

A Robust Queueing Network Analyzer Based on Indices of Dispersion

Wei You

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

©2019

Wei You

All Rights Reserved

ABSTRACT

A Robust Queueing Network Analyzer Based on Indices of Dispersion

Wei You

In post-industrial economies, modern service systems are dramatically changing the daily lives of many people. Such systems are often complicated by uncertainty: service providers usually cannot predict when a customer will arrive and how long the service will be. Fortunately, useful guidance can often be provided by exploiting stochastic models such as queueing networks. In iterating the design of service systems, decision makers usually favor analytical analysis of the models over simulation methods, due to the prohibitive computation time required to obtain optimal solutions for service operation problems involving multidimensional stochastic networks. However, queueing networks that can be solved analytically require strong assumptions that are rarely satisfied, whereas realistic models that exhibit complicated dependence structure are prohibitively hard to analyze exactly.

In this thesis, we continue the effort to develop useful analytical performance approximations for the single-class open queueing network with Markovian routing, unlimited waiting space and the first-come first-served service discipline. We focus on open queueing networks where the external arrival processes are not Poisson and the service times are not exponential.

We develop a new non-parametric robust queueing algorithm for the performance approximation in single-server queues. With robust optimization techniques, the underlying stochastic processes are replaced by samples from suitably defined uncertainty sets and the worst-case scenario is analyzed. We show that this worst-case characterization of the performance measure is asymptotically exact for approximating the mean steady-state workload in $G/G/1$ models in both the light-traffic and heavy-traffic limits, under mild regularity

conditions. In our non-parametric Robust Queueing formulation, we focus on the customer flows, defined as the continuous-time processes counting customers in or out of the network, or flowing from one queue to another. Each flow is partially characterized by a continuous function that measures the change of stochastic variability over time. This function is called the index of dispersion for counts. The Robust Queueing algorithm converts the index of dispersion for counts into approximations of the performance measures. We show the advantage of using index of dispersion for counts in queueing approximation by a renewal process characterization theorem and the ordering of the mean steady-state workload in $GI/M/1$ models.

To develop generalized algorithm for open queueing networks, we first establish the heavy-traffic limit theorem for the stationary departure flows from a $GI/GI/1$ model. We show that the index of dispersion for counts function of the stationary departure flow can be approximately characterized as the convex combination of the arrival index of dispersion for counts and service index of dispersion for counts with a time-dependent weight function, revealing the non-trivial impact of the traffic intensity on the departure processes. This heavy-traffic limit theorem is further generalized into a joint heavy-traffic limit for the stationary customer flows in generalized Jackson networks, where the external arrival are characterized by independent renewal processes and the service times are independent and identically distributed random variables, independent of the external arrival processes.

We show how these limiting theorems can be exploited to establish a set of linear equations, whose solution serves as approximations of the index of dispersion for counts of the flows in an open queueing network. We prove that this set of equations is asymptotically exact in approximating the index of dispersion for counts of the stationary flows. With the index of dispersion for counts available, the network is decomposed into single-server queues and the Robust Queueing algorithm can be applied to obtain performance approximation. This algorithm is referred to as the Robust Queueing Network Analyzer.

We perform extensive simulation study to validate the effectiveness of our algorithm. We show that our algorithm can be applied not only to models with non-exponential distributions but also to models with more complex arrival processes than renewal processes,

including those with Markovian arrival processes.

Table of Contents

List of Figures	v
List of Tables	vii
1 Introduction	1
1.1 Challenges in Analyzing the Open Queueing Networks	3
1.1.1 Jackson Networks	3
1.1.2 Generalized Jackson Networks	4
1.1.3 General Open Queueing Networks	5
1.2 Approximation Algorithms	6
1.2.1 Decomposition Approximations	6
1.2.2 Heavy-Traffic Limit Approximations	6
1.2.3 Robust Queueing Approximations	8
1.2.4 Approximations Based on Non-Parametric Traffic Descriptions . . .	8
1.3 Main Contributions	9
1.3.1 Non-Parametric Traffic Description for Queueing Networks	9
1.3.2 Heavy-Traffic Limits for Stationary Network Flows	10
1.3.3 The Robust Queueing Network Analyzer and Simulation Studies . .	12
1.4 Outline	12
2 Robust Queueing for the $G/G/1$ Model	14
2.1 Robust Queueing for the Steady-State Waiting Time	15

2.1.1	Parametric RQ for Waiting Time in the $GI/GI/1$ Model	16
2.1.2	Non-Parametric RQ for Waiting Time in the $G/G/1$ Model	20
2.2	Robust Queueing for the Steady-State Workload	21
2.2.1	The Workload Process and Its Reverse-Time Representation	21
2.2.2	Parametric and Non-Parametric RQ for the Workload	22
2.2.3	The Indices of Dispersion for Counts and Work	26
2.2.4	The Indices of Dispersion and the Mean Steady-State Workload	28
2.2.5	Robust Queueing with the IDW	31
2.2.6	The $RQ(b)$ Algorithm for Quantile Approximation	32
2.2.7	Heavy-Traffic and Light-Traffic Limits	33
2.2.8	Other Steady-State Performance Measures	34
2.2.9	An Illustrative Simulation Example	35
2.3	Estimating and Calculating the IDC	37
2.3.1	The IDC's for Renewal Processes.	37
2.3.2	The IDC of the Markovian Arrival Process	38
2.3.3	Calculation of the IDW and IDC in Some Queueing Networks	42
2.3.4	Numerical Estimation of the IDC from Data.	43
2.4	More on Non-Parametric RQ for Waiting Times	46
2.5	Supporting Functional Central Limit Theorems	47
2.5.1	The Basic FCLT for the Partial Sums	48
2.5.2	The FCLT for Other Basic Processes	49
2.5.3	Alternative Scaling in the CLT	51
2.5.4	The Heavy-Traffic FCLT	52
2.5.5	The IDW and the Normalized Workload	54
2.6	Proofs	54
3	Robust Queueing for Queues In Series	62
3.1	Review of Stationary Point Processes	63
3.1.1	Renewal Processes and the Laplace Transform	63
3.1.2	Revisiting the IDC	65

3.1.3	The Palm-Khintchine Equation	66
3.2	Heavy-Traffic Limit Theorem for $GI/GI/1$ Stationary Departure	68
3.2.1	The Departure Variance in the $GI/M/1$ Queue	68
3.2.2	The Departure Variance in the $M/GI/1$ Queue	74
3.2.3	Heavy-Traffic Limit for the Stationary Departure Process	76
3.2.4	Extensions	84
3.3	Robust Queueing Network Analyzer for Queues In Series	85
3.3.1	Approximation of the Departure IDC	85
3.3.2	The RQNA algorithm for Queues in Series	87
3.4	Proofs	88
4	Heavy-Traffic Limits for Stationary Flows	98
4.1	The OQN Model	100
4.2	Existence, Uniqueness and Convergence Via Harris Recurrence	101
4.3	Heavy-Traffic Limit Theorems for the Stationary Processes	106
4.3.1	Representation of the Centered Stationary Flows	106
4.3.2	Heavy-Traffic Limit with Any Subset of Bottlenecks	107
4.4	Heavy-Traffic Limits with One Bottleneck Queue	111
4.4.1	A Single-Server Queue with Customer Feedback	112
4.4.2	Networks with One Bottleneck Queue	115
4.5	Functional Central Limit Theorem for the Stationary Flows	118
4.6	Proofs	120
5	Robust Queueing Network Analyzer	127
5.1	The Departure Operation	128
5.2	The Splitting Operation	130
5.2.1	Dependent Splitting: One Queue with Immediate Feedback	131
5.2.2	The General Case	132
5.3	The Superposition Operation	136
5.3.1	Dependent Superposition: A Splitting and Recombining Example	136

5.3.2	The General Case	139
5.4	The IDC Equation System	143
5.5	Feedback Elimination	145
5.5.1	Immediate Feedback Elimination	145
5.5.2	Near-Immediate Feedback	147
5.6	The RQNA Algorithm	148
5.6.1	RQNA for Tree-Structured Queueing Networks	148
5.7	Proofs	151
6	Simulation Experiments	153
6.1	Notation and Simulation Methodology	153
6.2	Robust Queueing for Single-Server Queues	157
6.2.1	The $GI/GI/1$ Models	157
6.2.2	A Queue with a Superposition Arrival Process	159
6.2.3	A Ten-Queues-in-Series Example	168
6.2.4	A MAP/MAP/1 Example	171
6.2.5	The Queues in Series Models	173
6.2.6	The Limitation of IDC	175
6.3	Robust Queueing Network Analyzer for Tandem Queues	182
6.3.1	Departure IDC Approximation in $G/G/1$ Models	183
6.3.2	An Illustrative Example	185
6.3.3	Comparisons with Previous Algorithms for Queues in Series	188
6.3.4	RQNA Performance in Tandem Queueus	190
6.4	Robust Queueing Network Analyzer for Open Queueing Networks	190
6.4.1	Feedback Elimination: A Three-Station Example.	199
6.4.2	A Ten-Station Example with Feedback.	200
	Bibliography	204

List of Figures

1.1	A clinic modeled as an open network of queues.	2
2.1	The net-input process.	22
2.2	Comparing the simulation estimation to the RQ approximation for the workload, as a function of traffic intensity, at the fifth queue of a five-queues-in-series model. The workload function have four internal modes.	36
3.1	Approximation of the departure IDC in the $E_2/M/1$ model.	74
3.2	Approximations of the departure IDC's in two $GI/GI/1$ models.	87
4.1	A single-server queue with customer feedback.	112
5.1	Approximations of the IDC's in a single-server queue with feedback model.	132
5.2	A simple splitting and recombining example.	136
5.3	Approximation of the arrival IDC at queue 3 for the splitting and recombining model in Figure 5.2.	139
6.1	The normalized mean workload and the RQ approximation for $LN(4)/GI/1$ and $G(4)/GI/1$ models.	166
6.2	The normalized mean workload and the RQ approximation for $GI/E_4/1$ models.	167
6.3	A queue with superposition arrival process.	167
6.4	The RQ solution and approximation for the workload in the $\sum_i^n LN_i/H_2/1$ model.	168

6.5	A ten-queues-in-series model.	168
6.6	Comparing the RQ approximation, the $E_2/M/1$ approximation and the simulation estimation of the workload in the ten-queues-in-series model.	169
6.7	The IDW at the last queue of the ten-queues-in-series model: the continuous-time stationary version and the discrete-time Palm version.	170
6.8	Contrasting the discrete-time and continuous-time views: the analog of Figure 6.6 for the waiting time.	171
6.9	The IDW and RQ approximation for a MAP/MAP/1 model.	172
6.10	The performance of RQ at station 2 of the $LN(4)/E_4/1 \rightarrow \cdot/GI/1$ models.	177
6.11	Comparing the mean steady-state workload in single-server queues with identical arrival IDC and service-time distribution.	183
6.12	The RQNA approximation of the departure IDC of the $G(4)/GI/1$ and $E_4/GI/1$ models.	184
6.13	The RQNA approximation of the departure IDC of the models with MAP.	186
6.14	Comparing the simulation estimation to the RQ approximation for the workload, as a function of traffic intensity, at the last queue of a thirteen-queues-in-series model. The workload function have four internal modes.	187
6.15	A three-station example.	199
6.16	A ten-station with customer feedback example.	202

List of Tables

6.1	The sample mean, sample standard deviation (SD) and sample coefficient of variation (CV) of the mean estimator for the workload in various single-server queues.	156
6.2	The RQ approximation in various $GI/E_4/1$ models.	160
6.3	The RQ approximation in various $GI/LN(0.25)/1$ models.	161
6.4	The RQ approximation in various $GI/M/1$ models.	162
6.5	The RQ approximation in various $GI/H_2(4)/1$ models.	163
6.6	The RQ approximation in various $GI/G(4)/1$ models.	164
6.7	The performance of the RQ approximation in various $GI/GI/1$ models for $\rho > 0.5$	165
6.8	The maximum absolute error of the RQ approximation in various $GI/GI/1$ models.	165
6.9	Absolute error of the RQ-IDW approximation in the MAP/MAP/1 model.	173
6.10	The RQ approximation in various $GI/E_4/1 \rightarrow \cdot/E_4/1$ models.	176
6.11	The absolute relative error of the RQ approximation at station 2 of two queues in series models with E_4 service times at station 2.	177
6.12	The absolute relative error of the RQ approximation at station 2 of two queues in series models with $LN(0.25)$ service times at station 2.	178
6.13	The absolute relative error of the RQ approximation at station 2 of two queues in series models with M service times at station 2.	178
6.14	The absolute relative error of the RQ approximation at station 2 of two queues in series models with $H_2(4)$ service times at station 2.	179

6.15	The absolute relative error of the RQ approximation at station 2 of two queues in series models with $LN(4)$ service times at station 2.	179
6.16	The absolute relative error of the RQ approximation at station 2 of two queues in series models with $G(4)$ service times at station 2.	180
6.17	The absolute relative error of the RQ approximation at station 2 of two queues in series models with $\rho_1 = 0.7$ and E_4 service times at station 2. . .	180
6.18	The maximum absolute relative error of the RQ approximation at station 2 of two queues in series models with $\rho_1 = 0.7$ and M service times at station 2.	181
6.19	Comparing the RQ performance for $\rho_1 = 0.7$ and 0.9	181
6.20	A comparison of four approximation methods to simulation for 9 exponential (M) queues in series fed by a deterministic arrival process with $c_a^2 = 0$. . .	188
6.21	A comparison of four approximation methods to simulation for 9 exponential (M) queues in series fed by a highly-variable H_2 renewal arrival process with $c_a^2 = 8$	189
6.22	The RQNA approximation in various $LN(4)/GI/1 \rightarrow \cdot/E_4/1$ models with $\rho_1 = 0.9$	191
6.23	The RQNA approximation in various $E_4/GI/1 \rightarrow \cdot/LN(4)/1$ models with $\rho_1 = 0.9$	192
6.24	The RQNA approximation in various $E_4/GI/1 \rightarrow \cdot/E_4/1$ models with $\rho_1 = 0.9$.	193
6.25	The RQNA approximation in various $LN(4)/GI/1 \rightarrow \cdot/LN(4)/1$ models with $\rho_1 = 0.9$	194
6.26	The RQNA approximation in various $LN(4)/GI/1 \rightarrow \cdot/E_4/1$ models with $\rho_1 = 0.7$	195
6.27	The RQNA approximation in various $E_4/GI/1 \rightarrow \cdot/LN(4)/1$ models with $\rho_1 = 0.7$	196
6.28	The RQNA approximation in various $E_4/GI/1 \rightarrow \cdot/E_4/1$ models with $\rho_1 = 0.7$.	197
6.29	The RQNA approximation in various $LN(4)/GI/1 \rightarrow \cdot/LN(4)/1$ models with $\rho_1 = 0.7$	198
6.30	Traffic intensity of the four cases in the three-station example.	199

6.31	Variability of the service distributions of the four cases in the three-station example.	199
6.32	A comparison of six approximation methods for the total sojourn time in the three-station example in Figure 6.15 with parameters specified in Table 6.30 and 6.31.	201
6.33	A comparison of six approximation methods for the sojourn time at each station of the three-station example for Case D with parameters specified in Table 6.31.	201
6.34	A comparison of six approximation methods to simulation for the mean steady-state sojourn times at each station of the open queueing network in Figure 6.16.	203

Acknowledgments

This dissertation concludes five years of doctoral study and research. I am deeply indebted to many exceptional scholars and inspiring individuals, who guide me through an voyage of discovery that never meant to be a downwind sailing.

First and foremost, I would like to thank my advisor, Professor Ward Whitt, for the enlightening guidance throughout my life as a Ph.D. student. He is always a walking encyclopedia to me with his accurate pointers to the literature and his prompt and constructive feedback. His generosity in sharing his knowledge from both the mathematics and engineering point of view helped me greatly in understanding the big picture of the subject matter of this work. As a great educator and scholar, he has been and will always be a role model that I look up to.

I am very grateful to Professor David Yao, Assaf Zeevi, Martin Reiman and Henry Lam for serving as my dissertation committee. Their critical questions, insightful discussions and invaluable feedback helped me greatly in improving this dissertation and its presentation. Special thanks to Professor Zeevi for coming to my defense right after an overnight flight.

My research journey began with to the enlightening interactions with Professor Jim Dai, Jiheng Zhang, Ming Hu and Wotao Yin, before I came to Columbia Univeristy. I am privileged to have such brilliant undergraduate alumni from Nanjing University.

During my studies at Columbia University, I am fortunate enough to have taken many insightful and engaging courses from the exceptional faculty members. I thank Professor Donald Goldfarb, Karl Sigman, Garud Iyengar, David Yao, Daniel Bienstock, Adam Elmachtoub, Yuri Faenza, Tim Leung, Xuedong He, Martin Reiman, Jacob Leshno, John Yao, Philip Protter and Peter Orbanz for expanding my knowledge in probability, optimization, statistics and finance. I am also privileged to serve as teaching assistant for Professor

David Yao, Yuan Zhong, Ton Dieker and Doctor Soumyadip Ghosh. I am deeply indebted to them for their friendly advices, mentorship and support, which helped me tremendously in my academic and career development.

I am grateful to many friends and colleagues I met for the past five years. I received much help from my academic family, Guodong Pang, Yunan Liu, Ohad Parry, Song-Hee Kim, Itai Gurvich, Rouba Ibrahim, Ni Ma, Xu Sun, Xiaopei Zhang and Yan Chen. I am also indebted to my seniors who are kind enough to offer invaluable advice, Wotao Yin, Ming Hu, Ningyuan Chen, Jing Dong, Cun Mu, Xinyun Chen, Chun Wang, Liao Wang, Yanan Pei, Zhipeng Liu, Anran Li, Chun Ye, Jing Guo, Irene Lo, Francois Fagan, Di Xiao, Mingxian Zhong, Xinshang Wang and Fei He. I treasure the time I spend with follow Ph.D. students in the department, Shuangyu Wang, Mauro Escobar, Michael Hamilton, Brian Ward, Kavin Guo, Nourhan Sakr, Shuoguang Yang, Lin Chen, Jacob Bergquist, Omar El Housni, Fengpei Li, Ryan McNellis, Min-hwan Oh, Yunjie Sun, Xuan Zhang, Xinyu Zhang, Jingtong Zhao, Yeqing Zhou; and also with friends outside the department, Haoyuan Zhang, Yanchao Li, Zongxi Li, Jianwen Sun, Zhi Li, Boyuan Zhang, Chao Qin, Zhe Liu, Amine Allouah, Pengyu Qian, Francisco Castro, Pu He, Jiaqi Lu and Yang Kang. This is by no mean a comprehensive list, I would like to express my deep appreciation to every individual that shared this adventure with me.

I thank my advisor, the Columbia University and the NSF grants CMMI 1265070 and 1634133 for generous financial support.

Most importantly, I thank my family members for their unconditional support, especially my wife Huichen Zhu. Their faith in me has always encouraged me.

To my parents Jiangong, Yongling and my wife Huichen

Chapter 1

Introduction

This thesis contributes to analytical methods for designing and optimizing service systems. Such systems appear in a broad and diverse range of settings, including customer contact centers, hospitals, airlines, online marketplaces, ride-sharing platforms and cloud computing networks. In post-industrial economies, modern service systems are dramatically changing the daily lives of many people. Their rapid development leads to challenges in their design and operation, especially because such systems are often complicated by uncertainty: service providers usually cannot predict when a customer will arrive and how long the service will be. Hence, decision makers seek operating policies that adapt to the randomness of the customer flow and service requirements.

Fortunately, useful guidance can often be provided by exploiting mathematical models using stochastic processes. Prominent among these are stochastic queueing network models, because service is often provided in a sequence of steps. There is an extensive literature on the applications of queueing network models to service systems. For example, see [117] for a review of applications in computer networks, see [19; 64; 111] for examples in ride-sharing economies and see [33; 150; 45; 94] for healthcare-related applications.

For illustration, Figure 1.1 shows a queueing network view of a clinic. Patients are greeted at the reception desk and assigned to nurses for initial triage. Before the seeing a doctor, patients may have to take multiple tests conducted at different labs. Figure 1.1 also shows feedback flows, because patients might need medical procedures both before and

after seeing the doctor.

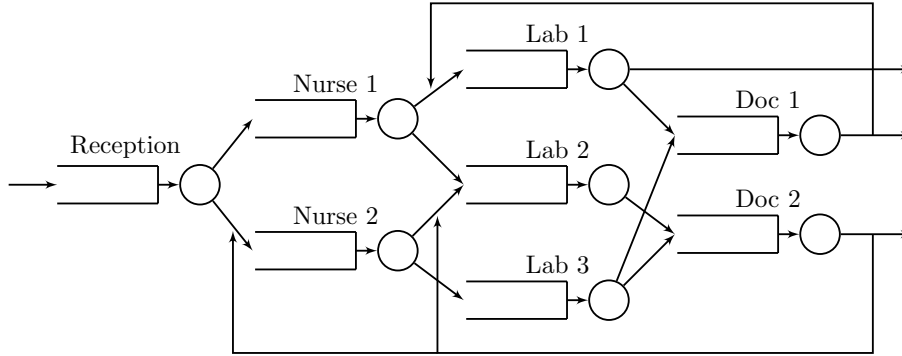


Figure 1.1: A clinic modeled as an open network of queues.

Service operation policies often rely on quantitative descriptions of the system performance, which are usually referred to as *performance measures*. For example, the *waiting time* characterizes the delay between joining the queue and entering service of a customer; the *queue length* counts the number of customers waiting in line; and the *workload* (virtual waiting time) measures the total amount of service requirements of the customers in the system.

A standard way to analyze the performance of complex queueing models is to employ computer simulation, see [123; 151] for examples. However, as noted in [54], a great disadvantage of simulation-based optimization methods is the often prohibitive computation time required to obtain optimal solutions for service operation problems involving multi-dimensional stochastic network, which is in large part due to the inherent combinatorial explosion of the decision space. Analytical analysis of the models can thus be very helpful. However, queueing networks that can be solved analytically require strong assumptions that are rarely satisfied, whereas realistic models are prohibitively hard to analyze exactly. Hence, analytical performance approximation of queueing networks remains an important tool, see Section 1.2 for a review.

In this thesis, we continue the effort to develop useful analytical performance approximations for the single-class open queueing network with Markovian routing, unlimited waiting space and the first-come first-served service discipline. We focus on open queueing networks where the external arrival processes are not Poisson and the service times are not

exponential. This thesis is based on [145; 144; 147; 146; 148].

1.1 Challenges in Analyzing the Open Queueing Networks

Exact analytical analysis of the open queueing networks is mostly limited to models with strong assumptions. The basis for most analysis is the theory of Jackson networks initiated by Jackson [85], which we discuss in Section 1.1.1. However, service systems are often complicated by significant deviations from the tractable structure of a Jackson network, usually resulted from complicated dependence in the network, as we discuss in Section 1.1.2.

1.1.1 Jackson Networks

A Jackson network arises when the external arrival processes are independent Poisson processes, the service time are mutually independent exponential random variables, independent of the external arrival processes. Customer routing in a Jackson network follows a *Markovian routing policy*: upon service completion at station i , the customer is directed to station j with probability $p_{i,j}$, which is independent of the system state and the past.

This model is especially tractable. For a Jackson network with K stations, let $Q_i(t)$ denote the queue length at queue i at time t . It is well known that $\mathbf{Q} \equiv \{Q_i : i = 1, 2, \dots, K\}$ forms a Markov process, so that the conditional probability distribution of the state $\mathbf{Q}(t)$ at a future time t , conditioning on the past values up to the current time s for $s < t$, depends only on the current state $\mathbf{Q}(s)$. Consequently, a Jackson network is also called a *Markov open queueing network*.

Jackson [85] showed that the steady-state vector for the number of customers at each queue in a Jackson network has a product-form distribution with independent geometric marginal distributions. Hence in steady-state the network can be viewed as if it is decomposed into mutually independent $M/M/1$ stations (in Kendall's notation), even though the queueing processes are not in fact independent.

This initial breakthrough was followed by vigorous research leading to an elaborate and useful theory, as can be seen from [36; 89; 119]. Due to its closed-form and product-form

solution, Jackson networks have been widely studied, e.g. in [115; 105; 13]. Jackson networks have also been applied to many service systems. For ride-sharing economy, [19] studied the optimal platform pricing, while [118] looked at the inventory rebalancing and vehicle routing problems. [103; 129; 149; 22] analyzed resource allocation and quality-of-service in cloud computing system. For healthcare related problems, [16] studies hospital staffing strategy to achieve optimal workflow efficiency under information security requirements; see also [67] for an overview.

1.1.2 Generalized Jackson Networks

A *generalized Jackson network* relaxes Jackson network's assumption on the distribution of the interarrival times and service times. Such a network assumes that the external arrival processes are independent renewal processes and the service times at each station are i.i.d. with general distributions, independent of the external arrival processes. The service policy is first-come first-served, while the routing policy is Markovian, as in Jackson networks. For theoretical analysis in this thesis, we mostly restrict to the setting of a generalized Jackson network.

Applications in communication, manufacturing and service systems are often complicated by significant deviations from the tractable structure of a Jackson network, in which cases generalized Jackson networks are more suitable.

In general, an external customer arrival process in a call center often is well modeled by a Poisson process, because it is generated by many separate people making decisions independently, at least approximately. But dependence in arrival processes may still be induced by over-dispersion, e.g., see [93] and references there. In most manufacturing systems, an external arrival process is often far less variable than a Poisson process by design. Even if external arrival processes can be regarded as Poisson processes, service-time distributions are often non-exponential, see [26; 56]. This is often resulted from complicated processing operations, such as those involving batching.

Non-exponential interarrival-time or service-time distributions produce complicated dependence structure in the departure processes, which will be inherited by the arrival processes at the subsequent stations. Then these processes cannot be renewal processes be-

cause (i) a departure process from any $GI/GI/1$ queue is necessarily non-renewal if the interarrival-time or service-time distribution is non-exponential and (ii) the superposition of independent renewal processes cannot be renewal unless all components are Poisson processes (in which case the superposition process is also Poisson); e.g., see [51; 52; 55].

Indeed, such dependence in departure processes is consistent with our heavy-traffic limit theorem for the stationary departure process in generalized Jackson network, see Theorem 5.1. It shows that the dependence structure in departure process depends on the traffic intensity and the interarrival-time and service-time distributions in a nontrivial manner.

Furthermore, dependence among different arrival and service processes are often observed in manufacturing/communication systems. Upon service completion, jobs are directed to subsequent stations. This corresponds to splitting the departure process, which introduces dependence among the sub-flows after splitting. In hospital settings, patients may revisit a doctor after completing several tests. In manufacturing lines, products may need rework after quality-control testings. This is referred to as *customer feedback*, which necessarily introduce dependence between the service and arrival processes.

1.1.3 General Open Queueing Networks

More general open queueing network models can be obtained from generalized Jackson network by further relaxing the assumption that the external arrival processes are independent renewal processes and the service times are i.i.d. In this general form, the external arrival processes can be non-renewal processes, such as those characterized by Markov arrival processes, where interarrival times may be dependent. Service times can also be dependent.

With the presence of these dependence structures, queue length process a general open queueing network is rarely a Markov process. These networks are often referred to as *non-Markov open queueing networks*. Such dependence have important performance implication, as we demonstrate in Section 2.2.9.

In this thesis, we develop approximation algorithms that expose the performance impact of dependence in non-Markov open queueing networks, using a novel non-parametric modeling approach; see Section 2.2.5.

1.2 Approximation Algorithms

In this section we briefly review approximation approaches for general open queueing networks.

1.2.1 Decomposition Approximations

Motivated by the product-form property of Markov OQNs, decomposition approximations for non-Markov OQNs have been widely investigated. In this approach, the network is decomposed into individual single-server queues, and the steady-state queue length processes are assumed to be approximately independent. For example, in [96] and [134] each queue is approximated by a $GI/GI/1$ model, where the arrival (service) process is approximated by a renewal process partially characterized by the mean and *squared coefficient of variation* (scv, variance divided by the square of the mean) of an interarrival (service) time.

While the decomposition approximations do often perform well, it was recognized that dependence in the arrival processes of the internal flows can be a significant problem. The approximation for superposition processes used in the QNA algorithm [134] attempts to address the dependence. Nevertheless, significant problems remained, as was dramatically illustrated by comparisons of QNA to model simulations in [124; 58; 125], as discussed in [142].

To address the dependence in arrival processes, decomposition methods based on Markov Arrival Process (MAP) have been developed. MAP was first suggested by Neuts [107]. Horváth et al. [81] approximates each station by a MAP/MAP/1 model. In Kim [91; 92], the queue is approximated by a MMPP(2)/ $GI/1$ model, where the arrival process is a Markov-modulated Poisson process with two states. MAP (and MMPP(2) as its special case) need not be a renewal process, hence are capable of modeling the autocorrelation in the arrival and service processes.

1.2.2 Heavy-Traffic Limit Approximations

The early decomposition approximation in [134] drew heavily on the central limit theorem (CLT) and heavy-traffic (HT) limit theorems. Approximations for a single queue follow from

[83; 84]. With these tools, approximations for general point processes and arrival processes were developed in [133; 135]. Heavy-traffic approximation of queues with superposition arrival processes in [138] helped capture the impact of dependence in such queues; see §4.3 of [134].

Another approach is to apply HT limit theorems for the entire network. Such HT limits were established for feedforward OQN's in Iglehart and Whitt [83; 84] and Harrison [70; 71], and then for general OQN's by Reiman [113]. These works showed that the queue length process converges to a multidimensional reflected Brownian motion (RBM) as every service station approaches full saturation simultaneously. A more general case with both strictly bottleneck and non-bottleneck queues and general initial conditions was studied in [35].

Approximations for the pre-limit OQN's were developed from these general heavy-traffic results, depending on the relatively tractable limiting RBM processes. Notably, we have the QNET algorithm in Harrison and Nguyen [73], where the steady-state mean queue length is approximated by the mean of the steady-state distribution of the limiting multi-dimensional RBM. Theoretical and numerical analysis of the stationary distribution of the multi-dimensional RBM is studied in [74; 75; 46]. It is worth noting that the process-level convergence of the queue length process to a RBM does not automatically imply the convergence of the steady-state distribution. Hence, these algorithms rely on the exchange-of-limit arguments to justify the steady-state approximation; see [66; 31; 27].

As a crucial step of the QNET algorithm, Dai and Harrison [46] proposed a numerical algorithm to calculate the steady-state density of a RBM, but it requires considerable computation time. The accuracy of that algorithm improves as the number of iterations n grows, and the authors there note that $n = 5$ generally gives satisfactory answers. For an OQN with d stations, the computational complexity is $O(d^{2n})$, see Section 6 of [46]. For practical application in large-scale systems, hybrid methods that combine decomposition approximation and heavy-traffic theory have been developed to reduce computation time. In particular, we have the Individual Bottleneck Decomposition (IBD) algorithm in Reiman [114] and the Sequential Bottleneck Decomposition (SBD) algorithm in Dai, Nguyen and Reiman [44].

1.2.3 Robust Queueing Approximations

Recently, a novel Robust Queueing (RQ) approach to analyze queueing performance in single-server queues has been proposed by Bandi et al. [18]. The key idea of Robust Queueing is to replace the underlying probability law by a suitable uncertainty set, and analyze the worst case performance. The authors there relied on the discrete-time Lindley's recursion to characterize the customer waiting times as a supremum over partial sums of the interarrival times and service times. Uncertainty sets for the sequence of partial sums are proposed based on central limit theorem and two-moment partial traffic descriptions of the arrival process and service process.

Although the general idea is simple, the challenge lies in identifying proper uncertainty sets and making connection to the original queueing system. The RQ approach is further studied in Section 2, where we develop a new non-parametric formulation of the RQ algorithm.

1.2.4 Approximations Based on Non-Parametric Traffic Descriptions

As a trade-off for mathematical tractability, most approximation approaches rely on incomplete traffic descriptions. For example, approximation approaches reviewed in Section 1.2.1-1.2.3 can be characterized as parametric approaches, where the general stochastic system is mapped into one of a parametric family of more structured models. Such approaches rely on a discrete set of parameters as traffic descriptions and a key step is to understand how these parameters evolve in the network.

Another stream of research model the temporal dependence in the stochastic processes by non-parametric traffic descriptions. In Jagerman et al. [86], the authors approximate a general stationary arrival process by a Peakness Matched Renewal Stream (PMRS). The key ingredient is the peakness function, which is determined by the arrival point process and the first two moments of the service-time distribution. As discussed there, given the same service-time distribution, this peakness function is equivalent to our index of dispersion traffic description, defined in (1.1). However, they relied on a two-parameter approximation for the peakness function of a stationary point process, where the parameters are estimated by simulation.

Similar non-parametric traffic descriptions has been studied in [99; 100; 86], but they only focus on single-station single-server queues.

1.3 Main Contributions

Despite many efforts in developing more sophisticated network analyzer to address the performance impact of dependence, early approximation approaches or Monte Carlo simulation remain to be the most popular choices in applications. This is largely due to the ease of implementation. For example, [9] identified the major bottleneck in a health center appointment clinic, where they applied the QNA algorithm to approximate the system performance. [41] studied the effect of service interrupts and hospital resources pooling on patient flow times, where parametric decomposition based on Kingman's formula is applied. [87] also applied decomposition method and two-moment approximations to analyze the impact of parallelization of care on customer sojourn time. [6] integrated simulation and optimization to find the optimal staffing allocation in an emergency department unit, where they considered a network of $M_t/G/1$ queues and the stochastic objective function is estimated by simulation. [54] also studied the resource allocation problem in general stochastic networks by simulation optimization.

This research is motivated by the practical need of high fidelity modeling tool for non-Markov open queueing networks, which is easy to implement and mathematically and computationally tractable. Towards this end, we contributes to the modeling and approximation of service systems by developing (1) the theories and applications of non-parametric traffic descriptions in open queueing networks; (2) an effective Robust Queueing Network Analyzer algorithm for performance approximations in open queueing networks; and (3) extensive simulation studies to demonstrate the performance of our approximation algorithm.

1.3.1 Non-Parametric Traffic Description for Queueing Networks

In this thesis, we follow a non-parametric approach to describe the arrival and service processes, see Section 1.2.4 for a brief review of literature in this line of research.

Let A be a *stationary* counting process, e.g. the arrival counting process at a queue.

We partially characterize A by its *Index of Dispersion for Counts* (IDC), a function of non-negative real numbers $I_A : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ defined as in §4.5 of [40],

$$I_A(t) \equiv \frac{\text{Var}(A(t))}{E[A(t)]}, \quad t \geq 0. \quad (1.1)$$

As regularity conditions, we assume that $E[A(t)]$ and $\text{Var}(A(t))$ are finite for all $t \geq 0$. For renewal processes, it suffices to assume that the inter-renewal time distribution have finite second moment.

Being a function of time t , IDC captures the variability in a point process over arbitrary timescale. The reference case is a Poisson process, for which the IDC is a constant function $I_A(t) \equiv 1$. This is consistent with the well known “memoryless” property of the Poisson process. The IDC is preferred for the same reason as that for the scv, because it separates variability from the scale. In this sense, IDC can be viewed as a continuous-time generalization of scv.

In compared with traditional parametric descriptions, the IDC encodes much more information of the underlying process. For examples, if A is a renewal process, then the inter-renewal-time distribution can be fully recovered from I_{A_e} , where A_e is the equilibrium renewal process associated with A ; see Theorem 3.1.

Fendick and Whitt [60] showed basic connection between arrival IDC and the normalized workload, see (2.29). However, that work did not yield systematic approximations.

In Section 2, we show how this non-parametric traffic description can be applied to develop a new RQ formulation for the continuous-time workload process. This RQ algorithm establish a bridge between the IDC traffic description and the performance measures in a single-server queue. It serves as the building block of our network algorithm.

To the best of our knowledge, we are the first to study the non-parametric traffic descriptions in a network setting.

1.3.2 Heavy-Traffic Limits for Stationary Network Flows

As reviewed in Section 1.2, the heavy-traffic literature has focused on the system state processes such as queue length, busy time, waiting time, workload and the sojourn time processes. Another approach is to focus on the customer *flow*, which is defined as the

continuous-time process counting customers in or out of the network, or flowing from one queue to another.

In working with customer flows, three network operations become essential. First, the *departure operation* as customers flow through a service station and an arrival process transforms into a departure process. Second, the *splitting operation* as a departure process split into multiple sub-processes and feed into different subsequent queues. Third, the *superposition operation* as departure flows from different queues combine together and feed into a queue.

The customer flows are of considerable interest in general. For example, the stationary departure process from a $GI/GI/1$ queue is remarkably complicated; e.g., it is only a stationary renewal process in the special case of an $M/M/1$ model, when it is Poisson, by Burke's [32] theorem.

However, the literature on heavy-traffic limits for network flows are rather limited. The heavy-traffic limit for departure process starting empty in the $GI/GI/1$ model and more general multi-channel models is an old result, being contained in Theorem 2 of [84]. The superposition of many i.i.d. copies of general renewal processes can often be well approximated by a Poisson process, as shown in [7], but the approximation quality depends strongly on the traffic intensity. The superposition operation has also been studied in [8; 138; 124].

In this thesis, we derive new heavy-traffic limit for the stationary flows and their IDCs in open queueing networks. In particular, we evidently derive the first heavy-traffic limits for the stationary departure process and its variance function for any station in any generalized Jackson network, except for the single-station $M/M/1$ queue. We exploit the $GI/M/1$ and $M/GI/1$ results here to directly establish heavy-traffic limits for the departure variance functions in generalized Jackson networks, see Theorem 3.8 and Theorem 5.1. Our heavy-traffic limits reveal the detailed interaction between different flows, granting us powerful tools for queueing approximation.

In Section 5.1 - 5.3, we discuss how these heavy-traffic limits can be applied to develop IDC equations that describes the effect of each network operation. Together with the RQ algorithm, the IDC equations provide intuitive explanation of the performance impact

of dependence. For example, we are able to answer questions such as: When is Poisson approximation suitable for superposition arrival processes? If it fails, how do we correct for the error?

1.3.3 The Robust Queueing Network Analyzer and Simulation Studies

We exploit the new non-parametric RQ method to propose a Robust Queueing Network Analyzer for the approximation of performance measures in a single-class non-Markov open queueing network.

The algorithm decomposes the network into individual $G/G/1$ models, where the arrival process and service process at each queue is partially specified by its rate and IDC, defined in (1.1). In Section 5.4, we discuss a set of linear equations (5.46), which we refer to as the *IDC equations*, to describe the effect of each of the three basic network operations. The IDCs of the total arrival flows at each queue is approximated by the solution to the IDC equations. The RQ algorithm (2.35) is then applied to generate approximation of the mean steady-state performance measures from the IDC of the total arrival process of each $G/GI/1$ queue in the network.

Our algorithm have analytical formulations as in (2.35) and (5.48), which makes it extremely easy to implement. The computational complexity is $O(K)$ or $O(K^2)$ if we apply the feedback elimination in Section 5.5, where K is the number of stations in the system.

In Section 6, we also conduct extensive simulation experiments to evaluate the effectiveness of the new Robust Queueing Network Analyzer and compare it to previous algorithms in [134; 73; 44; 81]. Our experiments indicate that our new algorithm performs as well or better than previous algorithms.

1.4 Outline

The rest of this documents is organized as follows. In Chapter 2, we develop the continuous-time Robust Queueing formulation for $G/G/1$ queues. We also show how that RQ algorithm can be used to approximate the mean and quantile of the steady-state performance measures. As a first step in developing our Robust Queueing Network Analyzer, in Chapter 3 we

establish heavy-traffic limits for the stationary departure flow from a $GI/GI/1$ queue and develop our approximation algorithm in the setting of queues in series models. In Chapter 4, we generalize the heavy-traffic limit theorem in the previous chapter to cover all stationary flows in generalized Jackson networks. In Chapter 5 we present the full Robust Queueing Network Analyzer algorithm. We do this by developing a framework for approximating the IDCs of the flows, where we develop IDC equations for three basic network operations: (i) flow through a queue (departure), (ii) splitting and (iii) superposition. Finally, Chapter 6 collects extensive simulation studies to demonstrate the performance of our algorithm.

Chapter 2

Robust Queueing for the $G/G/1$ Model

Robust optimization is proving to be a useful approach to complex optimization problems involving significant uncertainty; e.g., see [17; 20] and references therein. In that context, the primary goal is to create an efficient algorithm to produce useful practical solutions that appropriately capture the essential features of the uncertainty. [18] have applied this approach to create a Robust Queueing theory, which can be used to generate performance predictions in complex queueing systems, including networks of queues as well as single queues. Indeed, they construct a full robust queueing analyzer (RQNA) to develop relatively simple performance descriptions like those in the QNA [134]. But their network algorithm does not provide adequate characterization of the dependence arises in queueing networks. For example, they developed a Robust Burke's Theorem, which state that their uncertainty set for the departure process is asymptotically the same as that for the arrival process, see Theorem 4 and Section 4.3 there. However, in Theorem 5.1, we see that the dependence structure in the departure process depends non-trivially on both the arrival process and the service process at that station.

We make further progress in the Robust Queueing direction. Even though we only focus on one queue in this chapter, we ultimately develop methods that apply to complex networks of queues. Queues in the context of a network exhibit complex dependence, as

discussed in Section 1.1.3. To serve as an important building block in that direction, we introduce new RQ formulation for general $G/G/1$ model, where three types of dependence are allowed: (1) dependence among interarrival time; (2) dependence among service times; and (3) dependence between interarrival and service times.

The rest of this chapter is organized as follows. In Section 2.1, after reviewing RQ for the steady-state waiting time in the single-server queue from Section 2 and Section 3.1 of [18], we develop an alternative formulation whose solution coincides with the [95] bound and is asymptotically correct in heavy-traffic. We postpone the discussion of the non-parametric RQ for waiting times to Section 2.4. In Section 2.2 we introduce new parametric and non-parametric RQ formulations for the continuous-time workload process and characterize their solutions. We also develop closed-form RQ solutions and show that the non-parametric RQ is asymptotically correct in both heavy and light traffic. In Section 2.3 we discuss algorithms to calculate or estimate the IDC functions. In Section 2.2.9 we present an illustrative simulation study that demonstrates (i) the strong impact of dependence upon performance and (ii) the value of the new RQ in capturing the impact of that dependence. For systematic numerical analysis of the RQ performance, we refer the readers to Section 6.2. Finally, Section 2.5 collects supporting functional central limit theorems, while Section 2.6 collects the proofs.

2.1 Robust Queueing for the Steady-State Waiting Time

We start by reviewing the RQ developed in Section 2 and Section 3.1 of [18], which involves separate uncertainty sets for the arrival times and service times. We then construct an alternative formulation with a single uncertainty set and show, for the $GI/GI/1$ queue, that a natural version of the RQ solution coincides with the [95] bound and so is asymptotically correct in the heavy-traffic limit. We show that both formulations provide insight into the relaxation time for the $GI/GI/1$ queue, i.e. the approximate time required to reach steady state.

In formulating RQ, we use the representation of the waiting time (before receiving service) in a general single-server queue with unlimited waiting space and the first-come

first-served (FCFS) service discipline, without imposing any stochastic assumptions. The waiting time of arrival n satisfies the Lindley recursion [101]

$$\begin{aligned} W_n &= (W_{n-1} + V_{n-1} - U_{n-1})^+ \\ &\equiv \max \{W_{n-1} + V_{n-1} - U_{n-1}, 0\}, \end{aligned} \quad (2.1)$$

where V_{n-1} is the service time of arrival $n-1$, U_{n-1} is the interarrival time between arrivals $n-1$ and n , and $\equiv$ denotes equality by definition. If we initialize the system by having an arrival 0 finding an empty system, then W_n can be represented as the maximum of a sequence of partial sums, using the Loynes [102] reverse-time construction; i.e.,

$$W_n = \max_{0 \leq k \leq n} \{S_k\}, \quad n \geq 1, \quad (2.2)$$

using reverse-time indexing with $S_0 \equiv 0$, $S_k \equiv X_1 + \cdots + X_k$ and $X_k \equiv V_{n-k} - U_{n-k}$, $1 \leq k \leq n$. We note that [18] actually look at the system time, which is the sum of an arrival's waiting time and service time. These representations are essentially equivalent.

If we extend the reverse-time construction indefinitely into the past from a fixed present state, then

$$W_n \uparrow W \equiv \sup_{k \geq 0} \{S_k\} \text{ with probability 1 as } n \rightarrow \infty,$$

allowing for the possibility that W might be infinite. For the stable stationary $G/G/1$ stochastic model with $E[U_k] < \infty$, $E[V_k] < \infty$ and $\rho \equiv E[V_k]/E[U_k] < 1$, then

$$P(W < \infty) = 1,$$

e.g., see [102] or Section 6.2 of [122].

2.1.1 Parametric RQ for Waiting Time in the $GI/GI/1$ Model

Bandi et al. [18] propose an RQ approximation for the steady-state waiting time W by performing a deterministic optimization in (2.2) subject to deterministic constraints, where we can ignore the time reversal. Treating the partial sums S_k^a of the interarrival times U_k and the partial sums S_k^s of the service times V_k separately leads to the two uncertainty sets for W

$$\begin{aligned} \mathcal{U}^a &\equiv \{\mathbf{U} \in \mathbb{R}^\infty : S_k^a \geq km_a - b_a\sqrt{k}, k \geq 0\} \quad \text{and} \\ \mathcal{U}^s &= \{\mathbf{V} \in \mathbb{R}^\infty : S_k^s \leq km_s + b_s\sqrt{k}, k \geq 0\}, \end{aligned} \quad (2.3)$$

where $\mathbf{U} \equiv \{U_k : k \geq 1\}$ and $\mathbf{V} \equiv \{V_k : k \geq 1\}$ are arbitrary sequences of real numbers in $\mathbb{R}^\infty$, $S_0^a = S_0^s = 0$, $S_k^a \equiv U_1 + \cdots + U_k$ and $S_k^s \equiv V_1 + \cdots + V_k$, $k \geq 1$, and m_a, m_s, b_a and b_s are parameters to be specified. The constraints in (2.3) are one sided because that is what is required to bound the waiting times above, as we can see from (2.1) and (2.2). Thus, the RQ optimization can be expressed as

$$W^* \equiv \sup_{\tilde{U} \in \mathcal{U}^a} \sup_{\tilde{V} \in \mathcal{U}^s} \sup_{k \geq 0} \{S_k^s - S_k^a\}. \quad (2.4)$$

where S_k^a (S_k^s) is a function of $\mathbf{U}$ ($\mathbf{V}$) specified above. Versions of this formulation in (2.4) and others in this paper also apply to the transient waiting time W_n , but we will focus on the steady-state waiting time.

Thinking of the general stationary $G/G/1$ stochastic model, where the distributions of U_k and V_k are independent of k (but stochastic independence is not assumed), [18] assume that $1/\lambda \equiv m_a \equiv E[U_k]$ and $1/\mu \equiv m_s \equiv E[V_k]$ and assume that $\mu > \lambda$, so that the traffic intensity $\rho \equiv \lambda/\mu < 1$. The square-root terms in the constraints in (2.3) are motivated by the central limit theorem (CLT). Thinking of the $GI/GI/1$ model in which the interarrival times U_k and service times V_k come from independent sequences of independent and identically distributed (i.i.d.) random variables with finite variances σ_a^2 and σ_s^2 , the CLT suggests that $b_a = \beta_a \sigma_a$ and $b_s = \beta_s \sigma_s$ for some positive constants β_a and β_s , perhaps with $\beta = \beta_a = \beta_s$. With this choice, these new parameters measure the number of standard deviations away from the mean in a Gaussian approximation. [18] also provided an extension to cover the heavy-tailed case, where finite variances might not exist; then $\sqrt{k}$ in (2.3) is replaced by $k^{1/\alpha}$ for $0 < \alpha \leq 2$, as we would expect from Section 4.5, 8.5 and 9.7 of [143], but we will not discuss that extension here.

From (2.1), it is evident that the waiting times depend on the service times and interarrival times only through their difference X_n . Thus, instead of the two uncertainty sets in (2.3), we propose the single uncertainty set

$$\mathcal{U}_p^x \equiv \{\mathbf{X} \in \mathbb{R}^\infty : S_k^x \leq -mk + b_x \sqrt{k}, k \geq 0\}, \quad (2.5)$$

where $\mathbf{X} \equiv \{X_k : k \geq 1\} \in \mathbb{R}^\infty$, $S_0^x \equiv 0$ and $S_k^x \equiv X_1 + \cdots + X_k$, $k \geq 1$, $m = m_a - m_s$, while b_x is constant parameter to be specified. The subscript “ p ” in $\mathcal{U}_p^x$ indicates that this

is a uncertainty set for a parametric RQ formulation, because we are using the parameters (m, b_x) to characterize the stochastic system. Combining (2.2) and (2.5), we obtain the alternative RQ optimization

$$W^* \equiv \sup_{\mathbf{X} \in \mathcal{U}_p^x} \sup_{k \geq 0} \{S_k^x\}. \quad (2.6)$$

where S_k^x is the function of $\mathbf{X}$ specified above.

The RQ formulations in (2.4) and (2.6) are attractive because the optimization problems have simple solutions in which all constraints are satisfied as equalities. That follows easily from the fact that W is a nondecreasing (nonincreasing) function of V_k (of U_k) for all k . The simple closed-form solution follows from the triangular structure of the equations; see Section 3.1 of [18]. The following is a direct extension of Theorem 2 of [18] to include the new RQ formulation in (2.6). The final statement involves an interchange of suprema, which is justified by Lemma 2.1.

Theorem 2.1 (RQ solutions for the steady-state waiting time) *The RQ optimizations (2.4) with $m_a > m_s > 0$ and (2.6) with $m > 0$ have the solution*

$$\begin{aligned} W^* &= \sup_{k \geq 0} \{-mk + b\sqrt{k}\} \\ &\leq \sup_{x \geq 0} \{-mx + b\sqrt{x}\} = -mx^* + b\sqrt{x^*} = \frac{b^2}{4m} \quad \text{for } x^* = \frac{b^2}{4m^2}. \end{aligned} \quad (2.7)$$

For (2.4), $b \equiv b_s + b_a$; for (2.6), $b \equiv b_x$. In (2.7), W^* is maximized at one of the integers immediately above or below x^* .

We now establish implications for the $GI/GI/1$ and general stationary $G/G/1$ models. Since the CLT underlies the heavy-traffic limit theory as well as the RQ formulation, it should not be surprising that we can make strong connections to heavy-traffic approximations. The new formulation in (2.6) is attractive because, with a natural choice of the constant b_x there, it matches the [95] bound for the mean steady-state wait $E[W]$ in the $GI/GI/1$ stochastic model and so is asymptotically correct in heavy-traffic, whereas that is not the case for (2.4) with a natural choice of b . To quantify the variability independent of the scale, let $c_a^2 \equiv \text{Var}(U_1)/(E[U_1])^2 = \rho^2 \mu^2 \sigma_a^2$ and $c_s^2 \equiv \text{Var}(V_1)/(E[V_1])^2 = \mu^2 \sigma_s^2$ be the *squared coefficients of variation* (scv's).

Corollary 2.1 (*RQ yields the Kingman bound for GI/GI/1*) In the setting of (2.6), if we let $b_x \equiv \beta\sqrt{\text{Var}(X_1)}$ and $\beta \equiv \sqrt{2}$, then $b_x = \sqrt{2(c_s^2 + \rho^{-2}c_a^2)}/\mu^2$ for the GI/GI/1 model with traffic intensity ρ , so that

$$W^* \equiv W^*(\rho) = \frac{\text{Var}(X_1)}{2|E[X_1]|} = \frac{\rho(c_s^2 + \rho^{-2}c_a^2)}{2\mu(1-\rho)}, \quad (2.8)$$

which is the upper bound for $E[W]$ in Theorem 2 of [95], so that $\mu(1-\rho)W^*(\rho) \rightarrow (c_a^2 + c_s^2)/2$ for any $\mu > 0$, as $\rho \uparrow 1$, which supports the heavy-traffic approximation $W^*(\rho) \approx \rho(c_a^2 + c_s^2)/2\mu(1-\rho)$, just as for $E[W]$ in the stochastic model.

Remark 2.1 In the setting of (2.4) as in [18], if we let $b_s \equiv \beta\sqrt{\text{Var}(V_1)}$ and $b_a \equiv \beta\sqrt{\text{Var}(U_1)}$, then we obtain $b = b_s + b_a = \beta(c_s + \rho^{-1}c_a)/\mu$ instead of $b = \sqrt{b_s^2 + b_a^2} = \beta\sqrt{c_s^2 + \rho^{-2}c_a^2}/\mu$, as needed. The difference between the RQ solutions for (2.4) and (2.6) can have serious implications for approximations; e.g., if $c_a^2 = c_s^2 = x$, then $(c_a^2 + c_s^2)/2 = x$, while $(c_a + c_s)^2/2 = 2x$, a factor of 2 larger. Hence, if we apply (2.4) with $b_a = b_s$ to the simple M/M/1 queues, one is forced to have a 100% error in heavy traffic. These two coincide only when at least one of b_a and b_s is 0, i.e., in D/GI/1 or GI/D/1 models, and the percentage error is the largest when service times and arrival times have the same variability. Fortunately, robust optimization has flexibility that makes it possible to circumvent the difficulties in the form of the optimization in (2.4). For example, [18] use statistical regression in their Section 7 to refine their solution to (2.4). Of course, such refinements complicate algorithms. ■

These RQ formulations provide insight into the rate of approach to steady state for the GI/GI/1 model, as captured by the relaxation time; see Section III.7.3 of [38] and Section XIII.2 of [12]. For RQ, steady state is achieved at a fixed time, whereas in the stochastic model steady state is approached gradually, with the error $|E[W_n] - E[W]|$ typically being of order $O(n^{-3/2}e^{-n/r})$ as $n \rightarrow \infty$, where $r \equiv r(\rho)$ is called the relaxation time.

Corollary 2.2 (*relaxation time for the GI/GI/1 queue*) With both (2.4) and (2.6), the place where the RQ supremum is attained is $x^*(\rho) = O((1-\rho)^{-2})$ as $\rho \uparrow 1$, which is the same order as the relaxation time in the GI/GI/1 model.

2.1.2 Non-Parametric RQ for Waiting Time in the $G/G/1$ Model

The RQ problems in (2.4) and (2.6) can be considered instances of a *parametric RQ*, because they depend on the stochastic model only through a finite number of parameters, in particular, (m_a, m_s, b_a, b_s) in (2.4) and (m, b_x) in (2.6).

We can expose the impact of dependence among the interarrival times and service times on the steady-state waiting time in the general stationary $G/G/1$ model as a function of the traffic intensity ρ by introducing a new *non-parametric RQ* formulation. With the $G/G/1$ model, we assume stationarity, so that there is a well defined steady state, but we allow dependence among the interarrival times and service times.

To treat the $G/G/1$ model, we replace the uncertainty set in (2.5) by

$$\mathcal{U}_f^x \equiv \{\mathbf{X} : S_k^x \leq E[S_k^x] + b'_x \sqrt{\text{Var}(S_k^x)}, \quad k \geq 0\}, \quad (2.9)$$

and similarly for the two constraints in (2.4). The name non-parametric RQ here may be confusing, because the uncertainty set (2.9) depends on a sequence of parameters $\{\text{Var}(S_k^x) : k \in \mathbb{N}\}$. Nevertheless, we use “non-parametric” here to keep the name consistent with the one we use for our RQ formulation for the workload process, where the uncertainty set indeed depends on a non-parametric continuous-time function, see (2.16).

For the $GI/GI/1$ model, the new uncertainty set (2.9) is essentially equivalent to the previous one in (2.5), but they can be very different with dependence.

It is significant that there are CLT’s to motivate the form of the constraints in (2.9), just as there are in the i.i.d. case underlying (2.5). These supporting CLT’s are reviewed in Section 2.5. The CLT supports the spatial scaling by $\sqrt{\text{Var}(S_k)}$ instead of $\sqrt{k}$, as we show in Section 2.5.3. Of course, the non-parametric RQ produces a more complicated optimization problem, but it is potentially more useful, in part because it too can be analyzed.

For brevity, we discuss this non-parametric RQ for the waiting time in Section 2.4, see (2.72) and (2.73). For the rest of this thesis, we focus on developing such a non-parametric RQ formulation for the continuous-time workload. As discovered in [60], it is convenient to focus on the steady-state workload when we want to expose the performance impact of the dependence among interarrival times and service times.

2.2 Robust Queueing for the Steady-State Workload

We now develop both a parametric and a non-parametric RQ formulations for the continuous-time workload in the single-server queue, paralleling (2.9).

The workload at time t is the amount of unfinished work in the system at time t ; it is also called the virtual waiting time because it represents the waiting time a hypothetical arrival would experience at time t . The workload is more general than the virtual waiting time because it applies to any work-conserving service discipline. We consider the workload primarily because it can serve as a convenient more tractable alternative to the waiting time.

We start by developing a reverse-time representation of the workload process in Section 2.2.1, paralleling (2.2). Then we develop the RQ formulations and give their solutions, which closely parallel Theorem 2.1. We then show that natural versions of the RQ formulations for the workload are exact for the $M/GI/1$ model and are asymptotically correct in both light traffic and heavy-traffic for the general stationary $G/G/1$ model.

2.2.1 The Workload Process and Its Reverse-Time Representation

As before, we start with a sequence $\{(U_k, V_k)\}$ of interarrival times and service times. The *arrival counting process* can be defined by

$$A(t) \equiv \max \{k \geq 1 : U_1 + \cdots + U_k \leq t\} \quad \text{for } t \geq U_1 \quad (2.10)$$

and $A(t) \equiv 0$ for $0 \leq t < U_1$, while the *total input of work* over $[0, t]$ and the *net-input process* are, respectively,

$$Y(t) \equiv \sum_{k=1}^{A(t)} V_k \quad \text{and} \quad N(t) \equiv Y(t) - t, \quad t \geq 0. \quad (2.11)$$

The workload at time t , starting empty at time 0, is the reflection map Ψ applied to N , i.e.,

$$Z(t) = \Psi(N)(t) \equiv N(t) - \inf_{0 \leq s \leq t} \{N(s)\}, \quad t \geq 0. \quad (2.12)$$

For illustration, Figure 2.1 shows an example of the net-input process $N(t)$ and the associated lower regulator $\inf_{0 \leq s \leq t} \{N(s)\}$. The workload process is exactly the difference between the two curves.

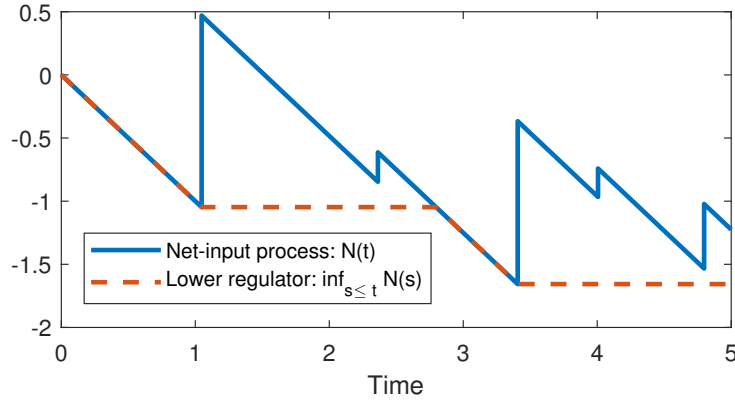


Figure 2.1: The net-input process.

