} G(t_1 - u)G^c(t_2 - u) \, dE(u),
\]

\( \tilde{g}_u(t) \) is the partial derivative of the cdf \( \tilde{G}_u(t) \) with respect to \( u \), and \( B(0, \cdot) \) and \( E(\cdot) \) are as in (5.4) and (5.10), respectively. For the next theorem, we further impose regularity conditions on
the service-time pdf \( g \) and the covariance function of the process \( \hat{B}(0,y) : y \geq 0 \).

**Assumption 5.6.1 (Further regularity conditions)** (i) The service-time pdf \( (g(t) : t \geq 0) \) is locally Hölder continuous of order \( \alpha \) for some 0 < \( \alpha < 1 \), i.e., for each \( N \in \mathbb{N} \) there exists 0 < \( \alpha = \alpha(N) < 1 \) and \( K = K(N) > 0 \) such that for 0 ≤ \( s, t \leq N \),

\[
|g(s) - g(t)| \leq K|s - t|^\alpha.
\] (5.45)

(ii) The FCLT limit \( \hat{B}(0,y) : y \geq 0 \) is Gaussian with mean 0, and for which the following condition holds:

\[
\int_0^\infty \int_0^\infty \mathbb{E}[\hat{B}(0,x)\hat{B}(0,y)](1 + h_G(x))\frac{\bar{g}_y(u)}{G^c(x)} \, dx \, dy < \infty, \quad \text{for all} \quad 0 \leq u \leq T,
\] (5.46)

where \( h_G \) is the hazard rate function associated with the cdf \( G \), \( \bar{g}_y(u) \) is the partial derivative of the cdf \( \bar{G}_y(\cdot) \) with respect to \( y \).

We have the following lemma.

**Lemma 5.6.1** The FCLT limit \( \hat{E}(t) : 0 \leq t \leq T \) satisfying (5.42) has a version with Hölder-continuous paths under Assumption 5.6.1.

**Proof of Lemma 5.6.1** From (i) of Assumption 5.6.1, we deduce that \( g \) is bounded on compact intervals of \([0, \infty)\). Then this implies that \((r(t) : 0 \leq t \leq T)\) satisfying (5.43) is bounded (see, e.g., Theorem 2.2.2(i) of [32]). Let \( M_r = \|r\|_T < \infty \) and \( g^\dagger_T = \sup_{0 < t \leq T} |g(t)| \). Then from (5.43) we obtain, for 0 ≤ \( s < t \leq T \), \( \alpha < 1 \),

\[
|r(s) - r(t)| \leq |g(s) - g(t)| + \int_0^s |g(s,u) - g(t,u)||r(u)| \, du + \int_s^t |g(t - v)||r(v)| \, dv
\]

\[
\leq K|s - t|^\alpha + M_r TK|s - t|^\alpha + M_r g^\dagger_T|s - t| \leq C_r|s - t|^\alpha
\] (5.47)

for some \( C_r > 0 \). The second inequality follows from (5.45) and that \( r \) is bounded on \([0, T]\) with \( M_r = \|r\|_T \). The third inequality follows from the basic inequality \(|s - t| = |s - t|^{1-\alpha}|s - t|^\alpha \leq T^{1-\alpha}|s - t|^\alpha \) for 0 < \( \alpha < 1 \) on the interval \([0, T]\). Similar arguments hold for 0 ≤ \( t < s \leq T \). Hence we conclude that \( r(t) \) satisfying (5.43) is Hölder continuous under Assumption 5.6.1.

We next prove, under Assumption 5.6.1, that \( \hat{E} \) satisfying (5.42) almost surely has Hölder-continuous sample paths whenever the covariance function of \( \hat{Y}, C_\hat{Y}(\cdot, \cdot) \), satisfies the following condition: For any 0 ≤ \( u \leq T \), there exist \( K_Y = K_Y(u) > 0 \) and 0 < \( \beta = \beta(u) < 1 \) such that

\[
|C_\hat{Y}(s,u) - C_\hat{Y}(t,u)| \leq K_Y|s - t|^\beta \quad \text{for} \quad 0 \leq s, t \leq T,
\] (5.48)
Also let $M_Y \equiv \sup_{(u,v) \in [0,T]^2} |C_Y(u,v)|$. From (5.44) and the fact that $r$ is Hölder continuous in the sense of (5.47) we obtain, for $0 \leq s < t \leq T$,  

$$
|C_E(s,t) - C_E(t,t)| \leq |C_Y(s,t) - C_Y(t,t)| + \int_0^t r(t-v)|C_Y(s,v) - C_Y(t,v)|
$$

$$
+ \int_0^t |r(t-u) - r(s-u)||C_Y(u,t)|
$$

$$
+ \int_s^t \int_0^t |r(s-u)r(t-v)||C_Y(u,v)|
$$

$$
+ \int_0^t \int_s^t |r(s-u) - r(t-u)||r(t-v)||C_Y(u,v)|
$$

$$
\leq C_1 |s-t|^\beta + C_2 |s-t|^\alpha + C_3 |s-t| \leq C_E |s-t|^{\alpha \vee \beta}
$$

for some $C_E > 0$.

To complete the proof of the lemma, it suffices to prove the following result.

**Lemma 5.6.2** If Assumption 5.6.1 holds, then the covariance function $C_Y(\cdot,\cdot)$ of the FCLT limit $\hat{Y}$ satisfies condition (5.48).

A proof of Lemma 5.6.2 is provided in Appendix D.1. Finally, the desired result follows from Corollary (25.6) of [94] (p.61) which provides a sufficient condition for centered Gaussian processes to have Hölder-continuous sample paths. This completes the proof of Lemma 5.6.1.

Having proved Lemma 5.6.1, we next have a result on the solution of the SDE for (5.37) in Theorem 5.5.3. A similar result is also presented in [3]. More specifically, Corollary 6.1. in [3] is on characterization of the FCLT ($\hat{W}(t) : t \geq 0$), and draws on a particular construction of a stochastic integral involving centered Gaussian processes with Hölder-continuous sample paths (§4.1 of [3]). Under Assumption 5.6.1, $\hat{E}$ turns out be a centered Gaussian process with Hölder-continuous sample paths that is independent of $\hat{N}$ and the Brownian motion in (5.38), and hence, we obtain an analogue of Corollary 6.1. in [3] involving time-varying arrivals.

**Corollary 5.6.1** Suppose that Assumption 5.2.1-5.6.1 hold. Then the stochastic differential equation (5.37) has a unique strong solution

$$
\hat{W}(t) \equiv \hat{W}_1(t) + \hat{W}_2(t) + \hat{W}_3(t)
$$

$$
= \int_0^t \frac{F^c(w(u))H(t,u)}{q(t,w(t))} d\hat{N}(u-w(u)) + \int_0^t \frac{\sqrt{F^c(v(u))F(v(u))H(t,u)}}{q(t,w(t))} dB(N(u))
$$

$$
- \int_0^t \frac{H(t,u)}{q(t,w(t))} d\hat{E}(u), \quad 0 \leq t \leq T,
$$

(5.49)
where, for $0 \leq r \leq t$ and $0 \leq u \leq t$,

$$H(t, u) \equiv e^{\int_0^t h(r) \, dr} \quad \text{with} \quad h(r) \equiv -\frac{\lambda(r - w(r))f(w(r))}{q(t, w(t))} = -\frac{\lambda(r - w(r))f(w(r))}{\lambda(t - w(t))F^c(w(t))},$$

(5.50)

the deterministic functions $w$ and $\lambda$ are given in (5.12) and in (5.4), respectively.

Fortunately, we can easily incorporate the time-varying arrival rate ($\lambda(t) : t \geq 0$) into the
proof of Corollary 6.1. of [3], and directly use Corollary 6.1. of [3] to deduce that

$$e^{-\int_0^t h(r) \, dr} W^*(t) = W^*(0) + \int_0^t F^c(w(u)) \frac{e^{-\int_0^u h(r) \, dr}}{q(t, w(t))} d\hat{N}(u - w(u))$$

$$+ \int_0^t \sqrt{\lambda F^c(v(u))F^c(v(u))} \frac{e^{-\int_0^u h(r) \, dr}}{q(t, w(t))} dB(u) - \int_0^t \frac{e^{-\int_0^u h(r) \, dr}}{q(t, w(t))} d\hat{E}(u).$$

Multiplying through by $H(t, 0) = e^{\int_0^t h(r) \, dr}$ yields (5.49). Note that $W^*(0) = 0$ is implied by
the assumption on the initial state of the system.

### 5.7 Proof of Theorem 5.5.1

We only prove the FCLT because the FWLLN follows as a corollary. We let $X_n$ denote the
number in system for the $G_t/GI/s_t + GI$ queueing system during the UL interval $[0, T]$ with no
critical loading at any time, and let $X_n^\infty$ denote the number in system for $G_t/GI/\infty$ queueing
system with the same arrival process, service times, and initial condition $X_n^\infty(0) = X_n(0)$ for
all $n$ under Assumption 5.2.2. We have established an FCLT in [4] for $X_n^\infty$ under the following
initial conditions:

$$(\hat{N}_n, \hat{B}_n(0, \cdot)) \Rightarrow (\hat{N}, \hat{B}(0, \cdot)) \quad \text{in} \quad \mathbb{D}([0, T]; \mathbb{R}) \times \mathbb{D}([0, \infty); \mathbb{R}) \quad \text{as} \quad n \to \infty,$$

where the limits $(\hat{N}(t) : t \geq 0)$ and $(\hat{B}(0, y) : y \geq 0)$ are as in Assumption 5.2.2. In Assumption
5.2.2 in this chapter, on the other hand, we further impose convergence for the initial number
in system, that is,

$$X_n(0) \Rightarrow X(0) \quad \text{and} \quad \hat{X}_n(0) \Rightarrow \hat{X}(0) \quad \text{in} \quad \mathbb{R} \quad \text{as} \quad n \to \infty$$

(5.51)

with $X(0) = s(0)$. We next prove that the impact of (5.51) is negligible under both LLN and
CLT scaling, and hence, the FCLT and FWLLN results in [4] prevails under Assumption 5.2.2
of this chapter. In particular, we prove that

$$\|\bar{X}_n - \bar{X}_n^\infty\|_T \Rightarrow 0 \quad \text{and} \quad \|\hat{X}_n - \hat{X}_n^\infty\|_T \Rightarrow 0 \quad \text{as} \quad n \to \infty.$$
The proof is similar to that in §8 of [65], and is provided here for completeness.

We first consider the interval \((0, T]\). Because the system is assumed to be strictly UL in fluid-scale, the net flow out \(D_n(t) - N_n(t) - (s_n(t) - s_n(0))\) is of order \(O(n)\) and is strictly positive over the interval \([u, T]\) for all \(u > 0\) due to the FWLLN established in [62]. In particular,

\[
(D_n(t) - N_n(t) - (s_n(t) - s_n(0))) = n(D(t) - \Lambda(t) - (s(t) - s(u))) + o(n) = O(n)
\]

where \(D, \Lambda,\) and \(s\) are bounded functions as in §5.3, and the first term on the right-hand side is strictly positive by definition of underloading. As a result, for \(u \leq t \leq T\), and sufficiently large \(n_0 > 0\),

\[
X_n(t) - s_n(t) = N_n(t) - D_n(t) - A_n(t) + X_n(u) - s_n(t)
\]

\[
\leq N_n(t) - D_n(t) + s_n(u) - s_n(t) \quad \text{for all } n \geq n_0,
\]

which implies from (5.53) that \(X_n(t) - s_n(t)\) is strictly negative and of order \(O(n)\). Hence, we deduce that \(\mathbb{P}(\sup_{u \leq t \leq T} \{X_n(t) - s_n(t)\} < 0) \to 1\) as \(n \to \infty\) for \(0 < u \leq T\).

Following the reasoning in the previous paragraph, the infinite-server queue with the same arrival process and service times are stochastically identical to the underloaded finite-server system in the interval \([u, T]\) given identical initial conditions \(X_n(u) = X_n(u)\). Hence the FCLT in [4] holds for the finite-server queue. We now extend the FCLT to cover the entire interval \([0, T]\). At the left endpoint 0, it is possible that \(X_n(0) > s_n(0)\), and therefore, the two systems, i.e., the finite-server and the infinite-server queuing systems, are not stochastically identical over \([0, t]\), for \(t > 0\), because \(X_n(0) - s_n(0)\) customers are waiting in line initially. From Assumption 5.2.2 and that \(X(0) = s(0)\), we observe that \(\dot{X}_n(0) = ns(0) + \sqrt{n}\dot{X}(0) + o(1)\). Hence the initial surplus \(X_n(0) - s_n(0) = \sqrt{n}X(0)^+ + o(1) = O(\sqrt{n})\).

Since the fluid model is UL with net flow out of \(O(n)\) in (5.53) over the interval \((0, T]\) and \(\lambda(t) \leq s'(t) + \sigma(t)\), the initial surplus \(\sqrt{n}\dot{X}(0)^+ = O(\sqrt{n})\) dissipates over a time interval of length \(O(1/\sqrt{n})\). In particular, let \(t_{1,n}\) be time net flow out equals the initial surplus for the first time in the \(n\)th system. Then, we have

\[
O(\sqrt{n}) = \sqrt{n}X(0)^+ = D_n(t_{1,n}) - N_n(t_{1,n}) - (s_n(t_{1,n}) - s_n(0))
\]

\[
= n(D(t_{1,n}) - \Lambda(t_{1,n}) - (s(t_{1,n}) - s(0))) + o(n) \leq nKt_{1,n} + o(n)
\]

where the second equality follows from the established FWLLN in [62], and the inequality follows by the implied Lipschitz continuity of the fluid limits (fluid limits are differentiable with bounded first derivatives). Hence \(t_{1,n}\) is of order \(O(1/\sqrt{n})\). Thus, we deduce that \(\|X_n - X_n^\infty\|_T = O(1)\) which is negligible after both LLN and CLT scaling. This completes the proof of (5.52).
5.8 Proof of Theorem 5.5.3

All proofs until the subsection “treating initial conditions” are done by assuming that the special initial condition in §5.5.2 is in full force. For notational convenience, we suppress the superscript * until that section. We use the superscript * in there to differentiate between processes associated with our two assumptions on initial conditions.

Given the joint convergence \((\hat{N}_n, \hat{D}_n) \Rightarrow (\hat{N}, \hat{D})\) in \(D^2([0,T];\mathbb{R})\), we first show that the CLT-scaled enter-service processes \((\hat{E}_n(t) : 0 \leq t \leq T)\) jointly converge with the two processes and that satisfies a convolution-type Volterra equation (CVE) almost surely for each \(n \geq 1\). We then show that the sequence of prelimit CVEs weakly converges to a limiting CVE in the space \(D([0,T];\mathbb{R})\) endowed with \(J_1\)-topology. We prove weak convergence using two different methods. First, we use compactness approach, i.e., we show that the sequence \(\{\hat{E}_n : n \geq 1\}\) is tight in \(D([0,T];\mathbb{R})\) and that every convergent subsequence must converge to the same limit. Second, we prove convergence by continuous mapping theorem. Then we derive a stochastic differential equation (SDE) for the scaled HOL-waiting-time process \((\hat{W}_n(t) : 0 \leq t \leq T)\) for each \(n \geq 1\).

We deduce convergence of the prelimit SDEs to a limiting SDE by using the arguments in [3] which draws on convergence \(\hat{E}_n \Rightarrow \hat{E}\) in \(D([0,T];\mathbb{R})\). Having established joint convergence \((\hat{N}_n, \hat{B}_n, \hat{E}_n, \hat{W}_n) \Rightarrow (\hat{N}, \hat{B}, \hat{E}, \hat{W})\) in an appropriate space, convergence of the other processes follows from continuous mapping theorem.

5.8.1 FCLT for \(\hat{E}_n(t)\)

We start with justifying the joint convergence \((\hat{N}_n, \hat{D}_n) \Rightarrow (\hat{N}, \hat{D})\) in \(D^2([0,T];\mathbb{R})\). The convergence holds because the system is assumed to be asymptotically overloaded, and therefore, the arrival process and the service-completion process is asymptotically independent given the special initial condition. In particular, it is possible that there might be idle servers initially, however, the impact of this is negligible after scaling (see [3, 65] for details). Then \((\hat{N}_n, \hat{D}_n, \hat{E}_n) \Rightarrow (\hat{N}, \hat{D}, \hat{E})\) in \(D^3([0,T];\mathbb{R})\) immediately follows from convergence-together theorem (see, e.g., Theorem 11.4.7. of [102]) provided \(\hat{E}_n \Rightarrow \hat{E}\) in \(D([0,T];\mathbb{R})\). We next prove the sequence \(\{\hat{E}_n : n \geq 1\}\) converge, and characterize its limit \(\hat{E}\).

We know from [62] that the LLN-scaled enter-service processes in (5.20) converge in distribution to a deterministic and continuous function. In particular, \(\bar{E}_n \equiv n^{-1} E_n \Rightarrow E\) in \(D([0,T];\mathbb{R})\) as \(n \to \infty\) where

\[
E(t) = s(t) - s(0) + \int_0^\infty \tilde{G}_x(t) dB(0, x) + \int_0^t G(t - u) dE(u), \quad 0 \leq t \leq T.
\] (5.54)
Then $\hat{E}_n(t) \equiv \sqrt{n}(\bar{E}_n(t) - E(t))$ reads

$$
\hat{E}_n(t) = \hat{s}_n(t) - \hat{s}_n(0) + \sqrt{n}\left(\frac{1}{n} \sum_{j=1}^{B_n(0)} \mathbf{1}(\nu_j - 1 \leq t) - \int_0^\infty \tilde{G}_x(t) dB(0, x)\right)
$$

$$
+ \sqrt{n}\left(\frac{1}{n} \sum_{i=1}^{E_n(t)} \mathbf{1}(\epsilon_i^n + \nu_i \leq t) - \int_0^t G(t - u) dE(u)\right)
$$

$$
= \hat{s}_n(t) - \hat{s}_n(0) + \sqrt{n}\left(\frac{1}{n} \sum_{j=1}^{B_n(0)} \mathbf{1}(\nu_j - 1 \leq t) - \tilde{G}_{\xi_j^n}(t)\right)
$$

$$
+ \sqrt{n}\left(\int_0^t \int_0^t \mathbf{1}(s + x \leq t) d\hat{U}_n(\bar{E}_n(s), G(x)) - \int_0^t G(t - u) dE(u)\right)
$$

$$
= \frac{1}{\sqrt{n}} \sum_{j=1}^{B_n(0)} \left(\mathbf{1}(\nu_j - 1 \leq t) - \tilde{G}_{\xi_j^n}(t)\right) + \int_0^\infty \tilde{G}_x(t) dB_n(0, x) + \int_0^t G^c(t - u) d\hat{E}_n(u)
$$

$$
+ \int_0^t \int_0^t \mathbf{1}(s + x \leq t) d\hat{U}_n(\bar{E}_n(s), G(x)) + O(1/\sqrt{n})
$$

$$
\equiv \hat{E}_{n,1}(t) + \hat{E}_{n,2}(t) + \hat{E}_{n,1}^\nu(t) + \hat{E}_{n,2}^\nu(t) + O(1/\sqrt{n}), \quad 0 \leq t \leq T. \tag{5.58}
$$

The third equality is obtained using the fact that the summation in (5.55) can be represented as an integral with respect to $\tilde{B}_n(0, \cdot)$, i.e.,

$$
\frac{1}{n} \sum_{j=1}^{B_n(0)} \tilde{G}_{\xi_j^n}(t) = \int_0^\infty \tilde{G}_x(t) dB_n(0, x), \quad 0 \leq t \leq T,
$$

where the integral is understood, for each $\omega \in \Omega$, as an integral with respect to the measure

$$
m_n(dx) = \frac{1}{n} \sum_{j=1}^\infty \delta_{\xi_j^n}(dx)
$$

where $\delta_{\xi_j^n}$ is the Dirac measure at $\xi_j^n$. With this alternative representation, the second term in (5.57) follows by definition. The integral with respect to $\hat{E}_n$ in (5.57) is understood in similar sense. The double integral in (5.56) is as in [54, 78], and can be represented as the sum of three terms one of which cancels out with the one-dimensional integral in (5.57). Hence (5.57) is represented as the sum of last two integrals in (5.57). Finally, $\hat{s}_n(t) \equiv n^{-1/2}(s_n(t) - s(t)) \leq n^{-1/2}$ for all $t \geq 0$. Hence the $O(1/\sqrt{n})$ term.
Using integration-by-parts formula for $\hat{E}_{n,1}^\nu$ in (5.58), and rearranging the terms yield
\[
\hat{E}_n(t) = \hat{Y}_n(t) + \int_0^t g(t-u)\hat{E}_n(u)\,du, \quad 0 \leq t \leq T,
\]
where
\[
\hat{Y}_n(t) \equiv \hat{E}_{n,1}^0(t) + \hat{E}_{n,2}^0(t) + \hat{E}_{n,2}^\nu(t) + O(1/\sqrt{n}).
\]

We prove convergence of the sequence $\{\hat{E}_n : n \geq 1\}$ by two different approaches.

**Approach I: The compactness approach.** We prove weak convergence in $\mathbb{D}([0,T];\mathbb{R})$ by first showing that the sequence $\{\hat{E}_n : n \geq 1\}$ is $C$-tight in $\mathbb{D}([0,T];\mathbb{R})$ which implies that every subsequence has a convergent subsequence. To establish full convergence, we then show that every convergent subsequence must converge to same limit.

**$C$-tightness of $\{\hat{E}_n\}$.** We prove $C$-tightness by using Gronwall’s inequality. In particular, from (5.59), we obtain
\[
|\hat{E}_n(t)| \leq |\hat{Y}_n(t)| + \int_0^t |\hat{E}_n(u)| g(t-u)\,du, \quad 0 \leq t \leq T,
\]
which implies, by Gronwall’s inequality,
\[
|\hat{E}_n(t)| \leq |\hat{Y}_n(t)| + \int_0^t \hat{Y}_n(u)\exp\left(\int_u^t g(t-s)\,ds\right) g(t-u)\,du \leq (1 + e^1)|\hat{Y}_n|_t
\]
for $0 \leq t \leq T$ and $n \geq 1$ where $|\hat{Y}_n|_t < \infty$ for all $0 \leq t \leq T$. Then the tightness of $\{\hat{E}_n : n \geq 1\}$ follows from the tightness of $\{\hat{Y}_n : n \geq 1\}$ by Proposition VI.3.35 of [42]. In particular, given the joint convergence
\[
(\hat{B}_n(0, \cdot), \hat{B}_n(0, \cdot), \hat{E}_n, \hat{U}_n) \Rightarrow (\hat{B}(0, \cdot), B(0, \cdot), E, \hat{U})
\]
in $\mathbb{D}^2([0,\infty),\mathbb{R}) \times \mathbb{D}([0,T],\mathbb{R}) \times \mathbb{D}([0,T];\mathbb{D}([0,\infty);\mathbb{R}))$ as $n \to \infty$, each component of $\hat{Y}_n$ converge in distribution to a Gaussian process with continuous sample paths (see [4]). Hence, by continuous mapping theorem, $\hat{Y}_n$ converges to a limit that is the sum of the limiting component processes. Then $C$-tightness of the sequence $\{\hat{Y}_n : n \geq 1\}$ is implied by convergence. Finally, (5.61) holds since $\hat{B}_n(0, \cdot)$ and $\hat{U}_n$ are independent, and the limit $E$ is deterministic.

**Characterizing the limit for convergent subsequences of $\{\hat{E}_n\}$.** Let $\{\hat{E}_{n_k} : k \geq 1\}$ be a convergent subsequence such that $\hat{E}_{n_k} \Rightarrow \hat{E}$ in $\mathbb{D}([0,T];\mathbb{R})$ as $n \to \infty$. Then $\hat{E}$ satisfies
\[
\hat{E}(t) = \hat{Y}(t) + \int_0^t g(t-u)\hat{E}(u)\,du, \quad 0 \leq t \leq T,
\]
almost surely, where

$$\hat{Y}(t) \equiv \hat{E}_1(t) + \hat{E}_2(t) + \hat{E}_3(t), \quad 0 \leq t \leq T.$$ 

To prove that the linear convolution-type Volterra equation in (5.62) has a unique solution, we directly apply the existing results. In particular, the existence and uniqueness of a solution to (5.62) follows from Theorem 2.3.1 and Theorem 2.3.5 in [32] because (5.62) is defined for each $\omega \in \Omega^o \subset \Omega$ with $\mathbb{P}(\Omega^o) = 1$. Then by Theorem 2.3.5 of [32], we know that for every almost-surely continuous $\hat{Y}$, there exists a unique almost-surely continuous process $\hat{E}$ to (5.62) provided $g \in L^1_{\text{loc}}(\mathbb{R}^+; \mathbb{C}^{n \times n})$. In our case, $\hat{Y}$ indeed has continuous sample paths for all $\omega \in \Omega^o$ since $\hat{B}(0, \cdot)$ has continuous sample paths by Assumption 5.2.2. Hence we conclude that every convergent subsequence must converge to the same unique limit.

**Approach II: The continuous mapping approach.** Let $\mathbb{D} \equiv \mathbb{D}([0, T]; \mathbb{R})$. The metric space $(\mathbb{D}, d_{J_1})$ with $d_{J_1}$ being the usual Skorohod metric is not complete. However, there exists an equivalent metric $d^o$ (i.e., gives the $J_1$-topology) under which the metric space $\mathbb{D}^o \equiv (\mathbb{D}([0, T]; \mathbb{R}), d^o)$ is complete. We make use of the complete space $\mathbb{D}^o$ to invoke Banach fixed-point theorem, and to prove existence and uniqueness of a fixed point $y \in \mathbb{D}$. In particular, we have the following lemma.

**Lemma 5.8.1** Let $y \in \mathbb{D}$, and the mapping $\psi$ be such that

$$y = \psi(y)$$

(5.63)

where, for a given $x \in \mathbb{D}$, the mapping $\psi : \mathbb{D} \to \mathbb{D}$ is given by

$$\psi(y)(t) \equiv x(t) + \int_0^t g(t - s)y(s) \, ds, \quad 0 \leq t \leq T.$$ 

(5.64)

Then the mapping $\psi$ is a contraction, and there exists a unique $y \in \mathbb{D}$ such that (5.63) holds.

**Proof of Lemma 5.8.1.** It follows that, for $u_1, u_2 \in \mathbb{D}$,

$$d^o(\psi(u_1), \psi(u_2)) \leq \int_0^t g(t - s)d^o(u_1, u_2) \, ds \leq G(T)d^o(u_1, u_2)$$

which implies that $\psi$ is a contraction mapping since $G(T) < 1$ for all $T \geq 0$ by Assumption 5.2.1. Hence we conclude, by Banach fixed-point theorem, that (5.63) has a unique solution in $\mathbb{D}$. This completes the proof.

We next define another mapping $\phi : \mathbb{D} \to \mathbb{D}$ such that $\phi(x) = y$, and $y$ uniquely solves (5.63). From Lemma 5.8.1, we see that the mapping $\phi$ is well-defined since $\phi(x)$ has one and only one image for all $x \in \mathbb{D}$. We next prove that the mapping $\phi$ is continuous.
Lemma 5.8.2 The mapping $\phi : \mathbb{D} \to \mathbb{D}$ defined as $\phi(x) = y$ where $y \in \mathbb{D}$ uniquely solves (5.63). Then the mapping $\phi : \mathbb{D} \to \mathbb{D}$ is continuous.

Proof of Lemma 5.8.2. Let $\{x_n\}$ be a sequence in $\mathbb{D}$ and $x$ is continuous. We prove $x_n \to x$ in $\mathbb{D}$ with respect to the topology induced by the metric $d^\circ$, by proving convergence in $\mathbb{D} \equiv (\mathbb{D}([0,T];\mathbb{R}), \| \cdot \|)$ endowed with uniform topology. Suppose $\|x_n - x\|_T \to 0$ as $n \to \infty$. Let $\delta > 0$ and $n_0 > 0$ be such that $\|x_n - x\|_T < G^c(T)\delta$ for $n \geq n_0$. Then, for $0 \leq t \leq T$,

$$|y_n(t) - y(t)| = |\phi(x_n)(t) - \phi(x)(t)| \leq \|x_n(t) - x(t)\| + \int_0^t g(t-s)|y_n(s) - y(s)| ds \leq \|x_n - x\|_T + G(T)\|y_n - y\|_T$$

which implies

$$\|y_n - y\|_T \leq \frac{\|x_n - x\|_T}{G^c(T)} < \delta \quad \text{for} \quad n \geq n_0$$

where $G^c(T) < 1$ for any $T \geq 0$ by Assumption 5.2.1. Hence $d^\circ(y_n, y) \to 0$ as $n \to \infty$ since $d^\circ(y_n, y) \leq \|y_n - y\|_T$. Moreover, the limit $y$ satisfying $\phi(x) = y$ is continuous because $x$ is continuous (see, e.g., Theorem 2.3.5. of [32]). Then, since $y$ is continuous, proving convergence in $(\mathbb{D}, \| \cdot \|)$ endowed with uniform topology is equivalent to proving convergence in $(\mathbb{D}^\circ, d^\circ)$ endowed with topology induced by $d^\circ$, i.e., no measurability issues arise because of the fact that the Borel $\sigma$-field does coincide with the Kolmogorov $\sigma$-field for $(\mathbb{D}, \| \cdot \|)$. Thus, the mapping $\phi$ is continuous in $(\mathbb{D}^\circ, d^\circ)$ with respect to the topology induced by $d^\circ$ which coincides with the usual $J_1$-topology (see [10]). This completes the proof the lemma.

Finally, we prove convergence for $\{\hat{E}_n : n \geq 1\}$ satisfying (5.59) by continuous mapping with composition by observing that (5.59) can be represented as $\hat{E}_n = \phi(\hat{Y}_n)$. Hence we conclude that

$$\hat{E}_n = \phi(\hat{Y}_n) \Rightarrow \hat{E} = \phi(\hat{Y}), \quad \text{in} \quad \mathbb{D}^\circ \quad \text{as} \quad n \to \infty,$$

where $\hat{E}$ is the unique solution to the limiting CVE $\hat{E} = \phi(\hat{Y})$, and has continuous sample paths since $\hat{Y}$ does.

5.8.2 SDE for $\hat{W}$ and FCLT for other processes

To derive an SDE for the limit $\hat{W}$, we make use of another representation for $(\hat{E}_n(t) : t \geq 0)$ that is valid under the special initial condition given in §5.5.2, i.e., all servers are busy and the
waiting line is empty. In particular, we use the representation in [3], that is,

\[ E_n(t) \equiv \sum_{i=1}^{N_n(t-W_n(t))} 1(\gamma_i^n > V_n(\tau_i^n)) \]

where \( V_n(\tau_i^n) \) is the offered waiting time of the \( i \)-th customer, i.e., time it takes for the \( i \)-th customer to start service assuming the customer is infinitely patient. The representation (5.66) draws on the fact that all customers that have entered service must be new arrivals under the special initial condition. Hence (5.66) is a function of the arrival process unlike (5.57). The representation in (5.57) is more general in the sense that it is still valid in case where there might be initial customers waiting in line. Representation (5.66), on the other hand, is valid only if the special initial condition holds. They are equivalent under the special initial condition. Then, just as in [3], we can derive an SDE for the prelimit HOL waiting time \( \hat{W}_n(t) : 0 \leq t \leq T \) for all \( n \geq 1 \). Following the arguments in [3], one obtains the limit in (5.37).

The FCLT limits for the other processes can be established just as in Chapter 4. Thus far, we have proved that \( (\hat{N}_n, \hat{D}_n, \hat{E}_n, \hat{W}_n) \Rightarrow (\hat{N}, \hat{E}, \hat{E}, \hat{W}) \) in \( D^4([0,T];\mathbb{R}) \) as \( n \to \infty \) where \( \hat{W} \) is as in (4.70). In particular, we start with \( (\hat{N}_n, \hat{D}_n) \Rightarrow (\hat{N}, \hat{D}) \) in \( D^2([0,T];\mathbb{R}) \) as \( n \to \infty \) because the system is assumed to operate in the ED regime, and the system starts with the special initial condition. In particular, the joint convergence holds since the arrival process and the service-completion process are asymptotically independent. Then, from §5.8.1, the joint convergence \( (\hat{N}_n, \hat{D}_n, \hat{E}_n) \Rightarrow (\hat{N}, \hat{E}, \hat{E}) \) in \( D^3([0,T];\mathbb{R}) \) immediately follows from convergence-together theorem (see Theorem 11.4.7. of [102]) since \( \|\hat{E}_n - \hat{D}_n\|_{[0,T]} \Rightarrow 0 \) in \( \mathbb{R} \). Finally, joint convergence of \( \{W_n\} \) follows from continuous mapping theorem. Similarly, the FCLT for the other processes follows from continuous mapping theorem.