As in Section 6.3 of [122], we again use a reverse-time construction to represent the workload in a single-server queue as a supremum, so that the RQ optimization problem becomes a maximization over constraints expressed in an uncertainty set, just as before, but now it is a continuous optimization problem. Using the same notation, but with a new meaning, let $Z(t)$ be the workload at time 0 of a system that started empty at time $-t$. Then $Z(t)$ can be represented as

$$Z(t) \equiv \sup_{0 \leq s \leq t} \{N(s)\}, \quad t \geq 0, \quad (2.13)$$

where N is defined in terms of Y as before, but Y is interpreted as the total work in service time to enter over the interval $[-s, 0]$. That is achieved by letting V_k be the k^{th} service time indexed going backwards from time 0 and $A(s)$ counting the number of arrivals in the interval $[-s, 0]$. Paralleling the waiting time in Section 2.1, $Z(t)$ increases monotonically as a function of t , there exist a Z (possibly infinite) such that

$$Z = \lim_{t \rightarrow \infty} Z(t) \quad (2.14)$$

For the stable stationary $G/G/1$ stochastic queue, Z corresponds to the steady-state workload and satisfies $P(Z < \infty) = 1$; see Section 6.3 of [122].

2.2.2 Parametric and Non-Parametric RQ for the Workload

Just as in Section 2.1, to create appropriate RQ formulations for the steady-state workload, it is helpful to have a reference stochastic model, which can be the stable stationary $G/G/1$

model, where such a steady-state workload is well defined.

In continuous time, we need to work with continuous-time stationarity instead of discrete-time stationarity; e.g., see [122]. Hence, we assume that $\{(A(t), Y(t)) : t \geq 0\}$ is a stationary process with $E[A(t)] = \lambda t$ and $E[Y(t)] = \rho t$ for all $t \geq 0$.

In the ordinary stochastic queueing model, $N(t)$ in (2.13) is a stochastic process and hence $Z(t)$ is a random variable. However, in Robust Queueing practice, $N(t)$ is viewed as a deterministic instance drawn from a pre-determined uncertainty set $\mathcal{U}$ of input functions, while the workload Z^* for a Robust Queue is regarded as the worst case workload over the uncertainty set, i.e.

$$Z^* \equiv \sup_{\tilde{N} \in \mathcal{U}} \sup_{x \geq 0} \{\tilde{N}(x)\}.$$

In our specific settings, we have the following uncertainty sets motivated from central limit theorem (CLT)

$$\mathcal{U}_p \equiv \left\{ \tilde{N} : \mathbb{R}^+ \rightarrow \mathbb{R} : \tilde{N}(s) \leq E[N(s)] + b_p \sqrt{\rho s / \mu}, s \geq 0 \right\}, \quad (2.15)$$

$$\mathcal{U} \equiv \left\{ \tilde{N} : \mathbb{R}^+ \rightarrow \mathbb{R} : \tilde{N}(s) \leq E[N(s)] + b \sqrt{\text{Var}(N(s))}, s \geq 0 \right\}, \quad (2.16)$$

$\mathcal{U}_p$ and $\mathcal{U}$ serve as the uncertainty sets for the parametric and non-parametric RQ, respectively. Here we regard $\tilde{N} \equiv \{\tilde{N}(s) : 0 \leq s \leq t\}$ as an arbitrary real-valued function on $\mathbb{R}^+ \equiv [0, \infty)$, while we regard $N \equiv \{N(s) : s \geq 0\}$ as the underlying stochastic process, and $\{\text{Var}(N(s)) : s \geq 0\} = \{\text{Var}(Y(s)) : s \geq 0\}$ as its variance-time function, which can either be calculated for a stochastic model or estimated from simulation or system data; see Section 2.3. In (2.16), b and b_p are parameters to be specified. Recall that, in working with continuous-time stationarity, we assume that $N(t)$ is the net input process associated with the stationary processes in the stochastic queue, so

$$E[N(t)] = E[Y(t) - t] = \rho t - t.$$

Paralleling Section 2.1, the uncertainty sets (2.15) and (2.16) are motivated by a CLT, but here for $Y(t)$ (and thus for $N(t)$), which we review in Section 2.5.3; in particular, see (2.88) and (2.90).

Remark 2.2 (choosing the parameters b_p and b) *The parameters b_p and b in (2.15) and (2.16) add a degree of freedom in the algorithm. In Section 2.2.6, we exploit the choice*

of b to develop quantile approximations for the steady-state workload processes. For the approximation of the mean, based on Corollary 2.3 below, we will let $b = \sqrt{2}$.

Paralleling Section 2.1, the associated RQ formulations are

$$Z_p^* \equiv \sup_{\tilde{N} \in \mathcal{U}_p} \sup_{s \geq 0} \left\{ \tilde{N}(s) \right\} \quad \text{and} \quad Z^* \equiv \sup_{\tilde{N} \in \mathcal{U}} \sup_{s \geq 0} \left\{ \tilde{N}(s) \right\}, \quad (2.17)$$

for $\mathcal{U}_p$ and $\mathcal{U}$ defined in (2.15) and (2.16), respectively. The same reasoning as before yields the following analog of Theorem 2.1. The proof can be found in Section 2.6.

Theorem 2.2 (*RQ solutions for the workload*) *The solutions of the RQ optimization problems in (2.17) are*

$$Z_p^* = -(1 - \rho)s^* + b_p \sqrt{\rho s^* / \mu} = \frac{\rho b_p^2}{4\mu(1 - \rho)} \quad \text{for} \quad s^* \equiv s^*(\rho) = \frac{\rho b_p^2}{4\mu(1 - \rho)^2} \quad (2.18)$$

and

$$Z^* = \sup_{s \geq 0} \left\{ -(1 - \rho)s + b \sqrt{\text{Var}(Y(s))} \right\}. \quad (2.19)$$

We immediately obtain the following corollary, which states that the RQ formulation in (2.17) yields the exact mean steady-state workload for the $M/GI/1$ model.

Corollary 2.3 (*exact for $M/GI/1$*) *For the $M/GI/1$ model, the total input process $\{Y(t) : t \geq 0\}$ is a compound Poisson process with $E[Y(t)] = \rho t$ and $\text{Var}(Y(t)) = \rho t(1 + c_s^2)/\mu$, so that $Z^* = Z_p^*$ if $b_p^2 = b^2(1 + c_s^2)$. If, in addition, $b \equiv \sqrt{2}$, then*

$$Z^* = Z_p^* = \frac{\rho(1 + c_s^2)}{2\mu(1 - \rho)} = E[Z], \quad (2.20)$$

where $E[Z]$ is the mean steady-state workload in the $M/GI/1$.

This corollary suggests a natural choice of b_p and b in (2.15) and (2.16).

The Variance-Time Function for the Total Input Process. For further progress, we focus on the variance-time function $\text{Var}(Y(t))$ in (2.19). As regularity conditions for $Y(t)$, we assume that $V_Y(t) \equiv \text{Var}(Y(t))$ is differentiable with derivative $\dot{V}_Y(t)$ having finite positive limits as $t \rightarrow \infty$ and $t \rightarrow 0$, i.e.,

$$\dot{V}_Y(t) \rightarrow \sigma_Y^2 \quad \text{as} \quad t \rightarrow \infty \quad \text{and} \quad \dot{V}_Y(t) \rightarrow \dot{V}_Y(0) > 0 \quad \text{as} \quad t \rightarrow 0, \quad (2.21)$$

for an appropriate constant σ_Y^2 . These assumptions are known to be reasonable; see Section 4.5 of [40], [60] and Section 2.3.

A common case in models for applications is to have positive dependence in the input process Y , which holds if

$$\text{Cov}(Y(t_2) - Y(t_1), Y(t_4) - Y(t_3)) \geq 0 \quad \text{for all} \quad 0 \leq t_1 < t_2 \leq t_3 < t_4. \quad (2.22)$$

Negative dependence holds if the inequality is reversed. These are strict if the inequality is a strict inequality. From (17) and (18) of Section 4.5 in [40], which is restated in (48) and (49) of [60], with positive (negative) dependence, under appropriate regularity conditions, $\dot{V}_Y(t) \geq 0$ and $\ddot{V}_Y(t) \geq (\leq) 0$.

Remark 2.3 (*example of negative dependence*) *Negative dependence in Y occurs if greater input in one interval tends to imply less input in another interval. Such negative dependence occurs when there is a specified number of arrivals in a long time interval, as in the $\Delta_{(i)}/GI/1$ model, where the arrival times (not interarrival times) are i.i.d. over an interval; see [80]. This phenomenon can also occur in queues with arrivals by appointment, where there are i.i.d. deviations about deterministic appointment times; e.g., see [94].*

Theorem 2.3 (*RQ exposing the impact of the dependence*) *Consider the non-parametric RQ optimization for the steady-state workload in the general stationary G/G/1 queue with $\rho < 1$ formulated in (2.17) and solved in (2.19). Assume that (5.8) holds for the variance-time function $V_Y(t)$.*

(a) *For each ρ , $0 < \rho < 1$, there exists (possibly not unique) $x^* \equiv x^*(\rho)$, such that a finite maximum is attained at x^* for all $t \geq s^*$. In addition, $0 < s^* < \infty$ and s^* satisfies the equation*

$$(1 - \rho) = \dot{h}(s) \quad \text{where} \quad h(s) \equiv b\sqrt{V_Y(s)}. \quad (2.23)$$

The time s^ is unique if $h(s)$ is strictly concave or strictly convex, i.e., if $\dot{h}(s)$ is strictly increasing or strictly decreasing.*

(b) *If there is positive (negative) dependence, as in (2.22) (with sign reversed), the variance function $V_Y(s)$ is convex (concave), so that the function $h(s)$ is concave. Moreover, a strict inequality is inherited. Thus, there exists a unique solution to the RQ if there is*

strict positive dependence or strict negative dependence. Moreover, the optimal time $s^*(\rho)$ is strictly increasing in ρ , approaching ∞ as $\rho \uparrow 1$.

2.2.3 The Indices of Dispersion for Counts and Work

The workload process is not only convenient because it leads to the continuous RQ optimization problem in (2.17) with solution in (2.19). It is also convenient to relate the variances of the arrival counting process $A(s)$ and the cumulative work input process $Y(s)$ to associated continuous-time indices of dispersion, studied in [60] and [59].

Consider a general single-server queue with a general arrival process A , i.e. $A(t)$ counts the total number of arrival before time t . The IDC defined in (1.1) is a continuous-time function associated with A . Being the variance function scaled by the mean function, the IDC exposes the variability over time, independent of the scale. For this reason, the IDC can be viewed as a continuous-time generalization of the squared coefficient of variation of a nonnegative random variable, i.e. the variance divided by the square of the mean. The IDC captures how the covariance in a point process changes over time, which extends the natural practice of including lag- k covariances in modeling the dependence in a point process.

The reference case is a Poisson arrival process, for which $I_A(t) = 1$, $t \geq 0$. However, for general arrival processes, the IDC is more complicated. Even the IDC for a deterministic D arrival process is complicated, because the IDC is for the stationary version of the arrival process, which lets the initial point be uniformly distributed over the constant interarrival time.

Similar to the IDC, the *Index of Dispersion for Work* (IDW) describes the variability associated with the cumulative input process Y in (2.11). The IDW is defined as in (1) of [60] by

$$I_w(t) \equiv \frac{\text{Var}(Y(t))}{E[V_1]E[Y(t)]}, \quad t \geq 0. \quad (2.24)$$

The IDW captures the variability of service requirement brought to the system as a function of time t . By re-arranging terms, we have the following representation of the variance function $V_Y(t)$ of the net-input process Y

$$V_Y(t) \equiv \text{Var}(Y(t)) = I_w(t)E[V_1]E[Y(t)] = \rho t I_w(t)/\mu, \quad t \geq 0. \quad (2.25)$$

We will see in Section 2.2.5 that it is convenient to express our non-parametric RQ (2.17) in terms of the IDW.

Since we are interested in the steady-state performance of the OQN, we assume that the processes A and Y have stationary increments. Given that arrival process and service times have constant determined rates, the mean functions $E[A(t)]$ and $E[Y(t)]$ are linear in time. Hence, much of the remaining behavior of the A and Y is determined by the variance-time function or index of dispersion. We are interested in the variance-time *function*, because it captures the dependence through the covariances; the processes (A, Y) have independent increments for the $M/GI/1$ model, but otherwise not.

To connect the IDC to the IDW, consider the special case where the service times V_i are i.i.d, independent of the arrival process A . The conditional variance formula gives a useful decomposition of the IDW

$$I_w(t) = I_a(t) + c_s^2, \quad t \geq 0, \quad (2.26)$$

where $I_a \equiv I_A$ denote the arrival IDC and $c_s^2 = \text{Var}(V_i)/E[V_i]^2$ is the scv of the service-time distribution.

The IDC and IDW are particularly convenient because of a more elementary time scaling convention.

Remark 2.4 (*time scaling convention*) Consider $A(t)$ with rate-1 and $A_\lambda(t) \equiv A(\lambda t)$ with rate- λ . Let $I_A(t)$ denote the IDC of $A(t)$, then we have

$$I_{A_\lambda}(t) = \frac{\text{Var}(A(\lambda t))}{E[A(\lambda t)]} = I_A(\lambda t).$$

For the IDW $I_{w,\lambda}(t)$ associated with $A_\lambda(t)$ and $\{V_i, i \in \mathbb{N}\}$, let $Y_\lambda(t) = \sum_{i=1}^{A_\lambda(t)} V_i = Y(\lambda t)$, then we have

$$I_{w,\lambda}(t) \equiv \frac{\text{Var}(Y_\lambda(t))}{E[V_1]E[Y_\lambda(t)]} = \frac{\text{Var}(Y(\lambda t))}{E[V_1]E[Y(\lambda t)]} \equiv I_w(\lambda t).$$

Remark 2.5 An important case for A is the renewal process; to have stationary increments, we assume that it is the equilibrium renewal process, as in Section 3.5 of [116]. Then $\text{Var}(A(t))$ can be expressed in terms of the renewal function, which in turn can be related to the interarrival-time distribution and its transform. The explicit formulas for

renewal processes appear in (14), (16) and (18) in Section 4.5 of [39]. The required Numerical transform inversion for the renewal function is discussed in Section 13 of [4]. The hyperexponential (H_2) and Erlang (E_2) special cases are described in Section III.G of [60].

It is also possible to carry out similar analyses for much more complicated arrival processes. [108] applies matrix-analytic methods to give explicit representations of the variance $\text{Var}(A(t))$ for the versatile Markovian point process or Neuts process; see Section 5.4, especially Theorem 5.4.1. Explicit formulas for the Markov modulated Poisson process (MMPP) are given on pp. 287-289.

All of these explicit formulas above have the asymptotic form

$$\text{Var}(A(t)) = \sigma_A^2 t + \zeta + O(e^{-\gamma t}) \quad \text{as } t \rightarrow \infty,$$

for σ_A^2 in (2.75). In terms of the IDC, let $c_A^2 = \sigma_A^2/\lambda$, we have

$$I_A(t) = c_A^2 + \zeta/\lambda t + O(e^{-\gamma t}/t) \quad \text{as } t \rightarrow \infty.$$

2.2.4 The Indices of Dispersion and the Mean Steady-State Workload

The IDC and IDW are important because of their close connection to the mean steady-state workload $E[Z]$. Recall that, the workload $Z(t)$ converges to the steady-state workload Z as t increases to infinity under regularity conditions, see (2.14). In [60] it was shown that the IDW I_w is intimately related to a scaled mean workload $c_Z^2(\rho)$, defined by

$$c_Z^2(\rho) \equiv \frac{E[Z_\rho]}{E[Z_\rho; M/D/1]}, \quad (2.27)$$

where $E[Z_\rho; M/D/1]$ is the mean steady-state workload in a M/D/1 model given by

$$E[Z_\rho; M/D/1] = \frac{E[V_1]\rho}{2(1-\rho)} = \frac{\rho}{2\mu(1-\rho)}. \quad (2.28)$$

As (2.28) suggests, the mean steady-state workload converges to 0 as $\rho \downarrow 0$ and diverges to infinity as $\rho \uparrow 1$. The normalization in (2.27) exposes the impact of variability separately from the traffic intensity. Under regularity conditions, the following finite positive limits exist and are equal:

$$\begin{aligned} \lim_{t \rightarrow \infty} \{I_w(t)\} &\equiv I_w(\infty) = c_Z^2(1) \equiv \lim_{\rho \rightarrow 1} \{c_Z^2(\rho)\}, \quad \text{and} \\ \lim_{t \rightarrow 0} \{I_w(t)\} &\equiv I_w(0) = c_Z^2(0) \equiv \lim_{\rho \rightarrow 0} \{c_Z^2(\rho)\}; \end{aligned} \quad (2.29)$$

see [60] and Section 2.5.5.

The reference case is the classical $M/GI/1$ queue, for which we have

$$c_Z^2(\rho) = 1 + c_s^2 = I_w(t) \quad \text{for all } 0 < \rho < 1, t \geq 0. \quad (2.30)$$

In great generality, we have

$$c_Z^2(0) = 1 + c_s^2 = I_w(0) \quad \text{and} \quad c_Z^2(1) = c_A^2 + c_s^2 = I_w(\infty), \quad (2.31)$$

where c_A^2 is the asymptotic variability parameter, i.e., the normalization constant in the functional central limit theorem (FCLT) for the arrival process. For a renewal process, c_A^2 coincides with the scv c_a^2 of an interarrival time.

The limits in (2.31) implies that, when c_A^2 is not nearly 1, $c_Z^2(\rho)$ varies significantly as a function of ρ . Hence, the impact of the variability in the arrival process upon the queue performance clearly depends on the traffic intensity. This important insight from [60] is the starting point for our analysis. In well-behaved models, $c_Z^2(\rho)$ as a function of ρ and $I_w(t)$ as a function of t tend to change smoothly and monotonically between those extremes, but OQN's can produce more complex behavior when both the traffic intensities at the queues and the levels of variability in the arrival and service processes at different queues vary. We illustrate in Section 2.2.9.

The challenge is to relate $c_Z^2(\rho)$ to the IDW $I_w(t)$ for $0 < \rho < 1$ and $t \geq 0$. As observed by [60], a simple connection would be $c_Z^2(\rho) \approx I_w(t_\rho)$ for some increasing function t_ρ , reflecting that the impact of the dependence among the interarrival times and service times has impact on the performance of a queue over some time interval $[0, t_\rho]$, where t_ρ should increase as ρ increases. The extreme cases are supported by (2.29), but we want more information about the cases in between.

The following theorem reveals more connections between the IDC and the mean steady-state workload.

Theorem 2.4 (Ordering of the mean steady-state workload in $GI/M/1$ models)

Consider two $GI/M/1$ queues with rate-1 arrival processes A_1 and A_2 , respectively. Let the service times be i.i.d. exponential random variables with mean $1/\mu = \rho$ so that the traffic intensity is ρ . Let I_{a_i} denote the IDC's associated with the arrival process A_i , for $i = 1, 2$.

Suppose we have

$$I_{a_1}(t) \geq I_{a_2}(t), \quad \text{for } t \geq 0,$$

then we have

$$E[Z_{1,\rho}] \geq E[Z_{2,\rho}], \quad \text{for } \rho \in (0, 1), \quad (2.32)$$

where $E[Z_{i,\rho}]$ is the mean steady-state workload in the i -th model, with traffic intensity ρ .

Proof. For $GI/M/1$ queue, the mean steady-state workload can be written as

$$E[Z] = \frac{\rho}{\mu(1 - \sigma)},$$

where σ is the unique root in $(0, 1)$ of the equation $\hat{f}(\mu(1 - \sigma)) = \sigma$, where $\hat{f}$ is the Laplace transform of the interarrival-time distribution, see Section 3.1.

With a change of variable $s = \mu(1 - \sigma)$, this is equivalent to finding the unique root in $(0, \rho^{-1})$ of

$$\hat{f}(s) = 1 - \rho/s.$$

Combining with (3.8), we can re-write the equation in terms of the Laplace transform of the variance function $V(t) = \text{Var}(A(t))$ as

$$\hat{V}(s) = \frac{2(1 - \rho)}{\rho s^3} - \frac{1}{s^2}. \quad (2.33)$$

Note that the right-hand side of (2.33) is positive on $(0, 2(1 - \rho)/\rho)$ and negative on $(2(1 - \rho)/\rho, \infty)$. Since the Laplace transform $\hat{V}(s)$ is positive for all s , the root must fall in $(0, 2(1 - \rho)/\rho)$, where the right-hand side of (2.33) is decreasing.

Now, fix a ρ and consider the variance function $V_i(t) = I_{a_i}(t)t$ induced by the IDC function I_{a_i} , we have

$$\begin{aligned} \hat{V}_1(s) &\equiv \int_0^\infty V_1(t)e^{-st}dt \\ &\geq \int_0^\infty V_2(t)e^{-st}dt \equiv \hat{V}_2(s). \end{aligned}$$

This implies that we must have $s_1^* \leq s_2^*$, where s_i^* is the root of (2.33) with $\hat{V}_i$. Otherwise, assume that $s_1^* > s_2^*$, we have

$$\frac{2(1 - \rho)}{\rho(s_1^*)^3} - \frac{1}{(s_1^*)^2} = \hat{V}_1(s_1^*) \geq \hat{V}_2(s_1^*) > \hat{V}_2(s_2^*) = \frac{2(1 - \rho)}{\rho(s_2^*)^3} - \frac{1}{(s_2^*)^2},$$

which contradicts the fact that right-hand side of (2.33) is decreasing on $(0, 2(1 - \rho)/\rho)$. Since $E[Z_i] = \rho/s_i^*$, we have $E[Z_{1,\rho}] \geq E[Z_{2,\rho}]$. ■

2.2.5 Robust Queueing with the IDW

To obtain more information, RQ can help. Using (2.25), we express the solution in (2.19) as

$$\begin{aligned} Z^* \equiv Z^*(b) &= \sup_{s \geq 0} \left\{ -(1 - \rho)s + b\sqrt{\text{Var}(Y(s))} \right\} \\ &= \sup_{s \geq 0} \left\{ -(1 - \rho)s + b\sqrt{\rho s I_w(s)/\mu} \right\} \end{aligned} \quad (2.34)$$

Our algorithm will exploit the one-dimensional optimization problem in (2.34), which is easy to solve given the IDW $I_w(x)$. In terms of the IDC, if we assume that the service times are i.i.d. and independent of the arrival process A , then by (2.26), we have

$$Z^* \equiv Z^*(b) = \sup_{s \geq 0} \left\{ -(1 - \rho)s + b\sqrt{\rho s (I_a(s) + c_s^2)/\mu} \right\}, \quad (2.35)$$

where $I_a \equiv I_A$ denote the IDC of the arrival process A . We will discuss methods of estimating and calculating the IDC and the IDW in Section 2.3.

To further relate the RQ solution in (2.34) to the steady-state workload in the G/G/1 queue, we define an RQ analog of the normalized mean workload in (2.27), in particular,

$$c_{Z^*}^2(\rho) \equiv \frac{2\mu(1 - \rho)Z_\rho^*}{\rho}. \quad (2.36)$$

The RQ approach allows us to establish versions of the variability fixed-point equation suggested in (9), (15) and (127) of [60].

Theorem 2.5 *Any optimal solution of the RQ in (2.34) is attained at s^* , which satisfies the equation*

$$s^* = \frac{b^2}{2} \cdot \frac{\rho I_w(s^*)}{2\mu(1 - \rho)^2} \left(1 + \frac{s^* \dot{I}_w(s^*)}{I_w(s^*)} \right)^2. \quad (2.37)$$

The associated RQ optimal workload in (2.34) can be expressed as

$$Z^* = \frac{b^2}{2} \cdot \frac{\rho I_w(s^*)}{2\mu(1 - \rho)} \left(1 - \left(\frac{s^* \dot{I}_w(s^*)}{I_w(s^*)} \right)^2 \right), \quad (2.38)$$

which is a valid nonnegative solution provided that $x^* \dot{I}_w(x^*) \leq I_w(x^*)$. If $b = \sqrt{2}$, then the associated scaled RQ workload satisfies

$$c_{Z^*}^2(\rho) = I_w(s^*) \left(1 - \left(\frac{s^* \dot{I}_w(s^*)}{I_w(s^*)} \right)^2 \right). \quad (2.39)$$

Proof. Note that $\rho s I_w(s)/\mu = V(s)$. Because we have assumed that $V_Y(s)$ is differentiable, so is I_w . We obtain (3.14) by differentiating with respect to s in (2.34) and setting the derivative equal to 0. After substituting (3.14) into (2.34), algebra yields (2.38). The limits in (5.8) imply that $s^* \dot{I}_w(s^*) \rightarrow 0$ and $I_w(s^*) \rightarrow I_w(\infty)$ as $\rho \rightarrow 1$. ■

Given that $s \dot{I}_w(s) \rightarrow 0$ as $s \rightarrow \infty$ and $s^* \rightarrow \infty$ as $\rho \uparrow 1$, if $b = \sqrt{2}$, then it is natural to consider the heavy-traffic approximation

$$s^* \approx \frac{\rho I_w(s^*)}{2\mu(1-\rho)^2} \quad \text{so that} \quad Z^* \approx \frac{\rho I_w(s^*)}{2\mu(1-\rho)} \quad \text{and} \quad c_{Z^*}^2 = I_w(s^*). \quad (2.40)$$

The first equation in (2.40) is a variability fixed-point equation of the form suggested in (15) of [60]. Hence, our RQ formulation in (2.17) can be viewed as a refined version of the approach in [60].

2.2.6 The RQ(b) Algorithm for Quantile Approximation

In this section, we show how RQ solution $Z^*(b)$ in (2.34) or (2.35) with parameter b can be used to approximate the full distribution of the stochastic steady-state workload Z , which we do via quantiles. Hence, we refer to it as the RQ(b) algorithm.

From the form of RQ(b), it is evident that as b increases, the approximation should apply more to the tail of the distribution. We find that a useful connection can be made between the parameter b and the quantiles of the distribution of the steady-state workload Z . For a nonnegative random variable Z and $0 < p < 1$, let the p^{th} quantile of (the distribution of) Z be

$$Z(p) \equiv \inf \{z \geq 0 : P(Z \leq z) = p\}, \quad 0 < p < 1, \quad (2.41)$$

i.e., the inverse of the cumulative distribution function (cdf). We propose the approximation

$$Z(\Pi(b)) \approx Z^*(b), \quad (2.42)$$

where $\Pi : (-\infty, \infty) \rightarrow (0, 1)$ is a one-to-one continuous function chosen to map the RQ parameter b into the quantile level p of Z .

For $GI/GI/1$ model, the standard heavy-traffic approximation implies that the steady-state workload Z should be approximately exponentially distributed; see Section 5.7 and Section 9.3 in [143]. In particular, for mean-1 service and traffic intensity ρ ,

$$P(Z > x) \approx e^{-x/m}, \quad x \geq 0, \quad \text{for } m \equiv \frac{\rho c_x^2}{2(1-\rho)}. \quad (2.43)$$

Thus, for quantile p of Z , denoted by $Z(p)$, we have $P(Z \leq W(p)) \approx 1 - e^{-Z(p)/m} = p$, so that

$$Z(p) \approx -\ln(1-p)m \quad (2.44)$$

for m in (2.43).

On the other hand, if we apply Theorem 2.2 to the $M/GI/1$ queue or the RBM approximation, then we get

$$Z^*(b) = \frac{b^2 m}{2}. \quad (2.45)$$

To match the actual *mean* in $M/GI/1$ for all ρ and to match the mean in heavy-traffic and light-traffic limits, we should chose $b^2 = \sqrt{2}$ as in Corollary 2.3, Theorem 2.5 and Theorem 2.6.

Further connection can be made by equating (2.44) and (2.45) to obtain an approximation for the desired function Π in (2.42), getting

$$p \approx \Pi(b) \equiv 1 - e^{-b^2/2}. \quad (2.46)$$

By (2.44), for an exponential random variable, the mean coincides with the $p = 1 - e^{-1} \approx 0.632$ quantile. By (2.46), this quantile corresponds to $b = \sqrt{2}$. Hence, the $RQ(\sqrt{2})$ is the RQ algorithm for the mean steady-state workload.

2.2.7 Heavy-Traffic and Light-Traffic Limits

The following result shows the great advantage of doing RQ with (i) the continuous-time workload and (ii) the non-parametric version of the RQ in (2.34). A proof is given in Section 2.6.

Theorem 2.6 (*heavy-traffic and light-traffic limits*) Under the regularity conditions assumed for the IDW $I_w(t)$, if $b_f \equiv \sqrt{2}$, then the non-parametric RQ solution in (2.34) is an asymptotically correct characterization of steady-state mean workload both in heavy traffic (as $\rho \uparrow 1$) and light traffic (as $\rho \downarrow 0$). Specifically, we have the following supplement to (2.29):

$$\lim_{\rho \uparrow 1} c_{Z^*}^2(\rho) = I_w(\infty) = \lim_{\rho \uparrow 1} c_Z^2(\rho) \quad \text{and} \quad \lim_{\rho \downarrow 0} c_{Z^*}^2(\rho) = I_w(0) = \lim_{\rho \downarrow 0} c_Z^2(\rho). \quad (2.47)$$

Remark 2.6 Theorem 2.6 greatly generalizes results in Theorem 2.3(b) with both light and heavy traffic addressed in the general case beyond positive or negative correlations. We also note that the parametric RQ solution can be made correct in heavy traffic or in light traffic, as above, by choosing the parameter b_p appropriately, but both cannot be achieved simultaneously unless $I_w(\infty) = I_w(0)$.

2.2.8 Other Steady-State Performance Measures

We develop approximations for other steady-state performance measures by applying exact relations for the $G/GI/1$ queue that follow from Little's law $L = \lambda W$ and its generalization $H = \lambda G$; e.g., see [141] and Chapter X of [12] for the $GI/GI/1$ special case. Let W, Q and X be the steady-state waiting time, queue length and the number in system (including the one in service, if any). By Little's law,

$$\begin{aligned} E[Q] &= \lambda E[W] = \rho E[W] \quad \text{and} \\ E[X] &= E[Q] + \rho = \rho(E[W] + 1). \end{aligned} \quad (2.48)$$

By Brumelle's formula [30] or $H = \lambda G$, (6.20) of [141],

$$E[Z] = \rho E[W] + \rho \frac{E[V^2]}{2\mu} = \rho E[W] + \rho \frac{(c_s^2 + 1)}{2\mu}, \quad (2.49)$$

Hence, given an approximation Z^* for $E[Z]$, we can use the approximations

$$\begin{aligned} E[W] &\approx \max\{0, Z^*/\rho - (c_s^2 + 1)/2\mu\} \quad \text{and} \\ E[Q] &\approx \lambda E[W]. \end{aligned} \quad (2.50)$$

Remark 2.7 (*network performance measures*) So far we only have discussed the performance measures for a single station. The total network performance measures, on the other hand, can also be derived. See Section 4.1 for the settings of the open queueing network models. For example, the expected value of the total sojourn time T_i^{tot} , i.e. the time needed to flow through the queueing network for a customer that enters the system from station i , is easily estimated from the obtained mean waiting time at each station. Assuming Markov routing with routing matrix P , a standard argument from discrete time Markov chain theory gives the mean total number of visits $\xi_{i,j}$ to station j by a customer entering the system at station i as

$$\xi_{i,j} = ((I - P)^{-1})_{i,j},$$

where $(I - P)^{-1}$ is the fundamental matrix of a absorbing Markov chain. Hence, the mean steady-state total sojourn time $E[T_i^{\text{tot}}]$ is approximated by

$$E[T_i^{\text{tot}}] \approx \sum_{j=1}^K \xi_{i,j} (E[W_j] + 1/\mu_j). \quad (2.51)$$

In real world applications, customers often experiences non-Markovian routing, where routes are customer-dependent. For ways to represent those scenarios and convert them (approximately) to the current framework, see §2.3 and §6 of [134]. ■

2.2.9 An Illustrative Simulation Example

In this section, we present an example that demonstrates (i) the strong impact of dependence upon performance and (ii) the value of the new RQ in capturing the impact of that dependence.

Consider an example of 5 single-server queues in series where the variability increases and then decreases 5 times, with the traffic intensities at successive queues decreasing. That makes the external arrival process and the earlier queues relevant only as the traffic intensity increases. Specifically, the example can be denoted by

$$E_{10}/H_2/1 \rightarrow \cdot/E_{10}/1 \rightarrow \cdot/H_2/1 \rightarrow \cdot/E_{10}/1 \rightarrow \cdot/M/1. \quad (2.52)$$

In particular, the external arrival process is a rate-1 renewal process with E_{10} interarrival times, thus $c_a^2 = 0.1$. The 1st queue has H_2 service times with mean 0.99 and $c_s^2 = 10$ (and

also balanced means, as before), thus the traffic intensity at this queue is 0.99. The 2nd queue has E_{10} service time with mean and thus traffic intensity 0.98. The 3rd queue has H_2 service times with mean 0.70 and $c_s^2 = 10$. The 4th queue has E_{10} service times with mean and thus traffic intensity 0.5. The last (5th) queue has an exponential service-time distribution with mean and traffic intensity ρ . We explore the impact of ρ on the performance of that last queue.

This example is designed so that the total arrival process at queue 5 have a very complicated dependence structure. Looking backwards starting from the 4th queue, i.e., the queue just before the last queue, the Erlang service act to smooth the arrival process at the last queue. Thus, for sufficiently low traffic intensities ρ at the last queue, the last queue should behave essentially the same as a $E_{10}/M/1$ queue, which has $c_a^2 = 0.1$, but as ρ increases, the arrival process at the last queue should inherit the variability of the previous service times and the external arrival process, and altering between $H_2/M/1$ and $E_{10}/M/1$ as the traffic intensity at the last queue increases. This implies that the normalized workload $c_Z^2(\rho)$ in (2.27) as a function of ρ should have four internal modes.

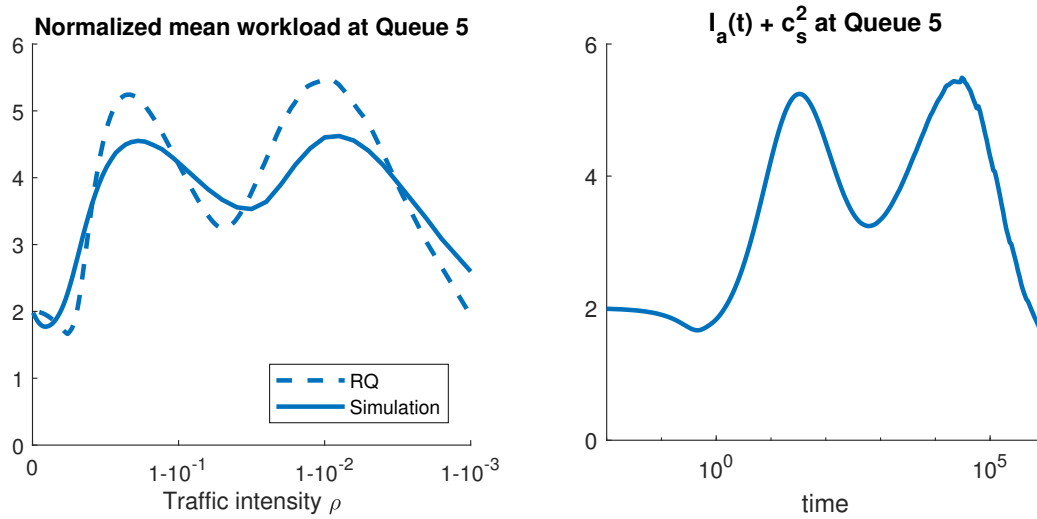


Figure 2.2: Comparing the simulation estimation to the RQ approximation for the workload, as a function of traffic intensity, at the fifth queue of a five-queues-in-series model. The workload function have four internal modes.

This behavior is substantiated by Figure 2.2 (left), which compares simulation estimates

of the normalized mean workload $c_Z^2(\rho)$ in (2.27) at the last queue with the RQ approximation $c_{Z*}^2(\rho)$ in (2.36). It shows that the normalized workload at the last queue fluctuates and each mode corresponds to a previous service process or the external arrival process. Figure 2.2 (left) also shows that RQ successfully captures all modes and provides a reasonably accurate approximation for all ρ . Note that a new scale in the horizontal x axis is used in Figure 2.2 (left), namely $-\ln(1 - \rho)$. Since 4 out of 6 modes lies in $\rho > 0.8$, the new scale act to stretch out the crowded plot under heavy traffic.

To conclude on this series-queue example, we show the IDW for the last queue in Figure 2.2 (right). The x axis of the figure is in log scale for easier display. Clearly, the IDW has the same qualitative property as the normalized workload as well as the RQ approximation, as we expect from equation (2.40).

2.3 Estimating and Calculating the IDC

For the applications of the RQ algorithm, it is significant that the IDW $I_w(t)$ and the IDC $I_A(t)$ can readily be estimated from data from system measurements or simulation, and calculated in a wide class of stochastic models. The time-dependent variance functions can be estimated from the time-dependent first and second moment functions, as discussed in Section III.B of [59]. Calculation depends on the specific model structure.

To start, we review the calculation of the indices of dispersion under several specific model structures in Section 2.3.3. In Section 2.3.1, we discuss the calculation of the IDC for the renewal process. In Section 2.3.4, we propose an algorithm to estimate the IDC from data.

2.3.1 The IDC's for Renewal Processes.

For renewal processes, the variance $\text{Var}(A(t))$ and thus the IDC $I_A(t)$ can either be calculated directly or can be characterized via their Laplace transforms and thus calculated by inverting those transforms and approximated by performing asymptotic analysis. Because we are interested in the steady-state behavior of the OQN, we are primarily interested in the equilibrium renewal process, as in Section 3.5 of [116]. For a stationary point process,

we have $E[A(t)] = \lambda t$. Hence, we only focus on $\text{Var}(A(t))$ here.

It turns out that the variance of the equilibrium arrival renewal process $V(t) \equiv \text{Var}(A(t))$ can be expressed in terms of the renewal function $m(t) \equiv E[A_0(t)]$, where A_0 is the corresponding ordinary renewal process, i.e. with an arrival at time 0. For a function f , let $\hat{f}$ denote the Laplace transform of f , defined by

$$\hat{f}(s) \equiv \mathcal{L}(f)(s) \equiv \int_0^\infty e^{-st} f(t) dt.$$

For the variance function $V(t)$, we have

$$\hat{V}(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s} \hat{m}(s) - \frac{2\lambda^2}{s^3} = \frac{\lambda}{s^2} + \frac{2\lambda}{s} \frac{\hat{g}(s)}{s(1 - \hat{g}(s))} - \frac{2\lambda^2}{s^3}, \quad (2.53)$$

where g is the density function of the interarrival-time distribution. For detailed derivation, see Section 3.1.

The variance function can then be obtained numerically, which is discussed in Section 13 of [4].

2.3.2 The IDC of the Markovian Arrival Process

Markovian arrival process (MAP) is an extremely versatile modeling tool for point processes. The MAP is initially introduced by Neuts in [107] as a versatile Markovian process. It was reformulated as the MAP in [104], which we follow as its definition. It contains a wide range of point processes as special cases, including the Poisson process, the Erlang and hyperexponential renewal processes, the Phase-type (PH) renewal process and the Markov Modulated Poisson process. Since the Phase-type distribution is dense in the field of all positive-valued distributions, the MAP can be applied to approximate any renewal process. But the MAP is also capable of modeling auto-correlation in point process, which do not appear in renewal processes. We summarize useful formula in this section.

The MAP is among one of the few general models that can be treated analytically, see for example [1; 97]. In this section, we review the exact formula for the IDC of the Markovian Arrival Process and its special cases; see also Section 5.4 of [108], especially Theorem 5.4.1.

The MAP is defined in terms of a continuous-time Markov chain (CTMC) with infinitesimal generator $\mathbf{D} = \mathbf{D}_0 + \mathbf{D}_1 \in \mathbb{R}^{n \times n}$, where $\mathbf{D}\mathbf{1} = \mathbf{0}$, $\mathbf{0}$ and $\mathbf{1}$ are the column vector of zeros

and ones, all the off-diagonal elements of $\mathbf{D}_0$ and all the elements of $\mathbf{D}_1$ are nonnegative. The transitions associated with $\mathbf{D}_1$ are called type 1 transitions. A MAP with parameters $(\mathbf{D}_0, \mathbf{D}_1)$, denoted by $\text{MAP}(\mathbf{D}_0, \mathbf{D}_1)$, is a point process where an event occurs if and only if a type 1 transition occurs in the CTMC.

Neuts [108] applies matrix-analytic methods to give explicit representations of the variance $\text{Var}(N(t))$ for the versatile Markovian point process or Neuts process; see Section 5.4, especially Theorem 5.4.1. We summarize the results below. Let $\boldsymbol{\alpha}$ be the steady-state probability (row) vector of the CTMC generated by the rate matrix $\mathbf{D}$, i.e. the solution to

$$\boldsymbol{\alpha}\mathbf{D} = \mathbf{0}, \quad \boldsymbol{\alpha}\mathbf{1} = 1. \quad (2.54)$$

Starting with the distribution of initial state specified by $\boldsymbol{\alpha}$, the resulting MAP will be a continuous-time stationary point process. We use $N(t)$ to denote such a stationary MAP. The mean function (for the stationary version) is

$$m(t) \equiv E[N(t)] = \lambda t, \quad \lambda = \boldsymbol{\alpha}\mathbf{D}_1\mathbf{1}. \quad (2.55)$$

The variance function $V(t) \equiv \text{Var}(N(t))$ of the stationary MAP $N(t)$ is

$$V(t) = (\lambda - 2\lambda^2 + 2\boldsymbol{\alpha}\mathbf{D}_1\mathbf{d}_1) t - 2\boldsymbol{\alpha}\mathbf{D}_1(\mathbf{I} - e^{\mathbf{D}t})\mathbf{d}_2, \quad (2.56)$$

where $\mathbf{d}_i \equiv (\mathbf{1}\boldsymbol{\alpha} - \mathbf{D})^{-i}\mathbf{D}_1\mathbf{1}$, $\mathbf{I}$ is the identity matrix of order n and $e^{\mathbf{D}t}$ is the matrix exponential. From (2.55) and (2.56), we obtain the IDC I_N of $N(t)$

$$I_N(t) = 1 - 2\lambda + 2\boldsymbol{\alpha}\mathbf{D}_1\mathbf{d}_1/\lambda - 2\boldsymbol{\alpha}\mathbf{D}_1(\mathbf{I} - e^{\mathbf{D}t})\mathbf{d}_2/\lambda t, \quad (2.57)$$

The limits of the IDC can then be easily derived. In particular, we have

Proposition 2.1 *For $\text{MAP}(\mathbf{D}_0, \mathbf{D}_1)$, we have*

$$I_N(0) = \lim_{t \rightarrow 0} I_N(t) = 1, \quad \text{and} \quad (2.58)$$

$$I_N(\infty) = \lim_{t \rightarrow \infty} I_N(t) = 1 - 2\lambda + 2\boldsymbol{\alpha}\mathbf{D}_1\mathbf{d}_1/\lambda. \quad (2.59)$$

Proof. The second statement follows from the fact that all eigenvalues of $\mathbf{D}$ have non-positive real parts, hence the matrix exponential $e^{\mathbf{D}t}$ converges to 0 as $t \rightarrow \infty$. ■

We remark that, for a non-renewal MAP, the limit $I_N(\infty)$ is not necessarily equal to the scv of the stationary inter-event time. For the interval-stationary version of the MAP, consider the Markov chain embedded at the arrival epochs, whose transition probability matrix $\mathbf{P}$ is given by

$$\mathbf{P} = \int_0^\infty e^{\mathbf{D}_0 t} \mathbf{D}_1 dt = -\mathbf{D}_0^{-1} \mathbf{D}_1.$$

The stationary vector $\boldsymbol{\alpha}_0$ of $\mathbf{P}$ is then obtained by solving $\boldsymbol{\alpha}_0 \mathbf{P} = \boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_0 \mathbf{1} = 1$. The cumulative distribution function (cdf) of a stationary inter-event time is then given by

$$F(t) = 1 - \boldsymbol{\alpha}_0 e^{\mathbf{D}_0 t} \mathbf{1},$$

and its scv is given by

$$\frac{2\boldsymbol{\alpha}_0(-\mathbf{D}_0)^{-2}\mathbf{1}}{(\boldsymbol{\alpha}_0(-\mathbf{D}_0)^{-1}\mathbf{1})^2} - 1,$$

which is in general different from (2.59).

However, the limiting variability parameter $I_N(\infty)$ coincide with the variability coefficient in the Brownian functional central limit theorem of the MAP. Let $\hat{N}_n(t) \equiv n^{-1/2}[N(nt) - \lambda nt]$ be the diffusion-scaled process, where λ is the rate defined in (2.55). A proof can be found in Section 2.6.

Theorem 2.7 (Functional central limit theorem) *Consider MAP($\mathbf{D}_0, \mathbf{D}_1$) with finite state and initial distribution $\boldsymbol{\alpha}$ in (2.54). Assume that the underlying Markov chain generated by $\mathbf{D} = \mathbf{D}_0 + \mathbf{D}_1$ is irreducible, then we have*

$$\hat{N} \Rightarrow cB \circ \lambda e,$$

where $c = \sqrt{I_N(\infty)}$ defined in (2.59), B is a standard Brownian motion, λ is the rate defined in (2.55) and e is the identity map.

We now summarize the exact formula for several useful special cases.

Example 2.1 (Erlang renewal process) The Erlang (E_k) random variable with shape parameter k and rate parameter λ is defined as the sum of k i.i.d. exponential random variables, each with rate λ , so that the mean is $m = k/\lambda$ and the scv is $c^2 = 1/k$. The

Erlang renewal process can be specified as a MAP with

$$\mathbf{D}_0 = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & 0 \\ 0 & -\lambda & \lambda & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda \end{bmatrix}, \quad \mathbf{D}_0 = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \lambda & 0 & 0 & \cdots & 0 \end{bmatrix}$$

For the case of $k = 2, 3$ and 4, we have

$$\begin{aligned} I_{E_2}(t) &= \frac{1}{2} + \frac{1}{4\lambda t} (1 - e^{-2\lambda t}), \\ I_{E_3}(t) &= \frac{1}{3} + \frac{4}{9\lambda t} (1 - e^{-3\lambda t/2} \cos(\sqrt{3}\lambda t/2)), \\ I_{E_4}(t) &= \frac{1}{4} + \frac{1}{8\lambda t} (1 - e^{-2\lambda t}) + \frac{1}{2\lambda t} (1 - e^{-\lambda t} \cos(\lambda t)). \end{aligned}$$

See also Section 4.3 and 4.5 of [39] and Section III.G of [60].

Example 2.2 (Hyperexponential renewal process) Consider the mixture of two independent exponential random variable, i.e. the hyperexponential (H_2) random variable, with probability density function defined as

$$f(t) = p\mu_1 e^{-\mu_1 t} + (1-p)\mu_2 e^{-\mu_2 t}, \quad t \geq 0.$$

By swapping μ_1 with μ_2 and p with $1-p$, we may assume $\mu_1 \geq \mu_2$ without loss of generality. The distribution can also be specified by the the rate $\lambda = \frac{\mu_1 \mu_2}{p\mu_2 + (1-p)\mu_1}$, the scv c^2 and the ratio between the mean of the component with the larger rate and the overall mean $r \equiv p/\mu_1 / (p/\mu_1 + (1-p)/\mu_2)$. Alternatively, the H_2 renewal process can be specified by the MAP with

$$\mathbf{D}_0 = \begin{bmatrix} -\mu_1 & 0 \\ 0 & -\mu_2 \end{bmatrix}, \quad \mathbf{D}_0 = \begin{bmatrix} p\mu_1 & (1-p)\mu_1 \\ p\mu_2 & (1-p)\mu_2 \end{bmatrix}$$

The IDC of the equilibrium H_2 renewal process is then

$$I_{H_2}(t) = c^2 - \frac{2\beta}{\gamma t} (1 - e^{-\gamma t}), \quad (2.60)$$

where $\gamma = (1-p)\mu_1 + p\mu_2$ and $\beta = p(1-p)(\mu_1 - \mu_2)^2/\gamma^2$, which is consistent with Section III.G of [60].

Example 2.3 (Markov modulated Poisson process) The Markov modulated Poisson process (MMPP) is a special case of MAP with $\mathbf{D}_1$ being restricted to a diagonal matrix with non-negative entries. $\mathbf{D}_0$ can be any matrix with nonnegative off-diagonal elements, as long as $(\mathbf{D}_0 + \mathbf{D}_1)\mathbf{1} = \mathbf{0}$. The MMPP(2), i.e. the MMPP with 2 states, is defined by

$$\mathbf{D}_0 = \begin{bmatrix} -r_1 - \lambda_1 & r_1 \\ r_2 & -r_2 - \lambda_2 \end{bmatrix}, \quad \mathbf{D}_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad (2.61)$$

The rate of this MMPP(2) is $\lambda = \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2}$ and the IDC is

$$I_{\text{MMPP}(2)}(t) = 1 + 2\alpha - \frac{2\alpha}{\gamma t} (1 - e^{-\gamma t}), \quad (2.62)$$

where $\gamma = r_1 + r_2$ and $\alpha = \frac{r_1 r_2 (\lambda_1 - \lambda_2)^2}{\gamma^2 (r_1 \lambda_2 + r_2 \lambda_1)}$.

2.3.3 Calculation of the IDW and IDC in Some Queueing Networks

The G/GI/1 Model. If the service times are i.i.d. with a general distribution having mean τ and scv c_s^2 and are independent of a general stationary arrival process, then as indicated in (58) and (59) in Section III.E of [60],

$$I_w(t) = c_s^2 + I_A(t), \quad t \geq 0, \quad (2.63)$$

where c_s^2 is the scv of a service time and I_A is the IDC of the general arrival process A .

The Multi-Class $\sum_i (G_i/G_i)/1$ Model. As indicated in (56) and (57) in Section III.E of [60], if the input comes from independent sources, each with their own arrival process and service times, then the overall IDC and IDW are revealing functions of the component ones. Let λ_i be the arrival rate, τ_i the mean service time of class i , and $\rho_i \equiv \lambda_i \tau_i$ be the traffic intensity for class i with $\lambda \equiv \sum_i \lambda_i$, $\tau \equiv \sum_i (\lambda_i / \lambda) \tau_i = 1$ so that $\rho = \lambda$. Then

$$I_A(t) \equiv \frac{\text{Var}(A(t))}{E[A(t)]} = \frac{\sum_i \text{Var}(A_i(t))}{\lambda t} = \sum_i \left(\frac{\lambda_i}{\lambda} \right) I_{A_i}(t) \quad (2.64)$$

and

$$I_w(\lambda t) \equiv \frac{\text{Var}(Y(t))}{\tau E[Y(t)]} = \frac{\sum_i \text{Var}(Y_i(t))}{\rho t} = \sum_i \left(\frac{\rho_i \tau_i}{\rho \tau} \right) I_{w,i}(t) \quad \text{for all } t \geq 0. \quad (2.65)$$

From (2.64) and (2.65), we see that I_A and I_w are convex combinations of the component I_{A_i} and $I_{w,i}$.

The Multi-Class $\sum_i G_i/GI/1$ Model. An important special case of the multi-class $\sum_i(G_i/G_i)/1$ model arising in open queueing networks is the $\sum_i G_i/GI/1$ model in which there are multiple general arrival streams coming to a queue where all arrivals experience common i.i.d. service times. We can combine (2.63) and (2.64) to get the expression for the IDW

$$I_w(t) \equiv I_A(t) + c_s^2, \quad t \geq 0, \quad (2.66)$$

where $I_A(t)$ is given by (2.64). Of course, if all the component arrival streams are Poisson processes, then $I_A(t) = 1$ for all $t \geq 0$, but otherwise the IDC $I_A(t)$ can be quite complicated.

The Balanced $\sum_i G_i/GI/1$ Model. We call the $\sum_i G_i/GI/1$ model *balanced*, if the arrival process is the superposition of n i.i.d. non-Poisson processes each with rate λ/n , so that the overall arrival rate is λ , and asymptotic variability parameter c_a^2 . From the results above, we obtain

$$I_{A,n}(t) = I_A(t/n) \quad \text{and} \quad I_{w,n}(t) = I_{w,1}(t/n), \quad t \geq 0, \quad (2.67)$$

so that the superposition IDC and IDW differ from those of a single component process only by the time scaling.

In particular, under regularity conditions, (i) the superposition arrival process is known to be non-Poisson and non-renewal unless the component arrival streams are Poisson. (ii) if we let $n \rightarrow \infty$ but keep the total rate fixed, then the superposition process approaches a Poisson process, but (iii) for any n , no matter how large, if we let $t \rightarrow \infty$, then the superposition process obeys the same CLT as a single component arrival process, and so has asymptotic variability parameter c_a^2 . Thus, we have $I_{A,n}(0) = 1$ and $I_{A,n}(\infty) = c_a^2$ for any $n \geq 1$.

We will show that RQ provides important insight when we conduct simulation experiments for this model in Section 6.2.2.

2.3.4 Numerical Estimation of the IDC from Data.

Now we present an algorithm to numerically estimate the variance $V(t) = \text{Var}(A(t))$ from a given realized sample path of the stationary point process $A(t)$. The main idea is based on Section 5.4 (iii) of [40].

Our goal is to estimate $V(t)$ for $0 < t < t_0$ using a realization of $A(t)$ for $0 < t < T$. The simplest way is to apply crude Monte Carlo method to estimate $V(t)$ for a fixed t and repeat over a finite grid of t 's. This method divide the sample path of $A(t)$ into non-overlapping intervals of length t and count the number of arrivals in each interval. The variance is then estimated by the sample variance of the counts. This method is simple to implement but can be slow to converge.

To accelerate the crude Monte Carlo method, we apply three techniques: (i) we use overlapping intervals instead of non-overlapping ones, which introduces bias but reduces sample variance; (ii) we calculate $V(t)$ only over a finite grid equally spaced in the logarithm scale instead of the linear scale; and (iii) we re-use the tallied number of events for shorter intervals to calculate the total number of events for longer interval, which avoids repetitive counting. We discuss the three techniques in turn:

To use overlapping intervals, consider first $k = T/t$ number of non-overlapping intervals, each with length t . Now, we further divide each intervals of length t in to r intervals of the same length $\tau = t/r$. Hence we have rk number of non-overlapping intervals of length τ . Let n_i be the number of events fall in the i -th interval, consider

$$U_i \equiv A(I_i) \equiv A[i\tau, (i+r)\tau) = n_i + n_{i+1} + \cdots + n_{i+r-1}, \quad i = 0, 1, \dots, rk - r + 1.$$

We estimate $V(t)$ with the sample variance $\bar{V}_l$ of $\{U_i\}_{i=1}^l$, where $l = rk - r + 1$. This estimator is in general biased but can achieve lower variance compared with the one obtained with crude Monte Carlo method. In Theorem 2.8 we show that this estimator of $V(t)$ is asymptotically consistent under mild conditions that $V(t)$ is differentiable with derivative $\dot{V}(t)$ having finite positive limits as $t \rightarrow \infty$.

For the third technique, we now present a algorithm to simultaneously estimate $V(2^i \tau)$ for some $\tau > 0$ and $i = 0, 1, \dots, l$. Let $\{I_i\}$ be the collection of non-overlapping intervals of length τ that covers $[0, T]$. Let $n_i = A(I_i)$ be the number of events on interval I_i . Then we have the following table from [40].

sample	time horizon t			
	τ	2τ	$2^2\tau$	$\dots$
1	n_1	$n_1 + n_2$	$n_1 + n_2 + n_3 + n_4$	$\dots$
2	n_2	$n_2 + n_3$	$n_3 + n_4 + n_5 + n_6$	$\dots$
3	n_3	$n_3 + n_4$	$n_5 + n_6 + n_7 + n_8$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

We find the estimation of $V(2^i\tau)$ by calculating the sample variance of the corresponding column.

Now that we have a efficient algorithm to estimate $V(2^i\tau)$ for fixed τ , we have obtained the estimations of a grid equally spaced in logarithm scale. To obtain estimations for finer grids we shift the crude grid by picking several $\tau \leq \tau_j \leq 2\tau$ equally spaced in log scale and, for each j , simultaneously estimate $V(2^i\tau_j)$ for all i .

Consistency of the estimator. We now show that the estimator of the variance function $V(t)$ is asymptotically consistent. Towards this end, we assume mild regularity conditions that $V(t)$ is differentiable with derivative $\dot{V}(t)$ having finite positive limits as $t \rightarrow \infty$, i.e.,

$$\dot{V}(t) \rightarrow \sigma^2 \text{ as } t \rightarrow \infty,$$

for an appropriate constant σ^2 , as in (5.8). The proof is postponed to Section 2.6.

Theorem 2.8 (Consistency of the estimator) *Let A be a time-stationary and ergodic point process with variance function $V(t)$ that is differentiable with derivative $\dot{V}(t)$ having finite positive limit as $t \rightarrow \infty$, i.e.,*

$$\dot{V}(t) \rightarrow \sigma^2 \text{ as } t \rightarrow \infty.$$

Then we have

$$\lim_{l \rightarrow \infty} \text{bias}(\bar{V}_l) = 0$$

for $l = rk - r + 1$, $r = t/\tau$, $k = T/t$ and $\bar{V}_k$ is the sample variance of $\{U_i\}_{i=1}^k$. Furthermore,

$$\lim_{l \rightarrow \infty} \bar{V}_l = V(t), \text{ w.p.1.}$$

Remark 2.8 (Implication for the choice of simulation run length) The bias comes from two kinds of cumulative correlation in $N(t)$. The first part of the bias comes from the correlation of overlapping intervals, which can be bounded by $4Ct/(k-1)$. This suggests that $\bar{V}$ is a reasonable estimator only when $k \gg t$. The second part of the bias comes from the correlation of the increments when $N(t)$ is non-Poisson. This part of the bias can be controlled by the convergence rate of $\dot{V}(t)$. This suggests that, regardless of our choice of t , $\bar{V}$ is a reasonable estimator only when T is much larger than the relaxation time it takes for $\dot{V}(t)$ to converge.

For example, if we are interested in $V(t)$ for $t = 10^5$ and we want a bias of less than ε , then we must use $T = tk \geq 4Ct^2/\varepsilon = O(t^2/\varepsilon)$ in order to control the first part of the bias. For the second part of the bias, in the queueing setting, the time it takes for $\dot{V}(t)$ to converge is roughly in the magnitude of $O(1/(1-\rho)^2)$. Hence, we need a simulation time of $T = O(t/(1-\rho)^2)$ to eliminate the bias. ■

2.4 More on Non-Parametric RQ for Waiting Times

We now discuss how Corollary 2.1 can be extended, with the aid of Section 2.5, to show that both the new parametric RQ in (2.6) and the new non-parametric RQ with uncertainty set in (2.9) are asymptotically correct in heavy traffic for the more general stationary $G/G/1$ model. For the general model, we regard $\{(U_k, V_k)\}$ as a stationary sequence. For notation simplicity, we assume $E[V_k] = 1$ and $E[U_k] = \rho^{-1} > 1$ for all k for the rest of this section.

Paralleling the parametric RQ optimization in (2.6), we have the non-parametric analog

$$W_{f,\rho}^* \equiv \sup_{\tilde{X} \in \mathcal{U}_f^x} \sup_{k \geq 0} \{S_k^x\}. \quad (2.68)$$

where $\mathcal{U}_f^x$ is defined in (2.9). For the $G/G/1$ model stationary in discrete time, the reasoning for Theorem 2.1 leads to the alternative representation as

$$W_{f,\rho}^* = \sup_{k \geq 0} \left\{ -mk + b_{f,d} \sqrt{\text{Var}(S_k^x)} \right\} \quad (2.69)$$

instead of (2.7), where $m \equiv (1-\rho)/\rho$ as before.