From the established convergence \( \hat{W}_n \Rightarrow \hat{W} \) in \( D([0,T];\mathbb{R}) \) as \( n \to \infty \), we deduce that \( \{V_n : n \geq 1\} \) jointly converges with the above process. In particular, \( \hat{V}_n \Rightarrow \hat{V} \) in \( D([0,T];\mathbb{R}) \) as \( n \to \infty \) with

\[ \hat{V}(t) = \frac{\hat{W}(t + v(t))}{1 - \hat{w}(t + v(t))}, \quad 0 \leq t \leq T. \]

We next prove the FCLT for the queue-length process using the FCLT for \( \{\hat{W}_n\} \) and continuous mapping theorem. Since waiting line is initially empty in each system, i.e., \( Q_n(0) = 0 \) for each \( n \), it suffices to establish convergence for the scaled versions of (5.25)-(5.27). Just as
in Chapter 4, we have
\begin{align*}
\dot{Q}_{n,1}(t) & \Rightarrow \dot{Q}_1(t) \equiv \int_{t-w(t)}^t F^c(t-s) \, d\hat{N}(s), \quad 0 \leq t \leq T, \\
\dot{Q}_{n,2}(t) & \Rightarrow \dot{Q}_2(t) \equiv \int_{t-w(t)}^t \int_0^1 1(x > F(t-s)) \, d\hat{U}(\Lambda(s), x), \quad 0 \leq t \leq T, \\
\dot{Q}_{n,3}(t) & \Rightarrow \dot{Q}_3(t) \equiv \dot{W}(t)F^c(w(t))\lambda(t-w(t)), \quad 0 \leq t \leq T,
\end{align*}

in $\mathbb{D}([0,T];\mathbb{R})$ as $n \to \infty$. Finally, given the above convergence, we immediately obtain, by continuous mapping theorem, the limit $\hat{A} = \hat{N} - \hat{Q} - \hat{E}$ and $\hat{X} = \hat{Q} + \hat{B}$ where $\hat{B}$ is trivially zero. The joint convergence in Theorem 4.5.2 then follows.

5.8.3 Treating initial conditions

We now relax the assumption on the initial condition. We denote the total number in system under the special initial condition by $X^*_n$, and that under the more general initial condition in Assumption 5.2.2 by $X_n$. We assume that the processes $X^*_n$ and $X_n$ are defined on the same probability space with the same arrival process, service and patience times, and that service times and patience times are assigned to each customer in order of entering service and the waiting line, respectively. The two processes differ by initial conditions. Under the special initial condition that all systems have empty waiting and full service capacity, it holds that $X^*_n(0) - s_n(0) = 0$ for all $n$, $X^*_n(0) \Rightarrow s(0)$, and $\dot{X}^*_n(0) \Rightarrow 0$ in $\mathbb{R}$. Under the more general conditions in Assumptions 5.2.2, on the other hand, it is possible that $X_n(0) \neq s(0)$ and $\dot{X}_n(0) \Rightarrow \dot{X}(0) \neq 0$.

The impact of the difference in initial conditions are carefully treated in [4] for the $GI/M/s+GI$ queueing system. Same arguments follows when service times are not exponentially distributed. We give the arguments in [4] here for completeness. We prove that
\begin{equation}
\|\hat{X}_n - \hat{X}_n^*\|_T \Rightarrow 0 \quad \text{and} \quad \|\hat{X}_n - (\hat{X}_n^* + \hat{X}(0)F^c_w(t))\|_T \Rightarrow 0 \quad \text{in} \quad \mathbb{D}([0, T]; \mathbb{R}), \tag{5.67}
\end{equation}
as $n \to \infty$, where $\hat{X}_n(0)$ is independent of $\hat{X}_n^*$ and $F^c_w(t)$ is given in (5.30).

Since the fluid model is OL, $\Lambda(t) > D(t)$ for all $t \in (0, T)$, and the initial net input rate into service $\lambda(t) > s'(t) + \sigma(t)$ where $\sigma(t)$ is the service-completion rate of the fluid queue, and $s'(t)$ is the rate of change in service capacity for the fluid queue. In the same vein as §5.7, there might be an initial deficit (i.e, idle servers) of magnitude $\sqrt{n}\hat{X}(0)^- = O(\sqrt{n})$, however, all servers become busy and remains so afterwards in an interval $[t_{1,n}, T]$ for $0 < t_{1,n} < T$ with $t_{1,n}$ being of order $O(1/\sqrt{n})$. However, we cannot immediately conclude that $\|X_n - X^*_n\| = O(1)$ as in §5.7. The reason is that initial deficit (or surplus) has an impact on the abandonment process
over the interval \([0, T]\).

Just as in [65], it holds that

\[
\|X_n - (X_n^* + U_n^+ - U_n^-)\|_T = O(1) \quad \text{as} \quad n \to \infty,
\]

where

\[
U_n^+(t) = (X_n(0) - s_n(0))^+ - A_{i,n,+}(t),
\]
\[
U_n^-(t) = -(X_n(0) - s_n(0))^-- A_{i,n,-}(t), \quad 0 \leq t \leq T.
\]

The process \(A_{i,n,+}(t)\) is the difference between the number of abandonments by time \(t\) and \(A_n^*(t)\) provided that the system has initial surplus \(X_n(0) - s_n(0) > 0\) whereas \(A_{i,n,-}(t)\) is the difference between the number of abandonments by time \(t\) and \(A_n^*(t)\) provided that the system has initial deficit \(X_n(0) - s_n(0) < 0\). By definition, only one of \(U_n^+\) and \(U_n^-\) can be positive at the same time.

We now prove the convergence of the LLN-scaled processes in (5.67). Since \(0 \leq (X_n(0) - s_n(0))^+ - A_{i,n,+}(t) \leq (X_n(0) - s_n(0))^+ = \sqrt{n}\hat{X}(0)^+ = O(\sqrt{n})\) and \(0 \leq -(X_n(0) - s_n(0))^-- A_{i,n,-}(t) \leq -(X_n(0) - s_n(0))^-- \sqrt{n}\hat{X}(0)^-- O(\sqrt{n})\), we deduce that \(\|\bar{X}_n - \bar{X}_n^*\|_T \Rightarrow 0\) as \(n \to \infty\) which is the desired result.

We next prove the convergence of the CLT-scaled processes in (5.67). First consider \(U_n^+(t)\). It is argued in [65] that \(U_n^+(t)\) is asymptotically equivalent to the process

\[
\tilde{U}_n^+(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{(X_n(0) - s_n(0))^+} 1(\zeta_i > t), \quad t \geq 0,
\]

where \(\{\zeta_i\}\) is a sequence of random variables each having a distribution with hazard rate \(h_F(w(u))\) at time \(u\). Then by the FCLT for the sequential empirical process, and continuous mapping theorem, we obtain

\[
\tilde{U}_n^+ \Rightarrow \hat{X}(0)^+ F^c_w(t) \quad \text{in} \quad \mathbb{D}([0, T]; \mathbb{R}) \quad \text{as} \quad n \to \infty.
\]

Treatment of \(U_n^-(t)\) is similar, consequently,

\[
\tilde{U}_n^- \Rightarrow \hat{X}(0)^- F^c_w(t) \quad \text{in} \quad \mathbb{D}([0, T]; \mathbb{R}) \quad \text{as} \quad n \to \infty.
\]

This completes the proof of (5.67).
5.9 Future Research

We next discuss the details of the remaining steps that needs to be done to have a full characterization of the FCLT for the alternating model.

5.9.1 Treating the Switching Points

Recall that the $G_t/GI/s_t + GI$ queueing system is assumed to alternate between UL and OL intervals where the system is critically loaded only at isolated regime switching points. These intervals and switching points are determined by the associated fluid queueing system which has the same parameters with the stochastic queueing system. In the alternating system, each interval $[t_i, t_{i+1}], i \geq 0$, is critically loaded at the interval endpoints $\{t_i, t_{i+1}\}$, and no critical loading is allowed in $(t_i, t_{i+1})$. The initial condition in Assumption 5.2.2 holds for the first interval $[0, t_1]$ which is a UL interval by assumption.

The FCLT in Theorem 5.5.1 is valid for UL intervals with critical loading allowed only at the left endpoint or strict underloading at the left endpoint. The FCLT in Theorem 5.5.3 is valid for OL intervals with critical loading allowed only at the left endpoint. If Assumption 5.2.2 holds at the left endpoint of each UL and OL interval, then the FCLT in Theorem 5.5.1 and Theorem 5.5.3 holds for each UL interval $[t_{2i}, t_{2i+1})$ for $i \geq 0$ and for each OL interval $[t_{2i+1}, t_{2i+2})$ for $i \geq 0$, of the alternating system. We cannot include the right endpoints because the FCLTs in the theorems in §5.5 are proved for closed intervals with critical loading only at the left endpoint. An extension of the FCLTs to intervals with critical loading at the right endpoint is left for future work.

To complete the analysis of the alternating system, we also need to prove that Assumption 5.2.2 holds not only at time 0 but at the subsequent switching points. In order for the FCLT in Theorem 5.5.1 and Theorem 5.5.3 to hold for every UL and OL interval, it must hold that

$$(\hat{N}_n, \hat{B}_n(t_i, \cdot), \hat{X}(t_i)) \Rightarrow (\hat{N}, \hat{B}(t_i, \cdot), \hat{X}(t_i)) \text{ in } D([t_i, t_{i+1}]; \mathbb{R}) \times \mathbb{R}^2 \text{ as } n \to \infty, \quad (5.68)$$

where $t_i$ and $t_{i+1}$ are switching points. However, for the alternating system, we only have joint convergence of the three processes at time 0. Therefore, we need to show that given the convergence in Assumption 5.2.2 at time 0, (5.68) holds at every switching point. Observe that $\hat{B}_n(t_i, \cdot) \Rightarrow \hat{B}(t_i, \cdot)$ in $D([0, \infty); \mathbb{R})$ as $n \to \infty$ trivially holds at every switching point when the service distribution is the exponential distribution due to the lack of memory property. However, when the service distribution is not the exponential distribution, it is more complicated to justify the convergence of the sequence $\{\hat{B}_n(\cdot, \cdot) : n \geq 1\}$ in the neighborhood of switching points. We run simulation experiments to gain better insight into the dynamics of $\{\hat{B}_n(\cdot, \cdot) : n \geq 1\}$ in the neighborhood of switching points. We conjecture that $\{\hat{B}_n(\cdot, \cdot) : n \geq 1\}$ is continuous in the
second argument for around switching points.

An FCLT for the other processes under the general initial condition requires a more careful treatment due to possible discontinuities that the limits for the other processes might have depending on the sign of \( \hat{X}(0) \). For instance, if \( \mathbb{P}(\hat{X}(0) < 0) > 0 \) and \( X(0) = s(0) \), then with positive probability, it might be the case that \( \hat{Q}(0) = \hat{X}(0)^+ = 0 \), and \( \hat{Q}(0+) > 0 \) because the system is strictly OL for all \( t > 0 \). This implies that \( \hat{Q}(0) \) is not right continuous, and hence, is not an element of \( D([0, T]; \mathbb{R}) \), i.e., no FCLT for \( \{\hat{Q}_n : n \geq 1\} \). Similarly, \( \hat{B}(0) = \hat{X}(0)^- < 0 \), and \( \hat{B}(0+) = 0 \) because all the servers are busy with probability 1 for all \( t > 0 \). Hence no FCLT for \( \{\hat{B}_n : n \geq 1\} \) in \( D([0, T]; \mathbb{R}) \). Although harder to prove, \( \hat{V}_n \) and \( \hat{W}_n \) fail to have an FCLT in \( D([0, T]; \mathbb{R}) \). In [4], the authors are able to prove an FCLT in \( D((0, T]; \mathbb{R}) \) (left endpoint excluded) for the other processes. They also prove that the other processes weakly converge in \( \mathbb{R} \) at the left endpoint 0.

Drawing on [65, 97], we conjecture the following FCLT that is an analogue of Theorem 4.3 in [65] for the remaining processes under general initial conditions.

**Conjecture 5.9.1** Suppose the assumptions in Theorem 5.5.3 hold. Then, as \( n \to \infty \),

\[
\left( \hat{B}_n, \hat{Q}_n, \hat{W}_n, \hat{V}_n, \hat{A}_n, \hat{X}_n \right) \Rightarrow \left( 0e, \hat{X}, \hat{W}, \hat{A}, \hat{X} \right) \quad \text{in} \quad D^6((0, T]; \mathbb{R}), \quad (5.69)
\]

where

\[
\hat{V}(t) = \hat{V}^*(t) + \frac{\hat{X}(0) F^c_w(t + v(t))}{D'(t + v(t))},
\]

\[
\hat{W}(t) = \hat{W}^*(t) + \frac{\hat{X}(0) F^c_w(t)}{\lambda(t - w(t)) F^c(w(t))},
\]

\[
\hat{A}(t) = \hat{N}^*(t) - \hat{D}^*(t) - \hat{X}(t) + \hat{X}(0),
\]

\( \hat{X}, \hat{V}^*, \hat{W}^*, \hat{N}^*, \hat{D}^*, \) and \( F^c_w \) are as in Theorem 5.5.3. The denominator \( D'(t + v(t)) \) in (5.70) is the derivative of the fluid service-completion process evaluated at \( t + v(t) \) where \( v \) is as in (5.13). The derivative \( D' \) is strictly positive by assumption, hence, (5.70) is well defined.

At the left endpoint 0, there exists a limit for the vector of random variables

\[
\left( \hat{X}_n(0), \hat{B}_n(0), \hat{Q}_n(0), \hat{V}_n(0) \right) \Rightarrow \left( \hat{X}(0), \hat{X}(0)^-, \hat{X}(0)^+, \frac{\hat{X}(0)^+}{D'(0)} \right). \quad (5.73)
\]

in \( \mathbb{R}^4 \) as \( n \to \infty \).
5.9.2 Simulation experiments

Finally, we will provide simulation comparisons to confirm effectiveness of the approximation formulas derived from the FCLT.
REFERENCES


[35] B. He and Y. Liu. A fluid approximation for the multi-class network of queues with

[36] B. He and Y. Liu. Staffing to stabilize the tail probability of delay in service systems with

[37] S. He. Diffusion approximation for efficiency-driven queues: A space-time scaling ap-

[38] C.-C. Huang and Y. Liu. Diffusion limit for the time-varying many-server queueing

  queues with applications to delay announcements and staffing. Submitted, 2015.


[41] D. L. Iglehart. Limit diffusion approximations for the many-server queue and the repair-


[44] O. Jennings, A. Mandelbaum, W. Massey, and W. Whitt. Server staffing to meet time-

  paper, 2014.


Appendix A

Appendix to Chapter 2

A.1 Overview

This appendix contains material supplementing Chapter 2. Appendix A.2 contains a proof of Lemma 2.5.1. Appendix A.3 contains a proof of the convergence result in (2.23). Next, in Appendix A.4, we show that the first term in (2.31) and (2.32) is asymptotically negligible using a law of large numbers for triangular arrays (see Theorem 9.1 of [22]; also see Theorem A.4.1 here). In §A.5, we provide the proof of Lemma 2.5.2.

A.2 Proof of Lemma 2.5.1

We observe that the service-completion process at each server in each queueing system whose indices are in the set $\mathcal{O}$ is asymptotically equivalent to a delayed renewal process. Therefore, the total service-completion process $D_n^{(i)}(t), i \in \mathcal{O}$, is asymptotically equivalent to superposition of delayed renewal processes. Of course it is possible that some of the queueing systems may be underloaded, i.e., some servers may be idling, for some time during overloaded intervals due to randomness. However, asymptotically ($n \to \infty$), no servers are idling any time during an overloaded period with probability 1. Mathematically speaking, if $D_n^{(i)}(t)$ is the total service-completion process associated with the overloaded queue $i \in \mathcal{O}$, and $Z_n^{(i)}(t)$ is superposition of $s_n^{(i)}(t)$ delayed renewal processes such that, in the $i$th process, time until first service completion has distribution $G_i$, and the remaining service-completion times have distribution $G_i$, then, we have

$$\|D_n^{(i)} - Z_n^{(i)}\|_{[t_1, t_2]} \Rightarrow 0 \quad \text{as} \quad n \to \infty, \quad \text{for all} \quad i \in \mathcal{O},$$

(A.1)
with probability 1, where \([t_1, t_2]\) is an interval during which none of the queues with indices in
the set \(\mathcal{O}\) switch regime. In our setting, \(G_i\) is the distribution of the remaining service time of
the customer initially in service in the \(i\)th queue, and \(G\) is the distribution of service times of
the customers who enter service after the initial customer completes service. The distribution
\(G_i\) may be different for each \(i \in \mathcal{O}\).

For each \(i \in \mathcal{O}\), we prove the tightness of the sequence of the total service-completion
processes \(\{D_n^{(i)}\}_{n \geq 1}\) by proving the tightness of the sequence \(\{Z_n^{(i)}\}_{n \geq 1}\), that is, sequence of
superposition of delayed renewal process. Once we prove the tightness of the latter, the tightness
of the former follows from (A.1) and Theorem 11.4.7. of [102]. To prove the tightness of the
sequence \(\{Z_n^{(i)}\}_{n \geq 1}\), we invoke the proof of Theorem 2 in [101]. Whitt [101] proves an FCLT for
superposition of stationary renewal processes, i.e., a renewal process with the first inter-renewal
time having density \(1 - G\) and all subsequent inter-renewal times having cdf \(G\), under a regularity
condition on the distribution \(G\). The author proves the FCLT in [101] by first proving that the
finite-dimensional distributions (fdds) of appropriately scaled superposition processes converges
distribution to multivariate Gaussian distribution, and then proving tightness in \(\mathbb{D}((0, \infty); \mathbb{R})\)
for the sequence superposition processes. Fortunately, the scaling in [101] is the same as ours,
therefore, we can directly adopt their proof of tightness for the sequence of superposition of
stationary renewal processes to prove the tightness of the sequence of superposition of delayed
renewal processes.

Consider a stationary renewal process with first inter-renewal time \(X^e\) and remaining inter-
renewal times \(X_1, X_2, \ldots\), and a delayed renewal process with first inter-renewal time \(X^d\), which
has a different distribution than \(X^e\), and the same remaining inter-renewal times \(X_1, X_2, \ldots\)
as the stationary renewal process. The only difference between the sample paths of these two
processes is time of the first renewal. The two processes have almost identical sample paths
except that there is a lag of magnitude \(|X^e - X^d|\) between the sample paths of these processes.
But, one can always “shift” to the left the sample path which has its first renewal later than the
other, so that the two sample paths coincide. Hence we deduce that the established tightness
in [101] for the CLT-scaled superposition of stationary renewal processes implies the tightness
of the CLT-scaled superposition delayed renewal processes here. The tightness of LLN-scaled
processes immediately follows and this completes the proof.

### A.3 Proof of the convergence result in (2.23).

Since the sequence \(\{\bar{E}_n^{(i)}\}\) is \(C\)-tight in \(\mathbb{D}\) by Lemma 2.5.1, every subsequence has a convergent
subsequence. Suppose we have such a convergent subsequence. We do not introduce a special
notation for the subsequence and, without loss of generality, we label that subsequence as
\(\{E_n^{(i)}(t)\}\) so that we have \(E_n^{(i)} \Rightarrow E^{(i)}\) in \(\mathbb{D}\) for all \(i \in \mathcal{O}\) where the limit \(E^{(i)}\) is yet to be
characterized. From (2.21), we see that the service-completion process of new customers has the same mathematical form as the service-completion process from an infinite-server queue with arrival process $E_n^{(i)}(t)$ and service times following cdf $G_i$. We can directly apply Theorem 3.1 of [78]. Consequently, we have

$$ D_n^{(i, \nu)}(t) \Rightarrow D^{(i, \nu)}(t) = \int_0^t G_i(t-s)b_i(s,0)ds, \quad t \in [0,T] \quad \text{for all } i \in O. \quad \text{(A.2)} $$

Combining (2.22), (A.2), (2.28) with $y = \infty$ and by applying the continuous mapping theorem with addition, we obtain weak convergence of the sequence $\{\tilde{E}_n^{(i)}(t)\}$ to an integral equation

$$ \tilde{E}_n^{(i)}(t) \Rightarrow s^{(i)}(t) - s^{(i)}(0) + B^{(i, \nu)}(0) - B^{(i, \nu)}(t) + \int_0^t G_i(t-s)b_i(s,0)ds. \quad \text{(A.3)} $$

For each $i \in O$, the derivative of (A.3) satisfies the fixed point equation (2.7), which has a unique solution (see [61]). Since the choice of the convergent subsequence is arbitrary, the derivative of the limit of every convergent subsequence of $\{\tilde{E}_n^{(i)}\}$ must satisfy (2.7). Hence, we have the full convergence of $\{\tilde{E}_n^{(i)}\}$ and $\{\tilde{D}_n^{(i, \nu)}\}$ for all $i \in O$.

### A.4 LLN for non-identically-distributed triangular arrays

We first review an LLN results for non-identically-distributed triangular arrays (e.g., see Theorem 9.1. of [22]).

Let $\{X_{k,n}\}, k = 1, \ldots, n$, be a general triangular array of random variables with cdf $F_{k,n}$. Assume that the random variables in each row of the triangular array are mutually independent. Define $S_n$ as the partial sum of $X_{k,n}$, i.e., $S_n = \sum_{k=1}^n X_{k,n}$. Also define $\tau_s(X_{k,n})$ as the truncated version of $X_{k,n}$, where $\tau_s(x) = x$ if $|x| \leq s$; $\tau_s(x) = -s$ if $x < -s$; $\tau_s(x) = s$ if $x > s$. Let $S_n^s = \sum_{k=1}^n \tau_s(X_{k,n})$. Consider the following condition: for arbitrary $\eta > 0$ and $\epsilon > 0$

$$ \mathbb{P}(|X_{k,n}| > \eta) < \epsilon, \quad k = 1, \ldots, n \quad \text{(A.4)} $$

for all $n$ sufficiently large. Now, we are ready to state the theorem.

**Theorem A.4.1** (LLN for non-i.i.d. random variables, Theorem 9.1 of [22]) If Condition (A.4) holds, then there exist constants $b_n$ such that $S_n - b_n \to 0$ in probability if and only if

$$ \sum_{i=1}^n \mathbb{P}(|X_{i,n}| > \eta) \to 0, \quad \sum_{i=1}^n \text{Var}(\tau_s(X_{i,n})) \to 0 \quad \text{as } n \to \infty \quad \text{(A.5)} $$

172
for each $\eta > 0$ and each truncation level $s$. In this case, one may take $b_n = \mathbb{E}[S^s_n]$.

We next make use of Theorem A.4.1 to prove that the first terms of (2.31) and (2.32) converge to 0. Conditioning on $\{\zeta_{n,i}(l)\}$, the first terms of (2.31), (2.32) are the LLN-scaled sum of independent non-identically-distributed zero-mean random variables with values in $[-1, 1]$. Therefore, we can use Theorem A.4.1 to prove convergence. We will later uncondition to obtain the desired result. The proof of the convergence of the first terms of (2.31) and (2.32) are similar. Therefore, we only provide a proof for the latter.

In our case, we have from (2.32)

$$X_{l,n} \equiv n^{-1} \left( \delta_{i,j}(\zeta_{n,i}^{(l)}) - P_{i,j}(\zeta_{n,i}^{(l)}) \right) \quad \text{and} \quad S_n = \sum_{l=1}^{n} X_{l,n} \quad \text{for all} \quad n \geq 1. \quad (A.6)$$

for fixed $i \in \mathcal{O}$, $j \in \{1, \ldots, m\}$. Using $\tau_s(\cdot)$, we define the truncation of $X_{l,n}$ and the partial sum of truncated variables accordingly.

Conditioning on the sequence $\{\zeta_{n,i}^{(l)}\}$, we have, by Markov inequality,

$$\mathbb{P}\{|X_{l,n}| > \eta\} \leq \frac{\mathbb{E}[|X_{l,n}|^2]}{\eta^2} = \frac{P_{i,j}(\zeta_{n,i}^{(l)})(1 - P_{i,j}(\zeta_{n,i}^{(l)}))}{n^2 \eta^2} \leq \frac{1}{n^2 \eta^2}$$

which implies that

$$\sum_{l=1}^{n} \mathbb{P}\{|X_{l,n}| > \eta\} \leq \frac{1}{n \eta^2} \to 0 \quad \text{as} \quad n \to \infty.$$

As for the second term in (A.5), we have that $\text{Var}(\tau_s(X_{l,n})) = \mathbb{E}[(\tau_s(X_{l,n}))^2]$ since $\mathbb{E}[X_{l,n}] = 0$ for $n \geq 1$, $1 \leq l \leq n$. The desired result easily follows because

$$\mathbb{E}[(\tau_s(X_{l,n}))^2] \leq \mathbb{E}[(X_{l,n})^2] \leq \frac{1}{n^2} \quad \text{for all} \quad s > 0$$

which implies

$$\sum_{l=1}^{n} \text{Var}(\tau_s(X_{l,n})) \leq \sum_{l=1}^{n} \frac{1}{n^2} = \frac{1}{n} \to 0 \quad \text{as} \quad n \to \infty.$$

for all $s > 0$. Since $\mathbb{E}[S^s_n] = 0$, we have $S_n \to 0$ in probability. More explicitly,

$$\sum_{l=1}^{n} \frac{\delta_{i,j}(\zeta_{n,i}^{(l)}) - P_{i,j}(\zeta_{n,i}^{(l)})}{n} \to 0 \quad \text{in probability} \quad \text{as} \quad n \to \infty. \quad (A.7)$$

The arguments of unconditioning follows from the arguments on p.255 of [81]. In particular, by Skorohod representation theorem, we may assume that the scaled enter-service process converges in $\mathbb{D}$ almost surely. Then we deduce that the above convergence holds whenever
the enter-service process converges almost surely in $\mathbb{D}$. Therefore, the conditional convergence is obtained by applying the generalized continuous mapping theorem, e.g., Theorem 3.4.4. of [102].

### A.5 Proof of Lemma 2.5.2

Let $\{x_n\}$ be a convergent sequence such that $d(x_n, x) \to 0$ where $d(\cdot, \cdot)$ is the Skorohod $J_1$ metric [10, 42, 102]. Then we have $\|x_n - x \circ \lambda_n\|_T \to 0$ and $\|\lambda_n - e\|_T \to 0$ where $e$ is the identity function, i.e., $e(t) = t$ for $t \geq 0$, and $\|\cdot\|$ is the uniform norm over the interval $[0, T]$. Let $M = \sup_{0 \leq t \leq T} |x(t)|$. Our goal is to show that $d(\phi(x_n), \phi(x)) \to 0$. Consider

$$
|\phi(x_n)(t) - \phi(x)(\lambda_n(t))| = \left| P_{i,j}(t)x_n(t) - \int_0^t x_n(u)dP_{i,j}(u) - P_{i,j}(\lambda_n(t))x(\lambda_n(t)) + \int_{\lambda_n(t)}^{\lambda(t)} x(u)dP_{i,j}(u) \right|
$$

$$
\leq \left| P_{i,j}(t)x_n(t) - P_{i,j}(\lambda_n(t))x(\lambda_n(t)) \right| + \left| \int_{\lambda_n(t)}^{\lambda(t)} x(u)dP_{i,j}(u) \right|
\leq \left| P_{i,j}(t)x_n(t) - P_{i,j}(\lambda_n(t))x(\lambda_n(t)) \right| + \left| \int_{\lambda_n(t)}^{\lambda(t)} x(u)dP_{i,j}(u) \right|
$$

$$
\leq \left| P_{i,j}(t)x_n(t) - P_{i,j}(\lambda_n(t))x(\lambda_n(t)) \right| + \left| \int_{\lambda_n(t)}^{\lambda(t)} x(u)dP_{i,j}(u) \right|
\leq \left| P_{i,j}(t)x_n(t) - P_{i,j}(\lambda_n(t))x(\lambda_n(t)) \right| + \left| \int_{\lambda_n(t)}^{\lambda(t)} x(u)dP_{i,j}(u) \right|
$$

The convergence of the first term follows from the convergence of $x_n \to x$ in $\mathbb{D}$. The convergence of the second and the third terms follows from the fact that $P_{i,j}(t)$ is continuous in $t$ and $\lambda_n \to e$ uniformly over the interval $[0, T]$.  

$$\begin{align*}
174
\end{align*}$$
Appendix B

Appendix to Chapter 3

B.1 Overview

This appendix has extra materials to supplement Chapter 3. In §B.2 we review useful results used in the proofs. In §B.3 we give proofs omitted in Chapter 3. In §B.4 we provide additional simulation results.

B.2 Useful Inequalities

In this section we review two useful inequalities. Both are used to prove (3.59) in §3.4.

**Theorem B.2.1 (Lévy’s inequality (symmetric case), Theorem 3.7.1 of Gut[33])** Let $X_1, X_2, \ldots, X_n$ be independent real-valued symmetric random variables (satisfying $-X_i \overset{d}{=} X_i$ for all $1 \leq i \leq n$) and let $S_n = \sum_{k=1}^{n} X_k$, $n \geq 1$ be the partial sums. Then, for any $x > 0$,

$$\mathbb{P}\left(\max_{1 \leq k \leq n} |S_k| > x\right) \leq 2\mathbb{P}(|S_n| > x). \quad (B.1)$$

**Theorem B.2.2 (Hoeffding’s inequality, Theorem 3.1.3 of Gut[33])** Let $X_1, X_2, \ldots, X_n$ be independent real-valued random variables such that $\mathbb{P}(a_k \leq X_k \leq b_k) = 1$ for $a_k, b_k \in \mathbb{R}$, $k = 1, \ldots, n$, and let $S_n = \sum_{k=1}^{n} X_k$, $n \geq 1$, denote the partial sums. Then

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| > x) \leq 2 \exp\left(-\frac{2x^2}{\sum_{k=1}^{n} (b_k - a_k)^2}\right). \quad (B.2)$$

B.3 Additional Proofs

We now provide proofs of Corollaries 3.3.4 and 3.5.1, and Lemma 3.4.2, which were omitted in Chapter 3.
Proof of Corollary 3.3.4. This follows from parts (ii) and (iii) of Theorem 3.3.2. We present the proof of the four-parameter covariance formulas in (i); the variance formulas in (i) and (ii) easily follow.

First, the covariances of $\hat{X}_1^{e,\nu}$ and $\hat{X}_2^{e,o}$ are

$$
\text{Cov}(\hat{X}_1^{e,\nu}(t_1, y_1), \hat{X}_1^{e,\nu}(t_2, y_2)) = \mathbb{E} \left[ \int_{(t_1-y_1)^+}^{t_1} G^c_{\nu}(t_1-s) \, d\tilde{N}(s) \times \int_{(t_2-y_2)^+}^{t_2} G^c_{\nu}(t_2-s) \, d\tilde{N}(s) \right] \\
= c^2 \lambda \int_{(t_1-y_1)^+ \lor (t_2-y_2)^+}^{t_1 \lor t_2} G^c_{\nu}(t_1-s)G^c_{\nu}(t_2-s) \, d\Lambda(s),
$$

and

$$
\text{Cov}(\hat{X}_2^{e,o}(t_1, y_1), \hat{X}_2^{e,o}(t_2, y_2)) = \mathbb{E} \left[ \int_{0}^{(y_1-t_1)^+} H^c_x(t_1) \, d\hat{X}_0(x) \times \int_{0}^{(y_2-t_2)^+} H^c_x(t_2) \, d\hat{X}_0(x) \right] \\
= \int_{0}^{(y_1-t_1)^+ \lor (y_2-t_2)^+} H^c_x(t_1)H^c_x(t_2) \, d\Sigma^c_{e,o}(x).
$$
Second, the covariance of $\hat{X}_2^{\nu}:

$$\text{Cov}(\hat{X}_2^{\nu}(t_1, y_1), \hat{X}_2^{\nu}(t_2, y_2)) = \mathbb{E} \left[ \left( \int_{(t_1 - y_1)^+}^{t_1} \int_0^\infty 1_{x + s > t_1} \, d\left( \hat{W}(\Lambda(s), G_\nu(x)) - G_\nu(x)\hat{W}(\Lambda(s), 1) \right) \right) \times \left( \int_{(t_2 - y_2)^+}^{t_2} \int_0^\infty 1_{x + s > t_2} \, d\left( \hat{W}(\Lambda(s), G_\nu(x)) - G_\nu(x)\hat{W}(\Lambda(s), 1) \right) \right) \right]$$

$$= \mathbb{E} \left[ \int_{(t_1 - y_1)^+}^{t_1} \int_0^\infty 1_{x + s > t_1} \, d\hat{W}(\Lambda(s), G_\nu(x)) \times \int_{(t_2 - y_2)^+}^{t_2} \int_0^\infty 1_{x + s > t_2} \, d\hat{W}(\Lambda(s), G_\nu(x)) \right] + \mathbb{E} \left[ \int_{(t_1 - y_1)^+}^{t_1} G_\nu^c(t_1 - s) \, d\hat{W}(\Lambda(s), 1) \times \int_{(t_2 - y_2)^+}^{t_2} G_\nu^c(t_2 - s) \, d\hat{W}(\Lambda(s), 1) \right]

- \mathbb{E} \left[ \int_{(t_1 - y_1)^+}^{t_1} \int_0^\infty 1_{x + s > t_1} \, d\hat{W}(\Lambda(s), G_\nu(x)) \times \int_{(t_2 - y_2)^+}^{t_2} G_\nu^c(t_2 - s) \, d\hat{W}(\Lambda(s), 1) \right]

- \mathbb{E} \left[ \int_{(t_1 - y_1)^+}^{t_1} G_\nu^c(t_1 - s) \, d\hat{W}(\Lambda(s), 1) \times \int_{(t_2 - y_2)^+}^{t_2} \int_0^\infty 1_{x + s > t_2} \, d\hat{W}(\Lambda(s), G_\nu(x)) \right]

= \int_{(t_1 - y_1)^+ \lor (t_2 - y_2)^+}^{t_1 \lor t_2} G_\nu^c(t_1 \lor t_2 - s) \, d\Lambda(s) + \int_{(t_1 - y_1)^+ \lor (t_2 - y_2)^+}^{t_1 \lor t_2} G_\nu^c(t_1 - s)G_\nu^c(t_2 - s) \, d\Lambda(s)

- 2 \int_{(t_1 - y_1)^+ \lor (t_2 - y_2)^+}^{t_1 \lor t_2} G_\nu^c(t_1 - s)G_\nu^c(t_2 - s) \, d\Lambda(s)

= \int_{(t_1 - y_1)^+ \lor (t_2 - y_2)^+}^{t_1 \lor t_2} G_\nu^c(t_1 \lor t_2 - s)G_\nu(t_1 \land t_2 - s) \, d\Lambda(s). \quad \blacksquare$$

**Proof of Corollary 3.5.1.** First, (3.68) and (3.69) easily follow from Theorems 3.1 and 3.2 of [78] by considering the performance of a system at the end of interval $[0, t_0]$ (that is, at time $t_0$) with the system being initially empty (at time 0). Then, it suffices to apply a time shift, that is, shifting the interval to the left by $t_0$ so that the interval becomes $[-t_0, 0]$.