We now recast the discrete-time RQ solution in (2.69) in terms of *indices of dispersion*

for intervals (IDI). As in Chapter 4 of [40], the IDI's are functions of $k \in \mathbb{N}$, defined by

$$I_k^a \equiv \frac{k \text{Var}(S_k^a)}{(E[S_k^a])^2}, \quad I_k^s \equiv \frac{k \text{Var}(S_k^s)}{(E[S_k^s])^2} \quad \text{and} \quad I_k^{a,s} \equiv \frac{k \text{Cov}(S_k^a, S_k^s)}{E[S_k^a]E[S_k^s]} \quad \text{for } k \in \mathbb{N}. \quad (2.70)$$

As the scaled versions of the discrete-time variance-time functions (sequences) $\text{Var}(S_k^x)$, $\text{Var}(S_k^a)$ and $\text{Var}(S_k^s)$, the IDI's measure the cumulative covariance in each partial sum. With (2.70),

$$\begin{aligned} \sqrt{\text{Var}(S_k^x)} &= E[U_1] \sqrt{k I_k^x}, \quad k \geq 1, \quad \text{and} \\ \sigma_X^2 &\equiv \lim_{k \rightarrow \infty} \{k^{-1} \text{Var}(S_k^x)\} = E[U_1]^2 I_\infty^x \end{aligned} \quad (2.71)$$

where

$$I_k^x \equiv I_k^a + \rho^2 I_k^s - 2\rho I_k^{a,s} \quad \text{for } \rho \equiv E[V_1]/E[U_1] < 1. \quad (2.72)$$

These three IDI's I_k^a , I_k^s and $I_k^{a,s}$ were used to develop queueing approximations in [58].

As a consequence, (2.69) can be rewritten as

$$W_{f,\rho}^* = \sup_{k \geq 0} \left\{ -(1-\rho)k/\rho + b_{f,d} \sqrt{k I_k^x} \right\}. \quad (2.73)$$

Similar to the continuous-time workload, we focus on the normalized mean waiting time and RQ approximation defined by

$$c_W^2(\rho) \equiv \frac{2(1-\rho)}{\rho} E[W_\rho], \quad \text{and} \quad c_{W^*}^2(\rho) \equiv \frac{2(1-\rho)}{\rho} W_{f,\rho}^*. \quad (2.74)$$

By essentially the same reasoning, we can show that both the parametric RQ and the non-parametric RQ for the steady-state waiting time W are asymptotically exact in heavy-traffic, with the same HT limit as for the continuous-time workload, if we choose the constant $b_{f,d}$ in (2.73) appropriately. The light-traffic behavior is much more complicated for the steady-state waiting time, as discussed in Section IV.A of [60] and Section 1 of [139]. That is a major reason for using the workload instead of the waiting time.

2.5 Supporting Functional Central Limit Theorems

In this section we establish (mostly review) the supporting Functional Central Limit Theorems (FCLT's) and the Central Limit Theorems (CLT's) that follow from them.

These are for the general stationary G/G/1 model, allowing stochastic dependence among the interarrival times and service times. Section 2.5.1 starts with a basic FCLT for partial sums of random variables from weakly dependent stationary sequences, as in Theorems 19.1-19.3 of [23] and Theorem 4.4.1 of [143].

To state the basic FCLT underlying the RQ approach to the waiting time and workload processes, we consider a sequence of models indexed by n with stationary sequence of interarrival times and service times. In Section 2.5.1 we establish the underlying FCLT for the partial sums of the interarrival times and service times. Then in 2.5.2 we establish a FCLT for other basic processes. In Section 2.5.3 we establish different ordinary CLT's that support the parametric RQ and non-parametric RQ. Finally, in Section 2.5.4 we establish heavy-traffic FCLT's for the waiting time and workload processes.

2.5.1 The Basic FCLT for the Partial Sums

Without loss of generality, we assume that the models are generated by a fixed sequence of mean-1 random variables $\{(U_k, V_k)\}$, with the interarrival times in model n being $U_{n,k} \equiv \rho_n^{-1} U_k$. For each n , let the sequence of pairs of partial sums be $\{(S_{n,k}^a, S_{n,k}^s : k \geq 1\}$. Let $\lambda_n = \rho_n$ and $\mu_n = 1$ denote the arrival rate and service rate in model n . Let $\lfloor x \rfloor$ denote the greatest integer less than or equal to the real number x . Let D^2 be the two-fold product space of the function space D and let $\Rightarrow$ denote convergence in distribution. For this initial FCLT, we let $\rho_n \rightarrow \rho$ as $n \rightarrow \infty$ for arbitrary $\rho > 0$. Let random elements in the function space D^2 be defined by

$$\left(\hat{S}_n^a(t), \hat{S}_n^s(t) \right) \equiv n^{-1/2} \left(\left[S_{n, \lfloor nt \rfloor}^a - \rho_n^{-1} nt \right], \left[S_{n, \lfloor nt \rfloor}^s - nt \right] \right), \quad t \geq 0.$$

Theorem 2.9 (*FCLT for partial sums of interarrival times and service times*) *Let $\{(U_k, V_k) : k \geq 1\}$ be a weakly dependent stationary sequence with $E[U_k] = E[V_k] = 1$. Let $U_{n,k} = \rho_n^{-1} U_k$ and $V_{n,k} = V_k$, $n \geq 1$, and assume that the variances and covariances satisfy*

$$\begin{aligned} 0 < \rho^{-2} \sigma_A^2 &\equiv \lim_{n \rightarrow \infty} \{n^{-1} \text{Var}(S_n^a)\} < \infty, \\ 0 < \sigma_S^2 &\equiv \lim_{n \rightarrow \infty} \{n^{-1} \text{Var}(S_n^s)\} < \infty \quad \text{and} \\ -\infty < \rho^{-1} \sigma_{A,S}^2 &\equiv \lim_{n \rightarrow \infty} \{n^{-1} \text{Cov}(S_n^a, S_n^s)\} < \infty. \end{aligned} \tag{2.75}$$

Then (under additional regularity conditions assumed, but not stated here)

$$\left(\hat{\mathbf{S}}_n^a, \hat{\mathbf{S}}_n^s\right) \Rightarrow \left(\hat{\mathbf{S}}^a, \hat{\mathbf{S}}^s\right) \quad \text{in } D^2 \quad \text{as } n \rightarrow \infty, \quad (2.76)$$

where $\left(\hat{\mathbf{S}}^a, \hat{\mathbf{S}}^s\right)$ is distributed as zero-drift two-dimensional Brownian motion (BM) with covariance matrix

$$\Sigma = \begin{pmatrix} \rho^{-2}\sigma_A^2 & \rho^{-1}\sigma_{A,S}^2 \\ \rho^{-1}\sigma_{A,S}^2 & \sigma_S^2 \end{pmatrix}.$$

Proof. The one-dimensional FCLT's for weakly dependent stationary sequences in D can be used to prove the two-dimensional version in Theorem 2.9. First, the limits for the individual processes $\hat{\mathbf{S}}_n^a$ and $\hat{\mathbf{S}}_n^s$ imply tightness of these processes in D , which in turn implies joint tightness in D^2 . Second, the Cramer-Wold device in Theorem 4.3.3 of [143] implies that limits for the finite-dimensional distributions for all linear combinations (which should be implied by the unstated regularity condition) implies the joint limit for the finite-dimensional distributions (fidi's). Finally, tightness plus convergence of the fidi's implies the desired weak convergence by Corollary 11.6.2 of [143]. ■

2.5.2 The FCLT for Other Basic Processes

As a consequence of Theorem 2.9, we also have an associated FCLT for scaled random elements associated with $S_{n,k}^x \equiv S_{n,k}^s - S_{a,k}^a$, $k \geq 1$, $A_n(s)$ and $Y_n(s) \equiv \sum_{i=1}^{A_n(s)} V_{n,i} = \sum_{i=1}^{A(\rho_n s)} V_i = Y(\rho_n s)$, $s \geq 0$, for A and Y in (2.10) and (2.11). Let the associated scaled processes be defined by

$$\left(\hat{\mathbf{S}}_n^x(t), \hat{\mathbf{A}}_n(t), \hat{\mathbf{Y}}_n(t)\right) \equiv n^{-1/2} \left(\left[S_{n, \lfloor nt \rfloor}^x - (1 - \rho_n^{-1})nt \right], [A_n(nt) - \rho_n nt], [Y_n(nt) - \rho_n nt] \right), \quad (2.77)$$

for $t \geq 0$. Let $\mathbf{B}(t)$ be standard (zero drift and unit variance) one-dimensional BM and let $\mathbf{e}$ be the identity function in D , i.e., $\mathbf{e}(t) = t$. Let $\stackrel{d}{=}$ mean equal in distribution, as processes if used for stochastic processes.

Corollary 2.4 (*joint FCLT for basic processes*) Under the conditions of Theorem 2.9,

$$\left(\hat{\mathbf{S}}_n^a, \hat{\mathbf{S}}_n^s, \hat{\mathbf{S}}_n^x, \hat{\mathbf{A}}_n, \hat{\mathbf{Y}}_n\right) \Rightarrow \left(\hat{\mathbf{S}}^a, \hat{\mathbf{S}}^s, \hat{\mathbf{S}}^x, \hat{\mathbf{A}}, \hat{\mathbf{Y}}\right) \quad \text{in } D^5 \quad \text{as } n \rightarrow \infty, \quad (2.78)$$

where $\hat{\mathbf{S}}^x = \hat{\mathbf{S}}^s - \hat{\mathbf{S}}^a \stackrel{d}{=} \sigma_X \mathbf{B}$, with variance function

$$\sigma_X^2 \equiv \sigma_X^2(\rho) = \rho^{-2}\sigma_A^2 + \sigma_S^2 - 2\rho^{-1}\sigma_{A,S}^2, \quad 0 < \sigma_X^2 < \infty, \quad (2.79)$$

for $\rho^{-2}\sigma_A^2$, σ_S^2 and $\rho^{-1}\sigma_{A,S}^2$ in (2.75), while

$$\begin{aligned} \hat{\mathbf{A}} &= -\rho\hat{\mathbf{S}}^a \circ \rho\mathbf{e} \stackrel{d}{=} -\rho\sigma_A\mathbf{B}_a \circ \rho\mathbf{e} \stackrel{d}{=} \rho^{3/2}\sigma_Y\mathbf{B}_a, \\ \hat{\mathbf{Y}} &= \hat{\mathbf{S}}^s \circ \rho\mathbf{e} - \rho\hat{\mathbf{S}}^a \circ \rho\mathbf{e} \stackrel{d}{=} \sigma_Y\mathbf{B} \circ \rho\mathbf{e} \stackrel{d}{=} \sqrt{\rho}\sigma_Y\mathbf{B}, \end{aligned} \quad (2.80)$$

where

$$\sigma_Y^2 \equiv \sigma_Y^2(\rho) = \sigma_A^2 + \sigma_S^2 - 2\sigma_{A,S}^2, \quad 0 < \sigma_Y^2 < \infty, \quad \text{for all } \rho. \quad (2.81)$$

Hence, $\hat{\mathbf{Y}} \stackrel{d}{=} \hat{\mathbf{S}}^x$ for $\rho = 1$, but not otherwise.

Proof. We apply the continuous mapping theorem (CMT) using several theorems from [143]. The CMT itself is Theorem 3.4.4. We treat the process $S_{n,k}^x$ using addition. We treat the counting processes A_n by apply the inverse map with centering to go from the FCLT for $S_{n,k}^a$ to the FCLT for the associated scaled counting processes, applying Theorem 7.3.2, which is a consequence of Corollary 13.8.1 to Theorem 13.8.2, which follows from Theorem 13.7.1. Then the limit for Y_n follows from Corollary 13.3.1. However, it is also possible to give a more elementary direct argument. First, let $\bar{A}_n(t) \equiv n^{-1}A_n(t)$, $t \geq 0$, and note that $\bar{A}_n \Rightarrow \rho\mathbf{e}$ as a consequence of the limit for $\mathbf{A}_n$. The initial limits all hold jointly by Theorems 11.4.4 and 11.4.5. Then observe that we can apply the continuous mapping theorem with composition and addition to treat $\mathbf{Y}_n$, because we can write

$$\mathbf{Y}_n = \mathbf{S}_n^s \circ \bar{A}_n + \mathbf{A}_n \quad (2.82)$$

i.e.,

$$\mathbf{Y}_n(t) \equiv n^{-1/2} \left(\sum_{k=1}^{A(nt)} -\rho nt \right), \quad t \geq 0, \quad (2.83)$$

while

$$(\mathbf{S}_n^s \circ \bar{A}_n)(t) = n^{-1/2} \left(\sum_{k=1}^{A(nt)} -A_n(nt) \right), \quad t \geq 0, \quad (2.84)$$

We then add to get (2.83), observing that two terms cancel.

We now derive alternative expressions for the limit process $\mathbf{Y}$. First, directly from (2.82) we obtain

$$\mathbf{Y} = \mathbf{S}^s \circ \rho \mathbf{e} + \mathbf{A} = \mathbf{S}^s \circ \rho \mathbf{e} - \rho \mathbf{S}^a \rho \mathbf{e} \stackrel{d}{=} \sigma_Y \mathbf{B} \circ \rho \mathbf{e} \stackrel{d}{=} \sqrt{\rho} \sigma_Y \mathbf{B}. \quad (2.85)$$

which justifies the expression for σ_Y^2 in (2.81). $\blacksquare$

Remark 2.9 (*uniform integrability*) Condition (2.75) implies that $k^{-1} \text{Var}(S_k^x) \rightarrow \sigma_X^2$ as $k \rightarrow \infty$ for σ_X^2 in (2.79). In addition to the conclusions of Theorem 2.76 and Corollary 2.4, we assume that the appropriate uniform integrability holds, so that we also have the continuous-time analog

$$s^{-1} \text{Var}(Y(s)) \rightarrow \sigma_Y^2 \quad \text{as } s \rightarrow \infty \quad (2.86)$$

for σ_Y^2 in (2.81).

2.5.3 Alternative Scaling in the CLT

Theorem 2.9 and Corollary 2.4 imply ordinary CLT's for the processes S_n^x and $Y_n(s)$ by simply applying the CMT with the projection map $\pi : D \rightarrow \mathbb{R}$ with $\pi(x) \equiv x(1)$.

Corollary 2.5 (*associated CLT's*) Under the assumptions of Theorem 2.9, there are CLT's for the partial sums S_n^x and the total input processes Y_n , stating

$$(S_n^x - nE[X_1]) / \sqrt{n\sigma_X^2} \Rightarrow N(0, 1) \quad \text{as } n \rightarrow \infty, \quad (2.87)$$

and

$$(Y_n - \rho n) / \sqrt{n\sigma_Y^2} \Rightarrow N(0, 1) \quad \text{as } n \rightarrow \infty, \quad (2.88)$$

where $N(0, 1)$ is a standard (mean-0, variance-1) normal random variable, σ_X^2 is the asymptotic variance constant in (2.75) and (2.79), and σ_Y^2 is the asymptotic variance constant in (5.8) and (2.81).

Clearly, Corollary 2.5 supports the parametric RQ formulations and indicates how to choose the parameters b_x and b_p in order to produce versions that should be asymptotically correct in heavy-traffic (see the next section). We now show that there are alternative

versions of these CLT's that support the non-parametric RQ formulations. First, instead of (2.87), we can also write

$$[S_n^x - E[S_n^x]]/\sqrt{\text{Var}(S_n^x)} \Rightarrow N(0, 1) \quad \text{as } n \rightarrow \infty. \quad (2.89)$$

Second, instead of (2.88), we can also write

$$[Y(t) - E[Y(t)]]/\sqrt{\text{Var}(Y(t))} \Rightarrow N(0, 1) \quad \text{as } t \rightarrow \infty. \quad (2.90)$$

The numerators in (2.87) and (2.89) are identical because $E[S_n^x] = nE[X_1]$ and $E[Y(t)] = \rho t$. The full statements in (2.87) and (2.89) are asymptotically equivalent as $n \rightarrow \infty$ by the CMT, because

$$\frac{S_n^x - nE[X_1]}{\sqrt{\text{Var}(S_n^x)}} = \frac{S_n^x - nE[X_1]}{\sqrt{n\sigma_X}} \times \frac{\sqrt{n\sigma_X}}{\sqrt{\text{Var}(S_n^x)}} \Rightarrow N(0, 1) \times 1 = N(0, 1).$$

The same is true for the CLT's in (2.88) and (2.90).

2.5.4 The Heavy-Traffic FCLT

Theorem 2.9 and Corollary 2.4 also provide a basis for heavy-traffic (HT) FCLT's for the waiting-time and workload processes. To state the HT FCLT, we let $\rho_n \rightarrow 1$ as $n \rightarrow \infty$ at the usual rate; see (2.92) below. Let $\hat{\mathbf{W}}^n$ and $\hat{\mathbf{Z}}^n$ be the random elements associated with the waiting time and workload processes, defined by

$$\left(\hat{\mathbf{W}}^n(t), \hat{\mathbf{Z}}^n(t) \right) = \left(n^{-1/2} W_{n, \lfloor nt \rfloor}, n^{-1/2} Z_n(nt) \right), \quad t \geq 0. \quad (2.91)$$

Let $\psi : D \rightarrow D$ be the one-dimensional reflection map with impenetrable barrier at the origin, assuming $x(0) = 0$, i.e., $\psi(x)(t) \equiv x(t) - \inf_{0 \leq s \leq t} x(s)$; see Section 13.5 of [143]. Here is the HT FCLT; it is a variant of Theorem 2 of [84]; see Section 5.7 and 9.6 in [143]. Given Corollary 2.4, it suffices to apply the Continuous Mapping Theorem (CMT) with the reflection map ψ .

Theorem 2.10 (*heavy-traffic FCLT*) *Consider the sequence of G/G/1 models specified above. If, in addition to the conditions of Theorem 2.9,*

$$n^{1/2}(1 - \rho_n) \rightarrow \eta, \quad 0 < \eta < \infty, \quad (2.92)$$

then

$$\left(\hat{\mathbf{W}}_n, \hat{\mathbf{Z}}_n\right) \Rightarrow \left(\psi(\hat{\mathbf{S}}^x - \eta \mathbf{e}), \psi(\hat{\mathbf{S}}^x - \eta \mathbf{e})\right) \quad \text{in } D^2 \quad \text{as } n \rightarrow \infty, \quad (2.93)$$

jointly with the limits in (2.78), where ψ is the reflection map and $\hat{\mathbf{S}}^x - \eta \mathbf{e} \stackrel{d}{=} \sigma_Y \mathbf{B} - \eta \mathbf{e}$ is BM with variance constant σ_Y^2 in (2.81) and drift $-\eta < 0$, so that $\psi(\hat{\mathbf{S}}^x - \eta \mathbf{e})$ is reflected BM (RBM).

The HT approximation for the mean steady-state wait and workload stemming from Theorem 2.10 is

$$E[W(\rho)] \approx E[Z_\rho] \approx \frac{\sqrt{n}\sigma_Y^2}{2\eta} \approx \frac{\sigma_Y^2}{2(1-\rho)} \quad (2.94)$$

for σ_Y^2 in (2.81), which is independent of ρ , using the mean of the exponential limiting distribution of the RBM $\psi(\sigma_x \mathbf{B} - \eta \mathbf{e})(t)$ as $t \rightarrow \infty$.

Remark 2.10 (the two forms of stationarity) *As discussed in the beginning of Section 2.2.2, there are two forms of stationarity, one for discrete time and the other for continuous time. When we focus on the waiting time, we use discrete-time stationarity; when we focus on the workload, we use continuous-time stationarity. So far in this section, we have built everything in the framework of discrete-time stationarity. However, in doing so, we automatically can get FCLT's in both settings. The theoretical basis is provided by [109].*

Remark 2.11 (the limit-interchange problem) *the standard HT limits for the processes do not directly imply limits for the steady-state distributions. Strong results have been obtained with i.i.d. assumptions, e.g., see [31], but the case with dependence is more difficult. Nevertheless, supporting results for the G/G/1 queue when dependence is allowed appear in [126; 127]. We assume that this interchange step is also justified.*

Remark 2.12 (the asymptotic method) *The RQ approach in Theorem 2.2 corresponds to approximating the arrival and service processes in the G/G/1 queue by the asymptotic method in [133], which develops approximations for the arrival and service processes using all the correlations. That is in contrast to the stationary-interval method discussed just before Section 2.5, which uses none of the correlations. Our RQ approach develops an intermediate methods in between those two extremes.*

2.5.5 The IDW and the Normalized Workload

We are motivated to develop the non-parametric RQ for the steady-state workload because of the close connection between the IDW $\{I_w(t) : t \geq 0\}$ and the normalized mean workload $\{c_Z^2(\rho) : 0 \leq \rho \leq 1\}$ established by [60]. The key asymptotic components are the heavy-traffic (HT) and light-traffic (LT) limits stated here in (2.29). Now that we have just developed the supporting HT FCLT, we review the theoretical support for (2.29).

First, the HT limit is supported by the FCLT for $\hat{\mathbf{Z}}_n$ in Theorem 2.10. We use the continuous-time stationarity, justified by Remark 2.10. For the FCLT's, we require weak dependence, which is specified by relatively complex mixing conditions. Given the weak dependence and the FCLT, we need extra regularity conditions to get to what is actually stated in (2.29). First we need the limit-interchange property discussed in Remark 2.11 to get associated limits for the steady-state distributions. Second, we need appropriate uniform integrability to get from convergence of random variables to convergence of their moments; see Remark 2.9.

The LT limit is established in Section IV.A of [60]. An important observation made there is that the LT limiting behavior is much more robust for the steady-state workload than for the steady-state waiting time. In particular, the LT limit for the steady-state waiting time depends more on the fine structure of the model. The LT limits provide theoretical insight into why it is easier to describe the mean steady-state workload than the mean steady-state waiting time, even though they agree in the HT limit.

2.6 Proofs

In this section we provide additional technical support for the main paper. First, a key step in obtaining tractable solutions of the RQ optimizations is an interchange of suprema. The following lemma shows that this interchange is justified in all cases.

Lemma 2.1 (*interchange of suprema*) *The interchange of suprema below holds for any real-valued function $f(x, y)$*

$$M := \sup_{\substack{x \in A \\ y \in B}} f(x, y) = \sup_{x \in A} \sup_{y \in B} f(x, y) = \sup_{y \in B} \sup_{x \in A} f(x, y),$$

where the joint supremum M is allowed to be infinite.

Proof By symmetry, we need only prove that

$$\sup_{\substack{x \in A \\ y \in B}} f(x, y) = \sup_{x \in A} \sup_{y \in B} f(x, y).$$

Suppose the joint supremum M is finite, then there exist a sequence $(x_n, y_n) \in A \times B$ such that $f(x_n, y_n) > M - 1/n$, where M is the finite joint supremum. Then, we have

$$\sup_{x \in A} \sup_{y \in B} f(x, y) \geq \sup_{y \in B} f(x_n, y) \geq f(x_n, y_n) \geq M - \frac{1}{n}, \quad \text{for all } n > 0.$$

This implies that

$$\sup_{x \in A} \sup_{y \in B} f(x, y) \geq M = \sup_{\substack{x \in A \\ y \in B}} f(x, y).$$

The other direction of inequality is trivial by noting that $M \geq f(x, y)$ and taking iterated supremum on both sides.

For the case where the joint supremum M is infinite, then there exist a sequence $(x_n, y_n) \in A \times B$ such that $f(x_n, y_n) > n$. Then

$$\sup_{x \in A} \sup_{y \in B} f(x, y) \geq \sup_{y \in B} f(x_n, y) \geq f(x_n, y_n) \geq n, \quad \text{for all } n > 0.$$

Hence the iterated supremum is also infinite, which completes the proof. ■

Proof of Theorem 2.2. The solutions of the RQ optimizations in (2.17) are

$$\begin{aligned} Z_p^* &\equiv \sup_{\tilde{N} \in \mathcal{U}_p} \sup_{s \geq 0} \left\{ \tilde{N}(t) \right\} = \sup_{s \geq 0} \sup_{\tilde{N} \in \mathcal{U}_p} \left\{ \tilde{N}(t) \right\} \\ &= \sup_{s \geq 0} \left\{ -(1 - \rho)s + b_p \sqrt{\rho s / \mu} \right\} \\ &= -(1 - \rho)x^* + b_p \sqrt{\rho x^* / \mu} = \frac{\rho b_p^2}{4\mu(1 - \rho)} \quad \text{for } x^* \equiv x^*(\rho) = \frac{\rho b_p^2}{4\mu(1 - \rho)^2} \end{aligned} \quad (2.95)$$

and

$$\begin{aligned} Z^* &\equiv \sup_{\tilde{N} \in \mathcal{U}} \sup_{s \geq 0} \left\{ \tilde{N}(t) \right\} = \sup_{s \geq 0} \sup_{\tilde{N} \in \mathcal{U}} \left\{ \tilde{N}(t) \right\} \\ &= \sup_{s \geq 0} \left\{ -(1 - \rho)s + b \sqrt{\text{Var}(N_\rho(s))} \right\} \\ &= \sup_{s \geq 0} \left\{ -(1 - \rho)s + b \sqrt{\text{Var}(Y_\rho(s))} \right\}. \end{aligned} \quad (2.96)$$

where the interchange of suprema is justified by Lemma 2.1. ■

Proof of Theorem 2.3. The inequalities can be satisfied as equalities just as before. There are finite values s_0 such that $\sqrt{V_Y(s)} \leq \sqrt{2\sigma_Y^2 s}$ for all $s \geq s_0$ by virtue of the limit in (5.8). (Also see (2.75) and (2.86).) That shows that the optimization can be regarded as being over closed bounded intervals. The assumed differentiability of V_Y implies that it is continuous, which implies that the supremum is attained over the compact interval. Because $\dot{V}_Y(x) \rightarrow \dot{V}_Y(0) > 0$, we see that there exists a small s' such that

$$-(1-\rho)s + b\sqrt{V_Y(s)} \geq -(1-\rho)s + b\sqrt{s\dot{V}_Y(0)/2} > 0 \quad \text{for all } s \leq s'.$$

As a consequence, the maximum in (2.19) must be strictly positive and must be attained at a strictly positive time.

The results for $\sqrt{V_Y(s)}$ with positive dependence follow from convexity properties of compositions. First, with positive dependence, $-\sqrt{V_Y(s)}$ is a convex function of an increasing convex function, and thus convex so that $\sqrt{V_Y(s)}$ is concave. Second, with negative dependence, we have $V_Y \geq 0$, $\dot{V}_Y(t) \geq 0$ and $\ddot{V}_Y(t) \leq (\leq) 0$. Thus, by direct differentiation

$$\ddot{h}(s) = \frac{b}{\sqrt{V_Y(s)}} \left(\frac{\ddot{V}_Y(s)}{2} - \frac{\dot{V}_Y(s)}{4V_Y(s)} \right) \leq 0,$$

with strictness implying a strict inequality. ■

We now prove Theorem 2.6, which follows from Theorem 2.11 and Theorem 2.12 here.

Theorem 2.11 (*RQ in heavy traffic*) Let $b'_z = \sqrt{2}$ and assume that $I_w(x)$ is non-negative, continuous and that $I_w(\infty) \equiv \lim_{x \rightarrow \infty} I_w(x)$ exist, then we have the following heavy-traffic limit for the normalized RQ optimal value

$$c_{Z^*}^2(1) \equiv \lim_{\rho \rightarrow 1} \frac{2(1-\rho)}{\rho} Z^*(\rho) = I_w(\infty). \quad (2.97)$$

To prove Theorem 2.11, we need two lemmas.

Lemma 2.2 (*order-preservation of the RQ solution*) Let f, g be two positive functions on non-negative real numbers, satisfying $f(x) \geq g(x)$ for all $x \geq 0$. Then we have

$$Z_f^* \geq Z_g^*,$$

where Z_f^* is the solution to the RQ problem with f replacing I_w .

Proof Let x_f^* denote the optimal solution to the RQ problem specified by f . Then

$$\begin{aligned} Z_f^* = -\frac{1-\rho}{\rho}x_f^* + b\sqrt{x_f^*f(x_f^*)} &\geq -\frac{1-\rho}{\rho}x_g^* + b\sqrt{x_g^*f(x_g^*)} \\ &\geq -\frac{1-\rho}{\rho}x_g^* + b\sqrt{x_g^*g(x_g^*)} = Z_g^*. \quad \blacksquare \end{aligned}$$

Lemma 2.3 (*continuity property of the normalized RQ solution*) Let $c_{Z^*}^2(\rho)(f)$ be the normalized solution to (2.34) with I_w replaced by f . Then $c_{Z^*}^2(\rho)$ is a continuous function from space $(C_b(\mathbb{R}^+, \mathbb{R}^+), \|\cdot\|_\infty)$ to $\mathbb{R}^+$, with the former one being the space of all continuous and bounded functions from $\mathbb{R}^+$ to $\mathbb{R}^+$ equipped with the supremum norm.

Proof Let $f, g \in (C_b(\mathbb{R}^+, \mathbb{R}^+), \|\cdot\|_\infty)$, satisfying $\|f - g\|_\infty \leq \epsilon$. Then we have

$$f(x) - \epsilon \leq g(x) \leq f(x) + \epsilon, \quad \text{for all } x \geq 0.$$

Since $f \in C_b(\mathbb{R}^+, \mathbb{R}^+)$, there exist $M > 0$ such that $f(x) < M$ for all $x \geq 0$. Then for all $x > M_\rho$, where $M_\rho \equiv (\rho b'_z/(1-\rho))^2 M$, we have

$$-\frac{1-\rho}{\rho}x + b'_z\sqrt{xf(x)} < -\frac{1-\rho}{\rho}x + b'_z\sqrt{xM} < 0$$

Hence,

$$\begin{aligned} c_{Z^*}(\rho)(g) &\leq c_{Z^*}(\rho)(f + \epsilon) = \frac{2(1-\rho)}{\rho} \sup_{0 \leq x \leq \tilde{M}_\rho} \left\{ -\frac{1-\rho}{\rho}x + b'_z\sqrt{x(f(x) + \epsilon)} \right\} \\ &\leq \frac{2(1-\rho)}{\rho} \sup_{0 \leq x \leq \tilde{M}_\rho} \left\{ -\frac{1-\rho}{\rho}x + b'_z\sqrt{xf(x)} + b'_z\sqrt{x\epsilon} \right\} \\ &\leq \frac{2(1-\rho)}{\rho} \sup_{0 \leq x \leq \tilde{M}_\rho} \left\{ -\frac{1-\rho}{\rho}x + b'_z\sqrt{xf(x)} \right\} + b'_z\sqrt{\tilde{M}_\rho\epsilon} \end{aligned} \quad (2.98)$$

$$\begin{aligned} &= c_{Z^*}(\rho)(f) + \frac{2(1-\rho)}{\rho}b'_z\sqrt{\tilde{M}_\rho\epsilon} \\ &= c_{Z^*}(\rho)(f) + 2(b'_z)^2\sqrt{(M + \epsilon)\epsilon}, \end{aligned} \quad (2.99)$$

where $\tilde{M}_\rho \equiv (\rho b'_z/(1-\rho))^2(M + \epsilon)$ and the first inequality follows from Lemma 2.2. Similarly, we can prove that

$$c_{Z^*}(\rho)(g) \geq c_{Z^*}(\rho)(f - \epsilon) \geq c_{Z^*}(\rho)(f) - 2(b'_z)^2\sqrt{(M + \epsilon)\epsilon}. \quad (2.100)$$

Combining (2.99) and (2.100), we have

$$|c_{Z^*}(\rho)(g) - c_{Z^*}(\rho)(f)| \leq 2(b'_z)^2\sqrt{(M + \epsilon)\epsilon}.$$

Hence the lemma holds. $\blacksquare$

Proof of Theorem 2.7. The FCLT of the MAP here is a special case of Theorem 3.2 and the following Remark 3.4 in [62]. We now verify the assumptions there, see (AS1)-(AS3) there. The stationarity of the MAP (AS1) and the second moment assumption (AS3) are directly assumed in our settings.

Now, we turn to the spectral gap assumption (AS2). Towards this end, the finite state space assumed here significantly simplifies the proof. First, as discussed in Section 2.4 of [62], for Markov jump processes, it suffices to consider the standard exponential ergodicity

$$|P_{i,j}(t) - \alpha_j| \leq Ce^{-\beta t}, \text{ for } t \geq 0, \text{ and any } \{i, j\},$$

for some finite and positive constants C and β . It is well-known that the exponential ergodicity is equivalent to that of any skeleton chain, i.e. the Markov chain obtained by evaluating the Markov process at $t_n = n\tau, n \in \mathbb{N}$ for any τ , see for example Theorem 1 in [130]. Note that the skeleton chain of a finite-state CTMC is always aperiodic. The required exponential ergodicity then follows from the fact that an irreducible and aperiodic discrete-time Markov Chain is always exponentially ergodic, see Theorem 4.9 of [98]. $\square$

Proof of Theorem 2.11. Recall that Theorem 2.5 suggest that the optimal solution is of order $O(\rho^2/(2(1-\rho)^2))$, we perform a change of variable $t = 2(1-\rho)^2 x/\rho^2$ in (2.34) and scale the space by a constant $\rho/(2(1-\rho))$. Hence, we have

$$c_{Z^*}^2(\rho) = \sup_{0 \leq t \leq \infty} \left\{ -t + 2\sqrt{t I_w \left(\frac{\rho^2}{2(1-\rho)^2} t \right)} \right\}. \quad (2.101)$$

Since $I_w(\infty) \equiv \lim_{x \rightarrow \infty} I_w(x)$ exist, there exist a T sufficiently large such that $|I_w(t) - I_w(\infty)| < \epsilon$ for all $t > T$. Now, we define

$$\tilde{I}_w(t) = \begin{cases} I_w(t), & t \leq T, \\ \text{linear}, & T - \epsilon < t \leq T, \\ I_w(\infty), & t > T. \end{cases}$$

By virtue of Lemma 2.3, we need only prove that $c_{Z^*}(1)(\tilde{I}_w) = \tilde{I}_w(\infty) = I_w(\infty)$.

Note that continuity and finite limit at $x = \infty$ implies that $I_x(x)$ is bounded, say $I_w(x) < M - \epsilon$ for all $x \geq 0$. Hence we have

$$-t + 2\sqrt{t \tilde{I}_w \left(\frac{\rho^2}{2(1-\rho)^2} t \right)} \leq -t + 2\sqrt{tM}. \quad (2.102)$$

We assume first that the limit $I_w(\infty)$ is strictly positive. The case where $I_w(\infty) = 0$ can be deduced by considering a sequence of functions $f_n(x)$ such that $f_n(\infty) > 0$ and $|I_w - f_n|_\infty < 1/n$, and applying Lemma 2.3.

Now, for the case where $I_w(\infty) > 0$, we can choose ρ_0 such that

$$T_\rho \equiv \frac{2(1-\rho_0)^2}{\rho_0^2} T < \min \left\{ I_w(\infty), 2M - I_w(\infty) - 2\sqrt{M^2 - I_w(\infty)M} \right\},$$

since the right-hand-side of the inequality will be strictly positive. Then for all $\rho > \rho_0$, we have

$$\begin{aligned} \sup_{0 \leq t \leq T_\rho} \left\{ -t + 2\sqrt{t\tilde{I}_w \left(\frac{\rho^2}{2(1-\rho)^2} t \right)} \right\} &\leq \sup_{0 \leq t \leq T_\rho} \left\{ -t + 2\sqrt{tM} \right\} \\ &\leq I_w(\infty). \end{aligned}$$

But plugging $I_w(\infty)$ into the objective function, we have the objective value $I_w(\infty)$ by the fact that $\frac{\rho^2}{2(1-\rho)^2} I_w(\infty) > T$ and that $\tilde{I}_w(t)$ is constant after $t > T$. This implies that

$$\begin{aligned} c_{Z^*}^2(\rho)(\tilde{I}_w) &= \sup_{T_\rho \leq t \leq \infty} \left\{ -t + 2\sqrt{t\tilde{I}_w \left(\frac{\rho^2}{2(1-\rho)^2} t \right)} \right\} \\ &= \sup_{T_\rho \leq t \leq \infty} \left\{ -t + 2\sqrt{tI_w(\infty)} \right\} \\ &= I_w(\infty), \quad \text{for all } \rho > \rho_0. \end{aligned}$$

Hence, we've proved that $c_{Z^*}(1)(\tilde{I}_w) = \tilde{I}_w(\infty) = I_w(\infty)$. ■

Next, we state the corresponding result for RQ in light traffic.

Theorem 2.12 (*RQ in light traffic*) *Let $b'_z = \sqrt{2}$ and assume that $I_w(x)$ is non-negative, continuous and that $I_w(0) \equiv \lim_{x \rightarrow 0} I_w(x)$ exist, then we have the following light-traffic limit for the normalized RQ optimal value*

$$c_{Z^*}^2(0) \equiv \lim_{\rho \rightarrow 0} \frac{2(1-\rho)}{\rho} Z^*(\rho) = I_w(0). \quad (2.103)$$

Proof As in the proof for heavy-traffic limit, we perform the same time and space scaling to get (2.101). For the same reason, we have (2.102), which implies that

$$-t + 2\sqrt{t\tilde{I}_w \left(\frac{\rho^2}{2(1-\rho)^2} t \right)} \leq -t + 2\sqrt{tM} < 0, \quad \text{for all } t > 4M.$$

Hence, we need only consider the supremum in (2.101) over bounded interval $[0, 4M]$. Note also that, since $I_w(0) \equiv \lim_{x \rightarrow 0} I_w(x)$ exist, for any $\epsilon > 0$, there exist a $\delta > 0$ such that $|I_w(t) - I_w(0)| < \epsilon$ for all $x \in [0, \delta]$. We now choose ρ_0 such that $2\rho_0^2 M / (1 - \rho_0)^2 < \delta$, and take a modification

$$\tilde{I}_w(t) = \begin{cases} I_w(0), & t < \delta, \\ \text{linear}, & \delta \leq t < \delta + \epsilon, \\ I_w(t), & t \geq \delta + \epsilon, \end{cases}$$

which satisfies $|I_w - \tilde{I}_w|_\infty < \epsilon$ and

$$c_{Z^*}^2(\rho)(\tilde{I}_w) = I_w(0), \quad \text{for all } \rho < \rho_0.$$

We then apply Lemma 2.3 to get the desired light-traffic limit. ■

Proof of Theorem 2.8. Let $K = rk - r + 1$ be the sample size, and assume that $V(t) = I(t)t < Ct$ for some constant C . Then

$$\begin{aligned} E[\bar{V}] &= \frac{1}{K-1} \sum_{i=1}^K E[U_i^2] - \frac{1}{K(K-1)} E \left[\left(\sum_{i=1}^K U_i \right)^2 \right] \\ &= \frac{1}{K-1} \left(\sum_{i=1}^K E[U_i^2] - \frac{1}{K} E \left[\sum_{i=1}^K U_i^2 + 2 \sum_{i>j} U_i U_j \right] \right) \\ &= E[U_1^2] - E[U_1]^2 - \frac{2}{K(K-1)} \sum_{i<j} \text{cov}(U_i, U_j) \\ &= V(t) - \frac{2}{K(K-1)} \left(\sum_{j<i<j+r} \text{cov}(U_i, U_j) + \sum_{i>j+r+1} \text{cov}(U_i, U_j) \right) \\ &= V(t) - \frac{2}{K(K-1)} \left(\sum_{i=1}^{r-1} (K-i) \text{cov}(U_1, U_{i+1}) + \sum_{i=r}^{K-1} (K-i) \text{cov}(U_1, U_{i+1}) \right) \\ &\equiv V(t) - (A + B) \end{aligned}$$

The covariance terms can be expressed as

$$\text{cov}(U_1, U_{1+i}) = \begin{cases} V(t - i\tau) + V(t + i\tau) - V(t) - V(i\tau), & i = 1, 2, \dots, r-1 \\ V(t + i\tau) - 2V(i\tau) + V(i\tau - t), & i = r, r+1, \dots, K-1 \end{cases} \quad (2.104)$$

Using the bound on $I(t)$, we have

$$\begin{aligned} A &= \frac{2}{K(K-1)} \sum_{i=1}^{r-1} (K-i) \text{cov}(U_1, U_{i+1}) \\ &\leq \frac{2}{K} \sum_{i=1}^{r-1} (V(t-i\tau) + V(t+i\tau)) \\ &\leq \frac{4Ct(r-1)}{K} \leq \frac{4Ct}{k-1}, \end{aligned}$$

and

$$\begin{aligned} B &= \frac{2}{K(K-1)} \sum_{i=r}^{K-1} (K-i) \text{cov}(U_1, U_{i+1}) \\ &\leq \frac{2}{K} \sum_{i=r}^{K-1} ((V(t+i\tau) - V(i\tau)) - (V(i\tau) - V(i\tau-t))) \\ &\leq \frac{2t}{K} \sum_{i=r}^{K-1} \left(\frac{V(t+i\tau) - V(i\tau)}{t} - \frac{V(i\tau) - V(i\tau-t)}{t} \right) \\ &\rightarrow 0, \text{ as } k \rightarrow \infty, \end{aligned}$$

where we used the regularity condition that $\dot{V}(t) \rightarrow \sigma^2$ as $t \rightarrow \infty$, and the fact that the average converges to 0 if the summands converge to 0.

Note that

$$\bar{V}_k \equiv \frac{1}{k-1} \sum_{i=1}^k U_i^2 - \frac{1}{k(k-1)} \left(\sum_{i=1}^k U_i \right)^2$$

By Continuous Mapping Theorem, we need only prove that both $\{U_i\}$ and $\{U_i^2\}$ follows Strong Law of Large Number (SLLN). This in turns is implied by the Strong Ergodic Theorem for stationary and ergodic sequence. The stationarity of both sequences are implied by the time-stationarity of the point process $N(t)$. The ergodicity of both sequence follows from the ergodicity of the underlying process $N(t)$. ■

Chapter 3

Robust Queueing for Queues In Series

In this Chapter, we generalize the RQ algorithm for single-server queues in (2.35) to a Robust Queueing Network Analyzer algorithm in the case of queues in series (tandem queues).

Towards this end, it is essential that the IDC at any station in the series of queues can be easily estimated or calculated. However, the departure process, which serves as the arrival process of the subsequent queue, can exhibit complicate dependence structure. For example, as discussed in [51; 55] and references there, the stationary departure process from a $GI/GI/1$ queue is a renewal process (ordinary or stationary) if and only if the queue is an $M/M/1$ queue, in which case it is a Poisson process. The major challenge is then to develop an effective approximation for the IDC of departure flow at each station.

We contribute to the understanding of the departure process by establishing the first heavy-traffic limit theorem for the stationary departure process from a $GI/GI/1$ queue. This heavy-traffic limit theorem allow us to justify the following approximation for the departure IDC $I_d(t)$ by the weighted average of the IDC's of the arrival and service processes, i.e.,

$$I_d(t) \approx w_\rho(t)I_a(t) + (1 - w_\rho(t))I_s(\rho t), \quad t \geq 0, \quad (3.1)$$

where $I_a(t)$ and $I_s(t)$ are the IDCs associated with the equilibrium arrival and service

renewal processes, w_ρ is a weight function, defined in (3.67), that depends on the interarrival-time and service-time distribution only through their first two moments.

The rest of this section is organized as follows. In Section 3.1, we review preliminary results for general stationary point processes, focusing especially on their variance functions. In Section 3.2, we establish the heavy-traffic limits for the stationary departure process and its variance function for a $GI/GI/1$ model. In Section 3.3, we develop approximation for the stationary departure IDC, and develop the RQNA algorithm for the queues-in-series models. Finally, we present postponed proofs in Section 3.4.

3.1 Review of Stationary Point Processes

In this section we review basic properties of stationary point processes; see [52] and [122] for more background. In Section 3.1.1 we review renewal processes and their Laplace transforms. In Section 3.1.3 we review the Palm-Khintchine equation and use it to express the variance function of a stationary point process in terms of the mean function of the Palm version.

3.1.1 Renewal Processes and the Laplace Transform

We start with a rate- λ renewal process $N \equiv \{N(t) : t \geq 0\}$. Let F be the *cumulative distribution function* (cdf) of the interval U between points (the interarrival time in a GI arrival process), having mean $E[U] = \lambda^{-1}$ and finite second moment. As a regularity condition for our queueing application, we also assume that F has a *probability density function* (pdf) f , where $F(t) = \int_0^t f(u) du$, $t \geq 0$. Throughout this paper, we assume that the interarrival-time distribution of our renewal arrival processes has a pdf. That pdf assumption ensures that the equilibrium renewal process arises as the time limit of the ordinary renewal process; e.g., see Section 3.4 and Section 3.5 of [116].

The stationary or equilibrium renewal process differs from the ordinary renewal process only by the distribution of the first interarrival times. Let F_e be the cdf of the equilibrium distribution, which has pdf $f_e(t) = \lambda(1 - F(t))$. Let $E^e[\cdot]$ denote the expectation under the stationary distribution (with first interval distributed according to F_e) and let $E^0[\cdot]$ denote

the expectation under *the Palm distribution* (with first interval distributed as F).

Conditioning on the first arrival, distributed as F under the Palm distribution or as F_e under stationary distribution, the renewal equations for the mean and second moment of $N(t)$, the number of points in an interval $[0, t]$, are:

$$\begin{aligned} m(t) &\equiv E^0[N(t)] = F(t) + \int_0^t m(t-s) dF(s), \\ m_e(t) &\equiv E^e[N(t)] = F_e(t) + \int_0^t m(t-s) dF_e(s), \\ \sigma(t) &\equiv E^0[N^2(t)] = F(t) + 2 \int_0^t m(t-s) dF(s) + \int_0^t \sigma(t-s) dF(s), \\ \sigma_e(t) &\equiv E^e[N^2(t)] = F_e(t) + 2 \int_0^t m(t-s) dF_e(s) + \int_0^t \sigma(t-s) dF_e(s). \end{aligned}$$

Throughout the paper, we use the Laplace Transform (LT) instead of the Laplace-Stieltjes Transform (LST). The LT of $f(t)$ and the LST of F , denoted by $\mathcal{L}(f)(s) \equiv \hat{f}(s)$, are

$$\hat{f}(s) \equiv \mathcal{L}(f)(s) \equiv \int_0^\infty e^{-st} f(t) dt = \int_0^\infty e^{-st} dF(t), \quad (3.2)$$

so that $f(t) = \mathcal{L}^{-1}(\hat{f})(t)$. Throughout the paper, we add a hat to either an LT or an item that appears in LT. The LT of f_e is then

$$\hat{f}_e(s) = \frac{\lambda(1 - \hat{f}(s))}{s} \text{ and } \hat{F}_e(s) = \frac{\hat{f}_e(s)}{s},$$

where $\lambda^{-1} \equiv \int_0^\infty t f(t) dt$ is the mean. Applying the LT to the renewal equations, we obtain

$$\hat{m}(s) = \frac{\hat{f}(s)}{s(1 - \hat{f}(s))}, \quad (3.3)$$

$$\hat{m}_e(s) = \frac{\hat{f}_e(s)}{s(1 - \hat{f}(s))} = \frac{\lambda}{s^2}, \quad (3.4)$$

$$\hat{\sigma}(s) = \frac{\hat{f}(s) + 2s\hat{m}(s)\hat{f}(s)}{s(1 - \hat{f}(s))} = \frac{\hat{f}(s)(1 + \hat{f}(s))}{s(1 - \hat{f}(s))^2}, \quad (3.5)$$

$$\hat{\sigma}_e(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s}\hat{m}(s) = \frac{\lambda(1 + \hat{f}(s))}{s^2(1 - \hat{f}(s))}. \quad (3.6)$$

From (3.4), we see that

$$E^e[N(t)] = \lambda t, \quad t \geq 0, \quad (3.7)$$

as must be true for any stationary point process.

Let $V(t) \equiv \text{Var}^e(N(t))$ be the variance process of $N(t)$ under time-stationary distribution. (We omit the e superscript on $V(t)$ because we will only discuss stationary variance functions.) Combining (3.6) and (3.7), we have

$$\hat{V}(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s} \hat{m}(s) - \frac{2\lambda^2}{s^3} = \frac{\lambda}{s^2} + \frac{2\lambda}{s} \frac{\hat{f}(s)}{s(1 - \hat{f}(s))} - \frac{2\lambda^2}{s^3}. \quad (3.8)$$

The variance function then can be obtained from the numerical inversion of the Laplace transform, e.g., see Section 13 of [4] and [5]. Term by term inversion shows that we can express $V(t)$ in terms of the renewal function $m(t)$

$$V(t) = \lambda \int_0^t (1 + 2m(u) - 2\lambda u) du. \quad (3.9)$$

3.1.2 Revisiting the IDC

We revisit the IDC by presenting the following renewal process characterization theorem. This theorem states that, for renewal process, the IDC function encodes the full information of the inter-renewal time distribution.

Theorem 3.1 (Renewal process characterization theorem) *A renewal process with an inter-renewal distribution having pdf f and cdf F having finite first two moments with positive mean λ^{-1} is fully characterized by any one of the following:*

1. the pdf $f(t)$ of the time between renewals;
2. the cdf $F(t)$ of the time between renewals;
3. the LT $\hat{f}(s)$;
4. the renewal function $m(t)$;
5. the LT $\hat{m}(s)$;
6. the rate λ and the variance function of the equilibrium renewal process $\sigma_e(t)$;
7. the rate λ and the LT $\hat{\sigma}_e(s)$;
8. the rate λ and the IDC $I_e(t) \equiv \sigma_e(t)/\lambda t$ of the equilibrium renewal process.

Proof The equivalence of the time functions and their transforms follows from the basic theory of Laplace transforms. Hence, we obtain the equivalence by explicit expressions in terms of the Laplace transforms, i.e.,

$$\hat{f}(s) = \frac{s\hat{m}(s)}{1 + s\hat{m}(s)} \quad \text{and} \quad \hat{f}(s) = \frac{s^2\hat{\sigma}_e(s) - \lambda}{s^2\hat{\sigma}_e(s) + \lambda}. \quad (3.10)$$

Then, from the definition of the IDC, we obtain $\sigma_e(t) = \lambda(t)I_e(t)$, $t \geq 0$. ■

Corollary 3.1 (Full characterization of a $GI/GI/1$ queue) *The $GI/GI/1$ queue with interarrival-time cdf F and service-time cdf G having finite second moments is fully characterized by the four-tuple $(\lambda, I_a(t), \tau, I_s(t))$, where τ is the mean service time and $I_a(t)$ ($I_s(t)$) is the IDC of the equilibrium renewal process associated with the interarrival (service) times.*

3.1.3 The Palm-Khintchine Equation

In this section, we show that the Palm-Khintchine equation can be used to derive a generalization of (3.9) for general stationary and ergodic point processes.

Consider a continuous-time stationary point process, i.e., having stationary increments. The main idea is the Palm transformation relating continuous-time stationary processes to the associated discrete-time stationary processes. An important manifestation of that relation is the Palm-Khintchine equation; see Theorem 3.4.II. of [52]. It is important here because it can be applied to generalize the variance formula discussed in Section 3.1.1; see Section 2.4 of [49] and Section 3.4 of [52].

We focus on orderly stationary ergodic point processes with finite intensity. (Orderly means that the points occur one at a time.) Let $N(s, t]$ denote the number of events in interval $(s, t]$, and $N(t) \equiv N(0, t]$.

Theorem 3.2 (Palm-Khintchine equation) *For an orderly stationary point process of finite intensity λ such that $P^e(N(-\infty, 0] = N(0, \infty) = \infty) = 1$, then*

$$\begin{aligned} P^e(N(t) \leq k) &= 1 - \lambda \int_0^t q_k(u) du \\ &= \lambda \int_t^\infty q_k(u) du, \quad \text{for } k = 0, 1, 2, \dots, \end{aligned} \quad (3.11)$$

where $q_k(t)$ is the probability of exactly k arrivals in $(0, t]$ under the Palm distribution, i.e.,

$$q_k(t) = \lim_{h \downarrow 0} P(N(t) = k | N(-h, 0] > 0). \quad (3.12)$$

Under ergodicity, the Palm distribution is equivalent to the event stationary distribution, so that $q_k(t) = P^0(N(t) = k)$.

We now apply Theorem 3.2 to generalize (3.9) and (3.8) to the case of orderly stationary ergodic point process.

Corollary 3.2 (Variance of a stationary ergodic point process) *For a general stationary ergodic point process with rate λ and finite second moment, the variance function is*

$$V(t) = \lambda \int_0^t (1 + 2m(u) - 2\lambda u) du, \quad t \geq 0, \quad (3.13)$$

where

$$m(t) \equiv E^0[N(t)] = \sum_{k=1}^{\infty} k q_k(t), \quad t \geq 0, \quad (3.14)$$

and its LT is

$$\hat{V}(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s} \hat{m}(s) - \frac{2\lambda^2}{s^3}, \quad (3.15)$$

where $\hat{m}(s)$ is the LT of $m(t)$.

Proof Let

$$p_k(t) = P^e(N(t) = k), \text{ for } k = 0, 1, 2, \dots \quad (3.16)$$

so that $\sum_{i=1}^k p_i(t) = P^e(N(t) \leq k)$. With Theorem 3.2, we can write

$$\begin{aligned} V(t) &= \sum_{k=1}^{\infty} k^2 p_k(t) - \lambda^2 t^2 = \sum_{k=1}^{\infty} k^2 \lambda \int_0^t (q_{k-1}(u) - q_k(u)) du - \lambda^2 t^2 \\ &= \lambda \int_0^t (1 + 2m(u) - 2\lambda u) du, \end{aligned} \quad (3.17)$$

where $m(t) \equiv E^0[N(t)]$ as in (3.14). Taking the Laplace Transform, we obtain

$$\hat{V}(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s} \hat{m}(s) - \frac{2\lambda^2}{s^3}. \quad \blacksquare$$

3.2 Heavy-Traffic Limit Theorem for $GI/GI/1$ Stationary Departure

The departure process is of considerable interest in general, because the stationary departure process from a $GI/GI/1$ queue is remarkably complicated; e.g., it is only a stationary renewal process in the special case of an $M/M/1$ model, when it is Poisson, by Burke's [32] theorem, also see [110] and references therein. Indeed, explicit transform expressions for the variance function of the stationary departure process are evidently only available for the $M/GI/1$ and $GI/M/1$ models, due to Takacs [128] and Daley [50; 51]; see [21] and [82] for related results on $GI/GI/1$. In [79; 152], the departure process from a $BMAP/MAP/1$ queue is shown to be a MAP of with an infinite number of states; and truncation approximation has been proposed for practical use.

In Section 3.2.1 we use Laplace transforms (LT's) of the stationary departure process in the $GI/M/1$ queue derived by [50; 51] to derive the HT limit of its variance function. In Section 3.2.2 we use the HT limit for the Palm version of the mean function derived by [128] to derive the HT limit of the stationary variance function. In Section 3.2.3 we establish the HT limit for the stationary departure process in the $GI/GI/1$ queue (Theorem 3.7) and its variance function (Theorem 3.8).

3.2.1 The Departure Variance in the $GI/M/1$ Queue

We now start considering the queueing models. In particular, we focus on the $GI/GI/1$ queue, which has unlimited waiting space, the first-come first-served service discipline and independent sequences of i.i.d. interarrival times and service times distributed as random variables U and V , respectively, where U has a pdf $f(t)$. Let $\lambda \equiv 1/E[U]$ be the arrival rate; let $\hat{f}(s) \equiv E[e^{-sU}]$ be the LT of the interarrival-time pdf $f(t)$; let $\mu \equiv 1/E[V]$ be the service rate; and let $\rho \equiv \lambda/\mu$ be the traffic intensity, assuming that $\rho < 1$.

Daley [50; 51] derived the LST of the variance $V_d(t)$ of the stationary departure process in a $GI/M/1$ queue. The associated LT of $V_d(t)$ is

$$\hat{V}_d(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s^3} \left(\mu\delta - \lambda + \frac{\mu^2(1-\delta)(1-\hat{\xi}(s))(\mu\delta(1-\hat{f}(s)) - s\hat{f}(s))}{(s + \mu(1-\hat{\xi}(s)))(s - \mu(1-\delta))(1-\hat{f}(s))} \right) \quad (3.18)$$

where $\hat{\xi}(s)$ is the root with the smallest absolute value in z of the equation

$$z = \hat{f}(s + \mu(1 - z)) \quad (3.19)$$

and $\delta = \hat{\xi}(0)$ is the unique root in $(0, 1)$ of the equation

$$\delta = \hat{f}(\mu(1 - \delta)), \quad (3.20)$$

which appears in the distribution of the stationary queue length in a $GI/M/1$ queue. Useful properties of $\hat{\xi}(s)$ and $\delta = \hat{\xi}(0)$ are contained in Lemma 3.2.

We now establish a HT limit for the departure variance function in the $GI/M/1$ model. To do so, we consider a family of $GI/M/1$ models parameterized by ρ , where $\lambda \equiv 1/E[U]$ and $\mu = \mu_\rho = 1/E[V] \equiv \lambda(1 + (1 - \rho)\gamma_\rho)$, where γ_ρ are positive constants such that $\lim_{\rho \uparrow 1} \gamma_\rho = \gamma > 0$. Note that if $\gamma_\rho = 1/\rho$, then we come to the usual case of $\lambda/\mu = \rho$. We allow this general scaling so that we can gain insight into reflected Brownian motion (RBM) with non-unit drift. Let the HT-scaled variance function be

$$V_{d,\rho}^*(t) \equiv (1 - \rho)^2 V_{d,\rho}((1 - \rho)^{-2}t), \quad t \geq 0. \quad (3.21)$$

Throughout this chapter, we use the star ($\star$) superscript with ρ subscript to denote HT-scaled items in the queueing model, as in (3.21), and the star superscript without the ρ subscript to denote the associated HT limit.

As should be expected from established HT limits, e.g., as in Section 5.7 and Chapter 9 in [143], the HT limit of the variance function $V_{d,\rho}^*(t)$ in (3.21) depends on properties of the normal distribution and RBM. Let $\phi(x)$ be the pdf and $\Phi(x)$ the cdf of the standard normal variable $N(0, 1)$. Let $\Phi^c(x) \equiv 1 - \Phi(x)$ be the complementary cdf (ccdf). Let $R(t)$ be canonical RBM (having drift -1 , diffusion coefficient 1) and let $R_e(t)$ be the stationary version, which has the exponential marginal distribution for each t with mean $1/2$. The correlation function $c^*(t)$ of R_e is defined as

$$\begin{aligned} c^*(t) &\equiv \text{cov}(R_e(0), R_e(t)) = 2(1 - 2t - t^2)\Phi^c(\sqrt{t}) + 2\sqrt{t}\phi(\sqrt{t})(1 + t) \\ &= 1 - H_2^*(t) \equiv 1 - \frac{E[R(t)^2 | R(0) = 0]}{E[R(\infty)^2]} \\ &= 1 - 2E[R(t)^2 | R(0) = 0], \quad t \geq 0, \end{aligned} \quad (3.22)$$

where $H_2^*(t)$ is the second-moment cdf of canonical RBM in [2], which has mean 1 and variance 2.5; see Corollaries 1.1.1 and 1.3.4 of [2] and Corollary 1 of [3]. The correlation function $c^*(t)$ has LT

$$\hat{c}^*(s) \equiv \frac{1}{s} - \frac{2}{s^2} \left(1 - \frac{\sqrt{1+2s} - 1}{s} \right); \quad (3.23)$$

see (1.10) of [2]. Equivalently, the Gaussian terms in (3.22) can be re-expressed as $\phi(\sqrt{t}) = e^{-t/2}/\sqrt{2\pi}$ and $\Phi^c(\sqrt{t}) = \left(1 - \text{erf}(\sqrt{t}/2)\right)/2$, where erf is the error function. This LT can be explicitly inverted, yielding

$$c^*(t) = -2(t^2 + 2t - 1)\Phi^c(\sqrt{t}) + 2\phi(\sqrt{t})\sqrt{t}(1+t). \quad (3.24)$$

By Corollary 1.3.5 of [2], the correlation function has tail asymptotics according to

$$c^*(t) = 1 - H_2^*(t) \sim \frac{16}{\sqrt{2\pi}t^3} e^{-(t/2)} \quad \text{as } t \rightarrow \infty. \quad (3.25)$$

From the correlation function, we define

$$w^*(t) \equiv 1 - \frac{1 - c^*(t)}{2t}, \quad t \geq 0. \quad (3.26)$$

It then follows from the explicit expression of $c^*(t)$ that

$$w^*(t) = \frac{1}{2t} \left((t^2 + 2t - 1) \left(1 - 2\Phi^c(\sqrt{t}) \right) + 2\phi(\sqrt{t})\sqrt{t}(1+t) - t^2 \right). \quad (3.27)$$

It can be easily verified that $w^*(t)$ is an increasing function satisfying $0 \leq w^*(t) \leq 1$. As we shall see in Theorem 3.3, this $w^*(t)$ serves as the weight function that appears in the limiting departure variance function.