When the system starts empty at $-t_0$, following (3.69) and Theorem 4.2 of [78], the variance function of $\hat{X}^e(0, y)$ is

$$\text{Var}(\hat{X}^e(0, y)) = \int_0^y \left[ (c_\lambda^2 - 1)G^e(s)^2 + G^e(s) \right] \lambda(-s) \, ds, \quad \text{for} \quad 0 \leq y \leq t_0. \quad \text{(B.3)}$$
Now plugging (3.68) and (B.3) into (3.36) yields that

$$\sigma^2_{\hat{X},o}(t) = \int_0^t H_u(t) H_u^c(t) G^c(u) \lambda(-u) du + \int_0^t (H_u^c(t))^2 \left[ (c^2 - 1) G^c(u)^2 + G^c(u) \right] \lambda(-u) du$$

$$= \int_0^t H_u(t) G^c(t + u) \lambda(-u) du + \int_0^t H_u^c(t) G^c(t + u) \left[ (c^2 - 1) G^c(u) + 1 \right] \lambda(-u) du$$

$$= \int_0^t G^c(t + u) \left[ (c^2 - 1) G^c(t + u) + 1 \right] \lambda(-u) du$$

$$= \int_{-t}^0 G^c(t - u) \left[ (c^2 - 1) G^c(t - u) + 1 \right] \lambda(u) du$$

where the last equality holds by a change of variable. Summing the above equation with $\sigma^2_{\hat{X},\nu}(t)$ in (3.35) yields (3.71).

Proof of Lemma 3.4.2. We mimic the proof of (A.14.15) in [95]. First for a deterministic $g \in D_D$, we have

$$\mathbb{E}\|g - X\|_{T,y}\uparrow \geq \mathbb{E}|g(t,y) - X(t,y)| \geq |g(t,y) - \mathbb{E}[X(t,y)]|, \quad \text{for } t \in [0,T], y \in [0,y\uparrow],$$

where the second inequality holds by Jensen’s inequality. Hence, we have

$$\mathbb{E}\|g - X\|_{T,y}\uparrow \geq \sup_{(t,y) \in [0,T] \times [0,y\uparrow]} |g(t,y) - \mathbb{E}[X(t,y)]| = \|g - \mathbb{E}[X]\|_{T,y\uparrow}. \quad \text{(B.4)}$$

By conditioning on $X$, we have

$$\mathbb{E}\left[ \phi \left( \|X - X^*\|_{T,y\uparrow} \right) \right] = \mathbb{E}_X \left[ \mathbb{E}_X \left[ \phi \left( \|X - X^*\|_{T,y\uparrow} \right) | X \right] \right]$$

$$\geq \mathbb{E}_X \left[ \phi \left( \mathbb{E}_X \left[ \|X - X^*\|_{T,y\uparrow} | X \right] \right) \right]$$

$$\geq \mathbb{E}_X \left[ \phi \left( \|X - \mathbb{E}[X]\|_{T,y\uparrow} \right) \right] = \mathbb{E} \left[ \phi \left( \|X - \mathbb{E}[X]\|_{T,y\uparrow} \right) \right],$$

where the first inequality holds by Jensen’s inequality and the second inequality holds by (B.4).

B.4 An Additional Example

Example B.4.1 (Simulation comparison for another example of Corollary 3.7.1) We now supplement Example 3 by considering that same $H^2_2(1,4)/LN(1,4)/\infty$ model with all parameters specified in Example 3, but with $b^* = b^*_2$ specified in (3.92). Comparisons with simulations are shown in Figure B.1.
Figure B.1: Example 4 with $b^*_2$ in (3.92): Simulation comparisons of the mean and variance for the number of customers in service of an $H^2_t(1, 4)/LN(1, 4)/\infty$ model, with the sinusoidal arrival rate (3.72) having parameters $a = c = 1$, $b = 0.6$, $\phi = 0$ and $n = 100$ and general initial conditions.
Appendix C

Appendix to Chapter 4

C.1 Proofs of Results in §5.4

C.1.1 Proof of Proposition 4.4.1

Suppose $\omega \in \Omega$ is such that $(Z(\omega; t) : t \geq 0)$ has H"older-continuous sample paths. For simplicity, we suppress $\omega$ hereafter. For $m > 0$, consider the partition $0 = u_0 < u_1 < \ldots < u_m = t$ and

$$L^{(m)}(t) = \sum_{i=0}^{m-1} J(t, u_i) (Z(u_{i+1}) - Z(u_i))$$

$$= \sum_{i=1}^{m} J(t, u_{i-1})Z(u_i) - \sum_{i=0}^{m-1} J(t, u_i)Z(u_i)$$

$$= J(t, u_{m-1})Z(u_m) - J(t, 0)Z(0) - \sum_{i=1}^{m-1} [J(t, u_i) - J(t, u_{i-1})] Z(u_i). \quad (C.1)$$

Because $Z(t)$ is continuous, the summation converges to the Riemann-Stieltjes integral as the partition mesh goes to 0 if $J(t, u)$ is monotone in the second component for each $t$. Moreover, if $J(t, u)$ is differentiable for each $t$, we can replace the integrator $dJ(t, u)$ of the Riemann-Stieltjes integral with $J_u(t, u)du$ where $J_u(t, u)$ is the derivative of $J(t, u)$ with respect to the second component. The Riemann-Stieltjes integral is well-defined if the derivative function is continuous for fixed $t$ (in general, finitely-many jumps are allowed). Therefore, for $t \geq 0$,

$$\int_0^t J(t, u) dZ(u) \equiv \lim_{m \to \infty} L^{(m)}(t) \overset{a.s.}{=} J(t, t)Z(\omega; t) - J(t, 0)Z(\omega; 0) - \int_0^t Z(\omega; u) dJ(t, u).$$
Moreover, with $\Delta \equiv \max \{ u_i - u_{i-1} : 1 \leq i \leq m \}$, we have

$$
\sum_{i=1}^{m-1} \left[ J(t, u_i) - J(t, u_{i-1}) \right] Z(u_i) - \int_0^t J_u(t, u) Z(u) \, du
= \sum_{i=1}^{m-1} \int_{u_{i-1}}^{u_i} \left[ \frac{J(t, u_i) - J(t, u_{i-1})}{u_i - u_{i-1}} (u_i - u_{i-1}) - J_u(t, u) \right] (Z(u) - Z(u_i)) \, du
\leq \frac{1}{\Delta} \cdot c_1 \Delta \cdot c_2 \Delta^\alpha \rightarrow 0 \quad \text{as} \quad \Delta \rightarrow 0
$$

where the inequality holds because $\hat{Z}$ has Hölder-continuous sample paths and $J(t, u)$ is differentiable with respect to the second component.

C.1.2 Proof of Theorem 4.4.1

We prove the assertion in two steps. First, we show in Lemma C.1.1 that if the sequence of covariance functions associated with the processes $\{ L^{(m)} : m \geq 1 \}$ converges to some function, then the sequence $\{ L^{(m)} : m \geq 1 \}$ converges in distribution to a Gaussian process. Moreover, the covariance function of the limiting Gaussian process coincides with the limit for the covariance functions associated with the sequence $\{ L^{(m)} : m \geq 1 \}$. Then, in the second step, we show that the covariance functions associated with $\{ L^{(m)} : m \geq 1 \}$ indeed converge.

Lemma C.1.1 Let $X^{(m)} \equiv \left( X_1^{(m)}, \ldots, X_l^{(m)} \right)$ be a sequence of centered Gaussian random vector in $\mathbb{R}^l$ and let $\Sigma^{(m)}$ be the covariance matrix of $X^{(m)}$. If $\Sigma^{(m)} \to \Sigma$ as $m \to \infty$, then $X^{(m)} \Rightarrow X$ where the limit $X$ is Gaussian with mean zero and covariance $\Sigma$.

Proof C.1.1 Consider the characteristic function $\phi_m(\theta) \equiv E \left[ e^{i\theta^T X^{(m)}} \right]$ of the vector $X^{(m)}$. The convergence $\Sigma^{(m)} \to \Sigma$ as $m \to \infty$ implies the convergence of the characteristic functions

$$
\phi_m(\theta) = e^{-\frac{1}{2} \theta^T \Sigma^{(m)} \theta} \to \phi(\theta) \equiv e^{-\frac{1}{2} \theta^T \Sigma \theta}
$$

due to continuity of $\phi_m$. Then the result follows from Lévy’s continuity theorem.

We next show that the covariance functions associated with the sequence $\{ L^{(m)} : m \geq 1 \}$ in (4.11) converge. We consider a partition of the interval $[0, t_2]$ such that there are a total of $m_2$ intervals partitioning $[0, t_2]$ and that $m_1$ of the intervals are partitioning $[0, t_1]$. We use the form in (C.1) to compute the covariance of $L^{(m)}(t)$. Let $C_Z(\cdot, \cdot)$ be the covariance function associated with the process $Z$. Then, for $0 \leq t_1 < t_2$ and the partition $0 = s_0 < s_1 < \ldots < s_{m_1-1} < s_{m_1} =$
Convergence of the first four terms follows from continuity of the map $u \mapsto J(t, u)$ for each fixed $t$ as $s_{m_1-1} \to t_1$ and $s_{m_2-1} \to t_2$ as $m \to \infty$. Convergence of the first four summations follows from the fact that $C_Z$ over is bounded over compact intervals, and $J(t, u)$ is bounded for each $t$ over compact intervals. Hence the limits for these terms are the Riemann-Stieltjes integrals given in (4.13). Finally, the last summation term converges to the two-dimensional Riemann-Stieltjes integral in (4.13) due to similar reasoning.

### C.2 Proof of Theorem 4.5.1

We first establish a FWLLN for $\{W_n : n \geq 1\}$ by using compactness approach, i.e., (i) we show that the sequence $\{W_n : n \geq 1\}$ is tight in $\mathbb{D}([0, T]; \mathbb{R})$, which implies that every subsequence has a convergent subsequence (see §C.2.1); (ii) every convergent subsequence converges to the same limit which in our case uniquely solves the ODE in (5.12) (see §C.2.2). Finally, in §C.2.3, we establish convergence for the other processes and characterize their limits.
Remark C.2.1 The proof provided in this section closely follows the arguments in §6.6 of [65]. We, hereby, redo the steps for the new representation of the enter-service process given in this paper. In particular, we use the decomposition of $E_n(t)$ in (4.16)-(4.19) that is different than the expressions for the enter-service process in [65]. The following proof involves some minor changes compared to the one in [65], and provided here for completeness.

C.2.1 Tightness of the sequence \{\text{W}_n\}

We prove tightness in two steps: First we show that \{W_n : n \geq 1\} is stochastically bounded in $\mathbb{D}([0, T]; \mathbb{R})$ and then show that the following criterion involving modulus of continuity is satisfied: For each $T > 0$ and $\epsilon > 0$,\n
$$\lim_{\delta \to 0} \lim_{n \to \infty} \sup_{\delta \mid 0} P(w(W_n, \delta, T) > \epsilon) = 0$$ \hspace{1cm} (C.2)

where $w(W_n, \delta, T)$ is the modulus of continuity of $W_n$, i.e., $\sup\{w(W_n, [t_1, t_2] : 0 \leq t_1 < t_2 \leq (t_1 + \delta) \wedge T)\}$ with $w(W_n, A) \equiv \sup\{W_n(s_1) - W_n(s_2) : s_1, s_2 \in A\}$.

**Stochastic boundedness.** Since we consider the system in the interval $[0, T]$ with special initial conditions, we immediately see that the HOL waiting time satisfies $0 \leq W_n(t) \leq T$ for all $n \geq 1, t \in [0, T]$. Hence we conclude that \{W_n : n \geq 1\} is stochastically bounded.

**Modulus of continuity.** The inequality $W_n(t + \delta) - W_n(t) \leq \delta$ for $\delta > 0$ and $t \geq 0$ holds because the HOL waiting time can increase at rate at most 1. Therefore, it remains to find a bound on $W_n(t) - W_n(t + \delta)$ to conclude that the modulus of continuity criterion (C.2) is satisfied.

To this end, let us define the $\delta$-increment of $\tilde{E}_{n,3}(t)$,\n
$$\tilde{E}_{n,3}(t, \delta) \equiv \tilde{E}_{n,3}(t + \delta) - \tilde{E}_{n,3}(t) = \int_{t-W_n(t)}^{t+\delta-W_n(t+\delta)} F^c(V_n(s)) \lambda(s) \, ds, \quad 0 \leq t \leq T.$$ \hspace{1cm} (C.3)

Because the complementary cdf $F^c$ is continuous and $F^c(x) > 0$ for all $x \geq 0$, it holds that $\inf_{x \in [0, T]} \{F^c(x)\} = c_1 > 0$ for any $T > 0$. Without loss of generality, we assume that $\lambda(x) > 0$ for all $x \in [t - W_n(t), t + \delta - W_n(t + \delta)]$ so that the integrand in (C.3) is bounded below by a constant $c > 0$. Then replacing the integrand with the constant $c$ yields a lower bound on $\tilde{E}_{n,3}(t, \delta)$. In particular,\n
$$W_n(t) - W_n(t + \delta) + \delta \leq \frac{\tilde{E}_{n,3}(t, \delta)}{c}, \quad 0 \leq t \leq T.$$ \hspace{1cm} (C.3)

From the FCLT in Theorem 2 of [101], we deduce that $\tilde{D}_n \Rightarrow D$ in $\mathbb{D}([0, T]; \mathbb{R})$ as $n \to \infty$, and by Assumption 4.2.4, we have $\sup_{t \in [0, T]} \{|E_n(t) - D_n(t)|\} \Rightarrow 0$ as $n \to \infty$. This implies
\( E(t) = D(t) \) for \( 0 \leq t \leq T \). Therefore, we have

\[
\lim_{n \to \infty} \{ W_n(t) - W_n(t + \delta) \} \leq \frac{D(t, \delta)}{c}, \quad 0 \leq t \leq T,
\]

which implies the modulus of continuity condition (C.2). Hence, \( \{ W_n : n \geq 1 \} \) is tight in \( \mathbb{D}([0,T]; \mathbb{R}) \). More specifically, \( \{ W_n : n \geq 1 \} \) is C-tight because (C.2) is a sufficient condition for C-tightness together with stochastic boundedness.

**C.2.2 Characterizing the limit of the sequence \( \{ W_n \} \).**

Due to C-tightness, we know that (i) every subsequence of \( \{ W_n : n \geq 1 \} \) has a convergent subsequence. Let \( \{ W_{n_k} : k \geq 1 \} \) be such a convergent subsequence with the limit \( w^* \), i.e., \( W_{n_k} \Rightarrow w^* \) in \( \mathbb{D}([0,T]; \mathbb{R}) \) as \( n \to \infty \). From (5.1), (5.2), and convergence of the subsequence \( \{ W_{n_k} : k \geq 1 \} \), we deduce that there exists a corresponding subsequence \( \{ V_{n_k} : k \geq 1 \} \) that converges to some \( v^* \) satisfying the fluid equations in (4.9), i.e.,

\[
v^*(t) = w^*(t + v^*(t)) \quad \text{and} \quad v^*(t - w^*(t)) = w^*(t), \quad 0 \leq t \leq T. \quad \text{(C.4)}
\]

Next we derive an ordinary differential equation (ODE) for \( w^* \) using the scaled enter-service process \( \bar{E}_n(t) : 0 \leq t \leq T \). First, from the FCLT in Theorem 2 of [101], we have \( \bar{D}_n \Rightarrow D \) in \( \mathbb{D}([0,T]; \mathbb{R}) \) with \( D(t) = \mu t \). This, together with Assumption 4.2.4, implies \( \sup_{t \in [0,T]} \{ |E_n(t) - D_n(t)| \} \Rightarrow 0 \) as \( n \to \infty \), and hence,

\[
\bar{E}_n(t) \Rightarrow E(t) = \int_0^t b(u,0) \, du = \mu t \quad \text{in} \quad \mathbb{D}([0,T]; \mathbb{R}) \quad \text{as} \quad n \to \infty \quad \text{(C.5)}
\]

where the right-hand side coincides with (4.6).

On the other hand, from (4.16) and that \( (W_{n_k}, V_{n_k}) \Rightarrow (w^*, v^*) \) in \( \mathbb{D}^2([0,T]; \mathbb{R}) \), the sequence \( \{ E_n : n \geq 1 \} \) along the subsequence associated with \( \{ W_{n_k} \} \) and \( \{ V_{n_k} \} \) converges to a limit \( E^* \) satisfying

\[
\bar{E}_n(t) \Rightarrow E^*(t) = E_n^*(t) \equiv \int_0^{t-w^*(t)} F_c(v^*(s)) \lambda(s) \, ds \quad \text{in} \quad \mathbb{D}([0,T]; \mathbb{R}) \quad \text{(C.6)}
\]

with

\[
\bar{E}_{n,1}^*(t) \equiv \bar{E}_{n,1}^*(t) \Rightarrow (0e)(t) \quad \text{and} \quad \bar{E}_{n,2}^*(t) \equiv \bar{E}_{n,2}^*(t) \Rightarrow (0e)(t) \quad \text{in} \quad \mathbb{D}([0,T]; \mathbb{R})
\]

as \( n \to \infty \). From [78], we know that LLN-scaled version of (4.17) and (4.18) vanish in the limit if \( W_n \) and \( V_n \) are replaced by the deterministic functions \( w \) and \( v \), respectively. Therefore, we
conclude that \( \{ \bar{E}_{n,1}^* \} \) and \( \{ \bar{E}_{n,2}^* \} \) converge to zero because the limits \( w^* \) and \( v^* \) in (C.6) are deterministic.

We now derive an ODE for the limit \( w^* \). Equating (C.5) and (C.6), and taking the derivative of both sides yield

\[
    b(t, 0) = (1 - \dot{w}(t)) F^c(w(t)) \lambda(t - w(t)) = (1 - \dot{w}(t)) F^c(w(t)) \lambda(t - w(t)), \quad (C.7)
\]

for \( 0 \leq t \leq T \) where the second equality holds by (C.4). This implies that

\[
    \frac{d}{dt} w^*(t) = 1 - b(t, 0) F^c(w(t)) \lambda(t - w(t)) = 1 - b(t, 0) q(t, w(t)), \quad 0 \leq t \leq T,
\]

which coincides with the ODE (5.12) that has a unique solution. This implies that any convergent subsequence \( \{ W^*_n \} \) must converge to the same limit. Hence full convergence of \( \{ W_n : n \geq 1 \} \).

**C.2.3 FWLLN of the other processes.**

First, we prove full convergence of \( \{ V_n : n \geq 1 \} \). In particular, for \( 0 \leq t \leq T \),

\[
    |V_n(t - W_n(t)) - v(t - w(t))| \leq |V_n(t - W_n(t)) - V_n(t - w(t))| + |V_n(t - w(t)) - v(t - w(t))| + |V_n(t - w(t)) - v(t - w(t))| \\
    = |W_n(t) - w(t) + O(1/n)| + |w(t) + O(1/n) - w(t)| \\
    \leq |W_n(t) - w(t)| + O(1/n), \quad (C.8)
\]

where the equality follows from (5.1) and (4.9). This implies convergence of \( V_n \Rightarrow v \) in \( D([0, T]; \mathbb{R}) \) as \( n \to \infty \) due to \( W_n \Rightarrow w \) in \( D([0, T]; \mathbb{R}) \) as \( n \to \infty \) in §C.2.2.

We further apply change of variable to (C.8) with \( u_n \equiv t - W_n(t) \) and \( u \equiv t - w(t) \) to obtain

\[
    \|V_n - v\| \leq \frac{\|W_n - w\|}{\gamma} + O(1/n) = O(1/n) \quad (C.9)
\]

for a constant \( \gamma > 0 \), where the equality holds because \( u_n = u + o(1) \). We will make use of (C.9) establishing an FCLT limit for \( \{ \bar{V}_n : n \geq 1 \} \) in §C.4.

We readily have the limit for the enter-service process \( \{ \bar{E}_n : n \geq 1 \} \) along the convergent subsequence in (C.6). Full convergence of enter-service follows from full convergence of \( \{ W_n \} \) and \( \{ V_n \} \) established above. Therefore, we can now drop the superscripts in (C.6). The limit of the sequence of service-completion processes must coincide with the limit of the sequence of enter-service processes because, in the limit, the system will be overloaded over the entire interval \([0, T]\).

The limit of the sequences of processes (5.25)-(5.27) can be obtained the same way it has
been done in [65] in which the authors make use of Theorem 3.1 of [78], and then apply continuous mapping theorem given $W_n \Rightarrow w$ in $D([0,T];\mathbb{R})$ as $n \to \infty$. From (6.17) of [65], we immediately write

\[ \bar{Q}_{n,i} \Rightarrow 0 \quad \text{for} \quad i = 1, 2; \quad \bar{Q}_{n,3} \Rightarrow Q_3(t) = \int_{t-w(t)}^{t} F^c(t-s) \lambda(s) \, ds \quad (C.10) \]

in $D([0,T];\mathbb{R})$ as $n \to \infty$.

C.3 Proofs of Results in §4.6

C.3.1 Proof of Theorem 4.6.2

The expressions for $C_{\tilde{W}_1}(t,t')$ and $C_{\tilde{W}_2}(t,t')$ are obtained by applying rules of Ito integral. Derivation of these functions follows from standard arguments and therefore the details are omitted.

To compute $C_{\tilde{W}_3}(t,t')$ we make use of (4.13) with $J(t,u) = \frac{H(t,u)}{q(u,w(u))}$ where $H(t,u)$ and $q(u,w(u))$ are as in (5.50). In particular, for $0 \leq t < t'$,

\[
C_{\tilde{W}_3}(t,t') = J(t,t)J(t',t')C_{\tilde{E}}(t,t') - \int_{0}^{t'} J(t,t)J_u(t',u)C_{\tilde{E}}(t,u) du - \int_{0}^{t} J(t',t')J_u(t,u)C_{\tilde{E}}(t',u) du + \int_{0}^{t} \int_{0}^{t'} J_u(t,u)J_v(t',v) C_{\tilde{E}}(u,v) dv du
\]

\[
= \frac{1}{\lambda^2 F^c(w(t))F^c(w'(t'))} C_{\tilde{E}}(t,t') - \frac{1}{\lambda F^c(w(t))} \int_{0}^{t'} J_u(t',u)C_{\tilde{E}}(t,u) du - \frac{1}{\lambda F^c(w(t))} \int_{0}^{t} J_u(t,u)C_{\tilde{E}}(t',u) du + \int_{0}^{t} \int_{0}^{t'} J_u(t,u)J_v(t',v) C_{\tilde{E}}(u,v) dv du
\]

where $J_u(t,u)$ is as in (4.43).

We next derive the covariance function for the limit queue-length process. First, $C_{\tilde{Q}_1}(t,t')$ can be obtained from isometry property of Ito integral. The function $C_{\tilde{Q}_2}(t,t')$ can be obtained by $\tilde{U}^\nu(\lambda s, y) = W(\lambda s, y) - yW(\lambda s, 1)$ where $W(\cdot, \cdot)$ is the standard two-dimensional Brownian motion (see preliminaries in Chapter 1). The last term easily follows by definition.

C.3.2 Proof of Theorem 4.6.3

To prove Theorem 4.6.3, we first prove Lemma 4.6.3.

Proof of Lemma 4.6.3. First, we prove the existence of $\tilde{E}(t)$. It suffices to show that for any $n \geq 1$ and $-t_1 < -t_2 < \ldots < -t_n \leq 0$, the matrix $M = (\tilde{C}(-t_i, -t_j))_{i,j=1}^{n}$ is nonnegative
definite. Let \( r_1 = t_1 \) and \( r_j = t_{j-1} - t_j \) for \( j = 2, \ldots, n \) and define \( N = (C_E(r_i, r_j))_{i,j=1}^n \). For any \( z = (z_1, z_2, \ldots, z_n)^T \in \mathbb{R}^n \), define \( y = (y_1, y_2, \ldots, y_n)^T \) such that \( y_1 = \sum_{i=1}^n z_i \) and \( y_j = -\sum_{i=j}^n z_i \) for \( j = 2, \ldots, n \). Given that \( \tilde{C}(t, s) = C_E(-t, -t) - C_E(-t, s - t) \), we can compute

\[
\begin{align*}
    z^T M z &= \sum_{i=1}^n \tilde{C}(-t_i, -t_i) z_i^2 + 2 \sum_{1 \leq i < j \leq n} \tilde{C}(-t_i, -t_j) z_i z_j = y^T N y.
\end{align*}
\]

We shall explain how to derive the above equation for \( n = 2 \).

\[
\begin{align*}
    z_1^2 \tilde{C}(-t_1, -t_1) + 2 z_1 z_2 \tilde{C}(-t_1, -t_2) + z_2^2 \tilde{C}(-t_2, -t_2) \\
    &= z_1^2 C_E(r_1, r_1) + 2 z_1 z_2 (C_E(r_1, r_1) - C_E(r_1, r_2)) + z_2^2 C_E(r_1 - r_2, r_1 - r_2) \\
    &= z_1^2 C_E(r_1, r_1) + 2 z_1 z_2 (C_E(r_1, r_1) - C_E(r_1, r_2)) + z_2^2 (C_E(r_1, r_1) - 2 C_E(r_1, r_2) + C_E(r_2, r_2)) \\
    &= (z_1 + z_2)^2 C_E(r_1, r_1) - 2 z_2 (z_1 + z_2) C_E(r_1, r_2) + z_2^2 C_E(r_2, r_2) = y^T N y.
\end{align*}
\]

Since \( C_E(\cdot, \cdot) \) is the covariance function of a Gaussian process, the matrix \( N \) is nonnegative definite and hence \( y^T N y \geq 0 \). As the vector \( z \) is any vector in \( \mathbb{R}^n \), we can conclude that \( M \) is also nonnegative definite and the existence of \( \tilde{E} \) follows. The argument is similar for \( n \geq 3 \), therefore, the details are omitted.

Next we show that (4.47) holds. Since a Gaussian process is fully characterized by its covariance function, it suffices to show that for any fixed \( t > 0 \) and \( 0 \leq r < s \leq t \),

\[
\text{Cov}(\tilde{E}(-t + r) - \tilde{E}(-t), \tilde{E}(-t + s) - \tilde{E}(-t)) = C_E(r, s).
\]

By our definition of \( \tilde{C}(t, s) \), we can compute

\[
\begin{align*}
    \text{Cov}(\tilde{E}(-t + r) - \tilde{E}(-t), \tilde{E}(-t + s) - \tilde{E}(-t)) \\
    &= C_E(t - r, t - r) - C_E(t - r, s - r) + C_E(t, s) + C_E(t, r) - C_E(t, t). \quad \text{(C.11)}
\end{align*}
\]

By the stationary increments of \( \tilde{E} \), we have

\[
\begin{align*}
    C_E(t - r, t - r) &= \text{Var}(\tilde{E}(t - r)) = \text{Var}(\tilde{E}(t) - \tilde{E}(r)) = C_E(t, t) - 2 C_E(t, r) + C_E(r, r), \\
    C_E(t - r, s - r) &= \text{Cov}(\tilde{E}(t - r), \tilde{E}(s - r)) = \text{Cov}(\tilde{E}(t) - \tilde{E}(r), \tilde{E}(s) - \tilde{E}(r)) \\
    &= C_E(t, s) - C_E(t, r) - C_E(r, s) + C_E(r, r),
\end{align*}
\]

which along with (C.11) implies that

\[
\text{Cov}(\tilde{E}(-t + r) - \tilde{E}(-t), \tilde{E}(-t + s) - \tilde{E}(-t)) = C_E(r, s).
\]

187
This completes the proof of the lemma.

**Back to proof of Theorem 4.6.3.** We characterize the steady-state limits for the processes \( \hat{W}_i \) for \( i = 1, 2, 3 \) in (5.49). To do so, we let \( t^* \geq 0 \) be a (sufficiently large) time instant where the performance of the associated fluid queue is very close to (fluid) steady-state behavior. As \( t^* \to \infty \), steady-state behavior becomes even more evident. We are interested in the behavior of the processes \( \hat{W}_i(t) : t > t^* \) for \( i = 1, 2, 3 \); and as \( t - t^* \to \infty \). In our treatment below, we assume \( t^* = 0 \). This does not cause any loss of generality because analysis on the interval \([0, t^*]\) as \( t^* \to \infty \) and that on \([t^*, t]\) as \( t - t^* \to \infty \) are identical. By letting \( t^* = 0 \) and \( t \to \infty \), we avoid repetition of arguments and we circumvent the treatment of extra terms that are negligible as \( t^* \to \infty \).

**Steady-state of \( \hat{W} \).** Let \( \mathcal{N}(0, \sigma^2) \) denote the normal random variable with mean 0 and variance \( \sigma^2 \). First, we treat \( \hat{W}_1(t) \) in (5.49) by applying a change of variable with \( u = s + v(s) \).

As a result, we obtain

\[
\hat{W}_1(t) = \int_{\kappa(0)}^{\kappa(t)} \frac{F^c(w(s + v(s)))H(t, s + v(s))}{\lambda F^c(w(t))} d\mathcal{N}(s + v(s) - w(s + v(s)))
\]

where \( \kappa(t) \) is the inverse function of the map \( s + v(s) \) evaluated at \( s = t \), i.e.,

\[
\kappa(t) = \inf\{s \geq 0 : s + v(s) > t\};
\]

and the second equality follows from (4.9).

Given the definitions in (4.44)–(4.45), the convention \( (t^* = 0) \) discussed above, and that the FCLT limit for the arrival process satisfies \( \mathcal{N}(t) \equiv \sqrt{\lambda c_\lambda B_\lambda(t)}, t \geq 0 \), (C.12) can be represented as

\[
\hat{W}_1(t) = \int_{\kappa(0)}^{\kappa(t)} \frac{c_\lambda}{\sqrt{\lambda}} e^{-h_F(w)(t-s-v(s))} dB_\lambda(s) + o(1)
\]

\[
\equiv \bar{B}_\lambda \left( \frac{c_\lambda^2}{\lambda} \int_{\kappa(0)}^{\kappa(t)} e^{-2h_F(w)(t-s-v(s))} ds \right) + o(1)
\]

\[
\equiv \bar{B}_\lambda \left( \frac{c_\lambda^2}{\lambda} \int_0^t e^{-2h_F(w)(t-s)} d\kappa(s) \right) + o(1)
\]

\[
\equiv \bar{B}_\lambda \left( \frac{c_\lambda^2}{2h_F(w)\lambda} \left( 1 - e^{-2h_F(w)t} \right) \right) + o(1)
\]

where \( \bar{B}_\lambda \) is a standard Brownian motion. The \( o(1) \) term in the first equality is obtained by replacing the integrand with the steady-state counterparts, and by using the fact that the inte-
and ˜

From (4.44)-(4.45) and (5.49), we have

\[ W_1(t) \Rightarrow W_1(\infty) \overset{d}{=} \mathcal{N}(0, \frac{c^2_{\lambda}}{2h_F(w)\lambda}). \]

Similarly, an application of Theorem 3.4.6 of [47] yields

\[ W_2(t) = \int_0^t \frac{\sqrt{F(w)}}{\sqrt{\lambda F_c(w)}} e^{-h_F(w)(t-u)} dB_a(u) + o(1) \overset{d}{=} \mathcal{B}_a \left( \frac{F(w)}{2\lambda f(w)} \left( 1 - e^{-2th_F(w)} \right) \right) \]

where \( \mathcal{B}_a \) is a standard Brownian motion. Hence, as \( t \to \infty \), we obtain

\[ W_2(t) \Rightarrow W_2(\infty) \overset{d}{=} \mathcal{N}(0, \frac{F(w)}{2\lambda f(w)}). \]