We now present the main result for the departure variance in the $GI/M/1$ special case. The idea of the proof is to exploit the explicit form of the LT $\hat{V}_{d,\rho}^*(t)$ of the scaled stationary departure variance and derive its HT limit. We then obtain the convergence of the HT-scaled variance function $V_{d,\rho}^*(t)$ by applying continuity theorem of the LT, see Theorem 2(a) in Chapter XIII of [57]. The proof of the theorem can be found in Section 3.4.

Theorem 3.3 (HT limit for the $GI/M/1$ departure variance) *Consider the $GI/M/1$ model with $1/E[U] = \lambda$ and $1/E[V] = \mu_\rho \equiv \lambda(1 + (1-\rho)\gamma_\rho)$, where γ_ρ are positive constants such that $\lim_{\rho \uparrow 1} \gamma_\rho = \gamma > 0$. Assume that $E[U^3] < \infty$ so that a two-term Taylor series*

expansion of the LT $\hat{f}(s)$ about the origin is valid with asymptotically negligible remainder. Then the HT-scaled variance function $V_{d,\rho}^*(t)$ defined in (3.21) converges as $\rho \uparrow 1$; i.e.,

$$V_{d,\rho}^*(t) \rightarrow V_d^*(t), \quad \text{as } \rho \uparrow 1 \quad \text{for all } t \geq 0, \quad (3.28)$$

where $V_d^*(t)$ is a continuous nonnegative real-valued function with LT

$$\begin{aligned} \hat{V}_d^*(s) &= \frac{\lambda}{s^2} + \frac{2\lambda}{s^2} \frac{c_a^2 - 1}{c_a^2 + 1} \frac{\gamma}{\hat{\xi}^*(s)} \\ &= \frac{\lambda c_a^2}{s^2} - \frac{\lambda(c_a^2 - 1)}{s^2} \left(1 - \frac{2}{c_a^2 + 1} \frac{\gamma}{\hat{\xi}^*(s)} \right), \end{aligned} \quad (3.29)$$

with $\hat{\xi}^*(s)$ being the unique root with non-negative real part of the quadratic equation

$$\left(\frac{c_a^2 + 1}{2} \right) \hat{\xi}^*(s)^2 - \gamma \hat{\xi}^*(s) - \frac{s}{\lambda} = 0. \quad (3.30)$$

In addition,

$$V_d^*(t) = w^* (\lambda \gamma^2 t / c_x^2) c_a^2 \lambda t + (1 - w^* (\lambda \gamma^2 t / c_x^2)) c_s^2 \lambda t \quad (3.31)$$

for $w^*(t)$ defined in (3.27), $c_x^2 \equiv c_a^2 + c_s^2$, $c_a^2 \equiv \text{Var}(U)/E[U]^2$ and $c_s^2 = 1$.

We shall want to relate our HT limit for the departure variance function to associated HT limits for the variance functions of the arrival and service processes. For that step, it is significant that the functional central limit theorem (FCLT) for any stationary point process has the same form as the FCLT for the associated Palm process, as was shown by [109]. Extra uniform integrability is required to get the associated limit for the variance function. To be relatively self-contained, we will directly derive the desired result from the transform of the equilibrium renewal process in (3.8).

For that purpose, let $A(t)$ denote the arrival renewal process and let $V_a(t) \equiv \text{Var}^e(A(t))$ denote its variance process under the stationary distribution. Similarly, we define $S(t)$ and $V_s(t)$ for the renewal process associated with the service-time distribution. The following lemma states that the terms $c_a^2 \lambda t$ and $c_s^2 \lambda t$ in (3.31) can be interpreted as the limiting variance function of the arrival and service renewal processes, respectively. This implies that the limiting departure variance function V_d^* is a convex combination of the arrival and service variance functions with a scaled version of the time-varying weight function $w^*(t)$. This convex combination result is consistent with the more elementary approximation used

in QNA; see [134; 136]; there the departure variability parameter is approximated by a convex combination of the arrival and service variability parameters.

Lemma 3.1 (*Limiting variance function of stationary renewal processes*) *Let $N(t)$ be a renewal process with rate λ and let c_N^2 be the scv of the inter-renewal-time distribution. Consider the HT-scaled stationary variance function*

$$V_{N,\rho}^*(t) \equiv \text{Var}^e((1-\rho)N((1-\rho)^{-2}t)),$$

then

$$V_{N,\rho}^*(t) \rightarrow V_N^*(t) \equiv \lambda c_N^2 t, \quad \text{as } \rho \uparrow 1.$$

Proof Let f denote the inter-renewal distribution. Recall the expression for the LT of a stationary renewal process in (3.8), we have

$$\begin{aligned} \hat{V}_N^*(s) &= \lim_{\rho \uparrow 1} \hat{V}_{N,\rho}^*(s) \equiv \lim_{\rho \uparrow 1} \mathcal{L}((1-\rho)^2 V_{N,\rho}((1-\rho)^{-2}t)) \\ &= \lim_{\rho \uparrow 1} (1-\rho)^4 \hat{V}_{N,\rho}((1-\rho)^2 t) \\ &= \lim_{\rho \uparrow 1} \left(\frac{\lambda}{s^2} + \frac{2\lambda}{s^2} \frac{\hat{f}((1-\rho)^2 t)}{1 - \hat{f}((1-\rho)^2 t)} - \frac{2\lambda^2}{(1-\rho)^2 s^3} \right) \\ &= \frac{\lambda}{s^2} + \frac{2\lambda^2}{s^2} \lim_{\rho \uparrow 1} \frac{1}{(1-\rho)^2 s} \left(\frac{\hat{f}((1-\rho)^2 t)}{\lambda \frac{1 - \hat{f}((1-\rho)^2 t)}{(1-\rho)^2 s}} - 1 \right) = \frac{\lambda c_N^2}{s^2}. \end{aligned}$$

The result follows from inverting the LT, i.e., $V_s^*(t) = \lambda c_N^2 t$. ■

To derive a pre-limit approximation, define the weight function

$$w_\rho(t) \equiv \frac{V_{d,\rho}(t) - V_s(\rho t)}{V_a(t) - V_s(\rho t)}, \quad (3.32)$$

where $V_a(t)$ and $V_s(t)$ are the variance functions associated with the equilibrium arrival process with rate λ and service renewal process with rate μ . Note that we have an additional scaling of ρ in $V_s(\rho t)$. By Remark 2.4, this is equivalent to consider a service renewal process with rate $\rho\mu = \lambda$. Of course, this does not change the heavy-traffic limit. But we add this extra scaling constant because it usually generate slightly more desirable approximation in light traffic. Define the HT-scaled weight function

$$w_\rho^*(t) = w_\rho((1-\rho)^{-2}t). \quad (3.33)$$

Combining Theorem 3.3 and Lemma 3.1, we obtain

Corollary 3.3 (Limiting weight function) *Under the assumptions in Theorem 3.3, we have*

$$w_\rho^*(t) \Rightarrow w^*(\lambda\gamma^2 t/c_x^2)$$

for w^* defined in (3.27).

This justifies the following approximation for the variance function of the stationary departure process from a $GI/M/1$ queue:

$$\begin{aligned} V_{d,\rho}(t) &= w_\rho(t)V_a(t) + (1 - w_\rho(t))V_s(t) \\ &\approx w^*((1 - \rho)^2\lambda\gamma^2 t/c_x^2)V_a(t) + (1 - w^*((1 - \rho)^2\lambda\gamma^2 t/c_x^2))V_s(t). \end{aligned} \quad (3.34)$$

We conclude this section with the tail asymptotic behavior of the variance function. To start, we re-write V_d^* in terms of c^* and H_2^* ,

$$\begin{aligned} V_d^*(t) &= c_a^2\lambda t + (1 - w^*(\lambda\gamma^2 t/c_x^2))(c_s^2 - c_a^2)\lambda t \\ &= c_a^2\lambda t + \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2}H_2^*(\lambda\gamma^2 t/c_x^2) \\ &= c_a^2\lambda t + \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2} - \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2}c^*(\lambda\gamma^2 t/c_x^2), \quad t \geq 0. \end{aligned} \quad (3.35)$$

Combining (3.25) and (3.35), we obtain the asymptotic behavior of the departure variance function.

Corollary 3.4 (Asymptotic behavior of the departure variance function) *Under the assumptions in Theorem 3.3,*

$$\begin{aligned} V_d^*(t) &= c_a^2\lambda t + \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2} - \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2}c^*(\lambda\gamma^2 t/c_x^2) \\ &\sim c_a^2\lambda t + \frac{(c_s^2 - c_a^2)c_x^2}{2\gamma^2} - \frac{8(c_s^2 - c_a^2)c_x^5}{\gamma^5} \frac{1}{\sqrt{2\pi\lambda^3 t^3}} e^{-\frac{\lambda\gamma^2 t}{2c_x^2}} \quad \text{as } t \rightarrow \infty, \end{aligned} \quad (3.36)$$

where $c_s^2 = 1$.

Example 3.1 To evaluate the approximation stemming from Theorem 3.3, consider the departure process from a $E_2/M/1$ model, where E_2 refers to the Erlang random variable with shape parameter 2, i.e. the sum of two i.i.d. exponential random variable. Figure

3.1 (left) reports $V_d^*(t) - \lambda c_a^2 t$ for four sets of parameters such that the limiting constant $(c_s^2 - c_a^2)c_x^2/2\gamma^2 = (1 - c_a^4)/2\gamma^2$ in Corollary 3.4 will be 1.5, 0.375, -1.5 and -1.5 , respectively. Figure 3.1 (right) confirms Theorem 3.3 by comparing simulation estimates of the HT-scaled and centered departure variance function $V_{d,\rho}^*(t) - \lambda c_a^2 t$ for $\rho = 0.8$ and 0.9 from simulation with the theoretical limit $V_d^*(t) - \lambda c_a^2 t$ for the $E_2/M/1$ model with $\lambda = 2$, $\gamma = 0.5$ and $c_a^2 = 0.5$, showing that the theoretical limit in (3.29) serves as a good approximation of the HT-scaled variance function.

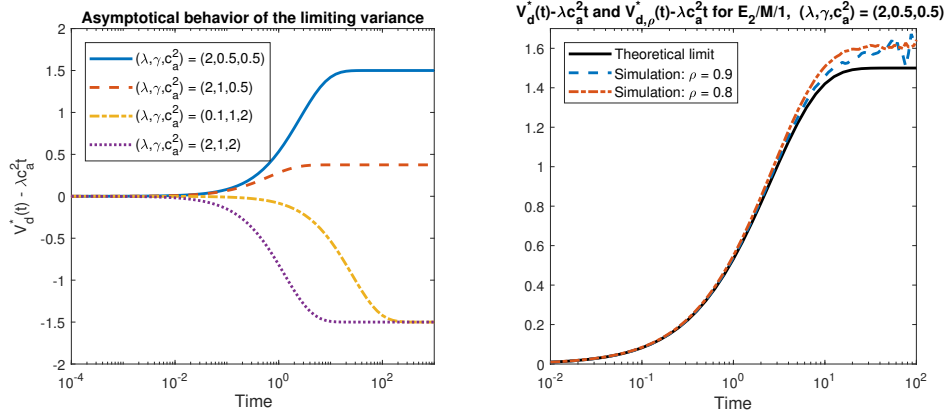


Figure 3.1: Approximation of the departure IDC in the $E_2/M/1$ model.

3.2.2 The Departure Variance in the $M/GI/1$ Queue

We now prove that the HT limit for the stationary departure variance in (3.31) also holds true for the $M/GI/1$ model. Of course, here we restrict our attention to $c_a^2 = 1$ instead of $c_s^2 = 1$ before. Theorem 3.8 will show that the same formula is valid for $GI/GI/1$ with general c_a^2 and c_s^2 .

Recall from (3.15) that the Laplace Transform of the variance function of a general stationary and ergodic point process is

$$\hat{V}(s) = \frac{\lambda}{s^2} + \frac{2\lambda}{s} \hat{m}(s) - \frac{2\lambda^2}{s^3}.$$

In the case of the $M/GI/1$ model, [128] (on p. 78) derived an expression for $\hat{m}_d(s)$.

Theorem 3.4 (Laplace transform of the Palm mean function) *The mean process of the Palm version (assuming a departure epoch at time 0) of the departure process from a $M/GI/1$ queue have the following Laplace transform*

$$\hat{m}_d(s) \equiv \int_0^\infty e^{-st} m_d(t) dt = \frac{\hat{g}(s)}{s(1 - \hat{g}(s))} \left(1 - \frac{s\Pi(\hat{\nu}(s))}{s + \lambda(1 - \hat{\nu}(s))} \right), \quad (3.37)$$

where $\hat{g}(s) = E[e^{-sV}]$ is the Laplace Transform of the service time pdf $g(t)$, $\hat{\nu}(s)$ is the root with the smallest absolute value in z of the equation

$$z = \hat{g}(s + \lambda(1 - z)) \quad (3.38)$$

and

$$\Pi(z) \equiv E[z^Q] = \frac{(1 - \lambda/\mu)(1 - z)\hat{g}(\lambda(1 - z))}{\hat{g}(\lambda(1 - z)) - z} \quad (3.39)$$

is the probability generating function of the distribution of the stationary queue length Q .

Note from (3.3) that the first part in (3.37), i.e.

$$\frac{\hat{g}(s)}{s(1 - \hat{g}(s))},$$

is exactly the Laplace Transform of the mean process of the service renewal process.

Now, we state the HT limit in terms of the HT-scaled variance function defined in (3.21) for the $M/GI/1$ special case. The result parallels that for the $GI/M/1$ case. The proof can be found in Section 3.4.

Theorem 3.5 (HT limit for the $M/GI/1$ departure variance) *Consider an $M/GI/1$ model with $1/E[V] = \mu$ and $1/E[U] = \lambda_\rho \equiv \mu(1 - (1 - \rho)\gamma_\rho)$, where γ_ρ are positive constants such that $\lim_{\rho \uparrow 1} \gamma_\rho = \gamma > 0$. Assume that $E[V^3] < \infty$ so that a two-term Taylor series expansion of the LT $\hat{g}(s)$ about the origin is valid with asymptotically negligible remainder. Then the HT-scaled variance function $V_{d,\rho}^*(t)$ defined in (3.21) converges as $\rho \uparrow 1$, i.e.,*

$$V_{d,\rho}^*(t) \rightarrow V_d^*(t) \quad \text{as } \rho \uparrow 1 \quad \text{for all } t \geq 0, \quad (3.40)$$

where the limit $V_d^*(t)$ is a continuous nonnegative function with LT

$$\hat{V}_d^*(s) = \frac{\mu c_s^2}{s^2} + \frac{\gamma \mu^2 (1 - c_s^2)}{s^3} \hat{\nu}^*(s), \quad (3.41)$$

with $\hat{\nu}^*(s)$ being the unique root with positive real part of the equation

$$\frac{1 + c_s^2}{2}(\hat{\nu}^*(s))^2 + \gamma\hat{\nu}^*(s) - \frac{s}{\mu} = 0. \quad (3.42)$$

In addition,

$$V_d^*(t) \equiv w^*(\mu\gamma^2 t/c_x^2)c_a^2\mu t + (1 - w^*(\mu\gamma^2 t/c_x^2))c_s^2\mu t, \quad (3.43)$$

where $w^*(t)$ in (3.27), $c_x^2 \equiv c_a^2 + c_s^2$, $c_a^2 = 1$ and $c_s^2 \equiv \text{Var}(V)/E[V]^2$.

With the same technique as in Corollary 3.4, one can prove the following corollary, which yields exactly the same asymptotic behavior.

Corollary 3.5 (Asymptotic behavior of the departure variance curve) *Under the assumptions in Theorem 3.5, we have the limit in (3.36), except now $c_a^2 = 1$ and c_s^2 is general.*

Proposition 6 of [78] developed a two-term asymptotic expansion for the variance function $V_{d,\rho}(t) \equiv \text{Var}(D_\rho(t))$ as $t \rightarrow \infty$ for the $M/GI/1$ queue and fixed $\rho < 1$. We discuss the connections between our result and this earlier result in Remark 3.3.

3.2.3 Heavy-Traffic Limit for the Stationary Departure Process

In this section, we establish an HT limit for the stationary departure process and its variance function in a $GI/GI/1$ queue. To do so, we apply the recent HT results for the stationary queue length (number in system) in [66] and [31] together with the HT limits for the general single-server queue in Section 9.3 of [143] and the general reflection mapping with non-zero initial conditions in Section 13.5 of [143]. As in [143], a major component of the proof is the continuous mapping theorem.

The corresponding limit starting out empty is contained in Theorem 2 of [84]. There has since been a substantial literature on that case; see [68; 88; 143]. As can be seen from Section 9.3 and Section 13.5 of [143], for the queue length, the key map is the reflection map Ψ applied to a potential net-input function x ,

$$\Psi(x)(t) \equiv x(t) - \zeta(x)(t), \quad t \geq 0, \quad (3.44)$$

where

$$\zeta(x) \equiv \inf \{x(s) : 0 \leq s \leq t\} \wedge 0, \quad t \geq 0 \quad (3.45)$$

with $a \wedge b \equiv \min \{a, b\}$, so that $\zeta(x) \leq 0$ and $\Psi(x)(t) \geq x(t)$ for all $t \geq 0$. The key point is that we now allow $x(0) \neq 0$.

3.2.3.1 A General Heavy-Traffic Limit for the $G/G/1$ Model

For the general $G/G/1$ single-server queue with unlimited waiting space and service provided in the order of arrival, we consider a family of processes indexed by the traffic intensity ρ , where $\rho \uparrow 1$. Let $Q_\rho(t)$ be the number of customers in the system at time t ; let $A_\rho(t)$ count the number of arrivals in the interval $[0, t]$, which we assume to have rate λ ; let $S_\rho(t)$ be a corresponding counting process for the successive service times, applied after time 0, to be applied to the initial $Q_\rho(0)$ customers and to all new arrivals; let $B_\rho(t)$ be the cumulative time that the server is busy in the interval $[0, t]$. Then the queue-length process can be expressed as

$$Q_\rho(t) \equiv Q_\rho(0) + A_\rho(t) - S_\rho(B_\rho(t)), \quad t \geq 0, \quad (3.46)$$

where the three components are typically dependent. (For simplicity, we assume that $A_\rho(0) = S_\rho(0) = B_\rho(0) = 0$ w.p.1.)

We have in mind that the system is starting in steady-state. Thus the triple

$$(Q_\rho(0), A_\rho(\cdot), S_\rho(\cdot))$$

is in general quite complicated for each ρ . Even in the relatively tractable $GI/GI/1$ cases, which we shall primarily treat, the residual interarrival time and service time at time 0 will be complicated, depending on ρ and $Q_\rho(0)$. We will need to make assumptions ensuring that these are uniformly asymptotically negligible in the HT limit.

By flow conservation, the departure (counting) process can be represented as

$$D_\rho(t) \equiv A_\rho(t) - Q_\rho(t) + Q_\rho(0), \quad t \geq 0. \quad (3.47)$$

Directly, or by combining (3.46) and (3.47),

$$D_\rho(t) \equiv S_\rho(B_\rho(t)), \quad t \geq 0. \quad (3.48)$$

Let

$$X_\rho(t) \equiv Q_\rho(0) + A_\rho(t) - S_\rho(t), \quad t \geq 0, \quad (3.49)$$

be a net-input process, acting as if the server is busy all the time, and thus allowing $X_\rho(t)$ to assume negative values. The cumulative busy time $B_\rho(t)$ is then relate to $X_\rho(t)$ by

$$B_\rho(t) = t + \zeta(X_\rho)(t), \quad t \geq 0. \quad (3.50)$$

As a consequence of the assumptions above, $X_\rho(0) = Q_\rho(0)$. Roughly,

$$Q_\rho(t) \approx \Psi(X_\rho)(t), \quad t \geq 0, \quad (3.51)$$

for Ψ in (3.44), but the exact relation breaks down because the service process shuts down when the system becomes idle, so that a new service time does not start until after the next arrival. While (3.51) does not hold exactly for each ρ , it holds in the HT limit, as shown in Theorem 9.3.4 of [143]. It would hold exactly if we used the modified system in which we let the continuous-time service process run continuously, so that equation (3.51) holds as an equality, as done by [24] and then again in Section 2 of [83]. Because the modified system has been shown to be asymptotically equivalent to the original system for these HT limits in [24] and [83], that is an alternative approach.

We now introduce HT-scaled versions of these processes, for that purpose, let

$$\begin{aligned} X_\rho^*(t) &\equiv (1 - \rho)X_\rho((1 - \rho)^{-2}t), \\ Q_\rho^*(t) &\equiv (1 - \rho)Q_\rho((1 - \rho)^{-2}t), \\ A_\rho^*(t) &\equiv (1 - \rho)[A_\rho((1 - \rho)^{-2}t) - (1 - \rho)^{-2}\lambda t], \\ S_\rho^*(t) &\equiv (1 - \rho)[S_\rho((1 - \rho)^{-2}t) - (1 - \rho)^{-2}\lambda t/\rho], \\ B_\rho^*(t) &\equiv (1 - \rho)[B_\rho((1 - \rho)^{-2}t) - (1 - \rho)^{-2}t], \\ D_\rho^*(t) &\equiv (1 - \rho)[D_\rho((1 - \rho)^{-2}t) - (1 - \rho)^{-2}\lambda t]. \end{aligned} \quad (3.52)$$

Let $\mathcal{D}$ be the function space of all right-continuous real-valued functions on $[0, \infty)$ with left limits, with the usual J_1 topology, which reduces to uniform convergence over all bounded intervals for continuous limit functions. Let $\mathcal{D}^k$ be the k -fold product space, using the product topology on all product spaces. Let $\Rightarrow$ denote convergence in distribution. Let e be the identity function in $\mathcal{D}$, i.e., $e(t) \equiv t, t \geq 0$.

Theorem 3.6 *If*

$$(Q_\rho^*(0), A_\rho^*, S_\rho^*) \Rightarrow (Q^*(0), A^*, S^*) \quad \text{in } \mathbb{R} \times \mathcal{D}^2 \quad \text{as } \rho \uparrow 1, \quad (3.53)$$

where A^* and S^* have continuous sample paths with $A^*(0) = S^*(0) = 0$ w.p.1., then

$$(A_\rho^*, S_\rho^*, B_\rho^*, X_\rho^*, Q_\rho^*, D_\rho^*) \Rightarrow (A^*, S^*, B^*, X^*, Q^*, D^*), \quad (3.54)$$

where the convergence is in $\mathcal{D}^6$ as $\rho \uparrow 1$ and

$$\begin{aligned} X^* &\equiv Q^*(0) + A^* - S^* - \lambda e, \\ B^* &\equiv \zeta(X^*) < 0, \\ Q^* &\equiv \Psi(X^*) = X^* - \zeta(X^*) \quad \text{and} \\ D^* &\equiv Q^*(0) + A^* - Q^* \\ &= Q^*(0) + A^* - \Psi(X^*) = S^* + \lambda e + \zeta(X^*) \end{aligned} \quad (3.55)$$

for Ψ and ζ in (3.44) and (3.45).

Proof First, note that

$$X_\rho^*(t) = Q_\rho^*(0) + A_\rho^*(t) - S_\rho^*(t) - \lambda t/\rho, \quad t \geq 0, \quad (3.56)$$

because A_ρ^* and S_ρ^* have different translation terms in (3.52), ensuring that the potential rate out is λ/ρ , which exceeds the rate in of λ , consistent with a stable model for each ρ , $0 < \rho < 1$. Hence, under the assumption, $X_\rho^* \Rightarrow X^* = Q^*(0) + A^* - S^* - \lambda e$ in $\mathcal{D}$. The limit $B_\rho^* \Rightarrow B^* = \zeta(X^*)$ is obtained by exploiting the relationship in (3.50). The limits for Q_ρ^* and D_ρ^* then follow from the continuous mapping theorem after carefully accounting for the busy and idle time of the server; see the proof of Theorem 9.3.4 and preceding material in [143]. ■

3.2.3.2 A Heavy-Traffic Limit for the Stationary Departure Process

Theorem 3.6 is not easy to apply to establish HT limits for stationary processes because condition (3.53) is not easy to check and the limit in (3.54) and (3.55) is not easy to evaluate.

In order to establish a tractable HT limit for the stationary departure process, we apply the recent HT limits for the stationary queue length in [66] and [31]. Their HT limits are for generalized open Jackson networks of queues, which for the single queue we consider reduce to the $GI/GI/1$ model. Following [31], we assume that the interarrival times and

service times come from independent sequences of i.i.d. random variables with uniformly bounded third moments ($2 + \epsilon$ would do).

Theorem 3.7 *For the GI/GI/1 model indexed by ρ , assume that (i) the interarrival-time cdf has a pdf as in Section 3.1.1 and (ii) the interarrival times and service times have means λ and λ/ρ , scv's c_a^2 and c_s^2 , without both being 0, and uniformly bounded third moments. Then:*

(a) *For each ρ , $0 < \rho < 1$, the process Q_ρ^* can be regarded as a stationary process, while the process D_ρ^* can be regarded as a stationary point process (with stationary increments).*

(b) *Condition (3.53) in Theorem 3.6 holds with*

$$A^* \equiv c_a B_a \circ \lambda e \quad \text{and} \quad S^* \equiv c_s B_s \circ \lambda e, \quad (3.57)$$

where B_a and B_s are independent standard (mean 0, variance 1) Brownian motions (BM's) that are independent of $Q^*(0)$, which is distributed as $R_e(0)$ with R_e being a stationary RBM with drift $-\lambda$ and variance $\lambda c_x^2 \equiv \lambda c_a^2 + \lambda c_s^2$, and so an exponential marginal distribution, i.e.,

$$P(Q^*(0) > x) = e^{-2x/c_x^2}, \quad x \geq 0. \quad (3.58)$$

(c) *The limits in Theorem 3.6 hold, where*

$$X^* \equiv Q^*(0) + c_a B_a \circ \lambda e - c_s B_s \circ \lambda e - \lambda e \quad (3.59)$$

with $Q^*(0)$, B_a and B_s being mutually independent.

(d) *We have*

$$\begin{aligned} D^* &= c_a B_a \circ \lambda e + Q^*(0) - Q^* \\ &= c_a B_a \circ \lambda e + Q^*(0) - \Psi(X^*) \end{aligned} \quad (3.60)$$

for Ψ in (3.44), or

$$D^* \equiv S^* + \lambda e + \zeta(X^*) \quad (3.61)$$

for ζ in (3.45).

Proof First, recall that the HT limit as $\rho \rightarrow 1$ starting empty is the RBM which converges as $t \rightarrow \infty$ to the exponential distribution in (3.58). We will be applying [66] and [31] to show that the two iterated limits involving $\rho \rightarrow 1$ and $t \rightarrow \infty$ are equal. Toward that end, we observe that, by Section X.3-X.4 of [12], the queue-length process has a proper steady-state distribution for each ρ . As on p. 63 of [66], we add the residual interarrival times and service times to the state description for $Q_\rho(t)$ to make it a Markov process that has a unique steady-state distribution for each ρ . These residual interarrival and service times are asymptotically negligible in the HT limit. The associated departure process $D_\rho(t)$ then necessarily is a stationary point process for each ρ . We then can apply Theorem 8 of [66] to have a limit for the scaled stationary distributions, so that condition (3.53) holds with (3.57). Since strong moment-generating-function-condition is imposed in (1) and (2) on p. 62 of [66], we apply the extension in Theorems 3.1 and 3.2 of [31] to cover our moment condition. Hence, we can apply Theorem 3.6 with these special initial distributions to get the associated process limits in the space $\mathcal{D}$. ■

We now establish an HT limit for the variance of the stationary departure process. The form of that limit is already given in Theorem 3.3. The proof can be found in Section 3.4.

Theorem 3.8 (Heavy-traffic limit for the $GI/GI/1$ departure variance function)

Under the conditions of Theorem 3.7 plus the usual uniform integrability conditions, for which it suffices for the interarrival times and service times to have uniformly bounded fourth moments,

$$\begin{aligned} V_{d,\rho}^*(t) &\equiv \text{Var}(D_\rho^*(t)) = E[D_\rho^*(t)^2] \\ &\rightarrow E[D^*(t)^2] = \text{Var}(D^*(t)) \equiv V_d^*(t) \quad \text{as } \rho \uparrow 1, \end{aligned} \quad (3.62)$$

where

$$V_d^*(t) = w^*(\lambda t/c_x^2) c_a^2 \lambda t + (1 - w^*(\lambda t/c_x^2)) c_s^2 \lambda t \quad (3.63)$$

with $c_x^2 = c_a^2 + c_s^2$, $c^*(t)$ is the correlation function in (3.22) and $w^*(t)$ is the weight function in (3.27); i.e., $V_d^*(t)$ is given in (3.31) with $\gamma = 1$, but allowing general c_s^2 . Moreover, we

have the covariance formulas

$$\begin{aligned}\text{Cov}(c_a B_a(\lambda t), Q^*(t)) &= (1 - w^*(\lambda t/c_x^2)) c_a^2 \lambda t, \\ \text{Cov}(c_s B_s(\lambda t), Q^*(t)) &= - (1 - w^*(\lambda t/c_x^2)) c_s^2 \lambda t.\end{aligned}\tag{3.64}$$

As a by-product, the covariance formulas in (3.64) can be generalized to describe the covariance of between a stationary RBM and a BM, where the underlying BMs are correlated.

Corollary 3.6 *Suppose $B = (B_1, B_2)$ is a 2-d Brownian motion with zero drift and covariance matrix $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{pmatrix}$. Let $Q = \Psi(B_1 + Q^*(0) - \lambda e)$ be the stationary RBM associated with the drifted BM $B_1 - \lambda e$ and $Q^*(0)$ has the stationary distribution of Q^* , which is independent of B_1 . Then*

$$\text{cov}(B_2, Q) = (1 - w^*(\lambda^2 t / \sigma_1^2)) \sigma_{1,2} t = \frac{\sigma_{1,2} \sigma_1^2}{2\lambda^2} (1 - c^*(\lambda^2 t / \sigma_1^2)).$$

Remark 3.1 (The quasireversible case) The limit process

$$(A^*, S^*, X^*, Q^*, D^*),$$

where

$$(A^*, S^*, X^*) = (c_a B_a \circ \lambda e, c_s B_s \circ \lambda e, Q^*(0) + c_a B_a \circ \lambda e - c_s B_s \circ \lambda e - \lambda e),$$

as in Theorem 3.7, can be called *the Brownian queue*; see [72; 76; 77; 110]. The Brownian queue is known to be quasireversible if and only if $c_a^2 = c_s^2$. In that case, the stationary departure process is a BM and the departures in the past are independent of the steady-state content. Consistent with that theory, $V_d^*(t) = c_a^2 \lambda t, t \geq 0$ in (5.9) if and only if $c_a^2 = c_s^2$. ■

Remark 3.2 We considered only the case where $\mu_\rho = \lambda/\rho$ in this section. Now, we list the results for a slightly more general case as in Section 3.2.1, where we have $\mu_\rho = \lambda(1+(1-\rho)\gamma_\rho)$ and $\lim_{\rho \uparrow 1} \gamma_\rho = \gamma$. One can easily check that Theorem 3.6 holds with

$$X^* = Q^*(0) + A^* - S^* - \lambda \gamma e;$$

Theorem 3.7 holds with

$$P(Q^*(0) > x) = e^{-2\gamma x/c_x^2} \quad \text{and} \quad X^* = Q^*(0) + c_a B_a \circ \lambda e - c_s B_s \circ \lambda e - \lambda \gamma e;$$

and Theorem 3.8 holds with

$$\begin{aligned} V_d^*(t) &= w^*(\lambda \gamma^2 t / c_x^2) c_a^2 \lambda t + (1 - w^*(\lambda \gamma^2 t / c_x^2)) c_s^2 \lambda t, \\ \text{Cov}(c_a B_a(\lambda t), Q^*(t)) &= (1 - w^*(\lambda \gamma^2 t / c_x^2)) c_a^2 \lambda t \quad \text{and} \\ \text{Cov}(c_s B_s(\lambda t), Q^*(t)) &= -(1 - w^*(\lambda \gamma^2 t / c_x^2)) c_s^2 \lambda t. \quad \blacksquare \end{aligned}$$

We conclude this section with the tail asymptotics of the departure variance function. Just as in Corollaries 3.4 and 3.5, we have

Corollary 3.7 (Tail Asymptotic of the departure variance function) *Under the assumptions in Theorem 3.8 and Remark 3.2,*

$$V_d^*(t) \sim c_a^2 \lambda t + \frac{(c_s^2 - c_a^2) c_x^2}{2\gamma^2} - \frac{8(c_s^2 - c_a^2) c_x^5}{\gamma^5} \frac{1}{\sqrt{2\pi} \lambda^3 t^3} e^{-\frac{\lambda \gamma^2 t}{2c_x^2}} \quad \text{as } t \rightarrow \infty. \quad (3.65)$$

Remark 3.3 Hautphenne et al.[78] developed explicit expressions for the y-intercept $\bar{b}_\theta$ of the linear asymptote for the variance of the stationary departure from M/GI/1

$$V(t) = \bar{v}t + \bar{b}_\theta + o(1) \quad \text{as } t \rightarrow \infty;$$

see Proposition 6 there. Their result (i) holds for $M/GI/1$ case; (ii) depends on the third moment of the service distribution; (iii) holds for general traffic intensity. Even though there is no direct heavy-traffic scaling in their result, the scaling parameter emerges in their expression, see the definition of $\bar{b}_\theta$ there. In specific, the scaling constant $\rho/(1-\rho)^2$ coincides (up to a multiple of ρ) with the one we use in (5.8).

On the other hand, our result here (i) coincides with their y-intercept (after scaling) in the HT limit in the $M/GI/1$ case, i.e., let $\rho = 1$, $\gamma = 1$ and $c_a = 1$; (ii) holds for general $GI/GI/1$ cases but only under HT limit; (iii) has explicit characterization of the remainder term, again only under HT limit. $\blacksquare$

3.2.4 Extensions

The approximation for the departure IDC $I_d(t)$ in (3.1) and (3.67) should be good for much more general models than $GI/GI/1$, with the independence conditions relaxed and more than 1 server.

In Section 4.4, we establish the same limit for the departure process at the bottleneck queue (the queue with the highest traffic intensity level in the network) in a generalized Jackson networks with a single bottleneck.

We also conjecture that the HT limit of the variance function in Theorem 3.8 extends to a larger class of models as well. Indeed, we conjecture that the limits established for $GI/GI/1$ extend in that way. First, Theorem 3.6 extends quite directly by exploiting [83; 84]. For the extension of Theorem 3.7, there is a large class of models for which the HT-scaled arrival and service processes have the limits

$$A^* \equiv c_a B_a \quad \text{and} \quad S^* \equiv c_s B_s, \quad (3.66)$$

where B_a and B_s are independent standard (mean 0, variance 1) Brownian motions (BM's) that are independent of the initial queue length. What is needed is the extension of [66] and [31] to more general models. We conjecture that can be done for $GI/GI/s$ and other models with regenerative structure in the arrival and service processes. For $GI/GI/s$ the queue-length process again becomes a Markov process if we append the s elapsed service times as well as the elapsed interarrival time, but it remains to do the hard technical analysis leading to an appropriate Lyapunov function. It is also of interest to establish related results for departure processes in models with non-renewal arrival processes, as in [61] and references therein.

The relevant approximation for the stationary departure process from a many-server $GI/GI/s$ queue evidently is quite different, being more like the service process than the arrival process. We conjecture that the relevant many-server heavy-traffic limit for the stationary departure process is a Gaussian process with the covariance function of the stationary renewal processes associated with the service times, as in the CLT for renewal processes in Theorems 7.2.1 and 7.2.4 of [143]. Partial support comes from [10], Appendix F of [11] and [65].

3.3 Robust Queueing Network Analyzer for Queues In Series

We conclude this chapter by explaining the important role that Theorem 3.8 plays in our *Robust Queueing Network Analyzer* (RQNA) based on the index of dispersion for counts. For queues in series models, the only relevant network operations is the departure operation, which will be the main focus of this section.

3.3.1 Approximation of the Departure IDC

We now propose an approximation for the IDC of the departure process and apply it to the queues-in-series models.

The main challenge in developing a full RQNA-IDC involving a decomposition approximation is calculating or approximating the required IDC for the arrival process at each queue. Inspired by (5.9), we propose to approximate the IDC of a departure process from a $G/GI/1$ queue by the weighted average of the IDC's of the arrival and service processes in (3.1). We thus require $(\lambda, \rho, I_a, I_s)$ as model data, where I_a is the arrival IDC and I_s is the service IDC, as defined in 1.1.

We now show how Theorem 3.8 suggests a weight function defined by

$$w_\rho(t) \equiv w^*((1 - \rho)^2 \lambda t / (\rho^2 c_x^2)), \quad (3.67)$$

where $c_x^2 \equiv c_a^2 + c_s^2$ and w^* is given in (3.27). The c_a^2 and c_s^2 here are the asymptotic variability parameters, i.e., the normalization constants in the FCLT for the arrival and service processes.

Let $I_{d,\rho}$ denote the departure IDC and define the weight function

$$w_\rho(t) \equiv \frac{I_{d,\rho}(t) - I_s(\rho t)}{I_a(t) - I_s(\rho t)} = \frac{V_{d,\rho}(t) - V_s(\rho t)}{V_a(t) - V_s(\rho t)}, \quad (3.68)$$

where $V_a(t)$ and $V_s(t)$ are the variance functions associated with the equilibrium arrival process with rate λ and service renewal process with rate μ . Note that this is exactly the same weight function we defined in (3.32), thus we have the same HT-scaled weight function w_ρ^* as in (3.33). Again, note that we have an additional scaling of ρ in $V_s(\rho t)$ to obtain slightly more desirable approximation in light traffic. We then apply Theorem 3.8 to obtain a generalization of Corollary 3.3.

Corollary 3.8 (Limiting weight function) *Under the assumptions in Theorem 3.8, we have*

$$w_\rho^*(t) \Rightarrow w^*(\lambda t/c_x^2)$$

for w^* defined in (3.27).

For the pre-limit approximation, we re-arrange terms in (3.68) and obtain

$$w_\rho(t) \approx w^*((1-\rho)^2 \lambda t/c_x^2). \quad (3.69)$$

One remaining issue is that the approximation (3.69) does not automatically yield a correct light traffic limit, in which case we must have $I_{d,0}(t) = I_a(t)$ since the service time is negligible. As a remedy, we propose to add a constant ρ^{-2} correction in the weight function, so that we have (3.67) as the final weight function. This specific choice of correction term is motivated from Remark 3.2, where we replace γ by γ_ρ in the pre-limit weight function and recall that the usual case of $\mu_\rho = \lambda/\rho$ corresponds to $\gamma_\rho = 1/\rho$.

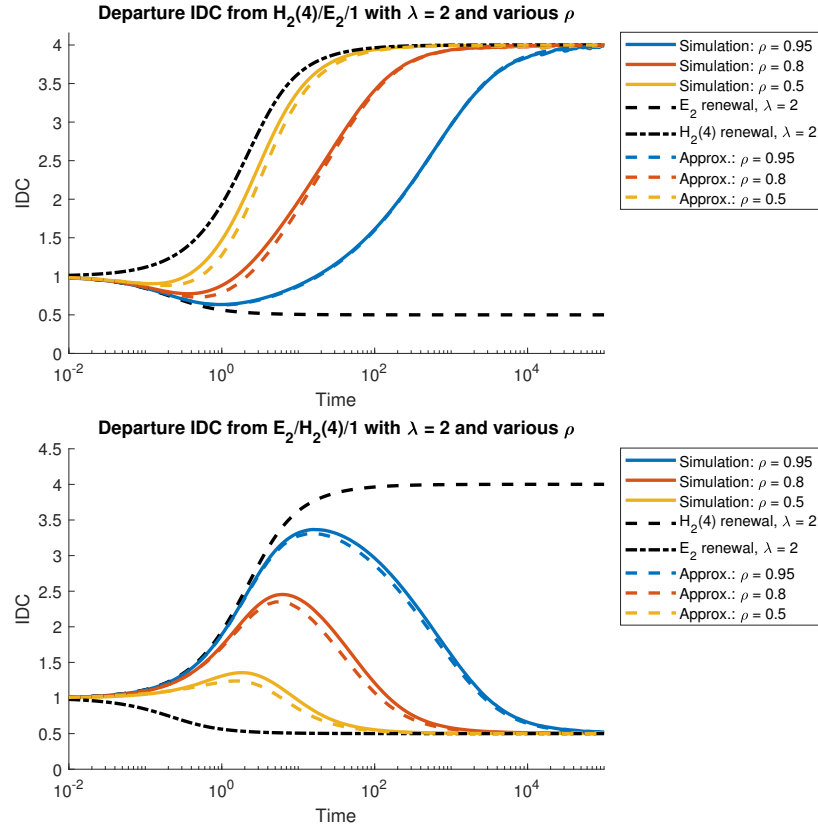
Remark 3.4 (Parallel to QNA) The convex combination in the approximation (3.1) is reminiscent of the convex combination for variability parameters in (38) of [134], i.e.,

$$c_{d,i} \approx (1 - \rho_i^2) c_{a,i}^2 + \rho_i^2 c_{s,i}^2, \quad (3.70)$$

which corresponds to a stationary-interval approximation, as discussed in [133; 134; 135]. Similar behavior can be seen in approximation (3.1). In particular, the canonical weight function w^* in (3.27) is a monotonically increasing function with $w^*(0) = 0$ and $w^*(\infty) = 1$. By the definition of $w_\rho(t)$ in (3.67), we see that for each t , the approximation places less weight on $I_{a,i}(t)$ and more weight on $I_{s,i}(t)$ as ρ_i increases. This makes sense intuitively, because the queue should be busy most of the time as ρ_i increases toward 1. Thus departure times tend to be minor variations of service times. In contrast, if ρ_i is very small, then the queue acts only as a minor perturbation of the arrival process.

However, (3.67) reveal a more subtle interaction between ρ_i and the variability of the departure process over different time scales. ■

We now demonstrate the effectiveness of our approximation with two simulation examples for $GI/GI/1$ queues, where neither the interarrival time nor service time has an

Figure 3.2: Approximations of the departure IDC's in two $GI/GI/1$ models.

exponential distribution. Let $H_2(c^2)$ be the H_2 (hyperexponential) distribution with scv c^2 and balanced means, as in (3.7) on p. 137 of [133]. Consider the $H_2(4)/E_2/1$ model with $\lambda = 2$, Figure 3.2 (top) reports the simulated departure IDCs for three different traffic intensities $\rho = 0.95, 0.8, 0.5$, as well as the approximation (3.1) with (3.67). The simulation estimation of the departure IDC is obtained from a single run of length 10^9 time units, with the first 10^6 time units are discarded in order for the system to approach steady-state. The reference IDCs I_a and I_s is calculated by numerically inverting the LT in (3.8). Figure 3.2 (bottom) is the corresponding plot for the $E_2/H_2(4)/1$ model with $\lambda = 2$.

3.3.2 The RQNA algorithm for Queues in Series

We conclude this section by briefly describe the RQNA algorithm for queues in series models, where the service times are assumed to be i.i.d. random variables, independent of the

external arrival process. To obtain the performance approximation at each queue, we start with the first queue, at which the external arrival process arrives. This first queue is simply regarded as a $G/GI/1$ queue as in Section 2.2.5, so we require the arrival IDC as model input. The external arrival process need not be renewal, as long as the IDC can be obtained from one of the approaches discussed in Section 2.3. With the external arrival IDC and the service scv, we simply apply the RQ algorithm in (2.35) to obtain the approximation of the steady-state mean workload. For other steady-state performance measures, see Section 2.2.8. Now, we proceed to the second queue. Note that the arrival process at the second queue is exactly the departure process from the first queue, whose IDC is approximated by (3.1) and (3.67). In revoking (3.1), we require the service IDC, which is again obtained as in Section 2.3. The next step is simply apply the RQ algorithm as for the first queue. The same procedure can be carried out in the same way for any more subsequent queues. We remark that the departure process from the first queue is not a renewal process unless the external arrival process is Poisson and the service times at queue 1 is exponential. Hence, the second queue (and any subsequent queues) can only be regarded as a $G/GI/1$ station, where the arrival process is a general point process. In this setting, the approximation cannot be justified by Theorem 3.8. However, we will show in Section 5.1 that the same approximation is supported by a similar HT limit theorem in generalized Jackson network, see Theorem 5.1. In fact, we conjecture that the same approximation holds for a wide range of $G/G/1$ models, for example the $MMPP/MMPP/1$ model with Markov modulated Poisson arrival and service processes, see Section 6.3.1 for an illustration.

For simulation studies of the performance of this RQNA algorithm, see Section 6.3.

3.4 Proofs

We now review a useful lemma on the properties of $\hat{\xi}(s)$ and δ , defined in (3.19) and (3.20); see p. 113 of [128] or Appendix 6 of [38]. (The notation here is slightly different.)

Lemma 3.2 (Takacs' root lemma) *If $Re(s) \geq 0$, then the root $\hat{\xi}(s)$ of the equation*

$$z = \hat{f}(s + \mu(1 - z))$$

that has the smallest absolute value is

$$\hat{\xi}(s) = \sum_{j=1}^{\infty} \frac{(-\mu)^{j-1}}{j!} \frac{d^{j-1}}{ds^{j-1}} \left(\hat{f}(\mu + s) \right)^j. \quad (3.71)$$

This root $\hat{\xi}(s)$ is a continuous function of s for $\text{Re}(s) \geq 0$. Furthermore, $z = \hat{\xi}(s)$ is the only root in the unit circle $|z| \leq 1$ if at least one of two conditions is satisfied (i) $\text{Re}(s) > 0$, or (ii) $\text{Re}(s) \geq 0$ and $\lambda/\mu < 1$. Specifically, $\delta = \hat{\xi}(0)$ is the smallest positive real root of the equation

$$\delta = \hat{f}(\mu(1 - \delta)).$$

If $\lambda/\mu < 1$, then $\delta < 1$ and if $\lambda/\mu \geq 1$ then $\delta = 1$.

Proof of Theorem 3.3. We let $\rho \uparrow 1$ by decreasing the service rate, so that $1/E[U] = \lambda$ is fixed. To allow general drift in the Brownian HT limit, we let $1/E[V] = \mu_\rho \equiv \lambda + (1 - \rho)\lambda\gamma_\rho$ in system ρ , for positive constants $\gamma_\rho \rightarrow \gamma$. Under this setting, we have $(\lambda - \mu_\rho)/(1 - \rho) \rightarrow -\lambda\gamma$ as $\rho \uparrow 1$. By (3.18) and (3.21), we have

$$\begin{aligned} \hat{V}_{d,\rho}^*(s) &= \mathcal{L} \left((1 - \rho)^2 V_{d,\rho}((1 - \rho)^{-2}t) \right) = (1 - \rho)^4 \hat{V}_{d,\rho}((1 - \rho)^2 s) \\ &= \frac{\lambda}{s^2} + \frac{\lambda}{s^2} \hat{W}(s) \end{aligned}$$

where

$$\hat{W}(s) \equiv \frac{2}{(1 - \rho)^2 s} \left(\mu_\rho \left(\delta - \frac{\lambda}{\mu_\rho} \right) + \frac{\mu_\rho \delta \frac{1 - \hat{f}((1 - \rho)^2 s)}{(1 - \rho)^2 s} - \hat{f}((1 - \rho)^2 s)}{\frac{(1 - \rho)^2 s + \mu_\rho (1 - \hat{\xi}((1 - \rho)^2 s))}{\mu_\rho (1 - \hat{\xi}((1 - \rho)^2 s))}} \cdot \frac{(1 - \rho)^2 s - \mu_\rho (1 - \delta)}{\mu_\rho (1 - \delta)} \cdot \frac{1 - \hat{f}((1 - \rho)^2 s)}{(1 - \rho)^2 s} \right).$$

Then, we write

$$\begin{aligned}
\hat{W}(s) &= \frac{2\mu_\rho}{(1-\rho)^2s} \frac{\left(\delta - \frac{\lambda}{\mu_\rho}\right) \hat{H}_\rho(s) \frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s} + \delta \frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s} - \frac{1}{\mu_\rho} \hat{f}((1-\rho)^2s)}{\hat{H}_\rho(s) \frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s}} \\
&= \frac{2\mu_\rho}{(1-\rho)^2s} \frac{\delta \left(\hat{H}_\rho(s) + 1\right) \frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s} - \frac{\lambda}{\mu_\rho} \hat{H}_\rho(s) \frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s} - \frac{1}{\mu_\rho} \hat{f}((1-\rho)^2s)}{\hat{H}_\rho(s) \frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s}} \\
&= \frac{1}{\hat{H}_\rho(s) \frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s}} \frac{2\mu_\rho}{(1-\rho)^2s} \left(\left(\delta - \frac{\lambda}{\mu_\rho}\right) \left(\hat{H}_\rho(s) + 1\right) \frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s} \right. \\
&\quad \left. + \frac{\lambda}{\mu_\rho} \left(\frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s} - \frac{1}{\lambda} \right) + \frac{1}{\mu_\rho} \left(1 - \hat{f}((1-\rho)^2s) \right) \right) \\
&= \frac{2\mu_\rho}{\hat{H}_\rho(s) \frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s}} \left(\frac{\delta - \frac{\lambda}{\mu_\rho}}{1-\rho} \frac{\hat{H}_\rho(s) + 1}{(1-\rho)s} \frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s} \right. \\
&\quad \left. + \frac{\lambda \frac{1-\hat{f}((1-\rho)^2s)}{(1-\rho)^2s} - 1}{\mu_\rho(1-\rho)^2s} + \frac{1-\hat{f}((1-\rho)^2s)}{\mu_\rho(1-\rho)^2s} \right)
\end{aligned}$$

where

$$\hat{H}_\rho(s) \equiv \left(\frac{1}{\mu_\rho} \frac{(1-\rho)^2s}{1-\hat{\xi}((1-\rho)^2s)} + 1 \right) \left(\frac{1}{\mu_\rho} \frac{(1-\rho)^2s}{1-\delta} - 1 \right).$$

By Lemma 3.2, we know that δ is positive and real, and $\delta < 1$ if $\rho < 1$ while $\delta = 1$ if $\rho = 1$. Hence, we may restrict the function $\hat{f}$ to the real axis. Then, expanding $\hat{f}$ in a Taylor series about 0, yields

$$\begin{aligned}
\delta = \hat{f}(\mu_\rho(1-\delta)) &\Rightarrow \delta = \hat{f}(0) + \hat{f}'(0)\mu_\rho(1-\delta) + \left(\frac{1}{2} \hat{f}''(0)\mu_\rho^2 + o(1) \right) (1-\delta)^2 \\
&\Rightarrow 0 = 1 - \delta - \frac{\mu_\rho}{\lambda} (1-\delta) + \left(\frac{1}{2} \hat{f}''(0)\mu_\rho^2 + o(1) \right) (1-\delta)^2 \\
&\Rightarrow 0 = \frac{1 - \frac{\mu_\rho}{\lambda}}{1-\rho} + \left(\frac{c_a^2 + 1}{2} \frac{\mu_\rho^2}{\lambda^2} + o(1) \right) \frac{1-\delta}{1-\rho} \\
&\Rightarrow \frac{1-\delta}{1-\rho} = \gamma_\rho \left(\frac{c_a^2 + 1}{2\rho} + o(1) \right)^{-1}
\end{aligned} \tag{3.72}$$

This implies that the following limit exist

$$\delta^* \equiv \lim_{\rho \uparrow 1} \frac{1-\delta}{1-\rho} = \frac{2\gamma}{c_a^2 + 1}. \tag{3.73}$$

Now, let $\hat{\xi}_{\rho,s} \equiv \hat{\xi}((1-\rho)^2s) = \hat{f}((1-\rho)^2s + \mu_\rho(1-\hat{\xi}_{\rho,s}))$, then similarly we have

$$0 = \gamma_\rho \frac{1 - \hat{\xi}_{\rho,s}}{1 - \rho} + \frac{s}{\lambda} - \left(\frac{c_a^2 + 1}{2\lambda^2} + o(1) \right) \left((1-\rho)s + \mu_\rho \frac{1 - \hat{\xi}_{\rho,s}}{1 - \rho} \right)^2. \quad (3.74)$$

Then (3.74) implies that the following limit exists

$$\hat{\xi}^\star(s) \equiv \lim_{\rho \uparrow 1} \frac{1 - \hat{\xi}_{\rho,s}}{1 - \rho}, \quad (3.75)$$

and

$$\frac{c_a^2 + 1}{2} \left(\hat{\xi}^\star(s) \right)^2 - \gamma \hat{\xi}^\star(s) - \frac{s}{\lambda} = 0. \quad (3.76)$$

Recall that $\hat{\xi}_{\rho,s}$ is defined to be the root of $z = \hat{f}((1-\rho)^2s + \mu_\rho(1-z))$ with smallest absolute value. By Lemma 3.2, this root is unique and lies in the unit circle unless $s = 0$ and $\rho = 1$, in which case $\hat{\xi}(0) = 1$. Furthermore, it can be proved by Weierstrass Preparation Theorem (see Theorem 6.2 of [37]) that $\hat{\xi}_{\rho,s}$ is continuous in (ρ, s) . Hence, we have

$$\operatorname{Re} \left(\frac{1 - \hat{\xi}_{\rho,s}}{1 - \rho} \right) > 0, \text{ for all } \rho < 1 \text{ and } s > 0.$$

By taking limit $\rho \uparrow 1$, we have $\operatorname{Re}(\hat{\xi}^\star(s)) \geq 0$ for all $s > 0$.

As a consequence, we pick the root of (3.76) with non-negative real part. In particular, for real s , we have

$$\hat{\xi}^\star(s) = \frac{\gamma + \sqrt{\gamma^2 + 2(c_a^2 + 1)s/\lambda}}{c_a^2 + 1}. \quad (3.77)$$

For complex s , the square root in (3.77) corresponds to two complex roots, which are also the roots of $(\gamma - \sqrt{\gamma^2 + 2(c_a^2 + 1)s/\lambda}) / (c_a^2 + 1)$, since the polynomial in (3.76) is of order 2. Hence, we may use the same expression (3.77) as in the real case, as long as we pick the one with non-negative real part.

Combining (3.73) and (3.75), we obtain

$$\lim_{\rho \uparrow 1} \hat{H}_\rho(s) = -1,$$

and

$$\begin{aligned} \hat{H}^\star(s) &\equiv \lim_{\rho \uparrow 1} \frac{\hat{H}_\rho(s) + 1}{(1-\rho)s} = \lim_{\rho \uparrow 1} \left(\frac{1}{\mu_\rho} \frac{1-\rho}{1-\delta} - \frac{1}{\mu_\rho} \frac{1-\rho}{1-\hat{\xi}((1-\rho)^2s)} + O(1-\rho) \right) \\ &= \frac{1}{\lambda} \left(\frac{c_a^2 + 1}{2\gamma} - \frac{1}{\hat{\xi}^\star(s)} \right) < \infty, \end{aligned} \quad (3.78)$$

where $\hat{\xi}^*$ is defined in (3.75). Moreover, we have

$$\lim_{\rho \uparrow 1} \frac{1 - \hat{f}((1-\rho)^2 s)}{(1-\rho)^2 s} = -\hat{f}'(0) = E[U] = 1/\lambda.$$

and

$$\lim_{\rho \uparrow 1} \frac{\frac{1 - \hat{f}((1-\rho)^2 s)}{(1-\rho)^2 s} - \frac{1}{\lambda}}{(1-\rho)^2 s} = -\frac{\hat{f}''(0)}{2} = -\frac{E[U^2]}{2} = -\frac{c_a^2 + 1}{2\lambda^2}.$$

Combining everything into the Laplace Transform of $(1-\rho)^2 V_{d,\rho}((1-\rho)^{-2}t)$, we have

$$\hat{V}_d^*(s) \equiv \lim_{\rho \uparrow 1} \hat{V}_{d,\rho}^*(s) = \frac{\lambda}{s^2} - \frac{\lambda(c_a^2 - 1)}{s^2} \left(\frac{2\gamma\lambda}{c_a^2 + 1} \hat{H}^*(s) - 1 \right) \quad (3.79)$$

$$= \frac{\lambda}{s^2} + \frac{2\lambda}{s^2} \frac{c_a^2 - 1}{c_a^2 + 1} \frac{\gamma}{\hat{\xi}^*(s)}. \quad (3.80)$$

Plugging in (3.77), we obtain

$$\begin{aligned} \hat{V}_d^*(s) &= \frac{\lambda}{s^2} + \frac{2\lambda}{s^2} \frac{c_a^2 - 1}{c_a^2 + 1} \frac{\gamma}{\hat{\xi}^*(s)} \\ &= \frac{\lambda}{s^2} + \frac{\lambda}{s^2} \frac{c_a^2 - 1}{c_a^2 + 1} \frac{\sqrt{1 + 2(c_a^2 + 1)s/(\lambda\gamma^2)} - 1}{s/(\lambda\gamma^2)}, \end{aligned}$$

where we pick the root such that $\left(\sqrt{1 + 2(c_a^2 + 1)s/(\lambda\gamma^2)} - 1 \right) / (s/(\lambda\gamma^2))$ has non-negative real part. We used the fact that $\text{Re}(z) \geq 0$ if and only if $\text{Re}(1/z) \geq 0$ for $z \neq 0$.