From (4.44)-(4.45) and (5.49), we have

\[
\Var(W_3(t)) = \frac{1}{\lambda^2 F_c(w)^2} \Var \left( \int_0^t e^{-h_F(w)(t-u)} d\tilde{E}(u) \right) \\
= \frac{1}{\lambda^2 F_c(w)^2} \Var \left( -e^{-h_F(w)t} \tilde{E}(-t) - \int_0^t h_F(w)e^{-h_F(w)s} \tilde{E}(-s) ds \right) \\
= \frac{1}{\lambda^2 F_c(w)^2} \left[ \int_0^t e^{-2h_F(w)} \tilde{C}(-t, -t) + 2h_F(w)e^{-h_F(w)t} \int_0^t e^{-h_F(w)s} \tilde{C}(-s, -t) ds \right] \\
+ 2h_F(w)^2 \int_0^t \int_0^x e^{-h_F(w)(x+y)} \tilde{C}(-x, -y) dy dx
\]

where the second equality follows from (4.47). Note that \( \tilde{C}(-t, -s) = \Cov(\tilde{E}(-s), \tilde{E}(-t)) \leq \sqrt{\Var(\tilde{E}(t)) \Var(\tilde{E}(s))} = \sqrt{\Var(\tilde{E}(t)) \Var(\tilde{E}(s))} \) as \( t \to \infty \). As \( \Var(\tilde{E}(t)) = O(t^2) \) as \( t \to \infty \), we can conclude that \( \tilde{C}(-t, -s) = O(st) \) as \( s, t \to \infty \). As a result, the first term in (C.13) is \( O(e^{-th_F(w)t^2}) \to 0 \) and the second term is \( O(e^{-th_F(w)t^2}) \to 0 \) as \( t \to \infty \). Besides the last double integral as converges as \( t \to \infty \) by definition. Hence we conclude that

\[ W_3(t) \Rightarrow W_3(\infty) \overset{d}{=} \mathcal{N}(0, \sigma_{W_3}^2) \quad \text{as} \quad t \to \infty \]

where

\[
\sigma_{W_3}^2 \equiv \frac{2}{\lambda^2 F_c(w)^2} \int_0^\infty \int_0^x e^{-h_F(w)(x+y)} \tilde{C}(-x, -y) dy dx,
\]

and \( \tilde{C}(\cdot, \cdot) \) is as defined in Lemma 4.6.3.
Finally, by independence, we conclude that
\[ \hat{W}(t) \Rightarrow \hat{W}(\infty) \equiv \hat{W}_1(\infty) + \hat{W}_2(\infty) + \hat{W}_3(\infty) \overset{d}{=} \mathcal{N}(0, \sigma^2_W), \quad \text{as} \quad t \to \infty, \]
where \( \sigma^2_W \equiv \frac{c^2}{2h_F(w)\lambda} + \frac{F(w)}{2\lambda f(w)} + \sigma^2_{W_3}. \)

**Steady-state of \( \hat{Q} \).** We next characterize the steady state for the queue-length process. We continue to use the convention \( t^* = 0 \). Then, for \( t > w \),
\[
\hat{Q}_1(t) = c_\lambda \int_{t-w}^t F^c(t-s) dB_\lambda(\lambda s) + o(1)
\]
\[
= c_\lambda \sqrt{\lambda} \int_w^w F^c(w-s) dB_\lambda(t-w+s) + o(1)
\]
\[
\overset{d}{=} c_\lambda \sqrt{\lambda} \int_w^w F^c(w-s) dB_\lambda(s) + o(1), \quad (C.15)
\]
where the \( o(1) \) term is obtained by replacing \( w(t) \) with \( w \). This implies that
\[
\hat{Q}_1(t) \Rightarrow \hat{Q}_1(\infty) \overset{d}{=} \mathcal{N}(0, \sigma^2_{Q_1}) \quad \text{as} \quad t \to \infty \quad \text{where} \quad \sigma^2_{Q_1} \equiv \lambda c^2 \int_0^w F^c(u)^2 du.
\]

**Remark C.3.1 (Extension to general Gaussian processes)** Note that the distributional equality in (C.15) follows from the fact that Brownian motion has stationary increments. Specifically, we make use of the fact that \( c = t - w > 0 \) and that the increments of Brownian motion satisfy \( B_\lambda(s_2 + c) - B_\lambda(s_1 + c) \overset{d}{=} B_\lambda(s_2) - B_\lambda(s_1) \) with \( s_1 < s_2 \) for any \( c \geq 0 \). We know that \( \hat{Q}_1 \) is characterized as integration-by-parts similar to the limit in Lemma 4.7.1 when \( \hat{N} \) is a more general process having the properties in Assumption 4.2.2. Therefore, we deduce that the same change of variable argument in (C.15) holds for general \( \hat{N} \) with continuous paths, and independent and stationary increments. However, assuming non-Gaussian \( \hat{N} \) reduces tractability of the steady state. Consideration of the special case \( \hat{N} \equiv c_\lambda \sqrt{\lambda} B_\lambda \) is to obtain more tractable results.

For each fixed \( t \geq 0 \), the double integral in (4.32) can be represented as an Ito integral. For \( t > w \), (4.44)-(4.45) apply, and consequently, (4.32) reduces to
\[
\hat{Q}_2(t) \overset{d}{=} - \int_{t-w}^t \sqrt{F(t-s)F^c(t-s)} d\tilde{B}_\alpha(\lambda s) + o(1)
\]
\[
\overset{d}{=} - \int_{w}^w \sqrt{\lambda F(w-s)F^c(w-s)} d\tilde{B}_\alpha(s) + o(1)
\]
where \( \tilde{B}_\alpha \) is a standard Brownian motion. The second equality follows from a similar argument.
becomes centered Gaussian process with covariance function \( C \) with constant rate distribution is exponential. Consequently, service completions at each server is a Poisson process. Remaining service times are exponentially distributed due to lack of memory if the service-time distribution is exponential.

### C.3.3 Proof of Corollary 4.6.2

Remaining service times are exponentially distributed due to lack of memory if the service-time distribution is exponential. Consequently, service completions at each server is a Poisson process with constant rate \( \mu > 0 \) which implies by [101] that the sequence \( \{ \tilde{E}_n : n \geq 1 \} \) converges to a centered Gaussian process with covariance function \( C_{\tilde{E}}(s, t) = \mu(s \wedge t) \) for \( s, t \geq 0 \). Then (4.46) becomes \( \tilde{C}(-x, -y) = \mu(x \lor y) - \mu |x - y| \) for \( x \geq 0, y \geq 0 \). Consequently, (C.14) becomes

\[
\sigma^2_{W_3} = 2 \frac{h_F(w)^2}{\lambda^2 F(w)^2} \int_0^\infty \int_0^x e^{-h_F(w)(x+y)} \mu y dy dx
\]

Finally, from (4.32) and (4.44)-(4.45), we immediately have

\[
\hat{Q}_3(t) = \lambda F^*(w) \hat{W}(t) \Rightarrow \hat{Q}_3(\infty) \equiv \lambda F^*(w) \hat{W}(\infty) \overset{d}{=} \mathcal{N}(0, \sigma^2_{Q_3}), \quad \text{as} \ t \to \infty,
\]

where \( \sigma^2_{Q_3} = \lambda^2 F^*(w)^2 \sigma^2_W \).

In conclusion, because the three limits \( \hat{Q}_1, \hat{Q}_2 \) and \( \hat{Q}_3 \) are independent, we have

\[
\hat{Q}(t) \Rightarrow \hat{Q}(\infty) \overset{d}{=} \mathcal{N}(0, \sigma^2_{Q}), \quad \text{where} \ \sigma^2_Q \equiv \sigma^2_{Q_1} + \sigma^2_{Q_2} + \sigma^2_{Q_3}.
\]
and \( h_F(w) = \theta \) in (4.50). Finally we obtain \( \sigma^2_Q \) in (4.51) as follows:

\[
\sigma^2_Q(\infty) = \lambda \int_0^w F^c(u)^2 \, du + \lambda \int_0^w F(u)F^c(u)^2 \, du + \lambda^2 F^c(w)^2 \sigma^2_W
\]

\[
= \lambda \int_0^w F^c(u) \, du + \lambda^2 F^c(w)^2 \sigma^2_W
\]

\[
= \frac{\lambda}{\theta} (1 - e^{-\theta w}) + \frac{\lambda}{\theta} \cdot \frac{1}{\rho} = \frac{\lambda}{\theta}
\]

which is the desired result.

### C.4 Proofs of Results in §4.7

#### C.4.1 Proof of Lemma 4.7.1

We consider the modified processes \( n^{-1/2}E_{n,1}'(t) \) given below. We first prove convergence for the sequence \( \{n^{-1/2}E_{n,1}'(t) : n \geq 1\} \), and then show that the difference between the modified sequences \( \{n^{-1/2}E_{n,1}' : n \geq 1\} \) and the original sequence \( \{n^{-1/2}E_{n,1}(t) : n \geq 1\} \) is asymptotically negligible (see (C.18)) which proves the desired convergence in (4.55).

Now consider, for \( 0 \leq t \leq T \),

\[
\frac{1}{\sqrt{n}} E_{n,1}'(t) = \int_0^{t-w(t)} F^c(v(s)) \, d\hat{N}_n(s)
\]

\[
= F^c(v(t-w(t)))\hat{N}_n(t-w(t)) - F^c(v(0))\hat{N}_n(0) - \int_0^{t-w(t)} \hat{N}_n(s) \, dF^c(v(s))
\]

\[
= F^c(w(t))\hat{N}_n(t-w(t)) - \hat{N}_n(0) - \int_0^t \hat{N}_n(s-w(s)) \, dF^c(w(s)). \quad (C.16)
\]

The second equality holds since \( \hat{N}_n(s) \) is of bounded variation, and therefore, the integral can be represented as the form after integration by parts for almost all \( \omega \in \Omega \). The last equality follows from (4.9), i.e., \( v(t-w(t)) = w(t) \), for \( 0 \leq t \leq T \), and that \( v(0) = w(0) = 0 \) as a consequence of Assumption 4.2.3. Next we define a mapping \( \psi : D([0,T]; \mathbb{R}) \rightarrow D([0,T]; \mathbb{R}) \) such that for \( z \in D([0,T]; \mathbb{R}) \),

\[
\psi(z)(t) = F^c(w(t))z(t) - z(0) - \int_0^t z(s) \, dF^c(w(s)), \quad 0 \leq t \leq T.
\]

We now prove that the mapping \( \psi \) is continuous in \( D([0,T]; \mathbb{R}) \). Let \( \{x_n\} \) be a sequence in
\( \mathbb{D}([0, T]; \mathbb{R}) \) such that \( \|x_n - x\|_T \to 0 \). Then

\[
|\psi(x_n)(t) - \psi(x)(t)| = \left| F^c(w(t))x_n(t) - x_n(0) - \int_0^t x_n(s) \, dF^c(w(s)) - F^c(w(t))x(t) + x(0) + \int_0^t x(s) \, dF^c(w(s)) \right| \\
\leq F^c(w(t))|x_n(t) - x(t)| + |x_n(0) - x(0)| + \|x_n - x\|_T \int_0^t dF^c(w(s)) \leq 4\|x_n - x\|_T.
\]

Hence the mapping \( \psi \) is continuous. In general, proving convergence with respect to the uniform topology does not necessarily imply \( J_1 \) convergence because there may be measurability issues (see e.g. [102, 10]). However, we will be interested in the case where the limit \( x \) is continuous, i.e., \( x \in \mathcal{C}([0, T]; \mathbb{R}) \). Therefore, we will not have any measurability issues and obtain the desired convergence in \( \mathbb{D}([0, T]; \mathbb{R}) \) with respect to Skorokhod’s \( J_1 \) metric.

Convergence of the modified processes in (C.16) follows by continuous mapping theorem. In particular, let \( Z_n(\cdot) \equiv \hat{\Lambda}_n(\cdot - W_n(\cdot)) \). Then \( Z_n : [0, T] \to \mathbb{R} \) and \( Z_n \Rightarrow Z \) in \( \mathbb{D}([0, T]; \mathbb{R}) \) as \( n \to \infty \) where \( Z(\cdot) \equiv \hat{\Lambda}(\cdot - w(\cdot)) \). Convergence of \( \{Z_n\} \) follows from continuous mapping theorem with composition. In particular, we apply Theorem 13.2.2 of [102] with the sequences \( \{\cdot - W_n(\cdot)\} \) and \( \{\hat{\Lambda}_n\} \) converging to the continuous limits \( \cdot - w(\cdot) \) and \( \hat{\Lambda} \), respectively. Then we have \( n^{-1/2} \hat{E}_{n,1}(\cdot) = \psi(Z_n)(\cdot) \Rightarrow \psi(Z)(\cdot) \) in \( \mathbb{D}([0, T]; \mathbb{R}) \) as \( n \to \infty \). We denote the limit \( \psi(Z) \) as

\[
\int_0^t F^c(w(s)) \, d\hat{N}(s - w(s)) \equiv F^c(w(t))\hat{N}(t - w(t)) - \hat{N}(0) - \int_0^t \hat{N}(s - w(s)) \, dF^c(w(s)).
\]

where the integral on the right-hand side is defined for almost all \( \omega \in \Omega \).

Finally, to establish (4.55), we show that the difference between the processes \( n^{-1/2}E_{n,1}(t) \) and \( n^{-1/2}E'_{n,1}(t) \) is asymptotically negligible as \( n \to \infty \). In particular,

\[
\frac{1}{\sqrt{n}} \left| E_{n,1}(t) - E'_{n,1}(t) \right| \tag{C.17}
\]

\[
= \frac{1}{\sqrt{n}} \left| \int_0^{t-W_n(t)} F^c(V_n(s)) \, d\hat{N}_n(s) - \int_0^{t-w(t)} F^c(v(s)) \, d\hat{N}_n(s) \right| \\
\leq \frac{1}{\sqrt{n}} \left| \int_0^{t-W_n(t)} F^c(V_n(s)) \, d\hat{N}_n(s) \right| + \frac{1}{\sqrt{n}} \int_0^{t-w(t)} F^c(V_n(s)) - F^c(v(s)) \, d\hat{N}_n(s) \\
\leq \frac{1}{\sqrt{n}} \left| \hat{N}_n(t - W_n(t)) - \hat{N}_n(t - w(t)) \right| + \frac{1}{\sqrt{n}} \left| \hat{N}_n(t - w(t)) - \hat{N}_n(0) \right|, \quad 0 \leq t \leq T, \tag{C.18}
\]

which converges to 0 due to the convergence \( (\hat{N}_n, W_n) \Rightarrow (\hat{N}, w) \) in \( \mathbb{D}([0, T]; \mathbb{R}) \) as \( n \to \infty \) to continuous limits, and by continuous mapping theorem with composition, i.e., \( \hat{N}_n(\cdot - W_n(\cdot)) \Rightarrow \hat{N}(\cdot - w(\cdot)) \Rightarrow \hat{N}(\cdot) \).

193
\[ \hat{N}(\cdot - w(\cdot)) \] in \( \mathbb{D}([0, T]; \mathbb{R}) \) as \( n \to \infty \). This completes the proof.

### C.4.2 Proof of Lemma 4.7.2

To prove convergence in (4.56), we apply the martingale FCLT in [81] (Also see [16, 39] for applications of the martingale FCLT). First we define a sequence of discrete-time processes (see (C.19)) and argue that they are martingales adapted to a specific filtration \( \mathcal{H}_k^n \) as defined below.

Next, we define continuous-time martingales using the discrete-time martingales in (C.19). Then we invoke Theorem 7.1.4. on p.339 in [20] to establish convergence, and characterize the limit.

Consider the discrete-time processes

\[
\hat{H}_k^n \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{k} (1(\gamma_i^n > w_i^n) - F^c(w_i^n)) \quad \text{for} \quad k = 1, 2, \ldots
\]  

(C.19)

Also, consider the filtration \( \mathcal{H}_k^n \equiv \sigma \{ \tau_{i+1}^n, \nu_i^n, \gamma_i^n : 1 \leq i \leq k \} \). Then \( \mathbb{E}[\hat{H}_k^n] \leq k/\sqrt{n} \) and

\[
\mathbb{E}[H_k^n - H_{k-1}^n | \mathcal{H}_{k-1}^n] = \frac{1}{\sqrt{n}} \left( \mathbb{E}[1(\gamma_k^n > w_k^n) | \mathcal{H}_{k-1}^n] - F^c(w_k^n) \right) = 0.
\]

which implies that the process \( \{ (\hat{H}_k^n, \mathcal{H}_k^n) : k \geq 1 \} \) is a discrete-time martingale for each \( n \geq 1 \).

Our next step is to replace \( k \) with \( \lfloor nt \rfloor \) for \( t \geq 0 \) to obtain a continuous-time martingale. By a direct application of Lemma 4.2 of [16], we deduce that, for each \( n \geq 1 \), the continuous-time process \( (\hat{H}_n(t), \mathcal{H}_n(t) : t \geq 0) \equiv (\hat{H}_{\lfloor nt \rfloor}^n, \mathcal{H}_{\lfloor nt \rfloor}^n : t \geq 0) \) is a martingale with quadratic variation

\[
\langle \hat{H}^n \rangle(t) = \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} (1(\gamma_i^n > w_i^n) - F^c(w_i^n))^2
\]

(C.20)

We next show that the sequence of martingales \( (\hat{H}^n(t), \mathcal{H}^n(t) : t \geq 0) \) satisfies the conditions of Theorem 7.1.4. of [20]. In particular, it is required that (i) jumps of the processes \( \hat{H}^n(t) \) are asymptotically negligible and (ii) quadratic variation of the processes converges in probability to the limit characterized in Theorem 7.1.1. of [20].

(i) **Negligibility of jumps.** We now show that condition (a) of Theorem 7.1.4. holds. Let \( \hat{H}^n(t-) \equiv \lim_{\epsilon \downarrow 0} \hat{H}^n(s) \). Then, for each \( T > 0 \), we have

\[
\sup_{0 \leq t \leq T} |\hat{H}^n(t) - \hat{H}^n(t-)| \leq 1/\sqrt{n}
\]

and hence

\[
\lim_{n \to \infty} \mathbb{E} \left[ \sup_{0 \leq t \leq T} |\hat{H}^n(t) - \hat{H}^n(t-)| \right] = 0.
\]

which is the desired condition.

(ii) **Convergence of quadratic variations.** We now prove that the quadratic variation pro-
Next we make use of the FWLLN for virtual waiting time process $V_n(t)$, i.e., $V_n \Rightarrow v$ in $\mathbb{D}([0, T]; \mathbb{R})$, and continuity of the function $F$ to show that (C.21) converges to 0. In particular,
for all $i \geq 1$,

$$F^c(w^n_i) = F^c(V_n(\tau^n_i -)) = F^c(v(\tau^n_i -) + o(1))$$

due to the fact that $V_n \Rightarrow v$ in $D([0,T];\mathbb{R})$. Combined with (C.23), this implies that the summands in (C.21) can be bounded above by

$$|F^c(v(\tau^n_i -) + o(1)) - F^c(v(\tau^n_i -))| + |F^c(v(\tau^n_i -) + o(1))^2 - F^c(v(\tau^n_i -))^2| \leq |o(1)|$$

where the inequality holds by continuity of the patience-time cdf $F$. This implies that the squared sum inside the expectation in (C.21) is bounded above by

$$\left| \frac{|o(1)| \lfloor nt \rfloor}{n^2} \right| \leq t^2 |o(1)| = o(1)$$

for all $t \geq 0$. Convergence of (C.21) to 0 then follows from dominated convergence theorem.

The summation in (C.22) can be alternatively represented as

$$\frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} F^c(v(\tau^n_i -))F(v(\tau^n_i -)) = \int_0^{\Lambda^{-1}(t)} F^c(v(u-))F(v(u-))d\tilde{A}_n(u)$$

$$\Rightarrow \int_0^{\Lambda^{-1}(t)} F^c(v(u-))F(v(u-))d\Lambda(u). \quad (C.24)$$

in $D([0,T];\mathbb{R})$ as $n \to \infty$ where the convergence (C.24) follows from continuous mapping theorem. Having established the convergence in (C.24), convergence in mean square is obtained by first applying continuous mapping theorem with the function $f(x) = x^2$, and then applying dominated convergence theorem by using the fact that both the summation and the limiting integral in (C.24) are bounded by $t$. Hence (C.22) converges to 0. That completes the proof of convergence of the quadratic variation (C.20).

Having proved conditions (i) and (ii) are indeed satisfied, by Theorem 7.1.4 of [20], we deduce that $\hat{H}^n \Rightarrow \hat{H}$ in $D([0,T];\mathbb{R})$ as $n \to \infty$ where $\hat{H}$ is a Gaussian process with independent increments and continuous sample paths. Moreover, as implied by the proof of Theorem 7.1.1. of [20], the limit $\hat{H}$ is indeed a time-changed Brownian motion where time-change is the limit for the quadratic variation, i.e.,

$$\hat{H}(t) = B(\langle \hat{H}(t) \rangle) = B \left( \int_0^{\Lambda^{-1}(t)} F^c(v(u-))F(v(u-))d\Lambda(u) \right), \quad t \geq 0,$$

where $B$ is the standard Brownian motion.

Finally, to complete the proof of convergence, we note that $\hat{E}_{n,2}(t) = \hat{H}^n(\hat{N}_n(t))$. Then, by continuous mapping theorem, we have $\hat{H}^n(\hat{N}_n) \Rightarrow \hat{H}(\Lambda)$ in $D([0,T];\mathbb{R})$ as $n \to \infty$. Consequently,
as $n \to \infty$,
\[
\frac{1}{\sqrt{n}} \sum_{i=1}^{N_n(t)} (1(\gamma_i > w_i^n) - F^c(w_i^n)) \Rightarrow B \left( \int_0^t F^c(v(u^-))F(v(u^-)) \, d\Lambda(u) \right) \tag{C.25}
\]
in $D([0,T];\mathbb{R})$.

C.5 Proofs of Results in §4.9

In this section we provide proofs of Theorem 4.9.1 and Corollary 4.9.1. The proof of both results closely follow the arguments in the proofs of Theorem 4.5.1, Theorem 4.5.2, Corollary 5.6.1 and Theorem 4.6.3. Therefore, we mostly refer to proofs of those results in the below proofs, and argue in what way the new staffing function $s_n = [ns_1 + \sqrt{ns_2}]$ make changes in arguments. We skip the lengthy details. Throughout the section, the processes with superscript $r$ correspond to those associated with staffing level $s_n$ whereas the processes without superscript $r$ correspond to those associated with staffing level $n$.

**Proof of Theorem 4.9.1.** We first argue joint convergence
\[
(\hat{N}_n, \hat{D}_n) \Rightarrow (\hat{N}, \hat{D}) \quad \text{in} \quad D^2([0,T];\mathbb{R}) \quad \text{as} \quad n \to \infty.
\]

We remark that the arrival process is not affected by the change of staffing level. Hence the assumed CLT for continue to holds, i.e, $\hat{N}^r(t) = \hat{N}(t)$, which implies that fluid limit remains unchanged, $\Lambda^r(t) = \Lambda(t)$, $t \geq 0$. The service-completion process, on the other hand, converges as $n \to \infty$
\[
\bar{D}^r_n(t) \equiv \frac{\sum_{j=1}^{s_n} D_j(t) - nD^r(t)}{\sqrt{n}} = \frac{s_n}{\sqrt{n}} \cdot \sum_{j=1}^{s_n} \frac{D_j(t) - s_n\mu t}{s_n} \Rightarrow D^r(t) \equiv s_1 D(t) = s_1 \mu t \quad \text{in} \quad D([0,T];\mathbb{R}) \tag{C.26}
\]
and
\[
\hat{D}^r_n(t) \equiv \frac{\sum_{j=1}^{s_n} D_j(t) - nD^r(t)}{\sqrt{n}} = \frac{s_n}{\sqrt{n}} \cdot \sum_{j=1}^{s_n} \frac{D_j(t) - s_n\mu t}{s_n} + \frac{s_n \mu t - nD^r(t)}{\sqrt{n}} = \sqrt{s_1} \hat{D}(t) - s_2 \mu t + O(1/\sqrt{n})
\]
\[
\Rightarrow \hat{D}^r(t) \equiv \sqrt{s_1} \hat{D}(t) - s_2 \mu t \quad \text{in} \quad D([0,T];\mathbb{R}) \tag{C.27}
\]
as $n \to \infty$ where $O(1/\sqrt{n})$ in the second equality accounts for the error caused by dropping $[\cdot]$ in $s_n$, and $\hat{D}(t)$ is the Gaussian process in Theorem 4.5.2. Hence we deduce from (C.27) that $\hat{D}^r(t)$ is a Gaussian process negative drift $-s_2 \mu t$ and covariance function $C^r(\cdot, \cdot) = s_1 C_E(\cdot, \cdot)$ with $C_E$
being the covariance function in Theorem 4.5.2. Also, from (C.27), joint convergence $(\hat{N}_n^r, \hat{D}_n^r)$ in $\mathbb{D}^2([0, T]; \mathbb{R})$ immediately follows from that of $(\hat{N}_n, \hat{D}_n) \Rightarrow (\hat{N}^r, \hat{D}^r)$ in $\mathbb{D}^2([0, T]; \mathbb{R})$ implied by Theorem 4.5.2. Furthermore, by Assumption 4.2.4, $\hat{E}_n^r = \hat{D}_n^r$ and hence $(\hat{N}_n^r, \hat{D}_n^r, \hat{E}_n^r) \Rightarrow (\hat{N}^r, \hat{D}^r, \hat{E}^r)$ in $\mathbb{D}^3([0, T]; \mathbb{R})$.

Having obtained the modified fluid limits in (C.26), and having established the joint convergence discussed in the above paragraph, we deduce that the proof in §C.2 continues to hold with minor modifications. In particular, (4.53)-(4.59) has the same mathematical form with fluid limits, and prelimit stochastic process replaced with their counterparts with superscript $r$. Moreover, by Assumption 4.2.4, $\hat{E}_n^r = \hat{D}_n^r$ and hence $(\hat{N}_n^r, \hat{D}_n^r, \hat{E}_n^r) \Rightarrow (\hat{N}^r, \hat{D}^r, \hat{E}^r)$ in $\mathbb{D}^3([0, T]; \mathbb{R})$.

In a similar fashion, one can modify the FCLT given the joint convergence $(\hat{N}_n^r, \hat{D}_n^r, \hat{E}_n^r) \Rightarrow (\hat{N}^r, \hat{D}^r, \hat{E}^r)$ in $\mathbb{D}^3([0, T]; \mathbb{R})$. The arguments in §4.7 continue to hold for modified fluid limits and cause only minor changes in the final expressions. In particular, (4.53)-(4.59) has the same mathematical form with fluid limits, and prelimit stochastic process replaced with their counterparts with superscript $r$. Moreover, (4.61)-(4.62) continues to hold because $s_n$ is $O(n)$. Hence the steps of proof in §4.7.1 can be replicated with counterpart processes. Only step that requires careful treatment is that of the last term in (4.70) where the limit enter-service process satisfies $\hat{E}(t) \equiv \sqrt{s_1} \hat{E}(t) - 2s_3\mu$, $0 \leq t \leq T$. Since the additional term $-s_2\mu$ is deterministic and $\sqrt{s_1} \hat{E}(t)$ is a centered Gaussian process, we can use similar arguments in proof of Corollary 5.6.1 to deduce that (4.77) is indeed the desired solution.

**Proof of Corollary 4.9.1.** We first derive the mean of $\hat{W}(\infty)$ from (4.77). Since the first three terms in (4.77) have mean equal to zero under the assumption that $\hat{N}(t) \equiv t_\lambda B(\lambda t)$, $t \geq 0$, with $B$ being a standard Brownian motion, $E[\hat{W}(\infty)]$ is the limit of the last term in (4.77) as $t \to \infty$. In particular,

$$E[\hat{W}^r(\infty)] = \lim_{t \to \infty} -2s_3 \mu \int_0^t \frac{H^r(t,u)}{q(t,w^r(t))} du = \lim_{t \to \infty} -2s_3 \mu \int_0^t \frac{e^{-hF(\lambda t)F(\lambda t)}}{\lambda F(\lambda t)F(\lambda t)} du + o(1)$$

where the $o(1)$ term accounts for the error by replacing $w^r$ with $w(t)$ and replacing $H^r(t, u)$ with the steady-state form as in (4.45). Having established the mean of $\hat{W}(\infty)$, it is easy to establish the mean for $\hat{V}(\infty)$ and $\hat{Q}(\infty)$ using the formulas

$$E[\hat{V}^r(\infty)] = E\left[\lim_{t \to \infty} \frac{\hat{W}^r(t)}{1 - \hat{w}^r(t + w^r(t))}\right] = E[\hat{W}^r(\infty)]$$

and

$$E[\hat{Q}^r(\infty)] = \lambda F(\lambda t)F(\lambda t)E[\hat{W}^r(\infty)].$$
Computation of variance is standard and as given in §C.3.2.
Appendix D

Appendix to Chapter 5

D.1 Proof of Lemma 5.6.2

We show that the covariance function $C_\hat{Y}(\cdot, \cdot)$ of $\hat{Y}$ in (5.32) satisfies condition (5.48). Due to independence, it suffices to show that the covariance function of components of $\hat{Y}$, i.e., $\hat{E}_1^\circ$, $\hat{E}_2^\circ$ and $\hat{E}_3^\circ$ in (5.33)-(5.34) satisfy condition (5.48). Let $0 \leq s < t \leq T$ and $0 \leq u \leq T$.

First consider $\hat{E}_1^\circ$ and its covariance function given in (5.33). It follows that

$$C_{\hat{E}_1^\circ}(s, u) - C_{\hat{E}_1^\circ}(t, u) = -\int_0^\infty \tilde{G}_v(s \land u) \tilde{G}_v^c(s \lor u) dB(0, v) + \int_0^\infty \tilde{G}_v(t \land u) \tilde{G}_v^c(t \lor u) dB(0, v)$$

$$\leq \int_0^\infty \left| \tilde{G}_v(s \land u) \tilde{G}_v^c(s \lor u) - \tilde{G}_v(t \land u) \tilde{G}_v^c(t \lor u) \right| dB(0, v)$$

$$\leq \int_0^\infty \left( \tilde{G}_v(s \land u) \tilde{G}_v^c(s \lor u) - \tilde{G}_v(s \land u) \tilde{G}_v^c(t \lor u) \right) dB(0, v) + \int_0^\infty \left( \tilde{G}_v(t \land u) \tilde{G}_v^c(t \lor u) - \tilde{G}_v(t \land u) \tilde{G}_v^c(t \lor u) \right) dB(0, v)$$

$$\leq C_1^\circ |s - t|^{\alpha_1^\circ}$$

for some $C_1^\circ > 0$ and $0 < \alpha_1^\circ < 1$ because $B(0, \infty) = B(0) = s(0) < \infty$ and $G$ is differentiable with bounded density on bounded intervals.

Secondly, consider the $\hat{E}_2^\circ$ in (5.34). Applying integration by parts yields

$$\hat{E}_2^\circ(t) = \hat{B}(0) \tilde{G}_\infty(t) - \hat{B}(0, 0) G_0(t) - \int_0^t \hat{B}(0, v) \tilde{g}_0(t) dv, \quad 0 \leq t \leq T.$$
where \( \tilde{g}_v(t) \) is the partial derivative of \( \tilde{G}_v(t) \) with respect to \( v \), i.e.,

\[
\tilde{g}_v(t) = \frac{\partial}{\partial v} \left[ 1 - \frac{G^c(t + v)}{G^c(v)} \right] = \frac{g(t + v)G^c(v) - g(v)G^c(t + v)}{G^c(v)^2}, \quad 0 \leq t \leq T. \tag{D.1}
\]

Then, because \( \hat{B}(0, \cdot) \) has mean 0, the covariance function can be computed as

\[
C_{\tilde{E}_2}(s, t) = \mathbb{E} \left[ \int_0^\infty \hat{B}(0, x)\tilde{g}_x(s) \, dx \int_0^\infty \hat{B}(0, y)\tilde{g}_y(t) \, dy \right] = \int_0^\infty \int_0^\infty \mathbb{E}[\hat{B}(0, x)\hat{B}(0, y)] \tilde{g}_x(s)\tilde{g}_y(t) \, dx \, dy, \quad 0 \leq s, t \leq T.
\]

Then it follows that, for \( 0 \leq u \leq T, \ 0 \leq s, t \leq T, \)

\[
C_{\tilde{E}_2}(s, u) - C_{\tilde{E}_2}(t, u) = \int_0^\infty \int_0^\infty \mathbb{E}[\hat{B}(0, x)\hat{B}(0, y)] (\tilde{g}_x(s) - \tilde{g}_x(t))\tilde{g}_y(u) \, dx \, dy

= \int_0^\infty \int_0^\infty \mathbb{E}[\hat{B}(0, x)\hat{B}(0, y)] \frac{(g(s + x) - g(t + x))G^c(x)}{G^c(x)^2} \tilde{g}_y(u) \, dx \, dy

- \int_0^\infty \int_0^\infty \mathbb{E}[\hat{B}(0, x)\hat{B}(0, y)] \frac{g(x)(G^c(s + x) - G^c(t + x))}{G^c(x)^2} \tilde{g}_y(u) \, dx \, dy

\leq K|s - t|^{\gamma} \int_0^\infty \int_0^\infty \mathbb{E}[\hat{B}(0, x)\hat{B}(0, y)] (1 + h_G(x)) \tilde{g}_y(u) \, dx \, dy
\]

for some \( K > 0 \) and \( 0 < \gamma < 1 \) where the boundedness of the last term is guaranteed by (5.46) in Assumption 5.6.1.

Finally we consider \( \hat{\dot{E}}_2^\nu \) in (5.34). Using the covariance formula given in [54, 78, 4], we obtain

\[
C_{\hat{\dot{E}}_2^\nu}(s, u) - C_{\hat{\dot{E}}_2^\nu}(t, u)

= \int_0^s G(s \wedge u - v)G^c(s \vee u - v) \, dE(v) - \int_t^\nu G(t \wedge u - v)G^c(t \vee u - v) \, dE(v)

\leq |E(s) - E(t)| \leq C_2^\nu |s - t|^{\alpha_2^\nu}
\]

for some \( C_2^\nu > 0 \) and \( 0 < \alpha_2^\nu < 1 \) where the last inequality holds since \( E(t) \) is differentiable with bounded first derivative and hence \( E(t) \) is Hölder continuous. This proves that the covariance function of \( \hat{\dot{E}}_2^\nu \) satisfies condition (5.48).