For the explicit inversion, one can exploit the LT of the correclation function in (3.23) and note that

$$\mathcal{L}(f(at))(s) = \frac{1}{a} \hat{f}(s/a)$$

for any constant $a \neq 0$ and any function f with LT $\hat{f}$. For our case here, we use $a = \lambda\gamma^2/(c_a^2 + 1)$. ■

Proof of Theorem 3.5. To simplify the proof, we consider the HT-scaled difference between departure variance function and service variance function. Let $1/E[V] = \mu$ and $1/E[U] = \lambda_\rho \equiv \mu(1 - (1-\rho)\gamma_\rho)$, where γ_ρ are positive constants such that $\lim_{\rho \uparrow 1} \gamma_\rho = \gamma > 0$. Under this setting, we have $(\lambda_\rho - \mu)/(1-\rho) \rightarrow -\mu\gamma$ as $\rho \uparrow 1$. Let $\hat{V}_{d,\rho}^*(s)$ and $\hat{V}_{s,\rho}^*(s)$ be the LT of $V_{d,\rho}^*(s)$ and $V_{s,\rho}^*(s)$, respectively. Recall that Π was defined in Theorem 3.4. By

(3.37), we have

$$\begin{aligned}
\hat{V}_d^*(s) - \hat{V}_s^*(s) &= \lim_{\rho \uparrow 1} \left(\hat{V}_{d,\rho}^*(s) - \hat{V}_{s,\rho}^*(s) \right) \\
&= \lim_{\rho \uparrow 1} (1 - \rho)^4 \left(\hat{V}_{d,\rho}((1 - \rho)^2 s) - \hat{V}_{s,\rho}((1 - \rho)^2 s) \right) \\
&= \lim_{\rho \uparrow 1} \left(\frac{\lambda_\rho - \mu}{s^2} + \frac{2(\lambda_\rho - \mu)}{s^2} \frac{\hat{g}((1 - \rho)^2 s)}{1 - \hat{g}((1 - \rho)^2 s)} - \frac{2(\lambda_\rho^2 - \mu^2)}{(1 - \rho)^2 s^3} \right) \\
&\quad - \lim_{\rho \uparrow 1} \left(\frac{2\lambda_\rho}{s^2} \frac{\hat{g}((1 - \rho)^2 s)}{1 - \hat{g}((1 - \rho)^2 s)} \frac{(1 - \rho)^2 s \Pi(\hat{\nu}((1 - \rho)^2 s))}{(1 - \rho)^2 s + \lambda_\rho (1 - \hat{\nu}((1 - \rho)^2 s))} \right) \\
&= \lim_{\rho \uparrow 1} \frac{\lambda_\rho - \mu}{s^2} \left(1 - 2(\lambda_\rho + \mu) \frac{1}{(1 - \rho)^2 s} \left(1 - \frac{\hat{g}((1 - \rho)^2 s)}{\mu \frac{1 - \hat{g}((1 - \rho)^2 s)}{(1 - \rho)^2 s}} \right) \right) \\
&\quad + \lim_{\rho \uparrow 1} \frac{2\lambda_\rho}{s^2} \frac{\hat{g}((1 - \rho)^2 s)}{\frac{1 - \hat{g}((1 - \rho)^2 s)}{(1 - \rho)^2 s}} \cdot \frac{1}{1 - \rho} \left(\frac{\gamma_\rho}{s} - \frac{\Pi(\hat{\nu}((1 - \rho)^2 s))}{(1 - \rho)s + \lambda_\rho \frac{1 - \hat{\nu}((1 - \rho)^2 s)}{1 - \rho}} \right) \\
&= \lim_{\rho \uparrow 1} \hat{F}_\rho^{(1)}(s) + \lim_{\rho \uparrow 1} \frac{2\lambda_\rho \mu}{s^2} \frac{\hat{g}((1 - \rho)^2 s)}{\mu \frac{1 - \hat{g}((1 - \rho)^2 s)}{(1 - \rho)^2 s}} \frac{1}{(1 - \rho)s + \lambda_\rho \frac{1 - \hat{\nu}((1 - \rho)^2 s)}{1 - \rho}} \cdot \hat{F}_\rho^{(2)}(s)
\end{aligned}$$

where

$$\hat{F}_\rho^{(1)}(s) \equiv \frac{\lambda_\rho - \mu}{s^2} \left(1 - 2(\lambda_\rho + \mu) \frac{1}{(1 - \rho)^2 s} \left(1 - \frac{\hat{g}((1 - \rho)^2 s)}{\mu \frac{1 - \hat{g}((1 - \rho)^2 s)}{(1 - \rho)^2 s}} \right) \right)$$

and

$$\hat{F}_\rho^{(2)}(s) \equiv \frac{\gamma_\rho}{1 - \rho} \left(1 - \rho + \frac{\lambda_\rho}{s} \frac{1 - \hat{\nu}((1 - \rho)^2 s)}{1 - \rho} - \frac{1}{\gamma_\rho} \Pi(\hat{\nu}((1 - \rho)^2 s)) \right).$$

One can easily show that $\hat{F}_\rho^{(1)}(s)$ converges to 0 as $\rho \uparrow 1$. Note also that $\hat{g}(0) = 1$ and $\hat{g}'(0) = -E[V] = -1/\mu$, then

$$\lim_{\rho \uparrow 1} \frac{\hat{g}((1 - \rho)^2 s)}{\mu \frac{1 - \hat{g}((1 - \rho)^2 s)}{(1 - \rho)^2 s}} = 1.$$

Furthermore, a Taylor series expansion around $s = 0$ yields

$$\begin{aligned}
\frac{\hat{\nu}((1 - \rho)^2 s) - 1}{1 - \rho} &= \frac{\hat{g}((1 - \rho)^2 s + \lambda_\rho (1 - \hat{\nu}((1 - \rho)^2 s))) - 1}{1 - \rho} \\
&= -\frac{1 - \rho}{\mu} s + \frac{\lambda_\rho}{\mu} \frac{\hat{\nu}((1 - \rho)^2 s) - 1}{1 - \rho} \\
&\quad + \frac{\hat{g}''(0) + o(1)}{2(1 - \rho)} ((1 - \rho)^2 s + \lambda_\rho (1 - \hat{\nu}((1 - \rho)^2 s)))^2,
\end{aligned}$$

which implies that

$$\lim_{\rho \uparrow 1} \hat{\nu}((1 - \rho)^2 s) = 1 \tag{3.81}$$

and

$$\begin{aligned}
0 &= -\frac{s}{\mu} + \frac{1 - \frac{\lambda_\rho}{\mu}}{1 - \rho} \frac{1 - \hat{\nu}((1 - \rho)^2 s)}{1 - \rho} + \frac{\hat{g}''(0) + o(1)}{2(1 - \rho)^2} ((1 - \rho)^2 s + \lambda_\rho (1 - \hat{\nu}((1 - \rho)^2 s)))^2 \\
&= -\frac{s}{\mu} + \gamma_\rho \frac{1 - \hat{\nu}((1 - \rho)^2 s)}{1 - \rho} + \frac{\hat{g}''(0) + o(1)}{2} \left((1 - \rho)s + \lambda_\rho \frac{1 - \hat{\nu}((1 - \rho)^2 s)}{1 - \rho} \right)^2 \\
&= -\frac{s}{\mu} + \gamma_\rho \frac{1 - \hat{\nu}((1 - \rho)^2 s)}{1 - \rho} + \frac{\lambda_\rho^2 c_s^2 + 1}{\mu^2} \frac{1}{2} \left(\frac{1 - \hat{\nu}((1 - \rho)^2 s)}{1 - \rho} \right)^2 + o(1),
\end{aligned}$$

where we used the fact that $\hat{g}''(0) = E[V^2] = (c_s^2 + 1)/\mu^2$. Hence,

$$\lim_{\rho \uparrow 1} \frac{1 - \hat{\nu}((1 - \rho)^2 s)}{1 - \rho} = \nu^*(s),$$

where

$$\frac{1 + c_s^2}{2} (\nu^*(s))^2 + \gamma \nu^*(s) - \frac{s}{\mu} = 0. \quad (3.82)$$

With essentially the same argument as in the proof of Theorem 3.3, one can also show that $\nu^*(s)$ is the only root of (3.82) with positive real part, furthermore

$$\nu^*(s) = \frac{-\gamma + \sqrt{\gamma^2 + 2(1 + c_s^2)s/\mu}}{1 + c_s^2}. \quad (3.83)$$

It remains to show that $\hat{F}_\rho^{(2)}(s)$ converges (pointwise) to a proper limit. To this end, we write

$$\begin{aligned}
&\hat{F}_\rho^{(2)}(s) \\
&= \frac{\gamma_\rho}{1 - \rho} \left(1 - \rho + \frac{\lambda_\rho}{s} \frac{1 - \hat{\nu}((1 - \rho)^2 s)}{1 - \rho} - \frac{1}{\gamma_\rho} \Pi(\hat{\nu}((1 - \rho)^2 s)) \right) \\
&= \gamma_\rho + \gamma_\rho \frac{1 - \hat{\nu}((1 - \rho)^2 s)}{1 - \rho} \frac{1}{1 - \rho} \left(\frac{\lambda_\rho}{s} - \frac{1}{\gamma_\rho} \frac{(1 - \lambda_\rho/\mu)(1 - \rho)\hat{g}(\lambda_\rho(1 - \hat{\nu}((1 - \rho)^2 s)))}{\hat{g}(\lambda_\rho(1 - \hat{\nu}((1 - \rho)^2 s)) - \hat{\nu}((1 - \rho)^2 s))} \right) \\
&= \gamma_\rho + \gamma_\rho \frac{1 - \hat{\nu}((1 - \rho)^2 s)}{1 - \rho} \frac{1}{1 - \rho} \left(\frac{\lambda_\rho}{s} - \frac{(1 - \rho)^2 \hat{g}(\lambda_\rho(1 - \hat{\nu}((1 - \rho)^2 s)))}{\hat{g}(\lambda_\rho(1 - \hat{\nu}((1 - \rho)^2 s)) - \hat{\nu}((1 - \rho)^2 s))} \right)
\end{aligned}$$

Note that

$$\begin{aligned}
&\hat{g}(\lambda_\rho(1 - \hat{\nu}((1 - \rho)^2 s)) - \hat{\nu}((1 - \rho)^2 s)) \\
&= \hat{g}(\lambda_\rho(1 - \hat{\nu}((1 - \rho)^2 s)) - \hat{g}((1 - \rho)^2 s + \lambda_\rho(1 - \hat{\nu}((1 - \rho)^2 s))) \\
&= (1 - \rho)^2 \frac{s}{\mu} - \hat{g}''(0)(1 - \rho)^2 s \lambda_\rho (1 - \hat{\nu}((1 - \rho)^2 s)) + O((1 - \rho)^4),
\end{aligned}$$

one can easily show that

$$\lim_{\rho \uparrow 1} \hat{F}_\rho^{(2)}(s) = \gamma - \gamma \hat{\nu}^*(s) \frac{\mu}{s} (1 + c_s^2 \hat{\nu}^*(s)). \quad (3.84)$$

Plugging everything into the Laplace Transform of the heavy-traffic scaled difference of the variance functions, we have

$$\begin{aligned} \hat{V}_d^*(s) &= \hat{V}_s^*(s) + \frac{2\mu^2}{s^2} \frac{1}{\mu \hat{\nu}^*(s)} \left(\gamma - \gamma \hat{\nu}^*(s) \frac{\mu}{s} (1 + c_s^2 \hat{\nu}^*(s)) \right) \\ &= \frac{\mu c_s^2}{s^2} + \frac{\gamma \mu^2 (1 - c_s^2)}{s^3} \hat{\nu}^*(s) \end{aligned} \quad (3.85)$$

where we apply (3.82) to obtain the simplified expression in (3.85).

To obtain the explicit inversion, we write

$$\hat{V}_d^*(s) = \frac{\mu c_s^2}{s^2} + \frac{\gamma \mu^2 (1 - c_s^2)}{s^3} \frac{-\gamma + \sqrt{\gamma^2 + 2(1 + c_s^2)s/\mu}}{1 + c_s^2}.$$

Then, one exploits the LT of the correlation function in (3.23) and note that $\mathcal{L}(f(at))(s) = \hat{f}(t/a)/a$, for any constant $a \neq 0$ and any function f with LT $\hat{f}$. For our case here, we use $a = \mu \gamma^2 / (1 + c_s^2)$. ■

Proof of Theorem 3.8. By combining Theorems 2.1 and 4.2 in Chapter X of [12], we deduce that the k^{th} moment of the steady-state queue length is finite if the $(k+1)^{st}$ moments of the interarrival time and service time are finite. We add the extra uniformly bounded fourth moment to provide the uniform integrability needed to get convergence of the moments in the HT limit. We use (3.47) to obtain the corresponding result for the departure process.

To get (5.9), combine (5.8) and (3.60). Note that

$$\text{Var}(Q^*(t)) = \text{Var}(Q^*(0)) = c_x^4/4,$$

so that

$$\text{Var}(D^*(t)) = c_a^2 t + \frac{c_x^4}{2} - 2\text{Cov}(Q^*(0), Q^*(t)) - 2\text{Cov}(c_a B_a(t), Q^*(t)), \quad (3.86)$$

where

$$\text{Cov}(Q^*(0), Q^*(t)) = \frac{c_x^4}{4} c^*(\lambda t / c_x^2), \quad t \geq 0; \quad (3.87)$$

see Section 2 of [3] or Theorem 5.7.11 of [143]. Inserting (3.87) into (3.86) yields the first line in (5.9) above. To establish the second limit, we do a space-time transformation of the limit, so that the limit is the same as one of the models analyzed directly.

Let us re-scale space and time so that the general result is in terms on B_a instead of $c_a B_a$ (assuming that $c_a > 0$), so that we can apply the established result for the $M/GI/1$ model. (Essentially the same argument works for $GI/M/1$.) The first step is to observe that the HT limit for the departure process $\{D^*(t) : t \geq 0\}$ can be written as a function $\Psi : \mathbb{R} \times \mathcal{D}^3 \rightarrow \mathcal{D}$ of the vector process $\{(Q^*(0), c_a B_a \circ \lambda e, c_s B_s \circ \lambda e, -\lambda e)\}$; i.e., by (3.60)

$$\begin{aligned} D^* &= \Psi((Q^*(0), c_a B_a \circ \lambda e, c_s B_s \circ \lambda e, -\lambda e)) \\ &= Q^*(0) + c_a B_a \circ \lambda e - \Psi(Q^*(0) + c_a B_a \circ \lambda e - c_s B_s \circ \lambda e - \lambda e)(t). \end{aligned}$$

If we replace the basic vector process $(Q^*(0), c_a B_a \circ \lambda e, c_s B_s \circ \lambda e, -\lambda e)$ by another that has the same distribution as a process, then the distribution of D^* will be unchanged.

By the basic time and space scaling of BM, for $c_a > 0$, the stochastic processes have equivalent distributions as follows

$$\begin{aligned} \{Q^*(0), c_a B_a(\lambda t), c_s B_s(\lambda t), -\lambda t\} &\stackrel{d}{=} c_a^2 \left\{ \frac{Q^*(0)}{c_a^2}, B_a(\lambda t/c_a^2), \frac{c_s}{c_a} B_s(\lambda t/c_a^2), -\frac{\lambda t}{c_a^2} \right\} \\ &\equiv c_a^2 \left\{ \frac{Q^*(0)}{c_a^2}, B_a(u), \frac{c_s}{c_a} B_s(u), -u \right\}, \end{aligned} \quad (3.88)$$

where $u = \lambda t/c_a^2$. After this transformation, to describe the system at time u , the associated RBM has drift -1 and variance coefficient $1 + (c_s^2/c_a^2) = c_x^2/c_a^2$. Note that the mean of the steady-state distribution associated with the new RBM is the diffusion coefficient divided by twice the absolute value of the drift, which is $c_x^2/(2c_a^2)$. As a result, $Q^*(0)/c_a^2$ is exactly the steady-state distribution needed for the new RBM. From above, we see that

$$\begin{aligned} D^*(t) &\stackrel{d}{=} \Psi \left(c_a^2 \left\{ Q^*(0)/c_a^2, B_a(u), (c_s/c_a) B_s(u), -u \right\} \right), \quad \text{for } u = \lambda t/c_a^2 \\ &= c_a^2 \Psi \left(\left\{ Q^*(0)/c_a^2, B_a(u), (c_s/c_a) B_s(u), -u \right\} \right) \\ &\equiv c_a^2 \tilde{D}^*(u) = c_a^2 \tilde{D}^*(\lambda t/c_a^2), \end{aligned}$$

where $\tilde{D}^*(u) \equiv \Psi \left(\left\{ Q^*(0)/c_a^2, B_a(u), (c_s/c_a) B_s(u), -u \right\} \right)$, corresponding to the $M/GI/1$ model with service scv c_s^2/c_a^2 . Now, let $\tilde{w}^*(t)$ denote the associated weight function in

(3.43) with $(\mu, \gamma, \tilde{c}_x^2) = (\lambda, 1, c_x^2/c_a^2)$, so that

$$\tilde{w}^*(t) = w^*(c_a^2 \lambda t / c_x^2).$$

We now turn to the variance. By applying (3.43), we obtain

$$\begin{aligned} V_d^*(t) &= c_a^4 \tilde{V}_d^*(u) = c_a^4 \tilde{V}_d^*(\lambda t / c_a^2) \\ &= c_a^4 \left(\tilde{w}^*(\lambda t / c_a^2) \frac{\lambda t}{c_a^2} + (1 - \tilde{w}^*(\lambda t / c_a^2)) \frac{c_s^2}{c_a^2} \frac{\lambda t}{c_a^2} \right) \\ &= w^*(\lambda t / c_x^2) c_a^2 \lambda t + (1 - w^*(\lambda t / c_x^2)) c_s^2 \lambda t, \end{aligned}$$

which agrees with the $GI/GI/1$ formula in (5.9). Thus, we have proved the variance formula for $GI/GI/1$.

Finally, it remains to establish the covariance formulas. First, by comparing the two lines in (5.9) and recall (3.26), we must have

$$\text{Cov}(c_a B_a(\lambda t), Q^*(t)) = (1 - w^*(\lambda t / c_x^2)) c_a^2 \lambda t.$$

Let $\tilde{B}_s(t) = -B_s(t)$, then we have

$$(Q^*(0), c_a B_a \circ \lambda e, c_s B_s \circ \lambda e) \stackrel{d}{=} (Q^*(0), c_a B_a \circ \lambda e, c_s \tilde{B}_s \circ \lambda e),$$

so

$$\begin{aligned} \text{Cov}(c_a B_a(t), Q^*(t)) &= \text{Cov}(c_a B_a(t), \Psi(Q^*(0) + c_a B_a + c_s \tilde{B}_s - e)(t)) \\ &= \text{Cov}(c_a B_a(t), \Psi(Q^*(0) + c_a B_a + c_s B_s - e)(t)), \end{aligned}$$

and

$$\begin{aligned} \text{Cov}(c_s B_s(\lambda t), Q^*(t)) &= \text{Cov}(-c_s \tilde{B}_s(\lambda t), \Psi(Q^*(0) + c_a B_a \circ \lambda e + c_s \tilde{B}_s \circ \lambda e - e)(t)) \\ &= -\text{Cov}(c_s B_s(\lambda t), \Psi(Q^*(0) + c_a B_a \circ \lambda e + c_s B_s \circ \lambda e - e)(t)). \end{aligned}$$

By symmetry, we thus have

$$\text{Cov}(c_s B_s(\lambda t), Q^*(t)) = - (1 - w^*(\lambda t / c_x^2)) c_s^2 \lambda t. \quad \blacksquare$$

Chapter 4

Heavy-Traffic Limits for Stationary Flows

In this chapter, we further study the IDC formulation of the RQ algorithm by establishing the heavy-traffic limit theorems for the stationary flows and their IDC's. A *customer flow* is defined as a continuous-time process, counting the number of customer in or out of the network, or flowing from one queue to another. A typical example is *the total arrival flow* at a queue, which is the superposition of the external arrival flow and internal arrival flows, with the later ones being (part of) the departure flows from other queues that are directed to the current queue. The IDC of the total arrival flow, together with the its rate and the service rate at the same station, serves as the input for the IDC formulation of the RQ algorithm, see (2.35).

Flows are special stochastic point processes, for which there is a well-developed general theory, as in [47; 48]. There also is a substantial literature on the general structure of stationary point processes in queueing systems, as in Chapter 1 of [15] and [122], but concrete results, such as explicit formulas describing the stochastic variability of the flows over time, are extremely rare. The familiar exception is the Markov OQN, for which there is a substantial theory, as in Chapter 4 of [132], but even in Markov networks, the flows can be quite complicated. First, by reversibility, for Jackson networks, the departure processes out of the network from the queues are independent Poisson processes, but the internal

flows need not be Poisson, even though the product-form property holds. In particular, the flows are Poisson if and only if they are not part of a loop; see [106; 131]. For non-Markov open networks, the flows are even more complicated.

This chapter contributes to the approximation of OQN's using the IDC's by establishing heavy-traffic limits for all the stationary flows in a OQN, allowing any subset of the stations to be bottleneck stations (critically loaded in the limit). The heavy-traffic limits are especially tractable in the case of a single bottleneck station, because they can be expressed in terms of one-dimensional reflected Brownian motion (RBM), so that the heavy-traffic approximation of the IDC can be calculated in closed-form just as in (3.1).

Our HT limit for the stationary flows rely heavily on the justification for interchanging the limits $t \rightarrow \infty$ and $\rho \uparrow 1$ in a OQN provided by Gamarnik and Zeevi [66] and Budharija and Lee [31]. By allowing an arbitrary subset of the queues to be bottleneck queues (have nondegenerate limits), while the rest have null limits, we follow Chen and Mandelbaum [34; 35]. Our main contributions here are the heavy-traffic limits for the stationary flows.

As a preliminary step for our heavy-traffic limit, we establish conditions for the existence of stationary flows in a GJN and for convergence to those stationary flows as time evolves. For that we rely heavily on the Harris recurrence that was used to establish the stability of a GJN under appropriate regularity, drawing on Sigman [120; 121] and Dai [42]; see Chapter VII of Asmussen [12].

This chapter is organized as follows. In Section 4.1, we formally introduce the open queueing network model and the continuous-time stochastic processes (flows) associated with it. This model goes beyond the assumptions that we make to establish Harris recurrence. Our RQNA algorithm and approximations of the IDC's are intended to work with this general model even without theoretical justification developed in this chapter. In Section 4.2, we establish the existence and convergence of the stationary flows. In Section 4.3, we show a general version of the heavy-traffic limit theorem for the stationary flows, allowing an arbitrary set of bottleneck queues that approaches heavy traffic simultaneously. Our joint HT functional central limit reveals the key connections among different flows. In Section 4.4, we focus on a special case, where we only have one bottleneck queue. This serves as a basis for our heavy-traffic approximation of the IDC's of the stationary flows.

4.1 The OQN Model

We start by formulating a general OQN model that goes beyond the assumptions we make to establish Harris recurrence. Let there be K single-server stations with unlimited waiting space and the first-come first-served (FCFS) discipline. We assume that the system starts empty at time 0, but that could be relaxed. We associate with each station i an external arrival point process $A_{0,i}$, which satisfies $A_{0,i}(t) < \infty$ with probability 1 for any t . Let $A_0 \equiv (A_{0,1}, \dots, A_{0,K})$ denote the vector of all external arrival processes.

Let $\{V_i^l : l \geq 1\}$ denote the sequence of service times at station i and define the (uninterrupted) service point (counting) process as

$$S_i(t) = \max \left\{ n \geq 0 : \sum_{l=1}^n V_i^l \leq t \right\}, \quad t \geq 0. \quad (4.1)$$

which we also assume to have finite sample path with probability 1.

In addition to external arrivals, departures from each station may be routed to other queues or out of the network. To specify the general routing (or splitting) process, let $\theta_i^l \in \{0, 1\}^K$ indicate the routing vector of the l -th departure from queue i . Following standard conventions, at most one component of θ_i^l is 1, and $\theta_i^l = e_j$ indicates that the l -th departure from the i -th queue is routed to station j for $1 \leq j \leq K$, where e_j is the j -th standard basis of the Euclidean space $\mathbb{R}^K$. The case $\theta_i^l = 0$ indicates that the l -th departure from the i -th queue exits the system. Finally, we define the routing decisions up to the n -th decision at station i by

$$\Theta_i(n) \equiv (\Theta_{i,1}(n), \dots, \Theta_{i,K}(n)) \equiv \sum_{l=1}^n \theta_i^l,$$

and let $\Theta_{i,0}(n)$ denote the number of customers that exit the system from station i in the first n departures.

For the internal arrival flows, let $A_{i,j}$ be the customer stream from i to j . Each internal arrival stream $A_{i,j}$ splits from the departure process D_i according to the splitting decision process $\Theta_{i,j}$, so that

$$A_{i,j}(t) = \Theta_{i,j}(D_i(t)), \quad t \geq 0, \quad 1 \leq i \leq K, \quad 0 \leq j \leq K. \quad (4.2)$$

Let $A_{\text{int}}(t) \equiv (A_{i,j}(t) : 1 \leq i, j \leq K)$ denote the matrix of all internal arrival flows.

For total arrival process at station i , let

$$A_i(t) = A_{0,i}(t) + \sum_{j=1}^K A_{j,i}(t)$$

and let $A(t) \equiv (A_1(t), \dots, A_K(t))$ be the vector of total arrival processes.

As observed in (7.1) and (7.2) in §7.2 of [34], the queue-length and departure processes at each queue are jointly uniquely characterized by the flow balance equations

$$Q_i(t) = Q_i(0) + A_i(t) - D_i(t) \quad \text{and} \quad D_i(t) = S_i(B_i(t)), \quad t \geq 0, \quad 1 \leq i \leq K, \quad (4.3)$$

where $B_i(t)$ is the cumulative busy time of server i up to time t , which by work conservation satisfies

$$B_i(t) = \int_0^t 1_{Q_i(u) > 0} du, \quad t \geq 0, \quad (4.4)$$

where 1_A is the indicator function with $1_A = 1$ on the set A and 0 elsewhere.

For the flow exiting the queueing system, let $D_{\text{ext},i}$ denote the flow that exits the system from station i . Hence

$$D_{\text{ext},i}(t) = \sum_{l=1}^{D_i(t)} \theta_{i,0}^l = \Theta_{i,0}(D_i(t)), \quad t \geq 0.$$

Finally, let $D_{\text{ext}}(t) \equiv (D_{\text{ext},1}(t), \dots, D_{\text{ext},K}(t))$ be the vector of external departure processes.

4.2 Existence, Uniqueness and Convergence Via Harris Recurrence

In this section we establish the existence of unique stationary flows and convergence to them as time increases for any initial state. Toward that end, we make three assumptions, the first one being

Assumption 4.1 *We assume that the OQN is a GJN, in particular:*

1. *The K external arrival processes are mutually independent (possibly null) renewal processes with finite rates λ_i , where the interarrival times have finite squared coefficient of variation (scv, variance divided by the square of the mean) $c_{a_0,i}^2$ for $1 \leq i \leq K$.*

2. The service times come from K mutually independent sequences of i.i.d. random variables with means $1/\mu_i$, $0 < \mu_i < \infty$, and finite scv $c_{s_i}^2$ for $1 \leq i \leq K$.
3. The routing is Markovian with a substochastic $K \times K$ routing matrix $P = (p_{i,j})_{1 \leq i,j \leq K}$ such that $p_{i,j} \geq 0$, $p_{i,0} \equiv 1 - \sum_{j=1}^K p_{i,j} \geq 0$ and $I - P'$ is invertible; For each $1 \leq i \leq K$, the sequence $\{\Theta_i(1), \Theta_i(2), \dots\}$ is i.i.d. with $P(\Theta_i(n) = e_j) = p_{i,j}$ and $P(\Theta_i(n) = 0) = p_{i,0} \equiv 1 - \sum_{j=1}^K p_{i,j}$.
4. The arrival, service and routing processes are mutually independent.

For completeness, we also assume that the network starts empty at time 0, so that no customer is in service or waiting, but this can be relaxed. The condition of finite scv's is used in the convergence of the distribution and in the next section; for relaxed assumptions, see the discussions below Theorem 4.1 and Theorem 4.2. Note that $I - P'$ is invertible if we assume that all customers eventually leave the system; see [36] or Theorem 3.2.1 of [90].

Let $U(t)$ denote the vector of residual external arrival times at time t ; let $V(t)$ be the vector of residual service times at time t , set to 0 when the server is idle; and let the *system state process* be

$$\mathcal{S}(t) \equiv (Q(t), U(t), V(t)), \quad t \geq 0. \quad (4.5)$$

Under our assumption, the initial condition is specified by $\mathcal{S}(0) = (0, 0, 0)$. The system state process $\mathcal{S}$ in (4.5) is an element of the function space $\mathcal{D}([0, \infty), \mathbb{R}^{3K})$ of real-valued functions on the half-line $[0, \infty)$ taking values in the Euclidean space $\mathbb{R}^{3K}$ that are right-continuous with left limits. As stated in §2.2 of [42], which draws on [53], Assumption 4.1 implies some basic regularity conditions.

Theorem 4.1 (strong Markov process) *Under Assumption 4.1, the system state process $\mathcal{S}$ is a strong Markov process.*

We remark that Assumption 4.1 is stronger than needed to ensure the strong Markov property. Since $\mathcal{S}$ is a piecewise-deterministic Markov process (defined in §3 of [53]), §4 of [53] showed that if the expected number of jumps on any interval $[0, t]$ is finite, then the process possesses the strong Markov property.

We now state the stability assumption in the sense of the traffic intensities. Let $\lambda_0 = (\lambda_{0,1}, \dots, \lambda_{0,K})$ be the external arrival rate vector and let $\lambda = (\lambda_1, \dots, \lambda_K)$ denote the vector of total arrival rate. We obtain λ by solving the *traffic-rate equations*

$$\lambda_i = \lambda_{0,i} + \sum_{j=1}^K \lambda_{j,i} = \lambda_{0,i} + \sum_{i=1}^K \lambda_j p_{j,i}, \quad (4.6)$$

or, in matrix form,

$$(I - P')\lambda = \lambda_0,$$

where I denotes the $K \times K$ identity matrix and P' is the transpose of P . Let $\lambda_{i,j} \equiv \lambda_i p_{i,j}$ be the rate of the internal arrival flow from i to j . Finally, let $\rho_i \equiv \lambda_i / \mu_i$ be the traffic intensity at station i .

Assumption 4.2 *The traffic intensities satisfy $\max_i \rho_i < 1$.*

Following convention, we say that the OQN is *stable* if the system state process in (4.5) is stable, i.e., if there exists a distribution π on $\mathbb{R}^{3K}$ for $\mathcal{S}(0)$ such that $\mathcal{S}(t)$ has that same distribution π for all $t \geq 0$. We now state the additional assumption to ensure the uniqueness of the stationary distribution π and the convergence of the distribution of $\mathcal{S}(t)$ to π .

Assumption 4.3 *Each non-null external arrival process has an interarrival-time distribution with a density that is positive for almost all t .*

Our assumption here implies the key assumption (A3) in both [42] and [43] that the distribution is unbounded and spread out, see also [42] and Chapter VII of [12]. This clearly avoids periodic behavior associated with the lattice case, but otherwise it is not restrictive for practical modeling.

The following theorem follows from Theorem 2 of [66] or Theorem 5.1 of [42] or Theorem 6.2 of [43], which extend earlier work on stability for OQNs in [25], [121] and [63].

Theorem 4.2 (existence, uniqueness and convergence) *Under Assumptions 4.1-4.3, the system state stochastic process $\mathcal{S}$ in (4.5) is a positive Harris recurrent Markov process. There exists a unique stationary distribution π and for every initial condition and the distribution of $\mathcal{S}(t)$ converges to π as $t \rightarrow \infty$.*

For a strong Markov process with right-continuous and left limit sample paths, the existence of a stationary distribution is shown in the early [14], which in turn draws on [69]. The uniqueness is shown in [42], which assumes that the interarrival times are unbounded, spreadout and have finite mean, and the service times have finite mean; see (1.2)-(1.5) there. The convergence follows from [43] under the additional assumption of finite second moment.

We now state the strong implications of Theorem 4.2. For that, we consider the system that starts at time s . For the system state processes, let $Q_s(t) = Q(s+t)$, $U_s(t) = U(s+t)$ and $V_s(t) = V(s+t)$, so that $\mathcal{S}_s \equiv (Q_s, U_s, V_s)$ is the system state process with initial condition $\mathcal{S}(s)$. Let $\Rightarrow$ denote weak convergence. Theorem 4.2 implies that

Corollary 4.1 *Under Assumptions 4.1-4.3, $Q_s(t)$ has unit (± 1) jumps and*

$$\mathcal{S}_s \Rightarrow \mathcal{S}_e \equiv (Q_e, U_e, V_e), \quad \text{as } s \rightarrow \infty, \quad (4.7)$$

where $\mathcal{S}_e$ is the system state process with initial condition $\mathcal{S}_e(0)$ distributed as the stationary distribution π and $\Rightarrow$ denote weak convergence in each coordinate.

Proof. Assumption 4.3 implies that with probability 1, there is at most 1 (internal or external) arrival at any station and that the arrival times do not coincide with departure times at any station. Hence, Q_s only has unit-jumps.

From Theorem 4.2, we have the convergence of one-dimensional distribution

$$\mathcal{S}_s(t_1) \Rightarrow \mathcal{S}_e(t_1), \quad \text{for all } t_1 \geq 0.$$

To extend the convergence to any finite-dimensional distribution, we utilize the Markov property of $\mathcal{S}(t)$ in Theorem 4.1. For any $t_2 = t_1 + \delta_1 > t_1$, the conditional probability distribution of the state $\mathcal{S}(t_2)$, conditioning on the past values up to the time t_1 , depends only on the current state $\mathcal{S}_s(t_1)$. Apply Theorem 4.2 again with initial state $\mathcal{S}_s(t_1)$, we have

$$(\mathcal{S}_s(t_1), \mathcal{S}_s(t_2)) \Rightarrow (\mathcal{S}_e(t_1), \mathcal{S}_e(t_2)), \quad \text{for all } 0 \leq t_1 < t_2.$$

By induction, the convergence can be extended to any finite-dimensional distribution. The weak convergence of the process $\mathcal{S}_s$ then follows from Theorem 12.6 in [23]. ■

Now, we turn to the existence of stationary flows. Define the auxiliary cumulative process $\mathcal{C}$, as in §VI.3 of [12], by

$$\mathcal{C}(t) \equiv (B(t), Y(t)), \quad (4.8)$$

where $B_i(t)$ is the cumulative busy times for server i over interval $[0, t]$ and

$$Y_i(t) \equiv \mu_i(t - B_i(t)) \quad (4.9)$$

is the cumulative idle time of station i , scaled by the service rate μ_i .

To focus on the flows, we describe the GJN by the aggregate process

$$\mathcal{M}(t) \equiv (\mathcal{S}(t), \mathcal{C}(t), \mathcal{F}(t)), \quad (4.10)$$

where

$$\mathcal{F}(t) \equiv (A_0(t), A_{\text{int}}(t), A(t), S(t), D(t), D_{\text{ext}}(t)) \quad (4.11)$$

is a vector of cumulative point processes, with the processes defined in §4.1. We refer to $\mathcal{F}$ in (4.11) as the *flows*. We say that a flow is *stationary* if it has stationary increments. We refer to [122] and Chapter 6 of [28] for background on stationary stochastic processes and ergodicity.

For the flows, let $A_{0,s}(t) = A_0(t+s) - A_0(s)$ be the external arrival counting process that starts at time s . Similarly, let $A_{\text{int},s}(t) = A_{\text{int}}(t+s) - A_{\text{int}}(s)$, $A_s(t) = A(t+s) - A(s)$, $D_s(t) = D(t+s) - D(s)$, $D_{\text{ext},s}(t) = D_{\text{ext}}(t+s) - D_{\text{ext}}(s)$, $B_s(t) = B(t+s) - B(s)$ and $Y_s(t) = Y(t+s) - Y(s)$ be the corresponding processes that starts at time s . The service processes $S_s(t)$ are more subtly defined by

$$S_{i,s}(t) \equiv S_i(B_i(s) + t) - S_i(B_i(s)), \quad \text{for } i = 1, 2, \dots, K, \quad (4.12)$$

which is a vector of delayed renewal processes with first intervals distributed as $V(s)$, the vector residual service time and at system time s (its i -th component is also the residual service time of the process S_i at time $B_i(s)$). This definition of the service process allow us to write the departure process as a composition of the two processes S_s and B_s via

$$\begin{aligned} D_s(t) \equiv D(s+t) - D(s) &= (S \odot B)(s+t) - (S \odot B)(s) \\ &= (S_s \odot B_s)(t), \quad t \geq 0., \end{aligned} \quad (4.13)$$

where $\odot$ is understood as component-wise composition, i.e. $D_{i,s} = S_{i,s} \circ B_{i,s}$ for all i . Finally, let $\mathcal{C}_s \equiv (B_s, Y_s)$ and $\mathcal{F}_s \equiv (A_{0,s}, A_{\text{int},s}, A_s, S_s, D_s, D_{\text{ext},s})$.

Theorem 4.3 (Existence and convergence of the stationary flows) *Under Assumptions 4.1-4.3, there exists unique stationary and ergodic cumulative processes (with stationary increments satisfying the LLN)*

$$\mathcal{C}_e \equiv (B_e, Y_e), \quad \mathcal{F}_e \equiv (A_{0,e}, A_{\text{int},e}, A_e, S_e, D_e, D_{\text{ext},e})$$

and a unique stationary process

$$\mathcal{S}_e \equiv (Q_e, U_e, V_e),$$

such that, as $s \rightarrow \infty$,

$$\mathcal{M}_s \equiv (\mathcal{S}_s, \mathcal{C}_s, \mathcal{F}_s) \Rightarrow (\mathcal{S}_e, \mathcal{C}_e, \mathcal{F}_e) \equiv \mathcal{M}_e, \quad (4.14)$$

where $\Rightarrow$ denote weak convergence in each coordinate. Furthermore, $A_{0,e}$ is the vector of equilibrium external arrival renewal processes, S_e is a vector of delayed renewal process with first interval distributed as $V_e(0)$.

4.3 Heavy-Traffic Limit Theorems for the Stationary Processes

To set the stage for our heavy-traffic limits, in Section 4.3.1 we present a centered representation of the flows. This representation parallels those used in [113; 34; 35; 42], but here we focus on the flows. Then in Section 4.3.2 we establish our main heavy-traffic limit.

4.3.1 Representation of the Centered Stationary Flows

Recall that the external arrival rate vector is λ_0 , so the total arrival rates are given by $\lambda = (I - P')^{-1} \lambda_0$ as in (4.6). For service, we start with rate-1 base service process S_i^0 for station i and scale it by μ_i so that the service process at station i is denoted by $S_i \equiv S_i^0 \circ \mu_i e$ with $e(t) = t$ being the identity function. Let the center processes be defined by

$$\begin{aligned} \tilde{A}_{0,i} &= A_{0,i} - \lambda_{0,i} e, \quad \tilde{A}_i = A_i - \lambda_i e, \quad \tilde{D}_i = D_i - \lambda_i e, \\ \tilde{\Theta}_{j,i} &= \Theta_{j,i} \circ (S_j \circ B_j) - p_{j,i} S_j \circ B_j, \quad \text{and} \quad \tilde{S}_i = S_i \circ B_i - \mu_i B_i. \end{aligned} \quad (4.15)$$

Furthermore, let $X(t)$ be the *net-input process*, allowing the service to run continuously, defined as

$$X \equiv Q(t) - (I - P')Y, \quad (4.16)$$

where Y is defined in (4.9).

The next theorem expresses the queue length processes, the centered total arrival and the centered departure flows in terms of the centered external arrival, service and routing processes. Let Ψ_P be the K -dimensional reflection map with reflection matrix P ; e.g., see Chapter 14 of [143]. A proof can be found in Section 4.6.

Theorem 4.4 (Centered representation) *The net-input process can be written as*

$$X = Q(0) + \tilde{A}_0 + \tilde{\Theta}'\mathbf{1} - (I - P')\tilde{S} + (\lambda_0 - (I - P')\mu)e, \quad (4.17)$$

while the queue length process can be written as

$$Q = X + (I - P')Y = \Psi_{I-P'}(X), \quad (4.18)$$

where $\Psi_{I-P'}$ is the K -dimensional reflection mapping with reflection matrix $I - P'$. In addition, the centered total arrival and departure processes can be written as

$$\tilde{A} = P'(I - P')^{-1}(Q(0) - Q) + (I - P')^{-1}(\tilde{A}_0 + \tilde{\Theta}'\mathbf{1}), \quad (4.19)$$

$$\tilde{D} = (I - P')^{-1}(Q(0) - Q + \tilde{A}_0 + \tilde{\Theta}'\mathbf{1}), \quad (4.20)$$

where the centered processes are defined in (4.15).

Remark 4.1 (Stationary flows) *Note that the representation in Theorem 4.4 does not impose any assumption on the initial condition of the open queueing network. As ensured by Theorem 4.3, there exists a stationary distribution π such that the flows are stationary if $S(0) \sim \pi$. With this specific initial condition, Theorem 4.4 applies to the stationary flows.*

4.3.2 Heavy-Traffic Limit with Any Subset of Bottlenecks

In this section, we establish the convergence of the stationary flows under HT limit. Throughout this section, we assume that the system is stationary in the sense of Theorem 4.3 and we suppress the subscript e to simplify the notation. We let an arbitrary pre-selected subset

$\mathcal{H}$ of the K stations be pushed into the HT limit while other stations stay unsaturated. Two important special cases are: (i) $|\mathcal{H}| = K$ so that all stations approaches HT at the same time, which corresponds to the original case in [113]; and (ii) $|\mathcal{H}| = 1$ so that only one station is in HT. This second case is appealing for applications because the RBM is only one-dimensional. We focus on it in detail later.

To start, consider a family of systems indexed by ρ . Let the ρ -dependent service rates be

$$\mu_{i,\rho} \equiv \lambda_i / (\kappa_i \rho), \quad 1 \leq i \leq K, \quad (4.21)$$

and set $\kappa_i = 1$ for all $i \in \mathcal{H}$ and $\kappa_i < 1$ for all $i \notin \mathcal{H}$. Equivalently, we have $\rho_i = \kappa_i \rho$. For the pre-limit systems we have the same representation of the flows as described in Theorem 4.4, with the only exception that μ_i in (4.17) is now replaced by the ρ -dependent version in (4.21).

We now define the HT-scaled processes. As in the usual HT scaling, we scale time by $(1 - \rho)^{-2}$ and scale space by $(1 - \rho)$. Thus we make the definitions

$$\begin{aligned} A_{0,i,\rho}^*(t) &\equiv (1 - \rho)[A_{0,i}((1 - \rho)^{-2}t) - (1 - \rho)^{-2}\lambda_{0,i}t], \\ A_{i,\rho}^*(t) &\equiv (1 - \rho)[A_{i,\rho}((1 - \rho)^{-2}t) - (1 - \rho)^{-2}\lambda_i t], \\ S_{i,\rho}^*(t) &\equiv (1 - \rho)[S_{i,\rho}((1 - \rho)^{-2}t) - (1 - \rho)^{-2}\mu_{i,\rho}t], \\ D_{i,\rho}^*(t) &\equiv (1 - \rho)[D_{i,\rho}((1 - \rho)^{-2}t) - (1 - \rho)^{-2}\lambda_i t], \\ D_{\text{ext},i,\rho}^*(t) &\equiv (1 - \rho)[D_{\text{ext},i,\rho}((1 - \rho)^{-2}t) - (1 - \rho)^{-2}\lambda_i p_{i,0}t], \\ A_{i,j,\rho}^*(t) &\equiv (1 - \rho)[A_{i,j,\rho}((1 - \rho)^{-2}t) - (1 - \rho)^{-2}\lambda_i p_{i,j}t], \\ \Theta_{i,j,\rho}^*(t) &\equiv (1 - \rho) \left[\sum_{l=1}^{\lfloor (1-\rho)^{-2}t \rfloor} \theta_{i,j}^l - p_{i,j}(1 - \rho)^{-2}t \right], \\ Q_{i,\rho}^*(t) &\equiv (1 - \rho)Q_{i,\rho}((1 - \rho)^{-2}t), \text{ for } 1 \leq i, j \leq K. \end{aligned} \quad (4.22)$$

Furthermore, let $\Theta_{i,\rho}^* \equiv (\Theta_{i,j,\rho}^* : 1 \leq j \leq K)$; let $\Theta_{\text{ext},\rho}^* \equiv (\Theta_{i,0,\rho}^* : 1 \leq i \leq K)$; and let $\mathcal{F}_\rho^*$ collects all the flows, defined as

$$\mathcal{F}_\rho^*(t) \equiv (A_{0,\rho}^*(t), A_{\text{int},\rho}^*(t), A_\rho^*(t), S_\rho^*(t), D_\rho^*(t), D_{\text{ext},\rho}^*(t)). \quad (4.23)$$

Finally, let $W_{i,\rho}^*(t) \equiv (1 - \rho)W_{i,\rho,\lfloor (1-\rho)^2 t \rfloor}$ denote the HT scaled waiting time process, where $W_{i,\rho,n}$ denotes the waiting time of the n -th customer at station i in the ρ -th system;

and let $Z_{i,\rho}^*(t) \equiv (1 - \rho)Z_{i,\rho}((1 - \rho)^2 t)$ denote the HT scaled workload process at station i in the ρ -th system.

Before presenting the HT limit of the systems, we introduce useful notation by discussing a modified and yet asymptotically equivalent system, where all service times at the nonbottleneck queues are set to zero.

Remark 4.2 (A reduced $\mathcal{H}$ -station network) Consider a $\mathcal{H}$ -station network, where all non-bottleneck queues have zero service times, so that they can be viewed as instantaneous switches. To obtain the rates and routing matrix in the equivalent network, we let $I_{\mathcal{A}}$ denote the $|\mathcal{A}| \times |\mathcal{A}|$ identity matrix for any index set $\mathcal{A}$; let $P_{\mathcal{H}}$ be the $|\mathcal{H}| \times |\mathcal{H}|$ submatrix of the original routing matrix P corresponding to the rows and columns in $\mathcal{H}$; similarly, let $P_{\mathcal{H}^c}$ be the $|\mathcal{H}^c| \times |\mathcal{H}^c|$ submatrix of P corresponding to $\mathcal{H}^c$; and let $P_{\mathcal{H}^c, \mathcal{H}}$ collect the routing probabilities from stations in $\mathcal{H}^c$ to the ones in $\mathcal{H}$, similarly, define $P_{\mathcal{H}, \mathcal{H}^c}$. Now the new $|\mathcal{H}| \times |\mathcal{H}|$ routing matrix, denoted by $\hat{P}_{\mathcal{H}}$, is

$$\begin{aligned} \hat{P}_{\mathcal{H}} &= P_{\mathcal{H}} + \sum_{l=0}^{\infty} P_{\mathcal{H}, \mathcal{H}^c} (P_{\mathcal{H}^c})^l P_{\mathcal{H}^c, \mathcal{H}} \\ &= P_{\mathcal{H}} + P_{\mathcal{H}, \mathcal{H}^c} \sum_{l=0}^{\infty} (P_{\mathcal{H}^c})^l P_{\mathcal{H}^c, \mathcal{H}} \\ &= P_{\mathcal{H}} + P_{\mathcal{H}, \mathcal{H}^c} (I_{\mathcal{H}^c} - P_{\mathcal{H}^c})^{-1} P_{\mathcal{H}^c, \mathcal{H}}. \end{aligned} \quad (4.24)$$

Note that the inverse $(I_{\mathcal{H}^c} - P_{\mathcal{H}^c})^{-1}$ appearing in (4.24) is the fundamental matrix associated with the transient finite Markov chain with transition matrix $P_{\mathcal{H}^c}$. If we let $\hat{P}_{\mathcal{H}^c, \mathcal{H}}$ denote the matrix of the probabilities that the first visit to a bottleneck queue of an external arrival at a non-bottleneck queue $i \in \mathcal{H}^c$ is at $j \in \mathcal{H}$, then we have

$$\hat{P}_{\mathcal{H}^c, \mathcal{H}} = \sum_{l=0}^{\infty} (P_{\mathcal{H}^c})^l P_{\mathcal{H}^c, \mathcal{H}} = (I_{\mathcal{H}^c} - P_{\mathcal{H}^c})^{-1} P_{\mathcal{H}^c, \mathcal{H}}. \quad (4.25)$$

Similarly, for the new external arrival rate $\hat{\lambda}_{0, \mathcal{H}}$, we write

$$\hat{\lambda}_{0, \mathcal{H}} = \lambda_{0, \mathcal{H}} + \hat{P}'_{\mathcal{H}^c, \mathcal{H}} \lambda_{0, \mathcal{H}^c} = \lambda_{0, \mathcal{H}} + P'_{\mathcal{H}^c, \mathcal{H}} (I_{\mathcal{H}^c} - P'_{\mathcal{H}^c})^{-1} \lambda_{0, \mathcal{H}^c}, \quad (4.26)$$

where $\lambda_{0, \mathcal{A}}$ denotes the column vector of the entries in λ_0 that corresponds to the index set $\mathcal{A}$. Since the total arrival rate in the modified system remains the same as the original

system, we have

$$\hat{\lambda}_{\mathcal{H}} = (I - \hat{P}'_{\mathcal{H}})^{-1} \hat{\lambda}_{0,\mathcal{H}} = \lambda_{\mathcal{H}}. \quad (4.27)$$

To simplify notation, we suppress the subscript used in the identity matrix I in the rest of the paper whenever there is no confusion on its dimension. ■

The following theorem states the joint heavy-traffic limit of the queue length process, the workload processes, the waiting time processes, the splitting-decision process and all the flows. As in [34; 35], we allow an arbitrary subset of nodes to be bottleneck queues (critically loaded) while the rest are sub-critically loaded. To treat the stationary processes, we apply [66] and [31], extended to include non-bottleneck queues. Because our basic model data involves only single arrival and service processes, with only the parameters being scaled, we do not need Assumption (A4) in [31]. The proof can be found in Section 4.6.

Theorem 4.5 (Heavy-traffic FCLT) *Under Assumption 4.1-4.2, consider a family of open queueing networks in stationarity, indexed by ρ . Let $\mathcal{H} \subset \{1, 2, \dots, K\}$ denote the index of the bottleneck stations: Assume that $\mu_{i,\rho} = \lambda_i/(\kappa_i \rho)$ for $1 \leq i \leq K$ and set $\kappa_i = 1$ for all $i \in \mathcal{H}$ and $\kappa_i < 1$ for all $i \notin \mathcal{H}$. Then, as $\rho \uparrow 1$,*

$$\begin{aligned} & (Q_{\rho}^*, W_{\rho}^*, Z_{\rho}^*, \Theta_{\rho}^*, \Theta_{\text{ext},\rho}^*, \mathcal{F}_{\rho}^*) \\ & \Rightarrow (Q^*, W^*, Z^*, \Theta^*, \Theta_{\text{ext}}^*, \mathcal{F}^*) \quad \text{in } \mathcal{D}^{9K+2K^2}, \end{aligned} \quad (4.28)$$

where:

1. For $0 \leq i \leq K$, $A_{0,i}^* = c_{a0,i} B_{a0,i} \circ \lambda_{0,i} e$ and $S_i^* = c_{s,i} B_{s,i} \circ \lambda_i e$, where $B_{a0,i}$ and $B_{s,i}$ are standard Brownian motions. $(\Theta_{i,j}^* : 0 \leq j \leq K)$ is a zero-drift $(K+1)$ -dimensional Brownian motion with covariance matrix $\Sigma_i = (\sigma_{j,k}^2 : 0 \leq j, k \leq K)$, where $\sigma_{j,j}^2 = p_{i,j}(1 - p_{i,j})\lambda_i$ and $\sigma_{j,k}^2 = -p_{i,j}p_{i,k}\lambda_i$ for $0 \leq i \neq j \leq K$. Furthermore, $B_{a0,i}$, $B_{s,i}$ and $(\Theta_{i,j}^* : 0 \leq j \leq K)$ are mutually independent, $1 \leq i \leq K$.
2. The queue length process Q^* consists of two parts. $Q_{\mathcal{H}^c}^* \equiv 0$ and $Q_{\mathcal{H}}^*$ is a stationary $|\mathcal{H}|$ -dimensional RBM

$$Q_{\mathcal{H}}^* \equiv \Psi_{\mathcal{H}} \left(\hat{X}_{\mathcal{H}}^* \right),$$

where $\Psi_{\mathcal{H}}$ is the $|\mathcal{H}|$ -dimensional reflection map with reflection matrix $R_{\mathcal{H}} \equiv I_{\mathcal{H}} - \hat{P}_{\mathcal{H}}$ and $\hat{X}_{\mathcal{H}}^*$ is the net-input process associated with the bottleneck queues, defined below. Furthermore, $Q_{\mathcal{H}}^*(0)$ has unique stationary distribution of the stationary RBM. $\hat{X}_{\mathcal{H}}^*$ is a $|\mathcal{H}|$ -dimensional Brownian motion

$$\hat{X}_{\mathcal{H}}^* = Q_{\mathcal{H}}^*(0) + \left(e'_{\mathcal{H}} + \hat{P}'_{\mathcal{H}^c, \mathcal{H}} e'_{\mathcal{H}^c} \right) (A_0^* + (\Theta^*)' \mathbf{1}) - (I_{\mathcal{H}} - \hat{P}_{\mathcal{H}}) S_{\mathcal{H}}^* - \hat{\lambda}_{0, \mathcal{H}} e, \quad (4.29)$$

where $e_{\mathcal{A}}$ collects columns in the K -dimensional identity matrix I that corresponds to index set $\mathcal{A}$; $\hat{P}_{\mathcal{H}}$, $\hat{P}_{\mathcal{H}^c, \mathcal{H}}$ and $\hat{\lambda}_{0, \mathcal{H}}$ are defined in (4.24), (4.25) and (4.26), respectively.

3. The total arrival process A^* can be regarded as a stationary process, having stationary increments, specified by

$$\begin{aligned} A^* &= (I - P')^{-1} (A_0^* + (\Theta^*)' \mathbf{1}) + P'(I - P')^{-1} (Q^*(0) - Q^*) \\ &= (I - P')^{-1} (A_0^* + (\Theta^*)' \mathbf{1}) + P'(I - P')^{-1} e_{\mathcal{H}} (Q_{\mathcal{H}}^*(0) - Q_{\mathcal{H}}^*). \end{aligned}$$

4. The stationary departure process D^* is specified as

$$D^* = Q^*(0) + A^* - Q^* = (I - P')^{-1} (Q^*(0) - Q^* + A_0^* + (\Theta^*)' \mathbf{1}) \quad (4.30)$$

In particular,

$$D_{\mathcal{H}^c}^* = Q_{\mathcal{H}^c}^* + A_{\mathcal{H}^c}^* - Q_{\mathcal{H}^c}^*(0) = A_{\mathcal{H}^c}^*.$$

5. The internal arrival flow $A_{i,j}^*$ can be expressed as

$$A_{i,j}^* = p_{i,j} D_i^* + \Theta_{i,j}^* \circ \lambda_i e, \quad \text{for } 1 \leq i, j \leq K$$

and the external departure flow can be expressed as

$$D_{\text{ext}, i}^* = p_{i,0} D_i^* + \Theta_{i,0}^* \circ \lambda_i e, \quad \text{for } 1 \leq i \leq K.$$

6. $Z_i^* = \lambda_i^{-1} Q_i^*$ and $W_i^* = Z_i^* \circ \lambda_i e$.

4.4 Heavy-Traffic Limits with One Bottleneck Queue

In this section we consider the special case in which there is only one bottleneck queue, which is useful for the IDC approximation and the RQNA applications because it is especially tractable, involving one-dimensional RBM instead of multi-dimensional RBM.

We start with the easiest special case: when $|\mathcal{H}| = K = 1$, which corresponds to the $GI/GI/1$ queue with i.i.d. customer feedback. But then we observe that the case of a single bottleneck is asymptotically equivalent to that except that the arrival process is generalized to include the immediate feedback associated with flows to all the other non-bottleneck queues.

As a consequence, we show that it is asymptotically correct in HT for a GJN with a single bottleneck queue to eliminate all feedback prior to analysis. Moreover, we show how to quantify feedback elimination.

4.4.1 A Single-Server Queue with Customer Feedback

Consider a single-server queue with customer feedback as depicted in Figure 4.1. Let A_0 denote the renewal external arrival process with rate λ_0 and scv $c_{a_0}^2$. Let the feedback probability be p , so that the effective arrival rate is $\lambda = \lambda_0/(1-p)$. Let service times be i.i.d. with rate $\mu_\rho = \lambda/\rho$ and scv c_s^2 , hence a traffic intensity of ρ . Let A denote the total arrival process; let A_{int} be the feedback flow; let S denote the service process; let D be the total departure process; and let D_{ext} denote the flow that exits the system.

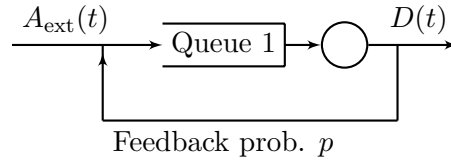


Figure 4.1: A single-server queue with customer feedback.

Corollary 4.2 (One $GI/GI/1$ queue with feedback) *Under Assumptions in Theorem 4.5, consider a family of single-server queues in stationarity, indexed by ρ . Assume that $\mu_\rho = \lambda/\rho$. Then, as $\rho \uparrow 1$,*

$$(Q_\rho^*, W_\rho^*, Z_\rho^*, \Theta_\rho^*, \Theta_{\text{ext},\rho}^*, \mathcal{F}_\rho^*) \Rightarrow (Q^*, W^*, Z^*, \Theta^*, \Theta_{\text{ext}}^*, \mathcal{F}^*) \quad \text{in } \mathcal{D}^{11},$$

where $\mathcal{F}_\rho^* = (A_{0,\rho}^*, A_\rho^*, A_{\text{int},\rho}^*, S_\rho^*, D_\rho^*, D_{\text{ext},\rho}^*)$, $\mathcal{F}^* = (A_0^*, A^*, A_{\text{int}}^*, S^*, D^*, D_{\text{ext}}^*)$ and:

1. $A_0^* = c_{a_0} B_{a_0} \circ \lambda_0 e$ and $S^* = c_s B_s \circ \lambda e$, where B_{a_0} and B_s are standard Brownian motions. $(\Theta^*, \Theta_{\text{ext}}^*)$ is a zero-drift two-dimensional Brownian motion with covariance

matrix $\Sigma = (\sigma_{i,j}^2 : 1 \leq i, j \leq 2)$, where $\sigma_{1,1}^2 = \sigma_{2,2}^2 = p(1-p)\lambda$ and $\sigma_{1,2}^2 = \sigma_{2,1}^2 = -p(1-p)\lambda$, so that

$$\Theta^* + \Theta_{\text{ext}}^* = 0.$$

Furthermore, B_{a_0} , B_s and $(\Theta^*, \Theta_{\text{ext}}^*)$ are mutually independent.

2. The queue length process Q^* is a stationary one-dimensional RBM

$$Q^* \equiv \Psi(X^*),$$

where Ψ is the one-dimensional reflection map and X^* is a one-dimensional Brownian motion

$$X^* = Q^*(0) + A_0^* + (\Theta^* - (1-p)S^*) - \lambda_0 e,$$

where $\lambda_0 = (1-p)\lambda$. Furthermore, $Q^*(0)$ has unique stationary distribution of the stationary one-dimensional RBM with drift $-\lambda_0$ and variance

$$\lambda_0 c_x^2 \equiv \lambda_0 (c_a^2 + p + (1-p)c_s^2),$$

so an exponential distribution with mean $c_x^2/2$.

3. The total arrival process A^* can be regarded as a stationary process, having stationary increments, specified by

$$A^* = \frac{1}{1-p} (A_0^* + \Theta^*) + \frac{p}{1-p} (Q^*(0) - Q^*).$$

4. The stationary total departure process D^* is specified as

$$\begin{aligned} D^* &= Q^*(0) + A^* - Q^* \\ &= \frac{1}{1-p} (Q^*(0) - Q^* + A_0^* + (\Theta^*)' \mathbf{1}) \end{aligned} \tag{4.31}$$

5. The internal arrival flow A_{int}^* can be expressed as

$$A_{\text{int}}^* = pD^* + \Theta^*$$

and the external departure flow can be expressed as

$$D_{\text{ext}}^* = (1-p)D^* + \Theta_{\text{ext}}^* = A_0^* + Q^*(0) - Q^*.$$

6. $Z^* = \lambda^{-1}Q^*$ and $W^* = Z^* \circ \lambda e$.

As observed in Section III of [134], to develop effective parametric-decomposition approximations for OQNs it is often helpful to preprocess the model data by eliminating immediate feedback for queues with feedback. The immediate feedback returns the customer to the end of the line. The approximation step is to put the customer instead back at the head of the line, so as to receive all its (geometrically random number of) service times at once. Clearly this does not alter the queue length process and the workload process.

The modified system does not have a feedback flow and the new service time will be the geometric random sum of the i.i.d. copies of the original service times, let $\tilde{S}$ denote the new service counting process.

This modification results in a change in the service rate and service scv. The new service rate is $(1-p)\mu = (1-p)\lambda/\rho = \lambda_0/\rho$ and, by conditional variance formula, the new scv is $\tilde{c}_s^2 = p + (1-p)c_s^2$. Hence, the heavy-traffic limit of the new service process is $\tilde{S}^* \equiv \tilde{c}_s^2 \tilde{B}_s \circ \lambda_0 e$. We now claim that $\tilde{S}^* \stackrel{dist.}{=} \Theta^* - (1-p)S^*$. To this end, note that $\Theta^* = \sqrt{p(1-p)}B_\Theta \circ \lambda e$ and $S^* = c_s B_s \circ \lambda e$, where B_Θ, B_s are independent standard Brownian motions (zero drift and unit variance) and $\lambda_0 = (1-p)\lambda$. Hence, from part (ii) of Corollary 4.2, we have

$$X^* \stackrel{dist.}{=} Q^*(0) + A_0^* + \tilde{S}^* - \lambda_0 e. \quad (4.32)$$

Let $\tilde{Q}^*, \tilde{Z}^*, \tilde{W}^*$ denote the HT limit of the queue length process, the workload process and the waiting time process in the modified single-server queue without feedback, having arrival process A_0 and service process $\tilde{S}$. Standard heavy-traffic theory implies that (4.32) is exactly the HT limit of the net-input process of a single-server queue so that $\tilde{Q}^* \stackrel{dist.}{=} Q^*$. Hence, we have

$$\begin{aligned} \tilde{Z}^* &\equiv \lambda_0^{-1} \tilde{Q}^* \stackrel{dist.}{=} (1-p)^{-1} \lambda^{-1} Q^* \equiv (1-p)^{-1} Z^*, \quad \text{and} \\ \tilde{W}^* &\equiv \tilde{Z}^* \circ \lambda_0 e \stackrel{dist.}{=} (1-p)^{-1} Z^* \circ \lambda_0 e \equiv (1-p)^{-1} W^* \circ (1-p)e. \end{aligned}$$

Note that the expected number of visit for the same customer is $(1-p)^{-1}$. This implies that for approximating the waiting time and workload in the original system, we need to adjust for per-visit version by multiplying the values in the modified system by $(1-p)$.

Theorem 4.6 (Eliminating immediate feedback) *For the single-server queue with feedback model in Corollary 4.2, consider the modified single-server queue, where immediate feedback are eliminated by placing the feedback customers at the head of the line. The joint heavy-traffic limit for the queue length process, the waiting time process, the workload process and the external departure process in the original model can be expressed in terms of those in the modified system as*

$$(Q^*, Z^*, W^*, D_{\text{ext}}^*) \stackrel{\text{dist.}}{=} (\tilde{Q}^*, (1-p)\tilde{Z}^*, (1-p)\tilde{W}^* \circ (1-p)^{-1}e, \tilde{D}_{\text{ext}}^*).$$

4.4.2 Networks with One Bottleneck Queue

We now consider the more general special case in which $K \geq 1$ but $|\mathcal{H}| = 1$. Without loss of generality, let $\mathcal{H} = \{h\}$, so that station h is the only bottleneck station. Then Theorem 4.5 can be restated as

Corollary 4.3 (Network with one bottleneck queue) *Under Assumption 4.1-4.2, consider a series of GJNs in stationarity, indexed by ρ . Assume that $\mu_{i,\rho} = \lambda_i/(\kappa_i\rho)$ for $1 \leq i \leq K$ and set $c_h = 1$ and $\kappa_i < 1$ for all $i \neq h$. Then, we have*

$$(Q_\rho^*, W_\rho^*, Z_\rho^*, \Theta_\rho^*, \Theta_{\text{ext},\rho}^*, \mathcal{F}_\rho^*) \Rightarrow (Q^*, W^*, Z^*, \Theta^*, \Theta_{\text{ext}}^*, \mathcal{F}^*)$$

as $\rho \uparrow 1$ in $\mathcal{D}^{9K+2K^2}$, where:

1. For $0 \leq i \leq K$, $A_{0,i}^* = c_{a_{0,i}}B_{a_{0,i}} \circ \lambda_{0,i}e$ and $S_i^* = c_{s_i}B_{s_i} \circ \lambda_i e$, where $B_{a_{0,i}}$ and B_{s_i} are standard Brownian motions. $(\Theta_{i,j}^* : 0 \leq j \leq K)$ is a zero-drift $(K+1)$ -dimensional Brownian motion with covariance matrix $\Sigma_i = (\sigma_{j,k}^2 : 0 \leq j, k \leq K)$, where $\sigma_{j,j}^2 = p_{i,j}(1-p_{i,j})\lambda_i$ and $\sigma_{j,k}^2 = -p_{i,j}p_{i,k}\lambda_i$ for $0 \leq i \neq j \leq K$. Furthermore, $B_{a_{0,i}}$, B_{s_i} and $(\Theta_{i,j}^* : 0 \leq j \leq K)$ are mutually independent, $1 \leq i \leq K$.
2. The queue length process Q^* consists of two parts. $Q_i^* \equiv 0$ for $i \neq h$ and Q_h^* is a stationary one-dimensional RBM

$$Q_h^* \equiv \Psi(\hat{X}_h^*),$$

where Ψ is the one-dimensional refelction map and $\hat{X}_h^*$ is the net-input process defined as

$$\hat{X}_h^* = Q_h^*(0) + \left(e'_h + \hat{P}'_{\mathcal{H}^c,h}e'_{\mathcal{H}^c}\right)(A_0^* + (\Theta^*)'\mathbf{1}) - (1 - \hat{P}_h)S_h^* - \hat{\lambda}_{0,h}e, \quad (4.33)$$

where $e_{\mathcal{A}}$ collects columns in the K -dimensional identity matrix I that corresponds to index set $\mathcal{A}$; $\hat{P}_h$, $\hat{P}_{\mathcal{H}^c, h}$ and $\hat{\lambda}_{0, h}$ are defined in (4.24), (4.25) and (4.26) with $\mathcal{H} = \{h\}$, respectively. Furthermore, $Q_h^*(0)$ has unique stationary distribution of the stationary RBM.

3. The total arrival process A^* can be regarded as a stationary process, having stationary increments, specified by

$$A^* = (I - P')^{-1} (A_0^* + (\Theta^*)' \mathbf{1}) + P'(I - P')^{-1} e_h (Q_h^*(0) - Q_h^*).$$

4. The stationary departure process D^* is specified as

$$D^* = Q^*(0) + A^* - Q^* = (I - P')^{-1} (Q^*(0) - Q^* + A_0^* + (\Theta^*)' \mathbf{1}).$$

In particular,

$$D_{\mathcal{H}^c}^* = Q_{\mathcal{H}^c}^* + A_{\mathcal{H}^c}^* - Q_{\mathcal{H}^c}^*(0) = A_{\mathcal{H}^c}^*.$$

5. The internal arrival flow $A_{i,j}^*$ can be expressed as

$$A_{i,j}^* = p_{i,j} D_i^* + \Theta_{i,j}^* \circ \lambda_i e, \quad \text{for } 1 \leq i, j \leq K$$

and the external departure flow can be expressed as

$$D_{\text{ext}, i}^* = p_{i,0} D_i^* + \Theta_{i,0}^* \circ \lambda_i e, \quad \text{for } 1 \leq i \leq K.$$

6. $Z_i^* = \lambda_i^{-1} Q_i^*$ and $W_i^* = Z_i^* \circ \lambda_i e$.

We conclude this section by observing that in a GJN with one bottleneck queue that the bottleneck queue is asymptotically equivalent to a $G/GI/1$ single-server queue with feedback in the HT limit, where the arrival process is a complex superposition of renewal arrival processes. We derive the explicit expression for the external arrival process and feedback probability in the equivalent network. We also show that feedback elimination is asymptotically correct for networks with one bottleneck.

We start with a convenient representation of the HT limit of the bottleneck queue. Let $\hat{p}_{i,h}$ be the (i, h) -th component of $\hat{P}_{\mathcal{H}^c, \mathcal{H}}$ in (4.25) and recall that $\hat{p} \equiv \hat{P}_h$ is the feedback probability defined in Remark 4.2.

Theorem 4.7 *The HT limit $\hat{X}_h^*$ in (4.33) can be expressed as the following one-dimensional Brownian motion*

$$\hat{X}_h^* = Q_h^*(0) + \hat{A}^* + \left(\hat{\Theta}_S^* - (1 - \hat{p})S_h^* \right) + \hat{\lambda}_{0,h}e, \quad (4.34)$$

where

$$\hat{A}^* = A_{0,h}^* + \sum_{i \in \mathcal{H}^c} \left(\hat{p}_{i,h} A_{0,i}^* + \hat{\Theta}_{i,h}^* \right), \quad (4.35)$$

and

$$\begin{aligned} \hat{\Theta}_{i,h}^* &= \sqrt{\hat{p}_{i,h}(1 - \hat{p}_{i,h})} B_{\hat{\Theta}_{i,h}} \circ \lambda_{0,i}e, \\ \hat{\Theta}_S^* &= \sqrt{\hat{p}(1 - \hat{p})} B_{\hat{\Theta}_S} \circ \lambda_i e, \end{aligned} \quad (4.36)$$

while $B_{\hat{\Theta}_{i,h}}$ and $B_{\hat{\Theta}_S}$ are independent standard Brownian motions.

Proof Since the drift term, the terms associated with A_0^* and S_h^* remain unchanged, it suffices to show that the terms related with the splitting decision processes share the same variance. In fact, by algebraic manipulation, one can check that

$$\begin{aligned} \text{Var} \left(\sum_{i \in \mathcal{H}^c} \hat{\Theta}_{i,h}^* + \hat{\Theta}_S^* \right) &= \sum_{i \in \mathcal{H}^c} \hat{p}_{i,h}(1 - \hat{p}_{i,h}) \lambda_{0,i}e + \hat{p}(1 - \hat{p}) \lambda_i e \\ &= \sum_{i=1}^K \left(e'_h + \hat{P}'_{\mathcal{H}^c,h} e'_{\mathcal{H}^c} \right) \Sigma_i \left(e_h + e_{\mathcal{H}^c} \hat{P}_{\mathcal{H}^c,h} \right) e \\ &= \text{Var} \left(e'_h (\Theta^*)' \mathbf{1} + \hat{P}'_{\mathcal{H}^c,h} e'_{\mathcal{H}^c} (\Theta^*)' \mathbf{1} \right) \end{aligned}$$

where Σ_i are the variance matrix defined in Theorem 4.5. ■

Now, consider a reduced one-station network consist of the only bottleneck queue, while all non-bottleneck queues have service times set to 0 so that they serve as instantaneous switches. In the reduced network, we define an external arrival $\hat{A}_0$ to the bottleneck queue to be any external arrival that arrive at the bottleneck queue for the first time. Hence, an external arrival may have visited one or multiple non-bottleneck queues before its first visit to the bottleneck queue. In particular, the external arrival process can be expressed as the superposition of (i) the original external arrival process $A_{0,h}$ at station h ; and (ii) the Markov splitting of the external arrival process $A_{0,i}$ at station i with probability $\hat{p}_{i,h}$, for $i \in \mathcal{H}^c$.

Theorem 4.7 implies that the reduced network is asymptotically equivalent to the original bottleneck queue in the sense of the stationary queue length process in the HT limit. Furthermore, comparing Theorem 4.7 with Corollary 4.2, we conclude that both the reduced network and the original bottleneck queue is asymptotically equivalent to a single-server queue with feedback, where the external arrival process is $\hat{A}$, the service times remain unchanged and the feedback probability is $\hat{p}$.

We then eliminate immediate feedback customers just as in Theorem 4.6, but with the extended interpretation of immediate feedback. Recalling that the non-bottleneck queues act as instantaneous switches, we recognize all customers that feed back to the bottleneck queue as immediate feedback, even after visiting non-bottleneck queues. The probability of feedback is then exactly $\hat{p} \equiv \hat{P}_h$ as in Remark 4.2. After feedback elimination, the new service process $\hat{S}$ is the renewal process associated with the new service times, i.e., a geometric sum of the original service times at the bottleneck queue. Note that the modified service process after feedback elimination have a HT limit $\hat{S}^* \equiv \hat{\Theta}_S^* - (1 - \hat{p})S_h^*$, where Θ_S^* is defined in (4.36), just as discussed in Section 4.4. This matches exactly with the “service” part in (4.34) of Theorem 4.7. Hence, we have the following theorem, extending Theorem 4.6.

Theorem 4.8 (Feedback elimination with one bottleneck queue) *For the bottleneck queue in the generalized Jackson network, consider the modified single-server queue with arrival process $\hat{A}$ and service process $\hat{S}$. The joint heavy-traffic limit for the queue length process, the waiting time process, the workload process and the external departure process in the original model can be expressed in terms of those in the modified system as*

$$(Q^*, Z^*, W^*, D_{\text{ext}}^*) \stackrel{\text{dist.}}{=} (\hat{Q}^*, (1 - p)\hat{Z}^*, (1 - p)\hat{W}^* \circ (1 - p)^{-1}e, \hat{D}_{\text{ext}}^*).$$

4.5 Functional Central Limit Theorem for the Stationary Flows

In this section, we focus on yet another important special case of Theorem 4.5 where we set $|\mathcal{H}| = 0$. In this special case, all stations are strictly non-bottleneck, i.e., $\mu_{i,\rho} = \lambda/(\kappa_i\rho)$ where $\kappa_i < 1$ for all i . As $\rho \uparrow 1$, the family of systems converges to a limiting system where

the traffic intensity at station i is $\rho_i = \kappa_i$. Hence, the scaling used in (4.22) corresponds to the diffusion scaling used in the usual FCLT. The following corollary describes the joint FCLT of the stationary flows.

Corollary 4.4 (FCLT for the stationary flows) *Under Assumption 4.1-4.2, consider a family of open queueing networks in stationarity, indexed by ρ . Assume that $\mu_{i,\rho} = \lambda_i/(\kappa_i\rho)$ with $\kappa_i < 1$ for $1 \leq i \leq K$. Then, as $\rho \uparrow 1$,*

$$\begin{aligned} (Q_\rho^*, W_\rho^*, Z_\rho^*, \Theta_\rho^*, \Theta_{\text{ext},\rho}^*, \mathcal{F}_\rho^*) \\ \Rightarrow (Q^*, W^*, Z^*, \Theta^*, \Theta_{\text{ext}}^*, \mathcal{F}^*) \quad \text{in } \mathcal{D}^{9K+2K^2}, \end{aligned} \quad (4.37)$$

where:

1. For $0 \leq i \leq K$, $A_{0,i}^* = c_{a_{0,i}} B_{a_{0,i}} \circ \lambda_{0,i} e$ and $S_i^* = c_{s_i} B_{s_i} \circ \lambda_i e$, where $B_{a_{0,i}}$ and B_{s_i} are standard Brownian motions. $(\Theta_{i,j}^* : 0 \leq j \leq K)$ is a zero-drift $(K+1)$ -dimensional Brownian motion with covariance matrix $\Sigma_i = (\sigma_{j,k}^2 : 0 \leq j, k \leq K)$, where $\sigma_{j,j}^2 = p_{i,j}(1-p_{i,j})\lambda_i$ and $\sigma_{j,k}^2 = -p_{i,j}p_{i,k}\lambda_i$ for $0 \leq i \neq j \leq K$. Furthermore, $B_{a_{0,i}}$, B_{s_i} and $(\Theta_{i,j}^* : 0 \leq j \leq K)$ are mutually independent, $1 \leq i \leq K$.
2. The queue length process $Q^* \equiv 0$.
3. The total arrival process A^* can be regarded as a stationary process, having stationary increments, specified by

$$A^* = (I - P')^{-1} (A_0^* + (\Theta^*)' \mathbf{1}).$$

4. The stationary departure process is the same as the stationary total arrival process, so that $D^* = A^*$.
5. The internal arrival flow $A_{i,j}^*$ can be expressed as

$$A_{i,j}^* = p_{i,j} D_i^* + \Theta_{i,j}^* \circ \lambda_i e, \quad \text{for } 1 \leq i, j \leq K$$

and the external departure flow can be expressed as

$$D_{\text{ext},i}^* = p_{i,0} D_i^* + \Theta_{i,0}^* \circ \lambda_i e, \quad \text{for } 1 \leq i \leq K.$$

6. Finally, $Z_i^* = W_i^* = 0$.

4.6 Proofs

Proof of Theorem 4.3

Proof By Corollary 4.1 and the definition of S_s in (4.12), the convergence of $V_s(0) = V(s)$ implies the convergence of S_s to S_e , with the later one being a delayed renewal process with first interval distributed as $V_e(0)$ and other intervals distributed as a generic service time. Similarly, the components of $A_{0,s}$ are delayed renewal process with the first interval distributed as the components of $U_s(0)$, which is converging to the vector $A_{0,e}$ of the equilibrium external arrival processes. By the convergence of $\mathcal{S}_s$, we have as $s \rightarrow \infty$

$$(Q_s, U_s, V_s, A_{0,s}, S_s) \Rightarrow (Q_e, U_e, V_e, A_{0,e}, S_e). \quad (4.38)$$

We now turn our focus to the cumulative busy time process defined in (4.4). Let $h : \mathbb{R}_+ \rightarrow \mathbb{R}$ be a continuous function defined by $h(t) = t \wedge 1 \equiv \min\{t, 1\}, t \geq 0$. Then the busy period process can be written as

$$B_{i,s}(t) = \int_s^{s+t} 1_{Q_i(u) > 0} du = \int_0^t 1_{Q_{i,s}(u) > 0} du = \int_0^t h(Q_{i,s}(u)) du, \quad \text{for } 1 \leq i \leq K. \quad (4.39)$$

The busy-period process thus has stationary increments because it is a measurable integrable function of $Q_{i,e}$, which is itself stationary. (Recall that general measurable functions of stationary process are stationary; see Proposition 6.6 of [28].) Let $\mathcal{C}(\mathbb{R}_+, \mathbb{R})$ denote the space of bounded continuous functions from $\mathbb{R}_+$ to $\mathbb{R}$, equipped with uniform norm. The mapping defined in (4.39) is a continuous mapping from $\mathcal{D}$ to $\mathcal{C}(\mathbb{R}_+, \mathbb{R})$; see Theorem 11.5.1 in [143]. The continuous mapping theorem then asserts that $B_s \Rightarrow B_e$, where $B_{i,e}(t) \equiv \int_0^t h(Q_{i,e}(u)) du$ for $t \geq 0$ and all i . For the cumulative idle-time process $Y_{i,s}(t) \equiv Y_i(t+s) - Y_i(s) = \mu_i(t - B_{i,s}(t))$, we note that t and $B_{i,s}(t)$ have continuous sample path, so that the linear function in (4.9) is continuous. Hence, we can extend the convergence as $s \rightarrow \infty$ in (4.38) to

$$(Q_s, U_s, V_s, A_{0,s}, S_s, B_s, Y_s) \Rightarrow (Q_e, U_e, V_e, A_{0,e}, S_e, B_e, Y_e). \quad (4.40)$$

The convergence established so far now implies associated convergence for the flows because the flow process $\mathcal{F}_s$ is determined by the state process $\mathcal{S}_s$. To make the connection, we introduce random vectors (T_s, J_s) , where T_s is the time of the first jump in Q_s and J_s

is the type of jump (external arrival to queue i , flow from queue i to queue j , or external departure from queue i), defined by

$$\begin{aligned} T_s &\equiv \min \{T_s^a, T_s^d\}, \quad \text{where} \\ T_s^a &\equiv \min \{U_{s,i}(0) : 1 \leq i \leq K\} \quad \text{and} \\ T_s^d &\equiv \min \{V_{s,i}(0) : Q_i(0) > 0, 1 \leq i \leq K\}. \end{aligned} \quad (4.41)$$

while $J_s = (0, i)$, (i, j) or $(i, 0)$ if the minimum in the definition of T_s is attained, respectively, by T_s^a with index i , T_s^d with index i and the routing is to j , T_s^d with index i and the routing is to outside the network.

We observe that we can regard $(T, J) : (s, \mathcal{S}_s) \rightarrow \mathbb{R} \times \mathcal{N}$, where $\mathcal{N}$ is a finite set, as a continuous map, so that $(T_s, J_s) \Rightarrow (T_e, J_e)$ as $s \rightarrow \infty$. We also observe that T_s is a stopping time with respect to the strong Markov process $\{\mathcal{S}_s(t) : t \geq 0\}$, so that we can repeat the construction for all successive jumps after time T_s .

In this way, we get convergence of the process of successive jump times and jump types (indexed by k)

$$\{(T_{s,k}, J_{s,k}) : k \geq 1\} \Rightarrow \{(T_{e,k}, J_{e,k}) : k \geq 1\} \quad \text{in } (\mathbb{R} \times \mathcal{N})^\infty \quad \text{as } s \rightarrow \infty. \quad (4.42)$$

That in turn implies convergence for the associated flow counting processes by applying the inverse map in §13.6 of [143] as stated. For example, we can write

$$\begin{aligned} N_s(t) &\equiv \min \{k \geq 0 : T_{s,1} + \cdots + T_{s,k} \leq t\} \quad \text{and} \\ A_{s,i,j}(t) &= \sum_{k=1}^{N_s(t)} 1_{J_{s,k}=(i,j)}. \quad \blacksquare \end{aligned}$$

Proof of Theorem 4.4. With the standard flow conservation law, we can write the queue length process in terms of the centered processes

$$\begin{aligned}
Q_i &= Q_i(0) + A_i - S_i \circ B_i \\
&= Q_i(0) + A_{0i} + \sum_{j=1}^K \Theta_{ji}(S_j \circ B_j) - S_i \circ B_i \\
&= Q_i(0) + (A_{0i} - \lambda_{0i}e) + \sum_{j=1}^K (\Theta_{ji}(S_j \circ B_j) - p_{ji}S_j \circ B_j) \\
&\quad - \sum_{j=1}^K (\delta_{ji} - p_{ji})(S_j \circ B_j - \mu_j B_j) + \sum_{j=1}^K (\delta_{ji} - p_{ji})\mu_j(e - B_j) \\
&\quad + \lambda_{0i}e - \sum_{j=1}^K (\delta_{ji} - p_{ji})\mu_j e.
\end{aligned}$$

Because $Y_i \equiv \mu_i(t - B_i)$ is the cumulative idle time, we can express Q in matrix form as

$$Q = Q(0) + A_0 + \tilde{\Theta}'\mathbf{1} - (I - P')\tilde{S} + (I - P')Y + (\lambda_0 - (I - P')\mu)e.$$

Furthermore, we have $Q = X + (I - P')Y$. Because Y is non-decreasing, $Y(0) = 0$ and Y_i increases only when $Q_i = 0$, (4.18) follows from the usual reflection argument.

Similarly, we can re-write the overall arrival process in terms of the centered processes

$$\begin{aligned}
A_i &= A_{0i} + \sum_{j=1}^K \Theta_{ji}(S_j \circ B_j) \\
&= (A_{0i} - \lambda_{0i}e) + \sum_{j=1}^K (\Theta_{ji}(S_j \circ B_j) - p_{ji}S_j \circ B_j) + \sum_{j=1}^K p_{ji}(S_j \circ B_j - \mu_j B_j) \\
&\quad - \sum_{j=1}^K p_{ji}\mu_j(e - B_j) + \lambda_{0i}e + \sum_{j=1}^K p_{ji}\mu_j e
\end{aligned}$$

or, in matrix notation, by

$$A = \tilde{A}_0 + \tilde{\Theta}'\mathbf{1} + P'\tilde{S} - P'Y + (\lambda_0 + P'\mu)e.$$

By (4.18), we have

$$\begin{aligned}
-P'Y &= P'(I - P')^{-1}(X - Q) \\
&= P'(I - P')^{-1}\left(Q(0) - Q + \tilde{A}_0 + \tilde{\Theta}'\mathbf{1} + \lambda_0 e\right) - P'\tilde{S} - P'\mu e.
\end{aligned}$$

Substituting into the matrix form of the arrival process, we have

$$\begin{aligned}
A &= \tilde{A}_0 + \tilde{\Theta}' \mathbf{1} + P' \tilde{S} - P' Y + (\lambda_0 + P' \mu) e \\
&= \tilde{A}_0 + \tilde{\Theta}' \mathbf{1} + P' \tilde{S} + (\lambda_0 + P' \mu) e \\
&\quad + P' (I - P')^{-1} \left(Q(0) - Q + \tilde{A}_0 + \tilde{\Theta}' \mathbf{1} + \lambda_0 e \right) - P' \tilde{S} - P' \mu e \\
&= P' (I - P')^{-1} (Q(0) - Q) + (I - P')^{-1} \left(\tilde{A}_0 + \tilde{\Theta}' \mathbf{1} \right) + \lambda e.
\end{aligned} \tag{4.43}$$

Finally, note that $D = Q(0) + A - Q$. ■

Proof of Theorem 4.5 Much of the statement follows from [34; 35] and [31]. First, the HT limit for the state process with an arbitrary subset $\mathcal{H}$ of critically loaded stations follows from [34; 35]. Second, the HT limit for the steady-state queue length follows from [31]. The papers [66] and [31] do not consider non-bottleneck stations, but their arguments extend to that more general setting. (See Remark 4.3 below for discussion.) We subsequently establish the heavy-traffic limits for the flows. We do so by exploiting the continuous mapping theorem with the direct representations of the stationary flows that we have established.

To carry out our proof, we work with the centered representation in Theorem 4.4, using the HT-scaling in (4.22). Thus, the HT-scaled net-input process is

$$X_\rho^* = Q_\rho^*(0) + A_{0,\rho}^* + \left(\tilde{\Theta}_\rho^* \right)' \mathbf{1} - (I - P') \tilde{S}_\rho^* + (\lambda_0 - (I - P') \mu_\rho) (1 - \rho)^{-1} e, \tag{4.44}$$

where $\tilde{S}_{i,\rho}^* \equiv S_{i,\rho}^* \circ \bar{\bar{B}}_{i,\rho}$, $\bar{\bar{B}}_{i,\rho} = (1 - \rho)^2 B_{i,\rho} \circ (1 - \rho)^{-2} e$, $\tilde{\Theta}_\rho^*$ is a matrix with its ij -th entry being $\Theta_{ij,\rho}^* \circ \bar{\bar{S}} \circ \bar{\bar{B}}_{i,\rho}$ and $\bar{\bar{S}} \circ \bar{\bar{B}}_\rho$ is a vector of length K with $\bar{\bar{S}} \circ \bar{\bar{B}}_{i,\rho} \equiv (1 - \rho)^2 S_{i,\rho} \circ B_{i,\rho} \circ (1 - \rho)^{-2} e$.

The HT-scaled queue length can be written as

$$Q_\rho^* = X_\rho^* + (I - P') Y_\rho^*.$$

We now re-write $Q_{\mathcal{H},\rho}^*$ and $Q_{\mathcal{H}^c,\rho}^*$ in block-wise matrix representation as follows

$$Q_{\mathcal{H},\rho}^* = X_{\mathcal{H},\rho}^* + (I - P'_{\mathcal{H},\mathcal{H}}) Y_{\mathcal{H},\rho}^* - P'_{\mathcal{H}^c,\mathcal{H}} Y_{\mathcal{H}^c,\rho}^* \tag{4.45}$$

$$Q_{\mathcal{H}^c,\rho}^* = X_{\mathcal{H}^c,\rho}^* + (I - P'_{\mathcal{H}^c,\mathcal{H}^c}) Y_{\mathcal{H}^c,\rho}^* - P'_{\mathcal{H},\mathcal{H}^c} Y_{\mathcal{H},\rho}^* \tag{4.46}$$

Solving for $Y_{\mathcal{H}^c,\rho}^*$ in (4.46) and substituting into (4.45), we have

$$Q_{\mathcal{H},\rho}^* = \hat{X}_{\mathcal{H},\rho}^* + (I - \hat{P}'_{\mathcal{H}}) Y_{\mathcal{H},\rho}^*, \tag{4.47}$$

where

$$\hat{X}_{\mathcal{H},\rho}^* = X_{\mathcal{H},\rho}^* - P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}(Q_{\mathcal{H}^c,\rho}^* - X_{\mathcal{H}^c,\rho}^*).$$

Now, we substitute into $\hat{X}_{\mathcal{H},\rho}^*$ the expression for X_{ρ}^* from (4.44), in block matrix notation, leaving a constant $\hat{\eta}_{\rho}$ in the final deterministic drift term initially unspecified, to obtain

$$\begin{aligned} \hat{X}_{\mathcal{H},\rho}^* &= Q_{\mathcal{H},\rho}^*(0) + A_{0,\mathcal{H},\rho}^* + e'_{\mathcal{H}}(\tilde{\Theta}_{\rho}^*)'\mathbf{1} - (I - P'_{\mathcal{H},\mathcal{H}})\tilde{S}_{\mathcal{H},\rho}^* + P'_{\mathcal{H}^c,\mathcal{H}}\tilde{S}_{\mathcal{H}^c,\rho}^* \\ &\quad - P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}Q_{\mathcal{H}^c,\rho}^* \\ &\quad + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}(Q_{\mathcal{H}^c,\rho}^*(0) + A_{0,\mathcal{H}^c,\rho}^* \\ &\quad + e'_{\mathcal{H}^c}(\tilde{\Theta}_{\rho}^*)'\mathbf{1} - (I - P'_{\mathcal{H}^c,\mathcal{H}^c})\tilde{S}_{\mathcal{H}^c,\rho}^* + P'_{\mathcal{H},\mathcal{H}^c}\tilde{S}_{\mathcal{H},\rho}^*) + \hat{\eta}_{\rho}(1 - \rho)^{-1}e \\ &= Q_{\mathcal{H},\rho}^*(0) + A_{0,\mathcal{H},\rho}^* + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}A_{0,\mathcal{H}^c,\rho}^* + (I - \hat{P}'_{\mathcal{H}})\tilde{S}_{\mathcal{H},\rho}^* \\ &\quad + e'_{\mathcal{H}}(\tilde{\Theta}_{\rho}^*)'\mathbf{1} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}e'_{\mathcal{H}^c}(\tilde{\Theta}_{\rho}^*)'\mathbf{1} \\ &\quad + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}(Q_{\mathcal{H}^c,\rho}^*(0) - Q_{\mathcal{H}^c,\rho}^*) + \hat{\eta}_{\rho}(1 - \rho)^{-1}e. \end{aligned}$$

Now we derive the drift term $\hat{\eta}_{\rho}$. To start, let

$$\eta_{\rho} = \lambda_0 - (I - P')\mu_{\rho}.$$

Just like how we treat the HT-scaled queue length process, we can re-write η_{ρ} into blocks

$$\eta_{\mathcal{H},\rho} = \lambda_{0,\mathcal{H}} - (I - P'_{\mathcal{H},\mathcal{H}})\mu_{\mathcal{H},\rho} + P'_{\mathcal{H}^c,\mathcal{H}}\mu_{\mathcal{H}^c,\rho}, \quad (4.48)$$

$$\eta_{\mathcal{H}^c,\rho} = \lambda_{0,\mathcal{H}^c} - (I - P'_{\mathcal{H}^c,\mathcal{H}^c})\mu_{\mathcal{H}^c,\rho} + P'_{\mathcal{H},\mathcal{H}^c}\mu_{\mathcal{H},\rho}. \quad (4.49)$$

Hence

$$\begin{aligned} \hat{\eta}_{\rho} &\equiv \eta_{\mathcal{H},\rho} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}\eta_{\mathcal{H}^c,\rho} \\ &= \lambda_{0,\mathcal{H}} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}\lambda_{0,\mathcal{H}^c} - (I - \hat{P}'_{\mathcal{H}})\mu_{\mathcal{H},\rho}. \end{aligned} \quad (4.50)$$

Note that the traffic-rate equation can be written as

$$\lambda_{0,\mathcal{H}} = (I - P'_{\mathcal{H},\mathcal{H}})\lambda_{\mathcal{H}} - P'_{\mathcal{H}^c,\mathcal{H}}\lambda_{\mathcal{H}^c},$$

$$\lambda_{0,\mathcal{H}^c} = (I - P'_{\mathcal{H}^c,\mathcal{H}^c})\lambda_{\mathcal{H}^c} - P'_{\mathcal{H},\mathcal{H}^c}\lambda_{\mathcal{H}}.$$

Substitute both $\lambda_{0,\mathcal{H}}$ and $\lambda_{0,\mathcal{H}^c}$ into (4.50), we have

$$\hat{\eta}_{\rho} = (I - \hat{P}'_{\mathcal{H}})(\lambda_{\mathcal{H}} - \mu_{\mathcal{H},\rho}). \quad (4.51)$$

To summarize, the HT-scaled net-input process associated with the bottleneck queues can be expressed as

$$\begin{aligned}
\hat{X}_{\mathcal{H},\rho}^* &= Q_{\mathcal{H},\rho}^*(0) + A_{0,\mathcal{H},\rho}^* + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}A_{0,\mathcal{H}^c,\rho}^* - (I - \hat{P}'_{\mathcal{H}})\tilde{S}_{\mathcal{H},\rho}^* \\
&\quad + e'_{\mathcal{H}}(\tilde{\Theta}_\rho^*)'\mathbf{1} + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}e'_{\mathcal{H}^c}(\tilde{\Theta}_\rho^*)'\mathbf{1} \\
&\quad + (I - \hat{P}_{\mathcal{H}})(\lambda_{\mathcal{H}} - \mu_{\mathcal{H},\rho})(1 - \rho)^{-1}e \\
&\quad + P'_{\mathcal{H}^c,\mathcal{H}}(I - P'_{\mathcal{H}^c,\mathcal{H}^c})^{-1}(Q_{\mathcal{H}^c,\rho}^*(0) - Q_{\mathcal{H}^c,\rho}^*). \tag{4.52}
\end{aligned}$$

Now we are ready to deduce the claimed conclusions. First for conclusion (i), most follows directly from Donsker's theorem, Theorem 4.3.2 of [143], and the GJN assumptions. The exception is the limit

$$(\tilde{S}_\rho^*, \tilde{\Theta}_\rho^*) \Rightarrow (S^*, \Theta^*)$$

which follows from the continuous mapping theorem by a random-time-change argument, as shown in [35].

For conclusion (ii), we apply [31] to get

$$(Q_{\mathcal{H},\rho}^*(0), Q_{\mathcal{H}^c,\rho}^*(0)) \Rightarrow (Q_{\mathcal{H}}^*(0), Q_{\mathcal{H}^c}^*(0)) \quad \text{as } \rho \uparrow 1.$$

Then the conclusion (ii) follows from Theorem 6.1 of [35]. In particular, there we see that $Q_{\mathcal{H}^c}^*$ is null, so that we can treat the two components of $(Q_{\mathcal{H},\rho}^*, Q_{\mathcal{H}^c,\rho}^*)$ separately. First, to treat $Q_{\mathcal{H},\rho}^*$, we apply the continuous mapping theorem with the reflection map using the representation above. To do so, we observe that, as $\rho \uparrow 1$,

$$(I - \hat{P}_{\mathcal{H}})(\lambda_{\mathcal{H}} - \mu_{\mathcal{H},\rho})(1 - \rho)^{-1}e \rightarrow -(I - \hat{P}_{\mathcal{H}})\lambda_{\mathcal{H}}e$$

and

$$Q_{\mathcal{H},\rho}^* = \hat{X}_{\mathcal{H},\rho}^* + (I - \hat{P}'_{\mathcal{H}})Y_{\mathcal{H},\rho}^* = \psi_{I - \hat{P}'_{\mathcal{H}}}(\hat{X}_{\mathcal{H},\rho}^*). \tag{4.53}$$

Conclusions (iii) and (iv) follows from the representations derived in Theorem 4.4, the continuous mapping theorem and the established convergence of the queue length process, the external arrival processes and the splitting-decision processes. To this end, we only need to apply diffusion scaling (accelerate time by $(1 - \rho)^{-2}$ and scale space by $(1 - \rho)$) to the

representations in Theorem 4.4 so that

$$\begin{aligned} A_\rho^* &= P'(I - P')^{-1} (Q_\rho^*(0) - Q_\rho^*) + (I - P')^{-1} (A_{0,\rho}^* + (\tilde{\Theta}_\rho^*)' \mathbf{1}), \\ D_\rho^* &= (I - P')^{-1} (Q_\rho^*(0) - Q_\rho^* + A_{0,\rho}^* + (\tilde{\Theta}_\rho^*)' \mathbf{1}). \end{aligned} \quad (4.54)$$

The second expression follows from the fact that $Q_{\mathcal{H}^c}^* = 0$.

Next, conclusions (v) follows from the limit of the departure process and the FCLT of the splitting operation in §9.5 of [143]. Finally, the associated limits for the workload can be related to the limit for the queue length as indicated in [35]. ■

Remark 4.3 (Elaboration on the application of [31]) We apply [31], but it must be extended to the model with non-bottleneck queues. We do not go through all details because we regard that step as minor, but we now briefly explain.

First, the main stability condition (A6) there holds in our setting here. Notice that our scaling convention here relies on the traffic intensity parameter ρ instead of the scaling parameter n used in [31]. Comparing (4.22) here with (A5) there, For the bottleneck queues, the two scaling conventions are connected by setting $n = (1 - \rho)^{-2}$, $\tilde{v}_i^n = 0$ and $\tilde{\beta}_i^n = -\lambda_i/\rho$. The stability condition here is then connected to that in [31] by setting $\theta_0 = -1$ in (13) there.

For the moment estimation in their Theorem 3.3, we treat $Q_{\mathcal{H}}$ and $Q_{\mathcal{H}^c}^*$ separately. For $Q_{\mathcal{H}}$, our representation (4.47) and (4.52) can be mapped to the representations (16) on p.51 of [31], but with slightly more complicated constant terms associated with the matrix multiplication we have in (4.52). Noting the expression of the drift term we have in (4.51), the rest of the proof is essentially the same. For $Q_{\mathcal{H}^c}^*$, by [34; 35], it is negligible in the sense of Theorem 3.3 of [31]. Theorem 3.4 of [31] relies only on the moment estimation as in their Theorem 3.3 and the strong Markov property of $\mathcal{S}(t)$ (which they denoted as $X(t)$). Finally, Theorem 3.5 and Theorem 3.2 of [31] remain unchanged.

Chapter 5

Robust Queueing Network Analyzer

In this chapter, we develop the Robust Queueing Network Analyzer (RQNA) algorithm in the network setting introduced in Section 4.1, which generalizes the RQ algorithm for single-server queues in Chapter 2 and for queues in series models in Chapter 3.3. In revoking the RQ algorithm in (2.35), our primary focus is to analyze and approximate the IDC's of the customer flows in a OQN. To start, we divided the flows into two groups, the *external flows* and the *internal flows*, and introduce the notations for the IDC's.

The external flows are the flows associated with the model primitives in Section 4.1, i.e., the external arrival flows and the flows associated with the (non-interrupted) service process. For the external arrival process $A_{0,i}$, we let $I_{a,0,i} \equiv \{I_{a,0,i}(t) : 0 \leq t \leq \infty\}$ denote the its IDC. For the service flows, let $I_{s,i} \equiv \{I_{s,i}(t); 0 \leq t \leq \infty\}$ be the IDC of the stationary renewal process associated with (4.1). For the case of renewal process, we necessarily have $I_{s,i}(\infty) = c_{s,i}^2$, as in Lemma 3.1. As regularity assumption, we assume that the IDC's $I_{a,0,i}$ and $I_{s,i}$ is continuous with finite limits at $t = 0$ and $+\infty$. The IDC's of the external flows are regarded as important input parameters of our RQNA algorithm, which is in stark contrast to the QNA algorithm in [134], where only to the means and scv's are required. In particular, we assume that we are given $(\lambda_{0,i}, I_{a,0,i}, \mu_i, I_{s,i})$ for each queue i and the routing matrix P . Practical methods to obtain the IDC's of the external flows are discussed in

Section 2.3.4.

The internal flows are all other customer flows in the network. In particular, we have the total arrival flows at each station, which is of particular interest for our RQNA algorithm. Let A_i denote the total arrival process at queue i and let $I_{a,i}$ be the associated IDC. Let D_i denote the departure process at queue i and let $I_{d,i}$ be the associated IDC. Finally, let $A_{i,j}$ denote the departing customer flow from queue i that are routed to queue j and let $I_{a,i,j}$ be the associated IDC.

The IDC's of the internal flows are regarded as unknown parameters to be approximated from the external IDC's. In Section 5.1 - 5.2, we demonstrate how the HT limits in Chapter 4 can be applied to develop IDC approximations for the three network operations: the departure operation, the superposition operation and the splitting operation, respectively. For each operation, our approximation is written as a linear equation, which we refer to as *the IDC equations*, see (5.1), (5.10) and (5.28). In Section 5.4, the IDC equations are combined into a general framework for approximating the IDC's of the flows.

One of the key assumptions in our IDC formulation of the RQ algorithm (2.35) is that the service times are i.i.d., independent of the arrival process at the same station. This assumption allow us to decompose the IDW as in (2.26) and work with the relatively simple arrival IDC. However, in the presence of customer feedback (so that a customer may re-visit a station), the service times is necessarily correlated with the arrival process at the feedback queue. As a mitigation, we propose a feedback elimination procedure in Section, which is supported by the heavy-traffic limit theorem in Theorem 4.6 and Theorem 4.8.

The full RQNA algorithm is presented in Section 5.6. We also present a more elementary version for tree-structured OQN's in Section 5.6.1.

5.1 The Departure Operation

In Section 3.3, we investigated the the departure IDC in the case of $GI/GI/1$ queue and proposed to approximate the departure IDC by (3.1) and (3.67). We now provide full support for this approximation in generalized Jackson network. In terms of the IDC's in

the network, we have

$$I_{d,i}(t) \approx w_i(t)I_{a,i}(t) + (1 - w_i(t))I_{s,i}(\rho_i t), \quad t \geq 0 \quad (5.1)$$

with

$$w_i(t) \equiv w^* \left((1 - \rho_i)^2 \lambda_i t / \rho_i^2 c_{x,i}^2 \right), \quad t \geq 0, \quad (5.2)$$

where $c_{x,i}^2 = c_{a,i}^2 + c_{s,i}^2$, $c_{a,i}^2$ is the limiting variability parameter of A_i , defined in (5.4), $c_{s,i}^2$ is the service scv at station i and $w^*(t)$ is the weight function in (3.27).

We start with a characterization of the *limiting variability parameter* of the total arrival process. Recall that $\Theta_i^* = (\Theta_{i,j}^* : 0 \leq j \leq K)$ is the collection of the Brownian limits of the decision processes at station i , see Theorem 4.5 for example. We have

$$\text{cov}(\Theta_{i,j}^*, \Theta_{i,k}^*) = \begin{cases} p_{i,j}(1 - p_{i,j})\lambda_i t, & j = k, \\ -p_{i,j}p_{i,k}\lambda_i t, & j \neq k. \end{cases}$$

Define

$$\Sigma_i \equiv (\text{cov}(\Theta_{i,j}^*, \Theta_{i,k}^*)/t)_{j,k=1}^K \in \mathbb{R}^{K \times K}$$

so that Σ_i is a constant matrix independent of t .

Lemma 5.1 (Limiting variability parameter) *Under the assumptions in Theorem 4.5 plus the usual uniform integrability conditions, for which it suffices for the interarrival times and service times to have uniformly bounded fourth moments,*

$$\lim_{\substack{t \rightarrow \infty \\ \rho \uparrow 1}} I_{a,\rho,i}(t) = \lim_{t \rightarrow \infty} \lim_{\rho \uparrow 1} I_{a,\rho,i}(t) = \lim_{\rho \uparrow 1} \lim_{t \rightarrow \infty} I_{a,\rho,i}(t) = c_{a,i}^2 \quad (5.3)$$

where

$$\begin{aligned} c_{a,i}^2 &= \text{Var} \left(e_i' (I - P')^{-1} (A_0^* + (\Theta^*)' \mathbf{1}) \right) \\ &= e_i' (I - P')^{-1} \left(\text{diag}(c_{a0,i}^2 \lambda_i) + \sum_{l=1}^K \Sigma_l \right) (I - P)^{-1} e_i. \end{aligned} \quad (5.4)$$

The following theorem generalizes Theorem 3.8.

Theorem 5.1 (Heavy-traffic limit for the departure variance function in GJN)

Under the assumptions in Corollary 4.3 plus the usual uniform integrability conditions, for

which it suffices for the interarrival times and service times to have uniformly bounded fourth moments, the HT limit of the departure process of the bottleneck station h can be written as

$$D_h^* = \tilde{Q}_h^*(0) + \tilde{A}_h^* - \tilde{Q}_h^*, \quad (5.5)$$

where

$$\tilde{A}_h^* = e'_h(I - P')^{-1} (A_0^* + (\Theta^*)'\mathbf{1}) \quad (5.6)$$

and

$$\tilde{Q}_h^* = \frac{1}{1 - \hat{P}_h} Q_h^* = \psi \left(\tilde{Q}_h^*(0) + \tilde{A}_h^* - S_h^* - \lambda_h e \right). \quad (5.7)$$

As a result, the limiting variance function of the departure process is

$$\begin{aligned} V_{d,h,\rho}^*(t) &\equiv \text{Var}(D_{h,\rho}^*(t)) = E[D_{h,\rho}^*(t)^2] \\ &\rightarrow E[D_h^*(t)^2] = \text{Var}(D_h^*(t)) \equiv V_{d,h}^*(t) \quad \text{as } \rho \uparrow 1, \end{aligned} \quad (5.8)$$

where

$$V_{d,h}^*(t) = w^*(\lambda_h t / c_{x,h}^2) c_{A,h}^2 \lambda_h t + (1 - w^*(\lambda_h t / c_{x,h}^2)) c_{s,h}^2 \lambda_h t \quad (5.9)$$

with $c_{x,h}^2 = c_{A,h}^2 + c_{s,h}^2$, $c_{A,h}^2$ is the limiting variability parameter of A_h , defined in (5.4), $c_{s,h}^2$ is the service scv at station h and $w^*(t)$ is the weight function in (3.27).

The approximation (5.1) is then justified by the exact same procedure as described in Section 3.3.1.

In Section 6.3.1, we provide numerical support for the following conjecture.

Conjecture 5.1 *Theorem 5.1 holds for general $G/G/1$ queue, where the arrival and service processes are stationary and ergodic point processes with finite IDC's and the service process is independent of the arrival process.*

5.2 The Splitting Operation

In this section, we derive our IDC equation for the splitting operation. In particular, we propose the following approximation

$$I_{a,i,j}(t) = p_{i,j} I_{d,i}(t) + (1 - p_{i,j}) + \alpha_{i,j}(t). \quad (5.10)$$

where α_i is defined in (5.16).

We start with a simple example in Section 5.2.1 as an illustration and then address the general case in Section 5.2.2.

5.2.1 Dependent Splitting: One Queue with Immediate Feedback

Consider the single-server queue with immediate customer feedback as in Section 4.4.1. This introduces dependence between the splitting decision process and the arrival process.

For the splitting operation, suppose that the splitting decision is independent of the departure process, then by the conditional variance formula, we have

$$\text{Var}(A_{\text{int}}(t)) = p^2 \text{Var}(D(t)) + p(1-p)\lambda t,$$

or equivalently, since $E[D(t)] = \lambda t$ and $E[A_{\text{int}}(t)] = p\lambda t = pE[D(t)]$,

$$I_{a,\text{int}}(t) = pI_d(t) + (1-p).$$

To address the impact of dependence on the IDC after the splitting operation, we propose to consider the correction term $\alpha(t)$ defined as

$$\alpha(t) \equiv I_{a,\text{int}}(t) - pI_d(t) - (1-p),$$

so that

$$I_{a,\text{int}}(t) = pI_d(t) + (1-p) + \alpha(t). \quad (5.11)$$

We propose to approximate the correction term $\alpha(t)$ by

$$\alpha(t) \approx \alpha^*((1-\rho)^2 t) \quad (5.12)$$

with

$$\alpha^*(t) \equiv 2\text{cov}(pD^*(t), \Theta^*(\lambda t))/p\lambda t = 2pw^*(t/c_x^2),$$

where $c_x^2 = \frac{1}{1-p}c_{a_0}^2 + \frac{p}{1-p} + c_s^2$ and the explicit expression is derived using Corollary 3.6.

We demonstrate the performance of the approximation (5.12) in Example 5.1 and postpone theoretical support to Corollary 5.1.

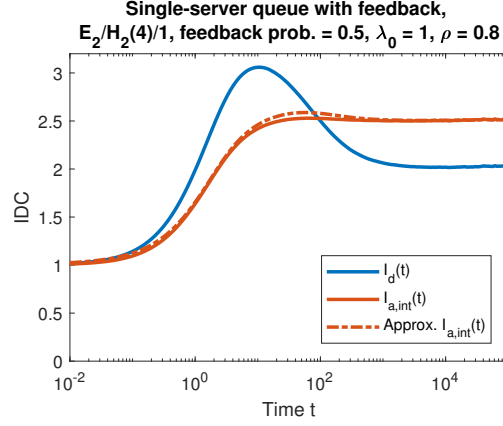


Figure 5.1: Approximations of the IDC's in a single-server queue with feedback model.

Example 5.1 (immediate feedback) Figure 5.1 compares the performance of the IDC approximation to simulations for the $E_2/H_2(4)/1$ single-server queue with feedback model, having feedback probability $p = 0.5$ and service scv $c_s^2 = 4$. Model parameters are described in the title. The simulation estimation of the IDC of the feedback flow is contrasted to the IDC approximation (5.11) with correction term (5.12) in dotted-and-dashed lines. The approximation matches simulation remarkably well.

5.2.2 The General Case

To treat splitting, we write the split process $A_{i,j}$ as a random sum. Recall from (4.2) that

$$A_{i,j}(t) = \sum_{l=1}^{D_i(t)} \theta_{i,j}^l, \quad t \geq 0.$$

We apply the conditional-variance formula to write the variance $V_{a,i,j}(t) \equiv \text{Var}(A_{i,j}(t))$ as

$$V_{a,i,j}(t) = E[\text{Var}(A_{i,j}(t)|D_i(t))] + \text{Var}(E[A_{i,j}(t)|D_i(t)]). \quad (5.13)$$

With the Markovian routing we have assumed, the routing decisions at each queue at each time are i.i.d. and independent of the history of the network. As a consequence, for feed-forward queueing networks, we can deduce that the collection of all routing decisions made at queue i up to time t is independent of $D_i(t)$. For the case in which independence holds, we can apply (5.13) to express $V_{a,i,j}(t)$ in terms of the variance of the departure

process, $V_{d,i}(t) \equiv \text{Var}(D_i(t))$; in particular,

$$V_{a,i,j}(t) = p_{i,j}^2 V_{d,i}(t) + p_{i,j}(1 - p_{i,j})\lambda_i t, \quad (5.14)$$

or, equivalently, since $E[D_i(t)] = \lambda_i t$ and $E[A_{i,j}(t)] = p_{i,j}\lambda_i t = p_{i,j}E[D_i(t)]$,

$$I_{a,i,j}(t) = p_{i,j}I_{d,i}(t) + (1 - p_{i,j}). \quad (5.15)$$

The formula (5.15) is an initial approximation, which parallels the approximation used for splitting in (40) of [134], i.e., $c_{a,i,j}^2 = p_{i,j}c_{d,i}^2 + (1 - p_{i,j})$.

However, the independence assumption will not hold in the existence of customer feedback, in which case there is a complicated dependence. we develop a more general formula to improve the approximation in general OQN's.

For that purpose, we apply the FCLT for split processes in Section 9.5 of [143] and the heavy-traffic limit theorems in Section 4.4. Based on the heavy-traffic analysis, we propose the splitting IDC equation in (5.10). To account for the dependence, we include a correction term $\alpha_{i,j}$, defined as

$$\alpha_{i,j,\rho_i}(t) \equiv I_{a,i,j}(t) - p_{i,j}I_{d,i}(t) - (1 - p_{i,j}), \quad (5.16)$$

which leads to (5.10). We propose to approximate α_{i,j,ρ_i} by

$$\alpha_{i,j,\rho_i}(t) \approx 2\xi_{i,j}p_{i,j}(1 - p_{i,j})w_i(t), \quad t \geq 0, \quad (5.17)$$

where $w_i(t)$ is the weight function in (5.2) and $\xi_{i,j}$ is the $(i, j)^{\text{th}}$ entry of the matrix $(I - P')^{-1}$.

We now provide theoretical support for the splitting approximation (5.17). Consider the diffusion-scaled processes indexed by ρ

$$\begin{aligned} D_{i,\rho}^*(t) &= (1 - \rho) [D_i((1 - \rho)^{-2}t) - \lambda_i(1 - \rho)^{-2}t], \\ \Theta_{i,\rho}^*(t) &= (1 - \rho) \left[\sum_{l=1}^{\lfloor (1-\rho)^{-2}t \rfloor} \theta^l - \mathbf{p}_i(1 - \rho)^{-2}t \right] \in \mathcal{D}^K, \\ \mathbf{A}_{i,\rho}^*(t) &= (1 - \rho) [\mathbf{A}_i((1 - \rho)^{-2}t) - \lambda_i \mathbf{p}_i(1 - \rho)^{-2}t] \in \mathcal{D}^K, \end{aligned} \quad (5.18)$$

for $t \geq 0$, where $\mathbf{p}_i \equiv E[\theta_i^l]$ is the i -th row of the routing matrix and $\mathbf{A}_{i,\rho} = (A_{i,j,\rho} : j = 1, 2, \dots, K)$ is the vector consists of all the streams after splitting. The following result rephrases Theorem 9.5.1 in Whitt (2002).

Theorem 5.2 (Theorem 9.5.1 of [143]) *Suppose that*

$$(D_{i,\rho}^*, \Theta_{i,\rho}^*) \Rightarrow (D_i^*, \Theta_i^*) \quad \text{as } \rho \uparrow 1 \quad \text{in } D^{K+1} \quad (5.19)$$

and that almost surely D^ and $\Theta^* \circ \lambda e$ have no common discontinuities of opposite sign. Then*

$$\mathbf{A}_{i,\rho}^* \Rightarrow \mathbf{A}_i^* \quad \text{in } D^K,$$

with

$$A_{i,j}^* \equiv p_{i,j} D^* + \Theta_{i,j}^* \circ \lambda_i e, \quad \text{for } 1 \leq j \leq K, \quad (5.20)$$

where $e(t) = t$ is the identity mapping.

Example 5.2 (Splitting the departures from a $G/GI/1$ queue) *If we split the departure process from the $GI/GI/1$ model with Markovian routing, then D^* is independent of Θ^* and Θ^* is a zero-drift K -dimensional Brownian motion with covariance matrix $\Sigma = (\sigma_{i,j}) \in \mathbb{R}^{K \times K}$, where $\sigma_{i,i}^2 = p_i(1 - p_i)$ and $\sigma_{i,j}^2 = -p_i p_j$ for $i \neq j$. Hence, from (5.20) we obtain*

$$\mathbf{A}^* = \mathbf{p} D^* + \Theta^* \circ \lambda e, \quad (5.21)$$

which is consistent with (5.14) and thus (5.15). ■

Theorem 5.2 assumes only a joint FCLT for the flow to split and the splitting decision process, so dependence is allowed. Thus it provides support for the general splitting equation in (5.10) for the case where $D_{i,j}$ and $\Theta_{i,j}$ are correlated. Define the HT-scaled correction term as

$$\alpha_{i,j,\rho}^*(t) \equiv \alpha_{i,j}((1 - \rho)^{-2}t) \quad (5.22)$$

and define the limiting correction term as

$$\alpha_{i,j}^*(t) \equiv 2\text{cov}(p_{i,j} D_i^*(t), \Theta_{i,j}^*(\lambda_i t)) / p_{i,j} \lambda_i t. \quad (5.23)$$

The following corollary follows from Theorem 5.2.

Corollary 5.1 *Under the assumptions in Theorem 5.2 plus the uniform integrability conditions, we have $\alpha_{i,j,\rho}^*(t) \rightarrow \alpha_{i,j}^*(t)$ as $\rho \uparrow 1$.*

Proof. By the definitions of the correction term in (5.16) and HT-scaled processes, we write

$$\begin{aligned}
\alpha_{i,j,\rho}^*(t) &= \alpha_{i,j}((1-\rho)^{-2}t) \\
&= I_{a,i,j}((1-\rho)^{-2}t) - p_{i,j}I_{d,i}((1-\rho)^{-2}t) - (1-p_{i,j}) \\
&= \frac{\text{Var}((1-\rho)A_{i,j}((1-\rho)^{-2}t))}{p_{i,j}\lambda_i t} - p_{i,j} \frac{\text{Var}((1-\rho)D_i((1-\rho)^{-2}t))}{\lambda_i t} - (1-p_{i,j}) \\
&= \frac{\text{Var}(A_{i,j,\rho}^*(t))}{p_{i,j}\lambda_i t} - p_{i,j} \frac{\text{Var}(D_{i,\rho}^*(t))}{\lambda_i t} - (1-p_{i,j}) \\
&\rightarrow \frac{\text{Var}(A_{i,j}^*(t))}{p_{i,j}\lambda_i t} - p_{i,j} \frac{\text{Var}(D_i^*(t))}{\lambda_i t} - (1-p_{i,j}) = \alpha_{i,j}^*(t). \quad \blacksquare
\end{aligned}$$

This corollary supports the following approximation for the correction term $\alpha_{i,j}$ in

$$\alpha_{i,j}(t) \approx \alpha_{i,j}^*((1-\rho)^2 t) \quad (5.24)$$

with $\alpha_{i,j}^*$ defined in (5.23).

It then remains to derive an explicit formula for $\alpha_{i,j}^*$. For any $\alpha_{i,j}$, the relevant routing flow is $A_{i,j}$ while the relevant departure flow is D_i . Naturally, we choose station i to be the HT station. So we let $\rho_i = \rho \uparrow 1$ and keep $\rho_j < 1$ for $j \neq i$. Define the HT scaled processes as in (4.22) and apply Theorem 5.1 to obtain

$$D_{i,\rho}^* \Rightarrow D_i^* = \tilde{A}_i^* + \tilde{Q}_i^*(0) - \tilde{Q}_i^*. \quad (5.25)$$

For the routing flow $A_{i,j}$, we apply Theorem 5.2 so that

$$A_{i,j,\rho}^* \Rightarrow A_{i,j}^* = p_{i,j}D_i^* + \Theta_{i,j} \circ \lambda_i e \quad \text{as } \rho \uparrow 1. \quad (5.26)$$

Define the correction term $\alpha_{i,j}^*$ as in (5.24), then Corollary 3.6 implies the following corollary, which leads to the correction term in (5.16).

Theorem 5.3 *Under the assumptions in Corollary 4.3 and Theorem 5.2 plus the uniform integrability conditions, we have*

$$\begin{aligned}
\alpha_{i,j,\rho}^*(t) &\rightarrow 2\text{cov}(p_{i,j}D_i^*(t), \Theta_{i,j}(\lambda_i t)) / (p_{i,j}\lambda_i t) \\
&= 2\xi_{i,j}p_{i,j}(1-p_{i,j})w^*(\lambda_i t/c_{x,i}^2), \quad \text{as } \rho \uparrow 1,
\end{aligned} \quad (5.27)$$

where $\xi_{i,j}$ is the $(i,j)^{\text{th}}$ entry of the matrix $(I-P')^{-1}$, $c_{x,i}^2 = c_{a,i}^2 + c_{s,i}^2$ and $c_{a,i}^2$ is the limiting variability parameter as solved from (5.46) and $c_{s,i}^2$ is the scv of the service distribution at station i .

Proof. Apply Corollary 4.3 to obtain expression for $D_i^*(t)$, then apply Corollary 3.6 for the explicit covariance in (5.27). ■

As a direct result of Theorem 5.3 and Corollary 5.1, we propose to approximate the correction term as in (5.17), which is asymptotically exact as $\rho \uparrow 1$.

5.3 The Superposition Operation

In this section, we derive our IDC equation for the superposition operation. In particular, we propose the following approximation

$$I_{a,i}(t) = \sum_{j=0}^K (\lambda_{j,i}/\lambda_i) I_{a,j,i}(t) + \beta_i(t), \quad (5.28)$$

where β_i is defined in (5.36).

We start with a simple example in Section 5.3.1 as an illustration and then address the general case in Section 5.3.2.

5.3.1 Dependent Superposition: A Splitting and Recombining Example

We consider a simple feed-forward network depicted in Figure 5.2, where an arrival process is first split into two streams according to Markovian routing, then sent to separate queues, and finally re-combine and enter a third queue. We aim to approximate the IDC of the superposition of the two stationary departure processes $A_3(t) \equiv D_1(t) + D_2(t)$. To do so, we establish the HT limit for the superposition arrival process at the third queue.

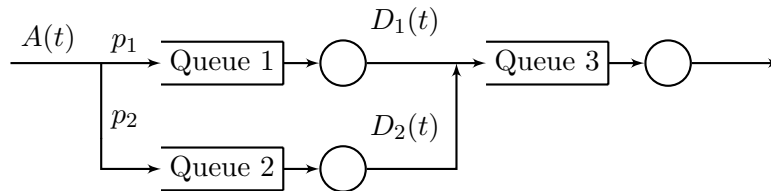


Figure 5.2: A simple splitting and recombining example.

Without loss of generality, assume that the traffic intensity ρ_1 at the first queue is larger than ρ_2 at the second queue. We then consider a family of systems indexed by ρ , where the traffic intensity at queue 1 is $\rho_1 = \rho$, which we will bring to heavy traffic, and the

traffic intensity at queue 2 is fixed at $\rho_2 \in [0, 1)$. Let $A_{i,\rho}$, $S_{i,\rho}$, $D_{i,\rho}$ and $Q_{i,\rho}$ denote the arrival process, the (uninterrupted) service renewal processes, the departure process and the queue length process at Queue i in the ρ -th system, respectively. Let the processes with superscript $\star$ be the heavy-traffic scaled processes as in (4.22). Theorem 4.5 implies the following

Corollary 5.2 (Heavy-traffic limit for splitting and recombining model) *Consider the system depicted in Figure 5.2. Assume that the external arrival process is renewal with rate λ and scv c_a^2 , the service times at queue 1 are i.i.d. with rate $p_1\lambda/\rho$ and scv $c_{s_1}^2$; the service times at queue 2 are i.i.d. with rate $p_2\lambda/\rho_2$ for $0 \leq \rho_2 < 1$ and scv $c_{s_2}^2$. Then*

$$\begin{aligned} & (A_\rho^\star, A_{1,\rho}^\star, A_{2,\rho}^\star, S_{1,\rho}^\star, S_{2,\rho}^\star, Q_{1,\rho}^\star, Q_{2,\rho}^\star, D_{1,\rho}^\star, D_{2,\rho}^\star, \Theta_{1,\rho}^\star, \Theta_{2,\rho}^\star) \\ & \Rightarrow (A^\star, A_1^\star, A_2^\star, S_1^\star, S_2^\star, Q_1^\star, Q_2^\star, D_1^\star, D_2^\star, \Theta_1^\star, \Theta_2^\star) \quad \text{in } \mathcal{D}^{11} \quad \text{as } \rho \rightarrow 1, \end{aligned}$$

where

$$\begin{aligned} A^\star & \equiv c_a B_a \circ \lambda e, \\ A_i^\star & \equiv p_i c_a B_a \circ \lambda e + \Theta_i^\star, \quad \text{for } i = 1, 2, \\ S_1^\star & \equiv c_{s_1} B_{s_1} \circ p_1 \lambda e, \quad S_2^\star \equiv c_{s_2} B_{s_2} \circ p_2 \lambda e / \rho_2, \\ Q_1^\star & \equiv \psi(Q_1^\star(0) + p_1 c_a B_a \circ \lambda e + \Theta_1^\star - c_{s_1} B_{s_1} \circ p_1 \lambda e - p_1 \lambda e), \quad Q_2^\star \equiv 0, \\ D_1^\star & \equiv p_1 c_a B_a \circ \lambda e + \Theta_1^\star + Q_1^\star(0) - Q_1^\star, \quad D_2^\star \equiv p_2 c_a B_a \circ \lambda e + \Theta_2^\star, \end{aligned} \tag{5.29}$$

with ψ being the one-dimensional reflection mapping and $(\Theta_1^\star, \Theta_2^\star)$ being a zero-drift two-dimensional Brownian motion with covariance matrix $\Sigma = (\sigma_{ij}) \in \mathbb{R}^{2 \times 2}$, where $\sigma_{ii}^2 = p_i(1 - p_i)\lambda$ and $\sigma_{ij}^2 = -p_i p_j \lambda$ for $i \neq j$.

To approximate the IDC of the total arrival process at queue 3, we write

$$\begin{aligned} I_{a,3,\rho}(t) & \equiv \frac{\text{Var}(A_{3,\rho}(t))}{E[A_{3,\rho}(t)]} = \frac{\text{Var}(D_{1,\rho}(t) + D_{2,\rho}(t))}{E[A_{3,\rho}(t)]} \\ & = \frac{\text{Var}(D_{1,\rho}(t))}{E[A_{3,\rho}(t)]} + \frac{\text{Var}(D_{2,\rho}(t))}{E[A_{3,\rho}(t)]} + 2\text{cov}(D_{1,\rho}(t), D_{2,\rho}(t)) / E[A_{3,\rho}(t)] \\ & = p_1 I_{d,1,\rho}(t) + p_2 I_{d,2,\rho}(t) + \beta_\rho(t), \end{aligned}$$

where

$$\beta_\rho(t) \equiv 2\text{cov}(D_{1,\rho}(t), D_{2,\rho}(t)) / E[A_{3,\rho}(t)]. \quad (5.30)$$

In general, exact characterization of β_ρ is not readily available. We propose the following approximation

$$\begin{aligned} \beta_\rho(t) &\approx 2\text{cov}(D_1^*((1-\rho)^2t), D_2^*((1-\rho)^2t)) / (\lambda(1-\rho)^2t) \\ &= 2p_1(1-p_1)(c_{a_0}^2 - 1)w^*((1-\rho)^2p_1\lambda t/c_{x_1}^2) \end{aligned} \quad (5.31)$$

with D_1^* and D_2^* being the diffusion limit in (5.29).

To justify the approximation (5.31), let $\beta_\rho^*(t) = \beta_\rho((1-\rho)^{-2}t)$ be the HT-scaled correction term. Corollary 5.2 implies the following limit.

Corollary 5.3 *Under the assumption in Theorem 5.2 and the uniform integrability conditions, we have*

$$\beta_\rho^* \rightarrow 2p_1(1-p_1)(c_{a_0}^2 - 1)w^*(p_1\lambda t/c_{x_1}^2). \quad (5.32)$$

Proof Note that Corollary 5.2 implies that

$$\begin{aligned} \text{cov}(D_{1,\rho}(t), D_{1,\rho}(t)) &= \text{cov}((1-\rho_1)^{-1}D_{1,\rho}^*((1-\rho_1)^2t), (1-\rho_1)^{-1}D_{2,\rho}^*((1-\rho_1)^2t)) \\ &\rightarrow (1-\rho_1)^{-2}\text{cov}(D_1^*((1-\rho_1)^2t), D_2^*((1-\rho_1)^2t)), \end{aligned}$$

as $\rho \uparrow 1$.

On the other hand, by applying Corollary 3.6, we have

$$\begin{aligned} \text{cov}(D_1^*(t), D_2^*(t)) &= \text{cov}(A_1^*(t), A_2^*(t)) - \text{cov}(Q_1^*(t), A_2^*(t)) \\ &= p_1(1-p_1)(c_{a_0}^2 - 1)\lambda t - \text{cov}(Q_1^*(t), A_2^*(t)) \\ &= p_1(1-p_1)(c_{a_0}^2 - 1)\lambda t w^*(p_1\lambda t/c_{x_1}^2), \end{aligned}$$

where $c_{x_1}^2 = c_{a_1}^2 + c_s^2$, $c_{a_1}^2 = p_1c_a^2 + (1-p_1)$. The limit then follows. ■

We demonstrate the performance of the approximation by making simulation comparisons in Example 5.3.

Example 5.3 Consider the queueing system in Figure 5.2 with rate-1 hyperexponential ($H_2(4)$) external arrival process and $c_a^2 = 4$, $p_1 = 0.25$, $p_2 = 0.75$ and i.i.d. Erlang (E_2)

service times with $c_{s_i}^2 = 0.5$. Figure 5.3 shows the results for two cases involving different traffic intensities: (i) $\rho_1 = \rho_2 = 0.7$ (left); and (ii) $\rho_1 = 0.8$ and $\rho_2 = 0.9$ (right). In each plot, we display, in solid lines, the IDC $I_{a,3}$ of the total arrival process at queue 3, the modified IDC's $p_i I_{d,i}$ of the departure processes from queue i , the simulated correction term β_ρ defined in (5.30). For approximations, we display, in broken lines, the approximated correction terms as in (5.31) and the approximated IDC using (5.31). Figure 5.3 shows remarkable agreement of the approximation and the simulation estimate.

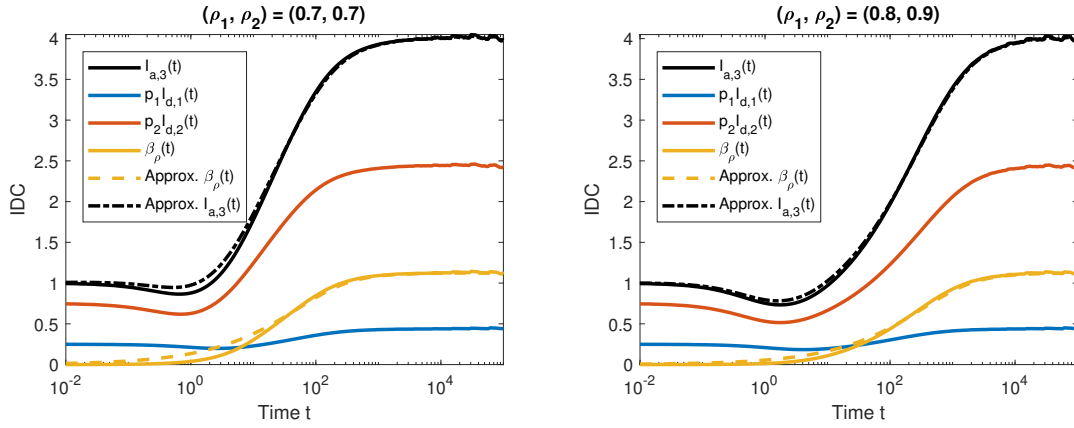


Figure 5.3: Approximation of the arrival IDC at queue 3 for the splitting and recombining model in Figure 5.2.

5.3.2 The General Case

We now investigate the impact of the superposition operation on the IDC's for general network. To start, consider the case where the individual streams are mutually independent, and hence we have

$$V_{a,i}(t) \equiv \text{Var}(A_i(t)) = \text{Var} \left(\sum_{j=0}^K A_{j,i}(t) \right) = \sum_{j=0}^K \text{Var}(A_{j,i}(t)),$$

so that

$$I_{a,i}(t) = \sum_{j=0}^K (\lambda_{j,i}/\lambda_i) I_{a,j,i}(t), \quad (5.33)$$

where $I_{a,j,i}(t) \equiv \text{Var}(A_{j,i}(t))/E[A_{j,i}(t)]$.

While (5.33) is exact when the streams are independent, it does not hold in general cases. Even for feed-forward networks, we may have a stream that splits and then recombines later, which introduces dependence.

For dependent streams, the variance of the superposition total arrival process at queue i can be written as

$$V_{a,i}(t) \equiv \text{Var} \left(\sum_{j=0}^K A_{j,i}(t) \right) = \sum_{j=0}^K \text{Var}(A_{j,i}(t)) + \beta_i(t) E[A_i(t)] \quad (5.34)$$

where $A_{0,i}$ denotes the external arrival process at station i ,

$$\beta_i(t) \equiv \sum_{j \neq k} \beta_{j,i;k,i}(t), \quad \text{and} \quad \beta_{j,i;k,i}(t) \equiv \frac{\text{cov}(A_{j,i}(t), A_{k,i}(t))}{E[A_i(t)]}. \quad (5.35)$$

Divide both sides of (5.34) by $E[A_i(t)] = \lambda_i t$, we arrive at our IDC equation in (5.28).

In general, exact characterization of the correction term $\beta_i(t)$ in (5.35) is not available. We now show that heavy-traffic limit theorem suggests the following approximation (assume without loss of generality that $\rho_j \geq \rho_i$)

$$\beta_{j,i;k,i}(t) = \beta_{k,i;j,i}(t) \approx (\zeta_{j,i;k,i}/\lambda_i) w^*((1 - \rho_j)^2 p_{j,i} \lambda_j t / \rho_j c_{x,j,i}^2), \quad (5.36)$$

where w^* is the weight function in (3.27), $c_{x,j,i}^2 = p_{j,i} c_{a,j}^2 + (1 - p_{j,i}) + p_{j,i} c_{s,j}^2$ and $c_{a,j}^2$ is solved from the variability equations for the asymptotic variability parameters in (5.49). The constant $\zeta_{j,i;k,i}$ is given by

$$\zeta_{j,i;k,i} = \nu'_j \left(\text{diag}(c_{a,0,i}^2 \lambda_i) + \sum_{l=1}^K \Sigma_l \right) \nu_k + \nu'_k \Sigma_j e_i + \nu'_j \Sigma_k e_i, \quad (5.37)$$

where $\nu_l \equiv p_{l,i} e'_l (I - P')^{-1}$ for $l = j, k$, e_i is the i -th unit vector, $\text{diag}(c_{a,0,i}^2 \lambda_i)$ is the diagonal matrix with $c_{a,0,i}^2 \lambda_i$ as the i -th diagonal entry, Σ_l is the covariance matrix of the splitting decision process at station l defined as $\Sigma_l \equiv (\sigma_{i,j}^l)$ with $\sigma_{i,i}^l = p_{l,i}(1 - p_{l,i})\lambda_l$ and $\sigma_{i,j}^l = -p_{l,i}p_{l,j}\lambda_l$ for $i \neq j$.

Next, we provide theoretical support the correction term β_i associated with dependent superposition. From (5.35), it suffices to specify $\beta_{k,i;j,i}$ for any station i and any pair of sub-flows $(A_{j,i}, A_{k,i})$ at that station. We assume without loss of generality that (i) $\rho_j \geq \rho_k$, or (ii) $\rho_j = \rho_k$ and $\lambda_{j,i} \geq \lambda_{k,i}$. In the case (ii), we break the tie by picking the index that

gives the larger rate $\lambda_{j,i}$. In both cases, we consider station j to be the HT station while keep all other stations unsaturated.

By Theorem 5.1, we have

$$\begin{aligned} A_\rho^\star &\Rightarrow A^\star = \tilde{A}^\star + \gamma_j \left(\tilde{Q}_j^\star(0) - \tilde{Q}_j^\star \right) \\ D_{j,\rho}^\star &\Rightarrow D_j^\star = \tilde{A}_j^\star + \tilde{Q}_j^\star(0) - \tilde{Q}_j^\star, \\ D_{l,\rho}^\star &\Rightarrow D_l^\star = A_l^\star, \quad \text{for } l \neq j, \end{aligned}$$

where $\tilde{A}^\star = (I - P')^{-1} (A_0^\star + (\Theta^\star)' \mathbf{1})$, $\tilde{Q}_j^\star$ is defined as in (5.7) with $h = j$ and $\gamma_j \in \mathbb{R}^K$ is defined as

$$\gamma_j = P'(I - P')^{-1} e'_j (1 - \hat{P}_j)$$

with $\hat{P}_j$ defined as in (4.24) with $\mathcal{H} = \{j\}$. Furthermore, Theorem 5.2 gives

$$\begin{aligned} A_{j,i}^\star &= p_{j,i} D_j^\star + \Theta_{j,i}^\star \circ \lambda_j e \\ &= p_{j,i} \tilde{A}_j^\star + \Theta_{j,i}^\star \circ \lambda_j e + p_{j,i} (\tilde{Q}_j^\star(0) - \tilde{Q}_j^\star) \end{aligned} \quad (5.38)$$

$$\begin{aligned} A_{k,i}^\star &= p_{k,i} D_k^\star + \Theta_{k,i}^\star \circ \lambda_k e \\ &= p_{k,i} \tilde{A}_k^\star + \Theta_{k,i}^\star \circ \lambda_k e + p_{k,i} \gamma_{j,k} (\tilde{Q}_j^\star(0) - \tilde{Q}_j^\star). \end{aligned} \quad (5.39)$$

We utilize the following approximations

$$A_{k,i}^\star \approx p_{k,i} \tilde{A}_k^\star + \Theta_{k,i}^\star \circ \lambda_k e \equiv \tilde{A}_{k,i}^\star \quad (5.40)$$

and

$$p_{j,i} \tilde{Q}_j^\star \approx \psi \left(p_{j,i} \tilde{Q}_j(0) + p_{j,i} A_j^\star + \Theta_{j,i}^\star \circ \lambda_j e - p_{j,i} S_j^\star - p_{j,i} \lambda_j e \right) \equiv \tilde{Q}_{j,i}^\star. \quad (5.41)$$

By Corollary 3.6

$$2\text{cov} \left(\tilde{A}_{k,i}^\star(t), \tilde{A}_{j,i}^\star(t) - \tilde{Q}_{j,i}^\star(t) \right) / (\lambda_i t) = 2 \frac{\zeta_{j,i;k,i}}{\lambda_i} w^\star(t/c_{x,j}^2), \quad (5.42)$$

where $\tilde{A}_{j,i}^\star \equiv p_{j,i} \tilde{A}_j^\star + \Theta_{j,i}^\star \circ \lambda_j e$ and $\zeta_{j,i;k,i}$ is the constant defined as

$$\zeta_{j,i;k,i} = \frac{1}{t} \text{cov} \left(\tilde{A}_{k,i}^\star(t), \tilde{A}_{j,i}^\star(t) \right). \quad (5.43)$$

Note that $\zeta_{j,i;k,i}$ is a constant independent of t since $\tilde{A}_{k,i}^\star(t)$ and $\tilde{A}_{j,i}^\star(t)$ are Brownian motions.

Finally, we define

$$\beta_{j,i;k,i}(t) = \beta_{k,i;j,i}(t) = 2 \frac{\zeta_{j,i;k,i}}{\lambda_i} w^*((1 - \rho_j)^2 p_{j,i} \lambda_j t / (\rho c_{x,j,i}^2)), \quad (5.44)$$

where $c_{x,j,i}^2 = p_{j,i} c_{a,j}^2 + (1 - p_{j,i}) + p_{j,i} c_{s,j}^2$ and $c_{a,j}^2$ is solved from (5.49).

The following lemma gives explicit formula for $\zeta_{j,i;k,i}$. Let $\nu_l \equiv p_{l,i} e'_l (I - P')^{-1}$ for $l = j, k$, where e_i is the i -th unit vector.

Lemma 5.2

$$\zeta_{j,i;k,i} = \nu'_j \left(\text{diag}(c_{a,0,i}^2 \lambda_i) + \sum_{l=1}^K \Sigma_l \right) \nu_k + \nu'_k \Sigma_j e_i + \nu'_j \Sigma_k e_i, \quad (5.45)$$

where $\text{diag}(c_{a,0,i}^2 \lambda_i)$ is the diagonal matrix with $c_{a,0,i}^2 \lambda_i$ as the i -th diagonal entry, Σ_l is the covariance matrix of Brownian limit of the splitting decision process $(\Theta_{l,i}^*)_{i=1}^K$ at station l defined as $\Sigma_l \equiv (\sigma_{i,j}^l)$ with $\sigma_{i,i}^l = p_{l,i}(1 - p_{l,i})\lambda_l$ and $\sigma_{i,j}^l = -p_{l,i}p_{l,j}\lambda_l$ for $i \neq j$.

Proof. By the definition of $\tilde{A}^*$ and $\tilde{A}_{j,i}^*$, we have

$$\begin{aligned} \tilde{A}_{j,i}^* &\equiv p_{j,i} \tilde{A}_j^* + \Theta_{j,i}^* = p_{j,i} e'_j (I - P')^{-1} (A_0^* + (\Theta^*)' \mathbf{1}) + \Theta_{j,i}^* \\ &= \nu_j \left(A_0^* + \sum_{l=1}^K \Theta_l^* \right) + e'_i \Theta_j^*, \\ \tilde{A}_{k,i}^* &\equiv p_{k,i} \tilde{A}_k^* + \Theta_{k,i}^* = p_{k,i} e'_k (I - P')^{-1} (A_0^* + (\Theta^*)' \mathbf{1}) + \Theta_{k,i}^*, \\ &= \nu_k \left(A_0^* + \sum_{l=1}^K \Theta_l^* \right) + e'_i \Theta_k^*, \end{aligned}$$

where A_0^* is the Brownian limit of the external arrival processes, i.e., $A_{0,i}^* \stackrel{d}{=} c_{a,0,i} B_{a,0,i} \circ \lambda_i e$ and $\Theta^* \equiv (\Theta_1^*, \dots, \Theta_K^*)' \in \mathbb{R}^{K \times K}$ with $\Theta_i^* = (\Theta_{i,1}^*, \dots, \Theta_{i,K}^*)$. Recall that Θ_i^* is the collection of the Brownian limits of the decision processes at station i , so that

$$\text{cov}(\Theta_{i,j}^*, \Theta_{i,k}^*) = \begin{cases} p_{i,j}(1 - p_{i,j})\lambda_i t, & j = k, \\ -p_{i,j}p_{i,k}\lambda_i t, & j \neq k. \end{cases}$$

Define

$$\Sigma_i \equiv (\text{cov}(\Theta_{i,j}^*, \Theta_{i,k}^*)/t)_{j,k=1}^K \in \mathbb{R}^{K \times K}$$

so that Σ_i is a constant matrix independent of t .

Notice that $A_{0,i}^*$, Θ_j^* for $1 \leq i, j \leq K$ are mutually independent, we have

$$\begin{aligned}
\zeta_{j,i;k,i} &\equiv \frac{1}{t} \text{cov} \left(\tilde{A}_{k,i}^*(t), \tilde{A}_{j,i}^*(t) \right) \\
&= \frac{1}{t} \text{cov} \left(\nu_j A_0^* + \sum_{l=1}^K (\nu_j + \delta_{l,j} e'_l) \Theta_l^*, \nu_k A_0^* + \sum_{l=1}^K (\nu_k + \delta_{l,k} e'_l) \Theta_l^* \right) \\
&= \frac{1}{t} \text{cov} (\nu_j A_0^*, \nu_k A_0^*) + \frac{1}{t} \sum_{l=1}^K \text{cov} ((\nu_j + \delta_{l,j} e'_l) \Theta_l^*, (\nu_k + \delta_{l,k} e'_l) \Theta_l^*) \\
&= \nu'_j \left(\text{diag}(c_{a_0,i}^2 \lambda_i) + \sum_{l=1}^K \Sigma_l \right) \nu_k + \nu'_k \Sigma_j e_i + \nu'_j \Sigma_k e_i. \quad \blacksquare
\end{aligned}$$

5.4 The IDC Equation System

We are now ready to assemble the building blocks into a system of linear equations (for each t) that describes the IDC's in the OQN. Combining (5.1), (5.10) and (5.28), we obtain *the IDC equations*. These are equations that should be satisfied by the unknown IDCs. For $1 \leq i, j \leq K$, the equations are

$$\begin{aligned}
I_{a,i}(t) &= \sum_{j=1}^K (\lambda_{j,i}/\lambda_i) I_{a,j,i}(t) + (\lambda_{0,i}/\lambda_i) I_{a,0,i}(t) + \beta_i(t), \\
I_{a,i,j}(t) &= p_{i,j} I_{d,i}(t) + (1 - p_{i,j}) + \alpha_{i,j}(t), \\
I_{d,i}(t) &= w_i(t) I_{a,i}(t) + (1 - w_i(t)) I_{s,i}(\rho_i t). \tag{5.46}
\end{aligned}$$

The parameters $p_{i,j}$, $\lambda_{i,j}$ and λ_i are determined by the model primitives and the traffic rate equations in Section 4.1. The IDC's of the external flows $I_{a_0,i}(t)$ and $I_{s,i}(t)$ are assumed to be calculated via exact or numerical inversion of Laplace Transforms, or estimated from data, as in Section 2.3.4. The weight functions $w_i(t)$ is defined in (5.2), which involves a limiting variability parameter $c_{a,i}^2$ as discussed in Lemma 5.1.

To solve for the limiting variability parameters $c_{a,i}^2$, we let $t \rightarrow \infty$ in (5.46) and denote $c_{a,i}^2 \equiv I_{a,i}(\infty)$, $c_{a,i,j}^2 \equiv I_{a,i,j}(\infty)$ and $c_{d,i}^2 \equiv I_{d,i}(\infty)$. Furthermore, we define

$$\begin{aligned}
c_{\alpha_{i,j}}^2 &\equiv \alpha_{i,j}(\infty) = 2\xi_{i,j} p_{i,j} (1 - p_{i,j}), \\
c_{\beta_i}^2 &\equiv \beta_i(\infty) = \frac{2}{\lambda_i} \sum_{j < k} \zeta_{j,i;k,i},
\end{aligned}$$

where we used $w^*(\infty) = 1$ in (5.16) and (5.36). Hence, we have the *limiting variability equations*:

$$\begin{aligned} c_{a,i}^2 &= \sum_{j=1}^K (\lambda_{j,i}/\lambda_i) c_{a,j,i}^2 + (\lambda_{0,i}/\lambda_i) c_{a,0,i}^2 + c_{\beta_i}^2, \\ c_{a,i,j}^2 &= p_{i,j} c_{d,i}^2 + (1 - p_{i,j}) + c_{\alpha_{i,j}}^2, \\ c_{d,i}^2 &= c_{a,i}^2, \quad 1 \leq i \leq K. \end{aligned} \tag{5.47}$$

where we used the fact that $w_i(t) \rightarrow 1$ as $t \rightarrow \infty$.

For a concise matrix notation, let

$$\begin{aligned} \mathbf{I}(t) &\equiv (I_{a,1}(t), \dots, I_{a,K}(t), I_{a,1,1}(t), \dots, I_{a,K,K}(t), I_{d,1}(t), \dots, I_{d,K}(t)), \\ \mathbf{b}(t) &\equiv (b_{a,1}(t), \dots, b_{a,K}(t), b_{a,1,1}(t), \dots, b_{a,K,K}(t), b_{d,1}(t), \dots, b_{d,K}(t)), \\ \mathbf{M}(t) &\equiv (M_{m,n}(t)) \in \mathbb{R}^{(2K+K^2)^2}, \quad m, n \in \{a_1, \dots, a_K, a_{1,1}, \dots, a_{K,K}, d_1, \dots, d_K\}, \\ \mathbf{c}^2 &\equiv (c_{a,1}^2, \dots, c_{a,K}^2, c_{a,1,1}^2, \dots, c_{a,K,K}^2, c_{d,1}^2, \dots, c_{d,K}^2), \end{aligned}$$

where

$$\begin{aligned} b_{a,i}(t) &\equiv \frac{\lambda_{0,i}}{\lambda_i} I_{a,0,i}(t) + \beta_i(t), \quad b_{a,i,j} \equiv (1 - p_{i,j}) + \alpha_{i,j}(t), \\ b_{d,i}(t) &\equiv (1 - w_i(t)) I_{s,i}(t); \quad M_{a_i, a_{j,i}(t)} = \frac{\lambda_{j,i}}{\lambda_i}, \\ M_{a_{i,j}, d_i}(t) &= p_{i,j}, \quad M_{d_i, a_i}(t) = w_i(t), \quad \text{and} \quad M_{m,n}(t) = 0 \text{ otherwise.} \end{aligned}$$

Then the IDC equations can be expressed concisely as

$$(\mathbf{E} - \mathbf{M}(t))\mathbf{I}(t) = \mathbf{b}(t), \tag{5.48}$$

while the limiting variability equations can be expressed as

$$(\mathbf{E} - \mathbf{M}(\infty))\mathbf{c}^2 = \mathbf{b}(\infty), \tag{5.49}$$

where $\mathbf{E} \in \mathbb{R}^{(2K+K^2)^2}$ is the identity matrix.

The following theorem states that these equations have unique solutions.

Theorem 5.4 *Assume that $I - P'$ is invertible. Then $\mathbf{E} - \mathbf{M}(t)$ is invertible for each fixed $t \in \mathbb{R}^+ \cup \{\infty\}$. Hence, for any given t and $\mathbf{b}$, the IDC equations in (5.48) have the unique solution*

$$\mathbf{I}(t) = (\mathbf{E} - \mathbf{M}(t))^{-1} \mathbf{b}(t)$$

and the limiting variability equations in (5.49) have the unique solution

$$\mathbf{c} = (\mathbf{E} - \mathbf{M}(\infty))^{-1} \mathbf{b}(\infty).$$

Proof. Let $\delta_{i,j}$ be the Kronecker delta function. Then substituting the equations for $I_{a,j,i}(t)$ and $I_{d,i}(t)$ into the equation for $I_{a,i}(t)$, we obtain an equation set for $I_{a,i}(t)$ with coefficient matrix $(\delta_{i,j} - (\lambda_{j,i}/\lambda_i)p_{j,i}w_j(t)) \in \mathbb{R}^{K^2}$. Note that $(\lambda_{j,i}/\lambda_i)w_j(t) \leq 1$ for $t \in \mathbb{R}^+ \cup \{\infty\}$, the invertibility of $I - P'$ implies that the equations for $I_{a,i}(t)$ have a unique solution. Substituting in the solution for $I_{a,i}(t)$, we obtain solutions for $I_{a,i,j}(t)$ and $I_{d,i}(t)$.

■

5.5 Feedback Elimination

In this section, we discuss the case in which customers can return (feedback) to a queue after receiving service there. Customer feedback introduces dependence between the arrival process and the service times, even when the service times themselves are mutually independent. As a result, the decomposition $I_w(t) = I_a(t) + c_s^2$ in (2.26) is no longer valid. Indeed, assuming that it is, as we do so far, can introduce serious errors, as we show in our simulation examples. We address this problem by introducing a feedback elimination procedure. We start with the so-called immediate feedback in Section 5.5.1 and generalize it into near-immediate feedback in Section 5.5.2.

5.5.1 Immediate Feedback Elimination

In Section III of [134] it is observed that it is often helpful to pre-process the model data by eliminating immediate feedback for queues with feedback. We now show how that can be done for the RQNA algorithm.

We consider a single queue with i.i.d. feedback. In this case, all feedback is *immediate feedback*, meaning that the customer feeds back to the same queue immediately after completing service, without first going through another service station. For a $GI/GI/1$ model allowing feedback, all feedback is necessarily immediate because there is only one queue.

Normally, the immediate feedback returns the customer back to the end of the queue.

However, in the immediate feedback elimination procedure, the approximation step is to put the customer back at the head of the line so that the customer receives a geometrically random number of service times all at once. Clearly this does not alter the queue length process or the workload process, because the approximation step is work-conserving.

The modified system is a single-server queue with a new service-time distribution and without feedback. Let N_p denote a geometric random variable with success probability $1-p$ and support $\mathbb{N}^+$, the positive natural numbers, then the new service time can be expressed as

$$S_p = \sum_{i=1}^{N_p} S_i, \quad (5.50)$$

where S_i 's are i.i.d. copies of the original service times. This modification in service times results in a change in the service scv. By the conditional variance formula, the scv of the total service time is $\tilde{c}_s^2 = p + (1-p)c_s^2$. The new service IDC in the modified system is the IDC of the stationary renewal process associated with the new service times. To obtain the new service IDC, we need only find the Laplace Transform of the new service distribution, then apply the algorithm in Section 2.3.1.

Let g_p denote the density function of the new service time, we have

$$\begin{aligned} \hat{g}_p(s) &\equiv E \left[\exp \left(-s \sum_{i=1}^{N_p} S_i \right) \right] = E \left[E \left[\exp \left(-s \sum_{i=1}^{N_p} S_i \right) \middle| N_p \right] \right] \\ &= E \left[\prod_{i=1}^{N_p} E [\exp(-s S_i)] \right] = E [\hat{g}^{N_p}(s)] = M_p(\hat{g}(s)), \end{aligned}$$

where $\hat{g}(s)$ is the Laplace transform of the original service distribution and M_p is the probability generating function of the geometric random variable described above.

For the mean waiting time, we need to adjust for per-visit waiting time by multiplying the waiting time in the modified system by $(1-p)$. Note that $(1-p)^{-1}$ is the mean number of visits by a customer in the original system.

In Theorem 4.8, it is shown that the modified system after the immediate feedback elimination procedure shares the same HT limits of the queue length process, the external departure process, the (per-visit) workload process and the (per-visit) waiting time process. Hence, the immediate feedback elimination procedure as an approximation is asymptotically

exact in the heavy-traffic limit.

5.5.2 Near-Immediate Feedback

Now, we consider general OQN's, where the feedback does not necessarily happen immediately, meaning that a departing customer may visit other queues before coming back to the feedback queue. To treat general OQN's, we extend the immediate feedback concept to the *near-immediate feedback*, which depends on the traffic intensities of the queues on the path the customer took before the feedback happens. The near-immediate feedback is defined as any feedback that does not go through any queue with higher traffic intensity.

By default, the RQNA algorithm eliminates all near-immediate feedback. To help understand near-immediate feedback, consider a modified OQN with one bottleneck queue, denoted by h . A *bottleneck queue* is a queue with the highest traffic intensity in the network. While all non-bottleneck queues have service times set to 0 so that they serve as instantaneous switches. In the reduced network, we define an external arrival $\hat{A}_0$ to the bottleneck queue to be any external arrival that arrive at the bottleneck queue for the first time. Hence, an external arrival may have visited one or multiple non-bottleneck queues before its first visit to the bottleneck queue. In particular, the external arrival process can be expressed as the superposition of (i) the original external arrival process $A_{0,h}$ at station h ; and (ii) the Markov splitting of the external arrival process $A_{0,i}$ at station i with probability $\hat{p}_{i,h}$, for $i \neq h$, where $\hat{p}_{i,h}$ denote the probability of a customer that enters the original system at station i ends up visiting the bottleneck station h , see (4.25).

In Theorem 4.8, we showed that this reduced network is asymptotically equivalent in the HT limit to the single-server queue with i.i.d. feedback that we considered in Section 5.5.1. In particular, the arrival process of the equivalent single-station system is $\hat{A}_0$ as described above, the service times remain unchanged and the feedback probability is $\hat{p}$, which is exactly the probability of a near-immediate feedback in the original system, see (4.24). Hence we showed that eliminating all feedback at the bottleneck queue as described above prior to analysis is asymptotically correct in HT for OQN's with a single bottleneck queue in terms of the queue length process, the external departure process, the (per-visit) workload process and the (per-visit) waiting time process. Moreover, the different variants of the

algorithm - eliminating all near immediate feedback or only the near-immediate feedback at the bottleneck queues - are asymptotically exact in the HT limit for an OQN with a single-bottleneck queue, because only the bottleneck queues have nondegenerate HT limit. In contrast, if there are multiple bottleneck queues, the HT limit requires multidimensional RBM, which is not yet incorporated in our RQNA algorithm.

5.6 The RQNA Algorithm

As input parameters, the RQNA algorithm requires the following model primitives

1. The network topology specified by the routing matrix P ;
2. External arrival processes specified by (i) the interarrival-time distribution, if renewal; or (ii) rate λ and IDC; or (iii) a realized sample path of the stationary external arrival process;
3. Service renewal process specified by (i) the service-time distribution; or (ii) the rate and IDC; or (iii) a realized sample path of the service renewal process.

Combining the traffic-rate equation, the limiting variability equation, the IDC equation and the feedback elimination procedure, we have obtained a general framework for the RQNA algorithm, which we summarize in Algorithm 1.

The general framework here allows different choices of (1) the correction terms $\alpha_{i,j}$ in Section 5.2; (2) the correction term β_i in Section 5.3; and (3) the feedback elimination procedure in Section 5.5.

The default settings are discussed in each of these sections. In particular, we use the correction terms in (5.16) and (5.36). For the feedback elimination procedure, we apply near-immediate feedback elimination to all stations.

5.6.1 RQNA for Tree-Structured Queueing Networks

A *tree-structured queueing network* is an OQN whose topology forms a directed tree. Recall that a directed tree is a connected directed graph whose underlying undirected graph is a tree. The queueing network in this setting contains either re-combining after splitting nor

Algorithm 1: A general framework of the RQNA algorithm for the approximation of the system performance measures.

Require: Specification of the correction terms $\alpha_{i,j}(t)$ in Section 5.2 and the correction term β_i in Section 5.3, a set of stations to perform feedback elimination.

Output : Approximation of the system performance measures.

- 1 Solve the traffic rate equations by $\lambda = (I - P')^{-1}\lambda_0$ and let $\rho_i = \lambda_i/\mu_i$;
 - 2 Solve the limiting variability equations by $\mathbf{c} = (\mathbf{E} - \mathbf{M}(\infty))^{-1}\mathbf{b}(\infty)$ specified in Section 5.4;
 - 3 Solve the IDC equations by $\mathbf{I}(t) = (\mathbf{E} - \mathbf{M}(t))^{-1}\mathbf{b}(t)$ for the total arrival IDCs, where we use $\mathbf{c}$ from Step 2 in the weight function w_i ;
 - 4 Select a set of stations to perform feedback elimination, as in Section 5.5. For each selected station, identify the flows to eliminate, then identify the corresponding feedback probability, the modified service IDC as in Section 5.5.1 as well as the reduced network. Repeat Step 1 to Step 3 on the reduced network to obtain the modified IDW (as the sum of the modified total arrival IDC and the modified service scv) at the selected station.
 - 5 Apply the RQ algorithm in (2.35) to obtain the approximations for the mean steady-state workload at each station.
 - 6 Apply the formulas in Section 2.2.8 to obtain approximations for the expected values of the steady-state queue length and waiting time at each queue and the total sojourn time for the system.
-

customer feedback. The tree-structured network is a special case of feed-forward network in which the superposed flows at each node have no common origin.

This special structure greatly simplifies the IDC-based RQNA algorithm. First, feedback elimination is unnecessary because there is no customer feedback. Second, for any internal flow $A_{i,j}$ that is non-zero, we must have $\alpha_{i,j} = 0$ for the correction term in (5.16), because the tree structure implies that the two processes D_i^* and $\Theta_{i,j}^*$ are mutually independent. In particular, by definition,

$$\alpha_{i,j}^*(t) \equiv 2\text{cov}(p_{i,j}D_i^*(t), \Theta_{i,j}^*(\lambda_i t))/E[A_{i,j}^*(t)] = 0.$$

Finally, the tree structure implies that $\beta_i = 0$ for the correction term for superposition because all superposed processes are independent.

With these simplifications of the correction terms, the equations in (5.46), yield, for $1 \leq i, j \leq K$,

$$\begin{aligned} I_{a_i}(t) &= \sum_{j=1}^K \frac{\lambda_{j,i}}{\lambda_i} I_{a_{j,i}}(t) + (\lambda_{0,i}/\lambda_i) I_{a_{0,i}}(t), \\ I_{a_{i,j}}(t) &= p_{i,j} I_{d_i}(t) + (1 - p_{i,j}), \\ I_{d_i}(t) &= w_i(t) I_{a_i}(t) + (1 - w_i(t)) I_{s_i}(t). \end{aligned}$$

The IDC equations in this setting inherit a special structure that allows a recursive algorithm. Note that the stations in the tree-structured network can be partitioned into disjoint layers $\{\mathcal{L}_1, \dots, \mathcal{L}_l\}$ such that for station $i \in \mathcal{L}_k$, it takes only the input flows from $j \in \bigcup_{j=1}^{k-1} \mathcal{L}_j$ for $1 \leq k \leq l$. To simplify the notation, we sort the node in the order of their layers and assign arbitrary order to nodes within the same layer. If $i \in \mathcal{L}_k$, then $\bigcup_{j=1}^{k-1} \mathcal{L}_j \subset \{1, 2, \dots, i-1\}$, so that $\lambda_{j,i} = 0$ for all $j \geq i$. Hence, by substituting in the equations for I_{d_i} and $I_{a_{i,j}}$ into that of I_{a_i} , we have

$$\begin{aligned} I_{a_i}(t) &= \sum_{j=1}^K \frac{\lambda_{j,i}}{\lambda_i} (p_{j,i} (w_j(t) I_{a_j}(t) + (1 - w_j(t)) I_{s_j}(t)) + (1 - p_{j,i})) + \frac{\lambda_{0,i}}{\lambda_i} I_{a_{0,i}}(t), \\ &= \sum_{j < i} \frac{\lambda_{j,i}}{\lambda_i} (p_{j,i} (w_j(t) I_{a_j}(t) + (1 - w_j(t)) I_{s_j}(t)) + (1 - p_{j,i})) + \frac{\lambda_{0,i}}{\lambda_i} I_{a_{0,i}}(t). \end{aligned} \quad (5.51)$$

Note that (5.51) exhibits a lower-triangular shape so that we can explicitly write down the solution in the order of the stations. We summarize the procedure in Algorithm 2.

With the total arrival IDCs, we simply continue to Step 5 and 6 in Algorithm 1 to obtain approximations to the system performance measures.

Algorithm 2: The RQNA algorithm for approximating the IDC's in a tree-structured queueing networks.

Require: The queueing network has tree structure.

Output : Solution to the IDC equations (5.48).

```

1 for  $i = 1$  to  $n$  do
2    $\lambda_i \leftarrow \lambda_{0,i} + \sum_{j < i} \lambda_j p_{j,i};$ 
3    $\rho_i \leftarrow \lambda_i / \mu_i;$ 
4    $c_{a,i}^2 \leftarrow \sum_{j < i} \frac{\lambda_{j,i}}{\lambda_i} c_{a,j,i}^2 + \frac{\lambda_{0,i}}{\lambda_i} c_{a,0,i}^2;$ 
5    $c_{x,i}^2 \leftarrow c_{a,i}^2 + c_{s,i}^2;$ 
6    $w_i(t) \leftarrow w^*((1 - \rho_i)^2 \lambda_i t / (\rho_i c_{x,i}^2));$ 
7    $I_{a,i}(t) \leftarrow$ 
       $\sum_{j < i} \frac{\lambda_{j,i}}{\lambda_i} (p_{j,i} (w_j(t) I_{a,j}(t) + (1 - w_j(t)) I_{s,j}(t)) + (1 - p_{j,i})) + \frac{\lambda_{0,i}}{\lambda_i} I_{a,0,i}(t);$ 
8    $I_{d,i}(t) \leftarrow w_i(t) I_{a,i}(t) + (1 - w_i(t)) I_{s,i}(t);$ 
9   for  $j < i$  do
10     $I_{a,i,j}(t) \leftarrow p_{i,j} I_{d,i}(t) + (1 - p_{i,j});$ 
11  end
12 end
13 return  $\mathbf{I}(t).$ 

```

5.7 Proofs

Proof of Theorem 5.1. We show that the same proof as for Theorem 3.8 can be carried out. To this end, first note that for $\hat{P}_{\mathcal{H}}$ defined in (4.24), one may verify that

$$e'_{\mathcal{H}}(I - P')^{-1} P' e_{\mathcal{H}} + I_{\mathcal{H}} = (I - \hat{P}'_{\mathcal{H}})^{-1}.$$

In particular, with $\mathcal{H} = \{h\}$, we have

$$e'_h(I - P')^{-1} P' e_h + 1 = (I - \hat{P}'_h)^{-1},$$

where $\hat{P}'_h$ is interpreted as the feedback probability at station h and $(I - \hat{P}'_h)^{-1}$ is interpreted as the expected number of visit to station h .

Then we have

$$\begin{aligned}
 D_h^* &= Q_h^*(0) + A_h^* - Q_h^* = Q_h^*(0) + e'_h A^* - Q_h^* \\
 &= e'_h (I - P')^{-1} (A_0^* + (\Theta^*)' \mathbf{1}) + (e'_h P' (I - P')^{-1} e_h + 1) (Q_h^*(0) - Q_h^*) \\
 &= e'_h (I - P')^{-1} (A_0^* + (\Theta^*)' \mathbf{1}) + \frac{1}{1 - \hat{P}'_h} (Q_h^*(0) - Q_h^*). \tag{5.52}
 \end{aligned}$$

Now, let $\tilde{Q}^* \equiv Q_h^*/(1 - \hat{P}'_h)$, then

$$\begin{aligned}
 \tilde{Q}_h^* &= \frac{1}{1 - \hat{P}'_h} Q_h^* \\
 &= \frac{1}{1 - \hat{P}'_h} \Psi \left(Q_h^*(0) + \left(e'_h + \hat{P}'_{\mathcal{H}^c, h} e'_{\mathcal{H}^c} \right) (A_0^* + (\Theta^*)' \mathbf{1}) - (1 - \hat{P}_h) S_h^* - \hat{\lambda}_{0, h} e \right) \\
 &= \Psi \left(\tilde{Q}_h^*(0) + \frac{1}{1 - \hat{P}'_h} \left(e'_h + \hat{P}'_{\mathcal{H}^c, h} e'_{\mathcal{H}^c} \right) (A_0^* + (\Theta^*)' \mathbf{1}) - S_h^* - \lambda_h e \right),
 \end{aligned}$$

where $\hat{\lambda}_{0, h}/(1 - \hat{P}'_h) = \lambda_h$, as in (4.27). Next, one may verify by block-wise inversion that

$$\frac{1}{1 - \hat{P}'_h} \left(e'_h + \hat{P}'_{\mathcal{H}^c, h} e'_{\mathcal{H}^c} \right) \equiv e'_h (I - P')^{-1}.$$

Then we have

$$D_h^* = \tilde{Q}_h^*(0) + \tilde{A}_h^* - \tilde{Q}_h^*$$

with

$$\tilde{Q}_h^* = \Psi \left(\tilde{Q}_h^*(0) + \tilde{A}_h^* - S_h^* - \lambda_h e \right).$$

The limiting variance function is derived in the exact same way as in Theorem 3.8 by noting that $\tilde{A}_h^*$ and S_h^* are two independent Brownian motions. ■

Chapter 6

Simulation Experiments

In this chapter, we present simulation experiments to showcase the performance of our RQ algorithm in 2.35 and (2.34) in Section 2.2 and the RQNA algorithms in Section 5.6 under various settings. We introduce the notation of the distributions and stochastic processes and our simulation methodology in Section 6.1. We study the RQ performance in the single-server $G/GI/1$ models and MAP/MAP/1 models in Section 6.2, the tandem queues models in Section 6.3 and the generalized Jackson networks in Section 6.4.

6.1 Notation and Simulation Methodology

Let us first introduce the notation for some common renewal and non-renewal processes used in this chapter. For renewal processes, we consider the following inter-renewal-time distributions

1. Exponential (M) distribution with mean $1/\lambda$ and scv $c^2 = 1$;
2. Erlang (E_k) distribution with mean $1/\lambda$ and scv $c^2 = 1/k$, i.e. the summation of k i.i.d. exponential random variables, each with mean $1/(\lambda k)$;
3. Hyperexponential ($H_2(c^2, r)$) distribution, i.e., a mixture of two exponential distributions with pdf

$$f(t) \equiv p\lambda_1 e^{-\lambda_1 t} + (1-p)\lambda_2 e^{-\lambda_2 t}, \quad t \geq 0,$$

which can be parameterized by its first three moments or the mean $1/\lambda$, scv c^2 and the ratio between the two components of the mean $r \equiv p_1/\lambda_1/(p_1/\lambda_1 + p_2/\lambda_2)$, where $\lambda_1 > \lambda_2$. If not specified explicitly, we consider the case $r = 0.5$ and reduce the notation to $H_2(c^2)$. This corresponds to stipulating balanced means as in (37) on p. 137 on [133]. The behavior as a function of the third parameter has been studied in [137].

4. Log-normal ($LN(c^2)$) distribution with mean 1 and scv c^2 ; and
5. Gamma ($G(4)$) distribution with mean 1 and scv $c^2 = 4$.

For non-renewal processes, we consider the Markovian Arrival Process (MAP). The MAP is defined in terms of a continuous-time Markov chain with infinitesimal generator $\mathbf{D} = \mathbf{D}_0 + \mathbf{D}_1$, where all the off-diagonal elements of $\mathbf{D}_0$ and all the elements of $\mathbf{D}_1$ are nonnegative. The transitions associated with $\mathbf{D}_1$ are called type 1 transitions. A MAP with parameters $(\mathbf{D}_0, \mathbf{D}_1)$, $\text{MAP}(\mathbf{D}_0, \mathbf{D}_1)$, is a point process where an event occurs when a type 1 transition occurs in the Markov chain. For more properties of the MAP and the calculation of its IDC, see Section 2.3.2.

Simulation estimation of the mean steady-state workload. We estimate the steady-state mean workload by the time-average of the workload process obtained from a single simulation run.

The simulation time required to achieve the same relative width of confidence interval (the width divided by the estimated mean) in the estimation of the steady-state mean workload is roughly $O((1 - \rho)^{-2})$, see [140] for more details. Hence it is significantly harder to obtain accurate simulation estimation for systems in heavy-traffic. In particular, the simulation run length we choose is $6.25 \times 10^6 / (1 - \rho)^2$ time units. This roughly amounts to the same number of arrivals if the arrival process has rate 1. For the system to approach steady-state, the first $1.25 \times 10^6 / (1 - \rho)^2$ time units were discarded.

Table 6.1 shows the statistical precision of our simulation estimation in three $GI/GI/1$ models and one MAP/MAP/1 model. The arrival MAP process is defined by generating

matrices specified below

$$\mathbf{D}_0 = \begin{pmatrix} -13.875 & 1.875 \\ 0.125 & -0.3916 \end{pmatrix}, \quad \mathbf{D}_1 = \begin{pmatrix} 12 & 0 \\ 0 & 0.2666 \end{pmatrix}. \quad (6.1)$$

So that the arrival rate, scv and the lag-1 correlation coefficient (the correlation between two consecutive interarrival times) are $\lambda = 1$, $c^2 = 4.1$ and $\rho = 0.23$. The scv here is defined as the scv of the interarrival time distribution in steady-state, i.e., when the underlying continuous-time Markov chain starts with its stationary distribution. It should be noted that in the presence of positive correlation among the interarrival times, the limiting variability parameter, defined by

$$c_A^2 \equiv I_A(\infty) = \lim_{t \rightarrow \infty} \text{Var}(A(t))/\lambda t, \quad (6.2)$$

is substantially larger than the scv of the interarrival time $c_a^2 = 4.1$, which is $c_A^2 = 9.07$. For the service process, we use the same MAP, but we assume that it is independent of the arrival MAP.

For each model and each ρ in Table 6.1, we perform 100 i.i.d. simulation run and collect the estimation of the mean (i.e., the time average of the workload process). We show the sample mean, the sample standard deviation (SD) and the sample coefficient of variation (CV, sample standard deviation divided by sample mean) of the 100 i.i.d. samples. The reason that we choose to report these instead of the confidence interval is that the distribution of the steady-state workload is not available, even in approximation. For We remark that the sample coefficient of variation is a biased estimation of the population coefficient of variation, see [29]. But we have carefully checked that the bias for the sample size of 100 is negligible in our cases here.

Table 6.1 implies that our choice of simulation run length grant us satisfactory statistical precision for demonstration purposes. Our result here is consistent with Table 1 and (10) in [140].

Simulation estimation of the IDC. For the numerical estimation of the arrival IDC from data, we use a single simulation run of 1.1×10^9 time units and discard the first 10^8 time units for the arrival process to approach stationarity. We then apply the algorithm in Section 2.3.4.

ρ	MAP/MAP/1			$H_2(4)/H_2(4)/1$			$M/M/1$			$E_4/E_4/1$		
	mean	SD	CV	mean	SD	CV	mean	SD	CV	mean	SD	CV
0.01	0.000	0.0000	2.6E-3	0.000	0.0000	3.2E-3	0.000	0.0000	1.1E-3	0.000	0.0000	0.5E-3
0.06	0.020	0.0001	3.2E-3	0.009	0.0000	2.9E-3	0.003	0.0000	1.2E-3	0.002	0.0000	0.5E-3
0.11	0.087	0.0003	3.6E-3	0.036	0.0001	2.9E-3	0.013	0.0000	1.0E-3	0.007	0.0000	0.5E-3
0.16	0.212	0.0008	3.8E-3	0.083	0.0002	2.9E-3	0.030	0.0000	1.0E-3	0.016	0.0000	0.5E-3
0.21	0.409	0.0014	3.4E-3	0.157	0.0004	2.8E-3	0.055	0.0001	1.0E-3	0.028	0.0000	0.4E-3
0.26	0.692	0.0023	3.3E-3	0.264	0.0008	3.0E-3	0.091	0.0001	1.0E-3	0.043	0.0000	0.4E-3
0.31	1.084	0.0038	3.5E-3	0.413	0.0012	2.8E-3	0.139	0.0001	1.0E-3	0.063	0.0000	0.4E-3
0.36	1.608	0.0054	3.3E-3	0.617	0.0018	2.9E-3	0.202	0.0002	1.1E-3	0.087	0.0000	0.4E-3
0.41	2.302	0.0081	3.5E-3	0.891	0.0022	2.5E-3	0.284	0.0003	1.1E-3	0.116	0.0000	0.4E-3
0.46	3.217	0.0124	3.8E-3	1.257	0.0029	2.3E-3	0.391	0.0004	1.0E-3	0.153	0.0001	0.4E-3
0.51	4.418	0.0130	2.9E-3	1.745	0.0042	2.4E-3	0.530	0.0005	1.0E-3	0.198	0.0001	0.4E-3
0.56	6.000	0.0171	2.8E-3	2.399	0.0053	2.2E-3	0.712	0.0007	1.0E-3	0.254	0.0001	0.4E-3
0.61	8.125	0.0240	3.0E-3	3.286	0.0070	2.1E-3	0.954	0.0010	1.0E-3	0.326	0.0001	0.4E-3
0.66	11.028	0.0359	3.3E-3	4.511	0.0090	2.0E-3	1.281	0.0015	1.2E-3	0.420	0.0002	0.4E-3
0.71	15.101	0.0419	2.8E-3	6.255	0.0111	1.8E-3	1.738	0.0019	1.1E-3	0.546	0.0002	0.4E-3
0.76	21.090	0.0627	3.0E-3	8.838	0.0183	2.1E-3	2.406	0.0024	1.0E-3	0.726	0.0003	0.4E-3
0.81	30.472	0.0916	3.0E-3	12.931	0.0290	2.2E-3	3.452	0.0032	0.9E-3	1.001	0.0005	0.5E-3
0.86	46.981	0.1233	2.6E-3	20.150	0.0386	1.9E-3	5.283	0.0046	0.9E-3	1.472	0.0006	0.4E-3
0.91	82.425	0.2338	2.8E-3	35.722	0.0617	1.7E-3	9.201	0.0085	0.9E-3	2.465	0.0010	0.4E-3
0.96	207.886	0.5351	2.6E-3	91.012	0.1699	1.9E-3	23.039	0.0207	0.9E-3	5.939	0.0027	0.5E-3

Table 6.1: The sample mean, sample standard deviation (SD) and sample coefficient of variation (CV) of the mean estimator for the workload in various single-server queues.

Note that the same arrival IDC is used for all service-time distributions and all choices of ρ as long as the arrival process remains the same, we can effectively maintain a list estimated/calculated renewal IDC for the RQ algorithm and assumes that the IDC is readily available. For processes with non-unit rates, the IDC's can be obtained as in Remark 2.4.

Simulation complexity. For the estimation of both the mean steady-state workload and the IDC, the simulation is implemented in C++ on a PC with a 4.8GHz Intel CPU.

For mean steady-state workload, it takes on average 30 minutes to obtain simulation estimations for 20 instances of $\rho \in \{0.01, 0.06, \dots, 0.91, 0.96\}$ for a single-station queue. About 70% of the simulation time is devoted to the case $\rho = 0.96$. The computation time scales roughly linearly with respect to the number of stations in the network.

For the simulation estimation of the IDC, the time required for each case is on average 8

minutes. Each case requires 8GB of RAM memory since we trade space for time by storing the entire sample path to speed up the estimation, see Section 2.3.4.

RQ and RQNA algorithms are implemented in MATLAB R2018b, both on the same PC with a 4.8GHz CPU. Assuming that the IDC of the external flows (see Chapter 5 for definitions) are available, both algorithms takes less than 1 second to obtain approximations for each network considered in this chapter, which is negligible in compare with the CPU time required for simulation estimations. As discussed above, we can maintain a list estimated/calculated renewal IDC for the RQ algorithm.

6.2 Robust Queueing for Single-Server Queues

In this section, we apply the RQ algorithm in (2.35) to single-server queues. The station under consideration may be any station within a larger queueing network. Towards this end, we assume that the arrival IDC $I_a(t) \equiv I_A(t)$ is obtained from simulation or numerical calculations as discussed in Section 2.3.

6.2.1 The $GI/GI/1$ Models

To start, we consider the $GI/GI/1$ models, where we have one single-server station with renewal arrival process and i.i.d. service times that are independent of the arrival process.

For the interarrival-time distribution, we consider the following cases: E_4 , $LN(0.25)$, $H_2(4)$, $LN(4)$ and $G(4)$ (see Section 6.1 for the definition of the distributions). For service-time distribution, we also consider exponential (M) distribution. We fix the arrival rate $\lambda = 1$. We do not include the Poisson arrival case here, because exact formula is available for performance measures in $M/GI/1$ queue, and in this case RQ produce exact values, see Corollary 2.3.

For each combination of the 5 interarrival-time distributions and 6 service-time distributions, we consider 20 traffic intensity levels $\rho \in \{0.01, 0.06, 0.11, \dots, 0.96\}$. Since we have fixed a rate-1 external arrival process, the service rate $\mu = 1/\rho$. For each traffic intensity, we compare the simulation estimation of the mean steady-state workload and the RQ approximation from (2.35). Note that the arrival processes are renewal, we calculate the arrival

IDC by the numerical inversion method for $G(4)$, by estimation from data for $LN(0.25)$ and $LN(4)$ and by exact formula for $E_4, H_2(4)$ and M , as discussed in Section 2.3.1.

We now compare the simulated mean steady-state workload and the RQ approximation (2.35). Let $Z(\rho)$ denote the steady-state workload and let $Z^*(\rho)$ denote the RQ solution from (2.17), for system at traffic intensity ρ . Let $c_{Z^*}^2(\rho)$ be the normalized RQ workload in (2.36) and let $c_Z^2(\rho)$ denote the normalized steady-state mean workload in (2.27), again for system at traffic intensity ρ . We define the relative error $RE(\rho)$ by

$$RE(\rho) \equiv \frac{c_{Z^*}^2(\rho) - c_Z^2(\rho)}{c_Z^2(\rho)}. \quad (6.3)$$

In Table 6.2 - 6.6, the top halves show the simulation estimation of $Z(\rho)$, the RQ solution $Z^*(\rho)$ and the relative error $RE(\rho)$ under various settings, whereas the bottom halves show the normalized version $c_Z^2(\rho)$ and $c_{Z^*}^2(\rho)$. For each table, we fix the service-time distribution and consider all 5 arrival processes and all traffic intensity levels.

Table 6.7 and Table 6.8 present the maximum (absolute) relative error and absolute error, defined as

$$\max_k |RE(\rho_k)| \quad \text{and} \quad \max_k |c_{Z^*}^2(\rho_k) - c_Z^2(\rho_k)|, \quad \text{for } \rho_k = -0.04 + 0.05k. \quad (6.4)$$

Figure 6.1 display two of the worst case arrival process in terms of the performance of the RQ approximation, i.e. the $LN(4)$ and $G(4)$ arrival process. Figure 6.2 displays the RQ approximation in the overall worst case service-time distribution, i.e. the E_4 service times.

We make the following observations.

1. For a fixed arrival process, the performance of RQ is not very sensitive to the service-time distribution beyond its scv. For example, compare Table 6.2 with Table 6.3 for service scv $c_s^2 = 0.25$, or compare Table 6.5 with Table 6.6 for service scv $c_s^2 = 4$. Recall that for $GI/GI/1$ model, the RQ solution depends on the service-time distribution only through its scv. This is demonstrate by Figure 6.1, where we plot the normalized mean workload $c_Z^2(\rho)$ and the corresponding RQ approximations $c_{Z^*}^2(\rho)$ as functions of ρ , for $LN(4)/GI/1$ and $G(4)/GI/1$ models with various service-time distributions. On the other hand, Figure 6.2 show that the arrival process have much more interesting impact on the system performance measure.

2. The RQ performance improves as the service scv increases. In general, the lower the scv of a service-time distribution, the smaller the normalized steady-state workload. But from Table 6.8, the absolute error remain similar for different service distribution, hence the relative error in cases with E_4 and $LN(0.25)$ service distributions tend to be larger, which can be as large as 30% (in the case of $E_4/E_4/1$ at $\rho = 0.41$). As an illustration, in Figure 6.1 we compare the simulated values and the RQ approximation in the $LN(4)/GI/1$ and $G(4)/GI/1$ model with various service-time distributions. We see that the absolute error remain the similar for different service-time distributions, but the relative error is higher for cases with lower service scv.
3. In Table 6.2 - 6.6, we see that the large relative errors are often accompanied by very small absolute error. In Table 6.7, we illustrate this observation by showing two cases: the full range of $\rho \in (0, 1)$ and the restricted range of $\rho > 0.5$. This comparison eliminates the unusually large relative error caused by the small denominator when ρ is small. In the reference case of $M/M/1$ the mean steady-state workload and waiting time at $\rho = 0.5$ is exactly the same as the mean service time. We see that when considering only the moderate to high traffic intensity levels, RQ produces a much better approximation. See Figure 6.1(top) for the cases that benefit the most.

6.2.2 A Queue with a Superposition Arrival Process

We now illustrate the performance of our RQ approach for a common but challenging network structure in Figure 6.3. This specific example is chosen to capture a known source of difficulty: the complex dependence in the arrival process to the queue, so that the relevant variability parameter of the arrival process at the queue can depend strongly on the traffic intensity of that queue, as discussed in [142].

Consider a balanced $\sum_i G_i/GI/1$ model from Section 2.3.3, where (2.67) can be applied. Let the rate-1 arrival process A be the superposition of $n = 10$ i.i.d. renewal processes, each with rate $1/n$, where the times between renewals have a lognormal distribution with mean n and $\text{scv } c_a^2 = 10$. Let the service-times distribution be hyperexponential (H_2), a mixture of two exponential distributions) with mean 1, $c_s^2 = 2$ and balanced means as on

	E_4 arrival		$LN(0.25)$ arrival		$H_2(4)$ arrival		$LN(4)$ arrival		G_4 arrival	
	Sim	RQ	Sim	RQ	Sim	RQ	Sim	RQ	Sim	RQ
Mean steady-state workload, E_4 service	0.01	0.00 0.00(1%)	0.00	0.00(1%)	0.00	0.00(-1%)	0.00	0.00(1%)	0.00	0.00(-24%)
	0.06	0.00 0.00(6%)	0.00	0.00(6%)	0.00	0.00(-4%)	0.00	0.00(0%)	0.00	0.00(-30%)
	0.11	0.01 0.01(11%)	0.01	0.01(11%)	0.01	0.01(-7%)	0.01	0.01(-3%)	0.02	0.01(-29%)
	0.16	0.02 0.02(16%)	0.02	0.02(17%)	0.02	0.02(-10%)	0.02	0.02(-8%)	0.04	0.03(-28%)
	0.21	0.03 0.03(21%)	0.03	0.03(22%)	0.04	0.04(-13%)	0.04	0.03(-13%)	0.08	0.06(-25%)
	0.26	0.04 0.05(24%)	0.04	0.05(26%)	0.07	0.06(-16%)	0.07	0.06(-17%)	0.14	0.11(-22%)
	0.31	0.06 0.08(27%)	0.06	0.08(29%)	0.11	0.09(-19%)	0.11	0.09(-19%)	0.22	0.18(-18%)
	0.36	0.09 0.11(28%)	0.09	0.11(30%)	0.18	0.14(-20%)	0.17	0.13(-20%)	0.33	0.28(-15%)
	0.41	0.12 0.15(27%)	0.11	0.15(30%)	0.27	0.21(-21%)	0.25	0.21(-18%)	0.48	0.43(-10%)
	0.46	0.15 0.19(25%)	0.15	0.19(28%)	0.40	0.33(-18%)	0.37	0.31(-14%)	0.68	0.63(-7%)
	0.51	0.20 0.24(22%)	0.19	0.24(24%)	0.59	0.53(-10%)	0.53	0.48(-8%)	0.95	0.91(-3%)
	0.56	0.25 0.30(17%)	0.25	0.29(19%)	0.86	0.87(1%)	0.75	0.74(-2%)	1.30	1.30(-0%)
	0.61	0.33 0.36(10%)	0.32	0.35(11%)	1.25	1.39(11%)	1.07	1.13(6%)	1.78	1.82(2%)
	0.66	0.42 0.43(3%)	0.41	0.42(3%)	1.82	2.14(17%)	1.54	1.72(12%)	2.44	2.53(4%)
	0.71	0.55 0.53(-4%)	0.53	0.51(-4%)	2.67	3.18(19%)	2.23	2.63(18%)	3.37	3.52(5%)
	0.76	0.73 0.68(-7%)	0.71	0.66(-6%)	3.98	4.69(18%)	3.31	4.04(22%)	4.76	4.97(4%)
	0.81	1.00 0.92(-8%)	0.98	0.91(-7%)	6.07	7.00(15%)	5.10	6.32(24%)	6.92	7.23(4%)
	0.86	1.47 1.36(-7%)	1.44	1.36(-6%)	9.85	10.98(11%)	8.44	10.37(23%)	10.79	11.14(3%)
	0.91	2.46 2.33(-6%)	2.43	2.32(-4%)	18.10	19.39(7%)	15.99	18.94(18%)	19.07	19.50(2%)
	0.96	5.94 5.77(-3%)	5.90	5.77(-2%)	47.32	48.89(3%)	44.27	48.54(10%)	48.47	48.94(1%)
Normalized mean steady-state workload, E_4 service	0.01	1.24 1.25(1%)	1.24	1.25(1%)	1.26	1.25(-1%)	1.24	1.25(1%)	1.80	1.37(-24%)
	0.06	1.18 1.25(6%)	1.18	1.25(6%)	1.30	1.25(-4%)	1.24	1.24(0%)	2.27	1.60(-30%)
	0.11	1.11 1.24(11%)	1.11	1.24(11%)	1.35	1.26(-7%)	1.28	1.24(-3%)	2.53	1.79(-29%)
	0.16	1.06 1.23(16%)	1.05	1.23(17%)	1.40	1.26(-10%)	1.34	1.23(-8%)	2.72	1.97(-28%)
	0.21	1.00 1.21(21%)	0.99	1.21(22%)	1.47	1.28(-13%)	1.41	1.23(-13%)	2.89	2.16(-25%)
	0.26	0.95 1.18(24%)	0.94	1.18(26%)	1.55	1.30(-16%)	1.48	1.24(-17%)	3.03	2.36(-22%)
	0.31	0.90 1.14(27%)	0.89	1.14(29%)	1.64	1.33(-19%)	1.57	1.27(-19%)	3.15	2.57(-18%)
	0.36	0.86 1.10(28%)	0.84	1.10(30%)	1.75	1.39(-20%)	1.67	1.33(-20%)	3.27	2.79(-15%)
	0.41	0.82 1.04(27%)	0.80	1.04(30%)	1.88	1.49(-21%)	1.76	1.44(-18%)	3.36	3.01(-10%)
	0.46	0.78 0.98(25%)	0.76	0.97(28%)	2.04	1.67(-18%)	1.87	1.60(-14%)	3.47	3.23(-7%)
	0.51	0.75 0.91(22%)	0.72	0.90(24%)	2.21	1.99(-10%)	1.98	1.82(-8%)	3.57	3.44(-3%)
	0.56	0.72 0.83(17%)	0.69	0.82(19%)	2.40	2.44(1%)	2.11	2.08(-2%)	3.65	3.64(-0%)
	0.61	0.69 0.75(10%)	0.66	0.74(11%)	2.63	2.91(11%)	2.25	2.37(6%)	3.73	3.81(2%)
	0.66	0.66 0.67(3%)	0.64	0.65(3%)	2.85	3.33(17%)	2.40	2.69(12%)	3.81	3.95(4%)
	0.71	0.63 0.60(-4%)	0.61	0.59(-4%)	3.07	3.66(19%)	2.57	3.03(18%)	3.88	4.05(5%)
	0.76	0.60 0.56(-7%)	0.59	0.55(-6%)	3.31	3.90(18%)	2.75	3.35(22%)	3.96	4.13(4%)
	0.81	0.58 0.53(-8%)	0.57	0.53(-7%)	3.51	4.06(15%)	2.96	3.66(24%)	4.01	4.19(4%)
	0.86	0.56 0.52(-7%)	0.55	0.51(-6%)	3.73	4.16(11%)	3.19	3.92(23%)	4.08	4.22(3%)
	0.91	0.54 0.51(-6%)	0.53	0.51(-4%)	3.94	4.22(7%)	3.48	4.12(18%)	4.15	4.24(2%)
	0.96	0.52 0.50(-3%)	0.51	0.50(-2%)	4.11	4.24(3%)	3.84	4.21(10%)	4.21	4.25(1%)

Table 6.2: The RQ approximation in various $GI/E_4/1$ models.

	E_4 arrival		$LN(0.25)$ arrival		$H_2(4)$ arrival		$LN(4)$ arrival		G_4 arrival	
ρ	Sim	RQ	Sim	RQ	Sim	RQ	Sim	RQ	Sim	RQ
Mean steady-state workload, $LN(0.25)$ service	0.01	0.00 0.00(1%)	0.00 0.00(1%)	0.00 0.00(1%)	0.00 0.00(-1%)	0.00 0.00(1%)	0.00 0.00(1%)	0.00 0.00(-24%)		
	0.06	0.00 0.00(6%)	0.00 0.00(6%)	0.00 0.00(6%)	0.00 0.00(-4%)	0.00 0.00(0%)	0.00 0.00(0%)	0.00 0.00(-30%)		
	0.11	0.01 0.01(11%)	0.01 0.01(11%)	0.01 0.01(11%)	0.01 0.01(-7%)	0.01 0.01(-3%)	0.01 0.01(-3%)	0.02 0.01(-29%)		
	0.16	0.02 0.02(16%)	0.02 0.02(17%)	0.02 0.02(17%)	0.02 0.02(-10%)	0.02 0.02(-8%)	0.02 0.02(-8%)	0.04 0.03(-28%)		
	0.21	0.03 0.03(20%)	0.03 0.03(21%)	0.03 0.03(21%)	0.04 0.04(-13%)	0.04 0.03(-13%)	0.04 0.03(-13%)	0.08 0.06(-25%)		
	0.26	0.04 0.05(24%)	0.04 0.05(25%)	0.04 0.05(25%)	0.07 0.06(-16%)	0.07 0.06(-17%)	0.07 0.06(-17%)	0.14 0.11(-22%)		
	0.31	0.06 0.08(26%)	0.06 0.08(28%)	0.06 0.08(28%)	0.11 0.09(-19%)	0.11 0.09(-19%)	0.11 0.09(-19%)	0.22 0.18(-18%)		
	0.36	0.09 0.11(27%)	0.09 0.11(30%)	0.09 0.11(30%)	0.18 0.14(-21%)	0.17 0.13(-20%)	0.17 0.13(-20%)	0.33 0.28(-14%)		
	0.41	0.12 0.15(26%)	0.11 0.15(29%)	0.11 0.15(29%)	0.27 0.21(-21%)	0.25 0.21(-18%)	0.25 0.21(-18%)	0.48 0.43(-11%)		
	0.46	0.15 0.19(24%)	0.15 0.19(27%)	0.15 0.19(27%)	0.40 0.33(-18%)	0.37 0.31(-14%)	0.37 0.31(-14%)	0.68 0.63(-7%)		
	0.51	0.20 0.24(21%)	0.19 0.24(23%)	0.19 0.24(23%)	0.59 0.53(-10%)	0.53 0.48(-9%)	0.53 0.48(-9%)	0.94 0.91(-3%)		
	0.56	0.26 0.30(16%)	0.25 0.29(18%)	0.25 0.29(18%)	0.86 0.87(1%)	0.75 0.74(-2%)	0.75 0.74(-2%)	1.30 1.30(-0%)		
	0.61	0.33 0.36(10%)	0.32 0.35(11%)	0.32 0.35(11%)	1.25 1.39(11%)	1.07 1.13(5%)	1.07 1.13(5%)	1.78 1.82(2%)		
	0.66	0.42 0.43(2%)	0.41 0.42(2%)	0.41 0.42(2%)	1.82 2.14(17%)	1.54 1.72(12%)	1.54 1.72(12%)	2.43 2.53(4%)		
	0.71	0.55 0.53(-4%)	0.53 0.51(-4%)	0.53 0.51(-4%)	2.67 3.18(19%)	2.23 2.63(18%)	2.23 2.63(18%)	3.38 3.52(4%)		
	0.76	0.73 0.68(-7%)	0.71 0.66(-6%)	0.71 0.66(-6%)	3.97 4.69(18%)	3.30 4.04(22%)	3.30 4.04(22%)	4.75 4.97(5%)		
	0.81	1.01 0.92(-8%)	0.98 0.91(-7%)	0.98 0.91(-7%)	6.06 7.00(16%)	5.10 6.32(24%)	5.10 6.32(24%)	6.95 7.23(4%)		
0.86	1.48 1.36(-8%)	1.45 1.36(-6%)	1.45 1.36(-6%)	9.82 10.98(12%)	8.44 10.37(23%)	8.44 10.37(23%)	10.75 11.14(4%)			
0.91	2.47 2.33(-6%)	2.44 2.32(-5%)	2.44 2.32(-5%)	18.03 19.39(8%)	16.01 18.94(18%)	16.01 18.94(18%)	19.06 19.50(2%)			
0.96	5.95 5.77(-3%)	5.90 5.77(-2%)	5.90 5.77(-2%)	47.28 48.89(3%)	44.20 48.54(10%)	44.20 48.54(10%)	48.49 48.94(1%)			
Normalized mean steady-state workload, $LN(0.25)$ service	0.01	1.24 1.25(1%)	1.24 1.25(1%)	1.24 1.25(1%)	1.26 1.25(-1%)	1.24 1.25(1%)	1.24 1.25(1%)	1.80 1.37(-24%)		
	0.06	1.18 1.25(6%)	1.18 1.25(6%)	1.18 1.25(6%)	1.30 1.25(-4%)	1.24 1.24(0%)	1.24 1.24(0%)	2.27 1.60(-30%)		
	0.11	1.12 1.24(11%)	1.11 1.24(11%)	1.11 1.24(11%)	1.35 1.26(-7%)	1.28 1.24(-3%)	1.28 1.24(-3%)	2.53 1.79(-29%)		
	0.16	1.06 1.23(16%)	1.05 1.23(17%)	1.05 1.23(17%)	1.40 1.26(-10%)	1.34 1.23(-8%)	1.34 1.23(-8%)	2.73 1.97(-28%)		
	0.21	1.00 1.21(20%)	0.99 1.21(21%)	0.99 1.21(21%)	1.47 1.28(-13%)	1.41 1.23(-13%)	1.41 1.23(-13%)	2.89 2.16(-25%)		
	0.26	0.95 1.18(24%)	0.94 1.18(25%)	0.94 1.18(25%)	1.55 1.30(-16%)	1.49 1.24(-17%)	1.49 1.24(-17%)	3.03 2.36(-22%)		
	0.31	0.91 1.14(26%)	0.89 1.14(28%)	0.89 1.14(28%)	1.64 1.33(-19%)	1.57 1.27(-19%)	1.57 1.27(-19%)	3.15 2.57(-18%)		
	0.36	0.86 1.10(27%)	0.85 1.10(30%)	0.85 1.10(30%)	1.75 1.39(-21%)	1.66 1.33(-20%)	1.66 1.33(-20%)	3.26 2.79(-14%)		
	0.41	0.82 1.04(26%)	0.80 1.04(29%)	0.80 1.04(29%)	1.88 1.49(-21%)	1.76 1.44(-18%)	1.76 1.44(-18%)	3.38 3.01(-11%)		
	0.46	0.79 0.98(24%)	0.77 0.97(27%)	0.77 0.97(27%)	2.03 1.67(-18%)	1.87 1.60(-14%)	1.87 1.60(-14%)	3.48 3.23(-7%)		
	0.51	0.75 0.91(21%)	0.73 0.90(23%)	0.73 0.90(23%)	2.21 1.99(-10%)	1.99 1.82(-9%)	1.99 1.82(-9%)	3.56 3.44(-3%)		
	0.56	0.72 0.83(16%)	0.70 0.82(18%)	0.70 0.82(18%)	2.41 2.44(1%)	2.11 2.08(-2%)	2.11 2.08(-2%)	3.65 3.64(-0%)		
	0.61	0.69 0.75(10%)	0.67 0.74(11%)	0.67 0.74(11%)	2.62 2.91(11%)	2.25 2.37(5%)	2.25 2.37(5%)	3.72 3.81(2%)		
	0.66	0.66 0.67(2%)	0.64 0.65(2%)	0.64 0.65(2%)	2.85 3.33(17%)	2.40 2.69(12%)	2.40 2.69(12%)	3.80 3.95(4%)		
	0.71	0.63 0.60(-4%)	0.61 0.59(-4%)	0.61 0.59(-4%)	3.08 3.66(19%)	2.56 3.03(18%)	2.56 3.03(18%)	3.89 4.05(4%)		
	0.76	0.61 0.56(-7%)	0.59 0.55(-6%)	0.59 0.55(-6%)	3.30 3.90(18%)	2.74 3.35(22%)	2.74 3.35(22%)	3.95 4.13(5%)		
	0.81	0.58 0.53(-8%)	0.57 0.53(-7%)	0.57 0.53(-7%)	3.51 4.06(16%)	2.95 3.66(24%)	2.95 3.66(24%)	4.03 4.19(4%)		
0.86	0.56 0.52(-8%)	0.55 0.51(-6%)	0.55 0.51(-6%)	3.72 4.16(12%)	3.20 3.92(23%)	3.20 3.92(23%)	4.07 4.22(4%)			
0.91	0.54 0.51(-6%)	0.53 0.51(-5%)	0.53 0.51(-5%)	3.92 4.22(8%)	3.48 4.12(18%)	3.48 4.12(18%)	4.14 4.24(2%)			
0.96	0.52 0.50(-3%)	0.51 0.50(-2%)	0.51 0.50(-2%)	4.10 4.24(3%)	3.84 4.21(10%)	3.84 4.21(10%)	4.21 4.25(1%)			

Table 6.3: The RQ approximation in various $GI/LN(0.25)/1$ models.

	E_4 arrival		$LN(0.25)$ arrival		$H_2(4)$ arrival		$LN(4)$ arrival		G_4 arrival	
ρ	Sim	RQ	Sim	RQ	Sim	RQ	Sim	RQ	Sim	RQ
Mean steady-state workload, M service	0.01	0.00 0.00(1%)	0.00 0.00(1%)	0.00 0.00(1%)	0.00 0.00(-1%)	0.00 0.00(1%)	0.00 0.00(1%)	0.00 0.00(-18%)		
	0.06	0.00 0.00(6%)	0.00 0.00(6%)	0.00 0.00(6%)	0.00 0.00(-4%)	0.00 0.00(-1%)	0.00 0.00(-1%)	0.01 0.00(-23%)		
	0.11	0.01 0.01(11%)	0.01 0.01(11%)	0.01 0.01(11%)	0.01 0.01(-6%)	0.01 0.01(-5%)	0.01 0.01(-5%)	0.02 0.02(-23%)		
	0.16	0.03 0.03(14%)	0.03 0.03(15%)	0.03 0.03(15%)	0.03 0.03(-10%)	0.03 0.03(-9%)	0.03 0.03(-9%)	0.05 0.04(-22%)		
	0.21	0.05 0.05(16%)	0.05 0.05(17%)	0.05 0.05(17%)	0.07 0.06(-12%)	0.06 0.06(-13%)	0.06 0.06(-13%)	0.10 0.08(-20%)		
	0.26	0.07 0.09(18%)	0.07 0.09(19%)	0.07 0.09(19%)	0.11 0.09(-15%)	0.11 0.09(-15%)	0.11 0.09(-15%)	0.18 0.15(-17%)		
	0.31	0.11 0.13(17%)	0.11 0.13(19%)	0.11 0.13(19%)	0.18 0.15(-17%)	0.17 0.14(-16%)	0.17 0.14(-16%)	0.28 0.24(-14%)		
	0.36	0.15 0.18(16%)	0.15 0.18(17%)	0.15 0.18(17%)	0.27 0.22(-18%)	0.26 0.22(-15%)	0.26 0.22(-15%)	0.42 0.37(-11%)		
	0.41	0.21 0.24(13%)	0.21 0.24(14%)	0.21 0.24(14%)	0.41 0.34(-17%)	0.38 0.33(-13%)	0.38 0.33(-13%)	0.60 0.55(-8%)		
	0.46	0.28 0.31(9%)	0.28 0.31(9%)	0.28 0.31(9%)	0.59 0.51(-14%)	0.55 0.49(-10%)	0.55 0.49(-10%)	0.84 0.80(-5%)		
	0.51	0.38 0.40(4%)	0.37 0.39(4%)	0.37 0.39(4%)	0.85 0.79(-7%)	0.77 0.72(-6%)	0.77 0.72(-6%)	1.16 1.14(-3%)		
	0.56	0.50 0.50(0%)	0.49 0.49(-0%)	0.49 0.49(-0%)	1.21 1.21(0%)	1.08 1.06(-1%)	1.08 1.06(-1%)	1.59 1.59(-0%)		
	0.61	0.66 0.64(-2%)	0.65 0.64(-2%)	0.65 0.64(-2%)	1.71 1.83(7%)	1.51 1.56(3%)	1.51 1.56(3%)	2.16 2.20(2%)		
	0.66	0.87 0.84(-3%)	0.86 0.84(-3%)	0.86 0.84(-3%)	2.43 2.70(11%)	2.12 2.30(8%)	2.12 2.30(8%)	2.96 3.04(3%)		
	0.71	1.16 1.12(-3%)	1.15 1.12(-3%)	1.15 1.12(-3%)	3.46 3.91(13%)	3.01 3.39(13%)	3.01 3.39(13%)	4.06 4.20(4%)		
	0.76	1.59 1.53(-3%)	1.58 1.53(-3%)	1.58 1.53(-3%)	5.02 5.66(13%)	4.37 5.06(16%)	4.37 5.06(16%)	5.68 5.90(4%)		
	0.81	2.25 2.18(-3%)	2.24 2.18(-3%)	2.24 2.18(-3%)	7.54 8.35(11%)	6.63 7.74(17%)	6.63 7.74(17%)	8.29 8.54(3%)		
	0.86	3.40 3.32(-2%)	3.39 3.32(-2%)	3.39 3.32(-2%)	12.00 13.00(8%)	10.70 12.46(16%)	10.70 12.46(16%)	12.82 13.14(2%)		
	0.91	5.86 5.76(-2%)	5.84 5.76(-1%)	5.84 5.76(-1%)	21.67 22.87(6%)	19.85 22.48(13%)	19.85 22.48(13%)	22.57 22.96(2%)		
	0.96	14.54 14.40(-1%)	14.51 14.40(-1%)	14.51 14.40(-1%)	56.25 57.54(2%)	53.45 57.21(7%)	53.45 57.21(7%)	57.16 57.58(1%)		
Normalized mean steady-state workload, M service	0.01	1.98 2.00(1%)	1.98 2.00(1%)	1.98 2.00(1%)	2.01 2.00(-1%)	1.99 2.00(1%)	1.99 2.00(1%)	2.61 2.14(-18%)		
	0.06	1.88 2.00(6%)	1.88 2.00(6%)	1.88 2.00(6%)	2.08 2.00(-4%)	2.02 1.99(-1%)	2.02 1.99(-1%)	3.10 2.39(-23%)		
	0.11	1.79 1.98(11%)	1.79 1.98(11%)	1.79 1.98(11%)	2.15 2.01(-6%)	2.09 1.99(-5%)	2.09 1.99(-5%)	3.37 2.60(-23%)		
	0.16	1.72 1.96(14%)	1.71 1.96(15%)	1.71 1.96(15%)	2.24 2.02(-10%)	2.18 1.98(-9%)	2.18 1.98(-9%)	3.57 2.79(-22%)		
	0.21	1.66 1.93(16%)	1.65 1.93(17%)	1.65 1.93(17%)	2.33 2.04(-12%)	2.27 1.98(-13%)	2.27 1.98(-13%)	3.74 3.00(-20%)		
	0.26	1.61 1.89(18%)	1.59 1.89(19%)	1.59 1.89(19%)	2.44 2.08(-15%)	2.36 2.01(-15%)	2.36 2.01(-15%)	3.88 3.21(-17%)		
	0.31	1.56 1.83(17%)	1.54 1.83(19%)	1.54 1.83(19%)	2.56 2.13(-17%)	2.46 2.07(-16%)	2.46 2.07(-16%)	3.99 3.42(-14%)		
	0.36	1.52 1.76(16%)	1.50 1.76(17%)	1.50 1.76(17%)	2.70 2.22(-18%)	2.56 2.17(-15%)	2.56 2.17(-15%)	4.10 3.64(-11%)		
	0.41	1.49 1.68(13%)	1.47 1.67(14%)	1.47 1.67(14%)	2.85 2.36(-17%)	2.67 2.31(-13%)	2.67 2.31(-13%)	4.20 3.86(-8%)		
	0.46	1.45 1.59(9%)	1.43 1.57(9%)	1.43 1.57(9%)	3.02 2.60(-14%)	2.78 2.50(-10%)	2.78 2.50(-10%)	4.30 4.08(-5%)		
	0.51	1.43 1.49(4%)	1.41 1.46(4%)	1.41 1.46(4%)	3.20 2.96(-7%)	2.90 2.72(-6%)	2.90 2.72(-6%)	4.39 4.28(-3%)		
	0.56	1.40 1.41(0%)	1.38 1.38(-0%)	1.38 1.38(-0%)	3.39 3.40(0%)	3.03 2.99(-1%)	3.03 2.99(-1%)	4.47 4.46(-0%)		
	0.61	1.38 1.35(-2%)	1.36 1.34(-2%)	1.36 1.34(-2%)	3.59 3.84(7%)	3.17 3.28(3%)	3.17 3.28(3%)	4.53 4.61(2%)		
	0.66	1.36 1.32(-3%)	1.34 1.31(-3%)	1.34 1.31(-3%)	3.79 4.21(11%)	3.31 3.58(8%)	3.31 3.58(8%)	4.62 4.74(3%)		
	0.71	1.34 1.29(-3%)	1.33 1.28(-3%)	1.33 1.28(-3%)	3.98 4.50(13%)	3.46 3.90(13%)	3.46 3.90(13%)	4.67 4.83(4%)		
	0.76	1.32 1.27(-3%)	1.31 1.27(-3%)	1.31 1.27(-3%)	4.17 4.70(13%)	3.63 4.20(16%)	3.63 4.20(16%)	4.72 4.90(4%)		
	0.81	1.30 1.26(-3%)	1.30 1.26(-3%)	1.30 1.26(-3%)	4.37 4.83(11%)	3.84 4.48(17%)	3.84 4.48(17%)	4.80 4.94(3%)		
	0.86	1.29 1.26(-2%)	1.28 1.26(-2%)	1.28 1.26(-2%)	4.54 4.92(8%)	4.05 4.72(16%)	4.05 4.72(16%)	4.85 4.97(2%)		
	0.91	1.27 1.25(-2%)	1.27 1.25(-1%)	1.27 1.25(-1%)	4.71 4.97(6%)	4.31 4.89(13%)	4.31 4.89(13%)	4.91 4.99(2%)		
	0.96	1.26 1.25(-1%)	1.26 1.25(-1%)	1.26 1.25(-1%)	4.88 4.99(2%)	4.64 4.97(7%)	4.64 4.97(7%)	4.96 5.00(1%)		

Table 6.4: The RQ approximation in various $GI/M/1$ models.

	E_4 arrival		$LN(0.25)$ arrival		$H_2(4)$ arrival		$LN(4)$ arrival		G_4 arrival		
ρ	Sim	RQ	Sim	RQ	Sim	RQ	Sim	RQ	Sim	RQ	
Mean steady-state workload, $H_2(4)$ service	0.01	0.00	0.00(1%)	0.00	0.00(1%)	0.00	0.00(-1%)	0.00	0.00(-0%)	0.00	0.00(-10%)
	0.06	0.01	0.01(5%)	0.01	0.01(5%)	0.01	0.01(-3%)	0.01	0.01(-3%)	0.01	0.01(-12%)
	0.11	0.03	0.03(6%)	0.03	0.03(6%)	0.04	0.03(-6%)	0.04	0.03(-5%)	0.04	0.04(-12%)
	0.16	0.07	0.07(7%)	0.07	0.07(7%)	0.08	0.08(-8%)	0.08	0.08(-7%)	0.10	0.09(-11%)
	0.21	0.13	0.13(6%)	0.13	0.13(6%)	0.16	0.14(-10%)	0.15	0.14(-8%)	0.19	0.17(-10%)
	0.26	0.21	0.22(4%)	0.21	0.22(4%)	0.27	0.24(-11%)	0.26	0.23(-9%)	0.32	0.30(-8%)
	0.31	0.31	0.32(3%)	0.31	0.32(2%)	0.41	0.37(-11%)	0.40	0.37(-8%)	0.50	0.47(-6%)
	0.36	0.45	0.46(1%)	0.45	0.45(0%)	0.62	0.56(-10%)	0.59	0.55(-7%)	0.73	0.70(-4%)
	0.41	0.64	0.63(-1%)	0.63	0.62(-2%)	0.89	0.82(-8%)	0.85	0.80(-6%)	1.05	1.02(-3%)
	0.46	0.86	0.85(-1%)	0.86	0.85(-1%)	1.25	1.20(-4%)	1.19	1.15(-3%)	1.44	1.43(-1%)
	0.51	1.17	1.15(-2%)	1.16	1.14(-2%)	1.75	1.73(-1%)	1.65	1.62(-1%)	1.98	1.99(0%)
	0.56	1.55	1.53(-2%)	1.55	1.53(-1%)	2.40	2.46(3%)	2.25	2.27(1%)	2.68	2.72(1%)
	0.61	2.07	2.04(-2%)	2.07	2.04(-1%)	3.28	3.46(5%)	3.07	3.17(3%)	3.62	3.70(2%)
	0.66	2.79	2.73(-2%)	2.78	2.73(-2%)	4.52	4.81(6%)	4.23	4.44(5%)	4.91	5.02(2%)
	0.71	3.76	3.70(-2%)	3.75	3.70(-1%)	6.25	6.68(7%)	5.85	6.25(7%)	6.71	6.86(2%)
	0.76	5.17	5.12(-1%)	5.16	5.12(-1%)	8.81	9.40(7%)	8.29	8.95(8%)	9.35	9.55(2%)
	0.81	7.41	7.34(-1%)	7.40	7.34(-1%)	12.95	13.63(5%)	12.19	13.20(8%)	13.54	13.75(2%)
	0.86	11.33	11.23(-1%)	11.33	11.23(-1%)	20.18	21.00(4%)	19.19	20.64(8%)	20.83	21.09(1%)
	0.91	19.65	19.56(-0%)	19.64	19.55(-0%)	35.69	36.72(3%)	34.33	36.47(6%)	36.58	36.78(1%)
	0.96	49.16	48.96(-0%)	49.12	48.96(-0%)	90.77	92.12(1%)	89.03	91.83(3%)	92.14	92.15(0%)
Normalized mean steady-state workload, $H_2(4)$ service	0.01	4.97	5.00(1%)	4.97	5.00(1%)	5.05	5.00(-1%)	5.01	5.00(-0%)	5.72	5.17(-10%)
	0.06	4.77	4.99(5%)	4.77	4.99(5%)	5.18	5.01(-3%)	5.13	4.99(-3%)	6.26	5.50(-12%)
	0.11	4.69	4.96(6%)	4.68	4.96(6%)	5.34	5.02(-6%)	5.27	4.98(-5%)	6.55	5.75(-12%)
	0.16	4.61	4.91(7%)	4.60	4.91(7%)	5.48	5.05(-8%)	5.37	4.99(-7%)	6.73	5.99(-11%)
	0.21	4.57	4.84(6%)	4.56	4.83(6%)	5.65	5.11(-10%)	5.51	5.05(-8%)	6.90	6.23(-10%)
	0.26	4.54	4.73(4%)	4.53	4.73(4%)	5.81	5.19(-11%)	5.62	5.14(-9%)	7.06	6.46(-8%)
	0.31	4.50	4.61(3%)	4.48	4.59(2%)	5.95	5.32(-11%)	5.73	5.27(-8%)	7.14	6.69(-6%)
	0.36	4.47	4.49(1%)	4.46	4.46(0%)	6.11	5.50(-10%)	5.86	5.44(-7%)	7.23	6.91(-4%)
	0.41	4.46	4.40(-1%)	4.45	4.38(-2%)	6.27	5.77(-8%)	5.97	5.64(-6%)	7.34	7.13(-3%)
	0.46	4.40	4.35(-1%)	4.39	4.33(-1%)	6.39	6.11(-4%)	6.07	5.86(-3%)	7.37	7.32(-1%)
	0.51	4.40	4.32(-2%)	4.39	4.31(-2%)	6.60	6.51(-1%)	6.20	6.11(-1%)	7.46	7.49(0%)
	0.56	4.36	4.30(-2%)	4.35	4.29(-1%)	6.73	6.91(3%)	6.31	6.38(1%)	7.53	7.63(1%)
	0.61	4.35	4.28(-2%)	4.34	4.27(-1%)	6.87	7.24(5%)	6.45	6.65(3%)	7.58	7.75(2%)
	0.66	4.35	4.27(-2%)	4.34	4.27(-2%)	7.06	7.50(6%)	6.60	6.92(5%)	7.67	7.83(2%)
	0.71	4.33	4.26(-2%)	4.32	4.26(-1%)	7.19	7.69(7%)	6.73	7.19(7%)	7.71	7.90(2%)
	0.76	4.30	4.26(-1%)	4.29	4.26(-1%)	7.32	7.81(7%)	6.89	7.44(8%)	7.77	7.94(2%)
	0.81	4.29	4.25(-1%)	4.29	4.25(-1%)	7.50	7.90(5%)	7.06	7.65(8%)	7.84	7.97(2%)
	0.86	4.29	4.25(-1%)	4.29	4.25(-1%)	7.64	7.95(4%)	7.27	7.81(8%)	7.89	7.98(1%)
	0.91	4.27	4.25(-0%)	4.27	4.25(-0%)	7.76	7.98(3%)	7.46	7.93(6%)	7.95	7.99(1%)
	0.96	4.27	4.25(-0%)	4.26	4.25(-0%)	7.88	8.00(1%)	7.73	7.97(3%)	8.00	8.00(0%)

Table 6.5: The RQ approximation in various $GI/H_2(4)/1$ models.

	E_4 arrival		$LN(0.25)$ arrival		$H_2(4)$ arrival		$LN(4)$ arrival		G_4 arrival	
ρ	Sim	RQ	Sim	RQ	Sim	RQ	Sim	RQ	Sim	RQ
Mean steady-state workload, $G(4)$ service	0.01	0.00 0.00(1%)	0.00 0.00(1%)	0.00 0.00(1%)	0.00 0.00(-1%)	0.00 0.00(0%)	0.00 0.00(-10%)	0.00 0.00(-10%)	0.00 0.00(-10%)	
	0.06	0.01 0.01(5%)	0.01 0.01(5%)	0.01 0.01(5%)	0.01 0.01(-3%)	0.01 0.01(-3%)	0.01 0.01(-13%)	0.01 0.01(-13%)	0.01 0.01(-13%)	
	0.11	0.03 0.03(7%)	0.03 0.03(7%)	0.03 0.03(7%)	0.04 0.03(-6%)	0.04 0.03(-6%)	0.04 0.04(-13%)	0.04 0.04(-13%)	0.04 0.04(-13%)	
	0.16	0.07 0.07(8%)	0.07 0.07(8%)	0.07 0.07(8%)	0.08 0.08(-8%)	0.08 0.08(-8%)	0.10 0.09(-12%)	0.10 0.09(-12%)	0.10 0.09(-12%)	
	0.21	0.13 0.13(8%)	0.12 0.13(8%)	0.12 0.13(8%)	0.16 0.14(-10%)	0.16 0.14(-9%)	0.19 0.17(-11%)	0.19 0.17(-11%)	0.19 0.17(-11%)	
	0.26	0.20 0.22(6%)	0.20 0.22(7%)	0.20 0.22(7%)	0.27 0.24(-11%)	0.26 0.23(-10%)	0.32 0.30(-9%)	0.32 0.30(-9%)	0.32 0.30(-9%)	
	0.31	0.31 0.32(4%)	0.31 0.32(4%)	0.31 0.32(4%)	0.42 0.37(-12%)	0.40 0.37(-8%)	0.50 0.47(-7%)	0.50 0.47(-7%)	0.50 0.47(-7%)	
	0.36	0.45 0.46(2%)	0.44 0.45(1%)	0.44 0.45(1%)	0.63 0.56(-11%)	0.60 0.55(-8%)	0.74 0.70(-5%)	0.74 0.70(-5%)	0.74 0.70(-5%)	
	0.41	0.62 0.63(1%)	0.62 0.62(0%)	0.62 0.62(0%)	0.90 0.82(-9%)	0.86 0.80(-7%)	1.05 1.02(-3%)	1.05 1.02(-3%)	1.05 1.02(-3%)	
	0.46	0.85 0.85(-0%)	0.85 0.85(-0%)	0.85 0.85(-0%)	1.28 1.20(-6%)	1.20 1.15(-5%)	1.46 1.43(-2%)	1.46 1.43(-2%)	1.46 1.43(-2%)	
	0.51	1.15 1.15(-1%)	1.15 1.14(-1%)	1.15 1.14(-1%)	1.78 1.73(-3%)	1.67 1.62(-3%)	2.00 1.99(-0%)	2.00 1.99(-0%)	2.00 1.99(-0%)	
	0.56	1.54 1.53(-1%)	1.54 1.53(-1%)	1.54 1.53(-1%)	2.44 2.46(1%)	2.28 2.27(-0%)	2.71 2.72(0%)	2.71 2.72(0%)	2.71 2.72(0%)	
	0.61	2.06 2.04(-1%)	2.05 2.04(-1%)	2.05 2.04(-1%)	3.34 3.46(4%)	3.12 3.17(2%)	3.65 3.70(1%)	3.65 3.70(1%)	3.65 3.70(1%)	
	0.66	2.75 2.73(-1%)	2.75 2.73(-1%)	2.75 2.73(-1%)	4.57 4.81(5%)	4.27 4.44(4%)	4.95 5.02(1%)	4.95 5.02(1%)	4.95 5.02(1%)	
	0.71	3.73 3.70(-1%)	3.72 3.70(-1%)	3.72 3.70(-1%)	6.34 6.68(5%)	5.91 6.25(6%)	6.77 6.86(1%)	6.77 6.86(1%)	6.77 6.86(1%)	
	0.76	5.15 5.12(-1%)	5.15 5.12(-0%)	5.15 5.12(-0%)	8.95 9.40(5%)	8.39 8.95(7%)	9.39 9.55(2%)	9.39 9.55(2%)	9.39 9.55(2%)	
	0.81	7.39 7.34(-1%)	7.39 7.34(-1%)	7.39 7.34(-1%)	13.08 13.63(4%)	12.31 13.20(7%)	13.56 13.75(1%)	13.56 13.75(1%)	13.56 13.75(1%)	
0.86	11.28 11.23(-0%)	11.29 11.23(-1%)	11.29 11.23(-1%)	20.27 21.00(4%)	19.28 20.64(7%)	20.82 21.09(1%)	20.82 21.09(1%)	20.82 21.09(1%)		
0.91	19.62 19.56(-0%)	19.61 19.55(-0%)	19.61 19.55(-0%)	35.93 36.72(2%)	34.55 36.47(6%)	36.52 36.78(1%)	36.52 36.78(1%)	36.52 36.78(1%)		
0.96	49.02 48.96(-0%)	48.86 48.96(0%)	48.86 48.96(0%)	91.14 92.12(1%)	89.06 91.83(3%)	91.89 92.15(0%)	91.89 92.15(0%)	91.89 92.15(0%)		
Normalized mean steady-state workload, $G(4)$ service	0.01	4.96 5.00(1%)	4.96 5.00(1%)	4.96 5.00(1%)	5.03 5.00(-1%)	4.98 5.00(0%)	5.76 5.17(-10%)	5.76 5.17(-10%)	5.76 5.17(-10%)	
	0.06	4.75 4.99(5%)	4.75 4.99(5%)	4.75 4.99(5%)	5.16 5.01(-3%)	5.15 4.99(-3%)	6.34 5.50(-13%)	6.34 5.50(-13%)	6.34 5.50(-13%)	
	0.11	4.63 4.96(7%)	4.63 4.96(7%)	4.63 4.96(7%)	5.34 5.02(-6%)	5.30 4.98(-6%)	6.61 5.75(-13%)	6.61 5.75(-13%)	6.61 5.75(-13%)	
	0.16	4.55 4.91(8%)	4.53 4.91(8%)	4.53 4.91(8%)	5.52 5.05(-8%)	5.40 4.99(-8%)	6.81 5.99(-12%)	6.81 5.99(-12%)	6.81 5.99(-12%)	
	0.21	4.50 4.84(8%)	4.47 4.83(8%)	4.47 4.83(8%)	5.67 5.11(-10%)	5.56 5.05(-9%)	6.97 6.23(-11%)	6.97 6.23(-11%)	6.97 6.23(-11%)	
	0.26	4.47 4.73(6%)	4.44 4.73(7%)	4.44 4.73(7%)	5.84 5.19(-11%)	5.70 5.14(-10%)	7.09 6.46(-9%)	7.09 6.46(-9%)	7.09 6.46(-9%)	
	0.31	4.43 4.61(4%)	4.40 4.59(4%)	4.40 4.59(4%)	6.05 5.32(-12%)	5.75 5.27(-8%)	7.20 6.69(-7%)	7.20 6.69(-7%)	7.20 6.69(-7%)	
	0.36	4.40 4.49(2%)	4.39 4.46(1%)	4.39 4.46(1%)	6.21 5.50(-11%)	5.89 5.44(-8%)	7.30 6.91(-5%)	7.30 6.91(-5%)	7.30 6.91(-5%)	
	0.41	4.38 4.40(1%)	4.37 4.38(0%)	4.37 4.38(0%)	6.35 5.77(-9%)	6.04 5.64(-7%)	7.38 7.13(-3%)	7.38 7.13(-3%)	7.38 7.13(-3%)	
	0.46	4.35 4.35(-0%)	4.35 4.33(-0%)	4.35 4.33(-0%)	6.53 6.11(-6%)	6.14 5.86(-5%)	7.47 7.32(-2%)	7.47 7.32(-2%)	7.47 7.32(-2%)	
	0.51	4.34 4.32(-1%)	4.34 4.31(-1%)	4.34 4.31(-1%)	6.69 6.51(-3%)	6.28 6.11(-3%)	7.52 7.49(-0%)	7.52 7.49(-0%)	7.52 7.49(-0%)	
	0.56	4.33 4.30(-1%)	4.31 4.29(-1%)	4.31 4.29(-1%)	6.85 6.91(1%)	6.40 6.38(-0%)	7.61 7.63(0%)	7.61 7.63(0%)	7.61 7.63(0%)	
	0.61	4.32 4.28(-1%)	4.30 4.27(-1%)	4.30 4.27(-1%)	6.99 7.24(4%)	6.55 6.65(2%)	7.66 7.75(1%)	7.66 7.75(1%)	7.66 7.75(1%)	
	0.66	4.30 4.27(-1%)	4.29 4.27(-1%)	4.29 4.27(-1%)	7.14 7.50(5%)	6.67 6.92(4%)	7.73 7.83(1%)	7.73 7.83(1%)	7.73 7.83(1%)	
	0.71	4.29 4.26(-1%)	4.28 4.26(-1%)	4.28 4.26(-1%)	7.30 7.69(5%)	6.80 7.19(6%)	7.79 7.90(1%)	7.79 7.90(1%)	7.79 7.90(1%)	
	0.76	4.28 4.26(-1%)	4.28 4.26(-0%)	4.28 4.26(-0%)	7.44 7.81(5%)	6.97 7.44(7%)	7.81 7.94(2%)	7.81 7.94(2%)	7.81 7.94(2%)	
	0.81	4.28 4.25(-1%)	4.28 4.25(-1%)	4.28 4.25(-1%)	7.58 7.90(4%)	7.13 7.65(7%)	7.86 7.97(1%)	7.86 7.97(1%)	7.86 7.97(1%)	
0.86	4.27 4.25(-0%)	4.27 4.25(-1%)	4.27 4.25(-1%)	7.68 7.95(4%)	7.30 7.81(7%)	7.88 7.98(1%)	7.88 7.98(1%)	7.88 7.98(1%)		
0.91	4.26 4.25(-0%)	4.26 4.25(-0%)	4.26 4.25(-0%)	7.81 7.98(2%)	7.51 7.93(6%)	7.94 7.99(1%)	7.94 7.99(1%)	7.94 7.99(1%)		
0.96	4.26 4.25(-0%)	4.24 4.25(0%)	4.24 4.25(0%)	7.91 8.00(1%)	7.73 7.97(3%)	7.98 8.00(0%)	7.98 8.00(0%)	7.98 8.00(0%)		

Table 6.6: The RQ approximation in various $GI/G(4)/1$ models.

		Service						
		E_4	$LN(0.25)$	M	$H_2(4)$	$LN(4)$	$G(4)$	
Max RE	Arrival	E_4	27.56%	27.02%	17.55%	6.53%	6.26%	7.86%
		$LN(0.25)$	30.45%	29.70%	18.94%	6.84%	6.50%	8.43%
		$H_2(4)$	20.82%	20.79%	17.68%	10.71%	10.16%	12.05%
		$LN(4)$	23.91%	24.09%	16.76%	8.57%	9.24%	9.81%
		$G(4)$	29.51%	29.55%	22.89%	12.25%	13.40%	13.25%
Max RE, $\rho > 0.5$	Arrival	E_4	21.54%	20.78%	4.42%	1.86%	2.41%	1.00%
		$LN(0.25)$	24.33%	23.35%	3.68%	1.82%	2.25%	0.75%
		$H_2(4)$	19.20%	19.03%	13.17%	6.97%	7.87%	5.33%
		$LN(4)$	23.91%	24.09%	16.76%	8.29%	9.24%	7.27%
		$G(4)$	4.52%	4.56%	3.67%	2.31%	2.34%	1.64%

Table 6.7: The performance of the RQ approximation in various $GI/GI/1$ models for $\rho > 0.5$.

		Service					
		E_4	$LN(0.25)$	M	$H_2(4)$	$LN(4)$	$G(4)$
Max absolute error	E_4	0.2404	0.2375	0.2823	0.3012	0.2896	0.3582
	$LN(0.25)$	0.2565	0.2526	0.3009	0.3145	0.2999	0.3821
	$H_2(4)$	0.5909	0.6014	0.5260	0.6379	0.5874	0.7290
	$LN(4)$	0.7315	0.7293	0.6658	0.5856	0.6474	0.5591
	$G(4)$	0.7537	0.7557	0.7732	0.8033	0.8512	0.8642

Table 6.8: The maximum absolute error of the RQ approximation in various $GI/GI/1$ models.

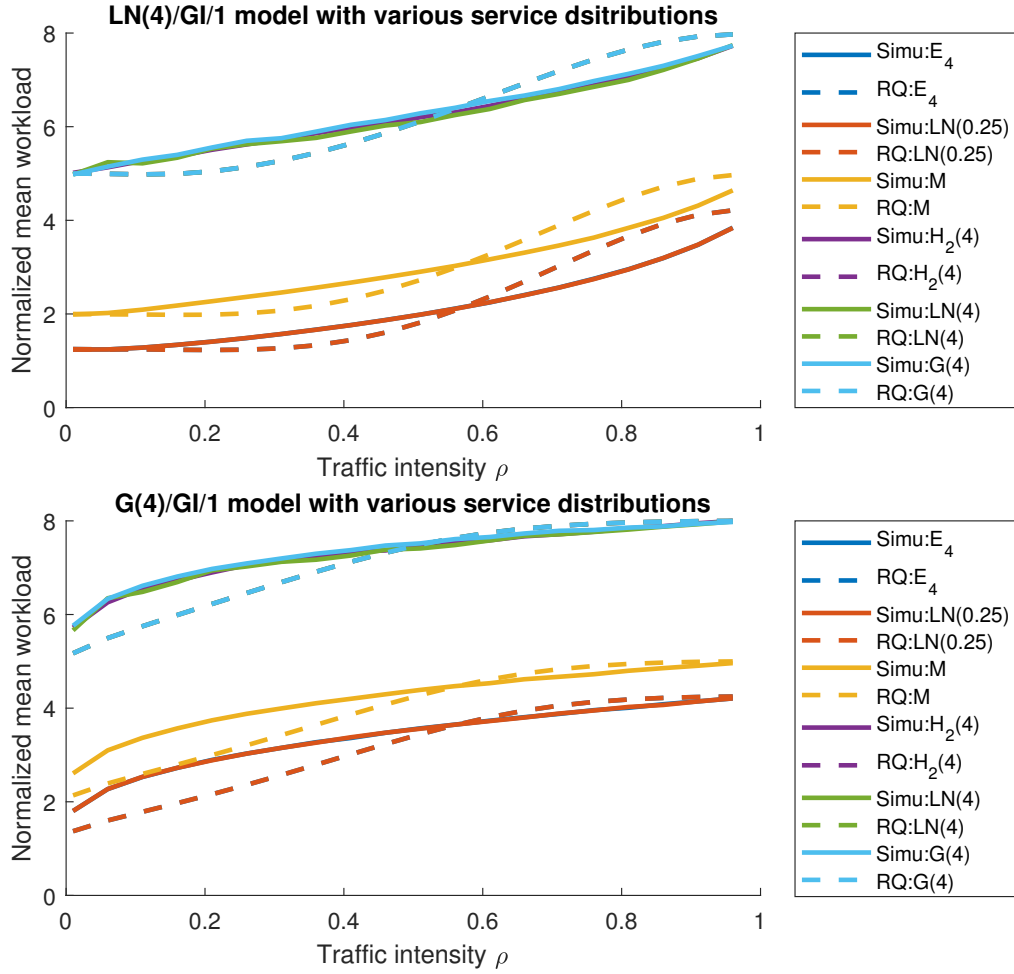


Figure 6.1: The normalized mean workload and the RQ approximation for $LN(4)/GI/1$ and $G(4)/GI/1$ models.

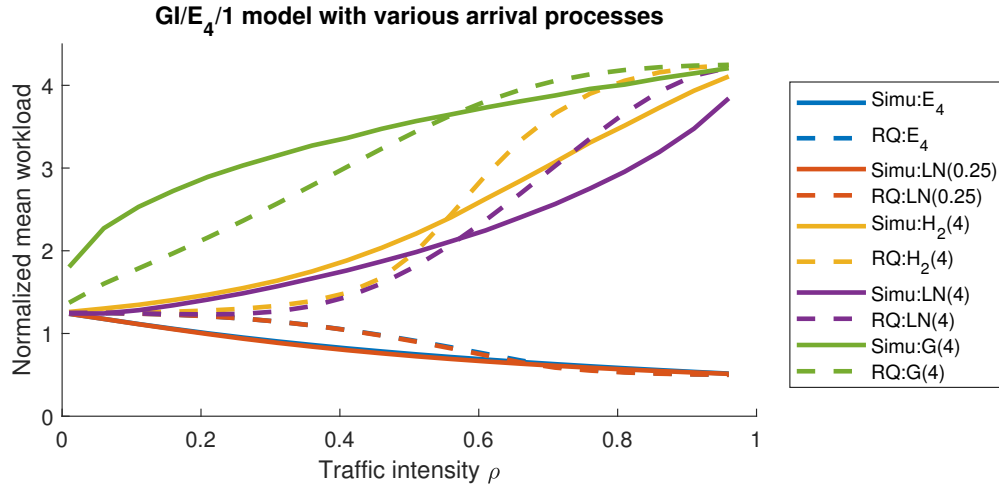


Figure 6.2: The normalized mean workload and the RQ approximation for $GI/E_4/1$ models.

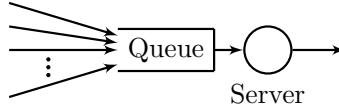


Figure 6.3: A queue with superposition arrival process.

p. 137 of [133]. Then (2.67) and (2.29) imply that the IDW has limits $I_w(0) = 1 + c_s^2 = 3$ and $I_w(\infty) = c_a^2 + c_s^2 = 12$, so that the IDW is not nearly constant.

Figure 6.4 (left) shows a comparison between the simulation estimate of the normalized workload $c_Z^2(\rho)$ in (2.27) and the approximation $c_{Z^*}^2(\rho)$ in (2.36) for this example. Two important observations are: (i) the normalized mean workload $c_Z^2(\rho)$ in (2.27) as a function of ρ is not nearly constant, and (ii) there is a close agreement between the RQ approximation $c_{Z^*}^2(\rho)$ in (2.36) and the direct simulation estimate; the close agreement for all traffic intensities is striking. It is important to note that the parametric RQ approximations produce constant approximations, and so cannot be simultaneously good for all traffic intensities.

For this example, we see that $c_Z^2(\rho) \approx 3$ for $\rho \leq 0.5$, which is consistent with the Poisson approximation for the arrival process and the associated $M/G/1$ queue, where $c_Z^2(\rho) = 3$ for all ρ , but the normalized workload increases steadily to 12 after $\rho = 0.5$, as explained in Section 9.8 of [143].

The estimates for Figure 6.4 were obtained for ρ over a grid of 99 values, evenly spaced between 0.01 and 0.99. Similarly, the RQ optimization was performed using (2.34) with a

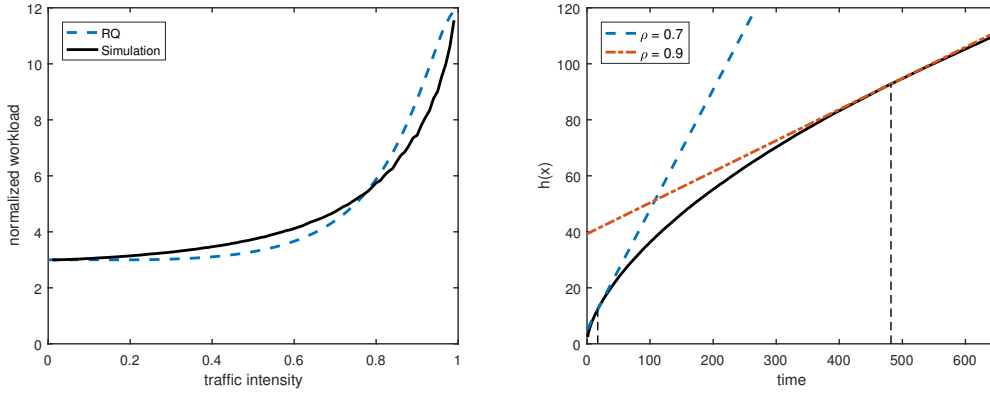


Figure 6.4: The RQ solution and approximation for the workload in the $\sum_i^n LN_i/H_2/1$ model.

discrete-time estimate of the IDW. By doing multiple independent runs, we ensured that the statistical variation was not an issue. For the main simulation of the arrival process we used 5×10^6 time units, discarding a large initial portion of the workload process to ensure that the system is approximately in steady state. (The component renewal arrival processes thus can be regarded as equilibrium renewal processes, as in Section 3.5 of Section [116].) For the simulation of the mean workload, we let the run length and amount discarded be proportional to $(1 - \rho)^{-2}$, as dictated by [140] and observed in Corollary 2.2.

6.2.3 A Ten-Queues-in-Series Example

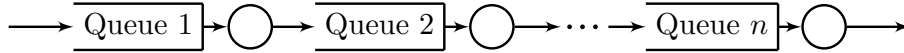


Figure 6.5: A ten-queues-in-series model.

This example is a variant of examples in [125], exposing the complex impact of variability on performance in a series of queues if the external arrival process and service times at a previous queue have very different levels of variability. This example has 10 single-server queues in series. The external arrival process is a rate-1 renewal process with H_2 interarrival times having $c_a^2 = 5$. The first 9 queues all have Erlang service times with $c_a^2 = 0.5$ denoted

by E_2 , i.e., the sum of 2 i.i.d. exponential random variables. The first 8 queues have mean service time and thus traffic intensity 0.6, while the 9th queue has mean service time and thus traffic intensity 0.95. The last (10th) queue has an exponential service-time distribution with mean and traffic intensity ρ ; we explore the impact of ρ on the performance of that last queue.

The Erlang services act to smooth the arrival process at the last queue. Thus, for sufficiently low traffic intensities ρ at the last queue, the last queue should behave essentially the same as a $E_2/M/1$ queue, which has $c_a^2 = 0.5$, but as ρ increases, the arrival process at the last queue should inherit the variability of the external arrival process, and behave like an $H_2/M/1$ queue with scv $c_a^2 = 5$.

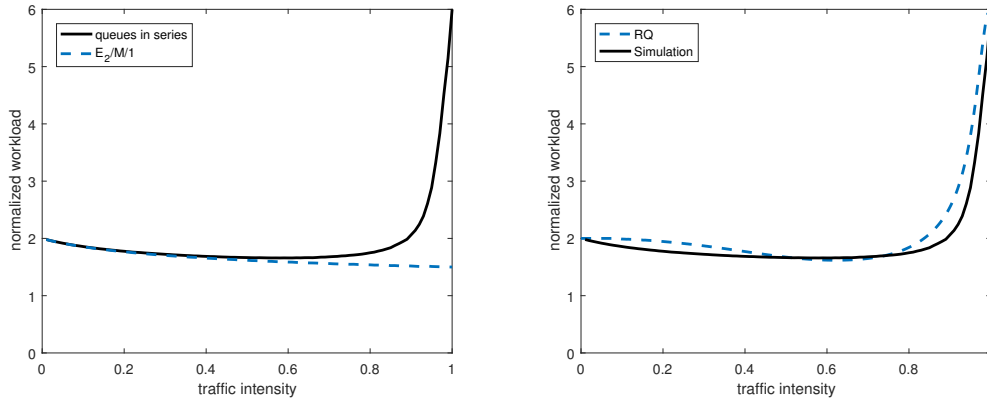


Figure 6.6: Comparing the RQ approximation, the $E_2/M/1$ approximation and the simulation estimation of the workload in the ten-queues-in-series model.

This behavior is substantiated by Figure 6.6, which compares simulation estimates of the normalized mean workload $c_Z^2(\rho)$ in (2.27) at the last queue of ten queues in series as a function of the mean service time and traffic intensity ρ there with the corresponding values in the $E_2/M/1$ queue (left) and with the RQ approximation $c_{Z*}^2(\rho)$ in (2.36) (right).

Figure 6.6 (left) shows that the last queue behaves like a $E_2/M/1$ queue for all traffic intensities ≤ 0.8 , but then starts behaving more like an $H_2/M/1$ queue as the traffic intensity approaches the value 0.95 at the 9th queue. Figure 6.6 (right) shows that RQ successfully captures this phenomenon and provides an accurate approximation for all ρ .

To elaborate on this series-queue example, we show the IDW for the last queue in

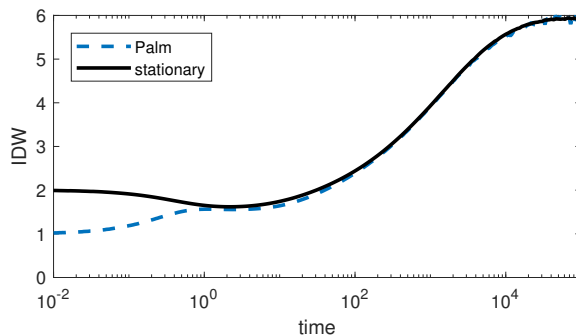


Figure 6.7: The IDW at the last queue of the ten-queues-in-series model: the continuous-time stationary version and the discrete-time Palm version.

Figure 6.7. The plot shows the IDW assuming continuous-time stationarity (which we use) together with the plot using the discrete-time Palm stationarity (see [122]) over the long interval $[10^{-2}, 10^5]$ in log scale. The good performance in Figure 6.6 for small values of ρ depends on using the proper (continuous-time) version.

We conclude this example by illustrating the discrete-time approach for approximating the expected steady-state waiting time $E[W]$ using the RQ optimization in (2.6) with uncertainty set in (2.9). Figure 6.8 is the discrete analog of Figure 6.6. Figure 6.8 compares simulation estimates of the normalized mean waiting time $c_W^2(\rho)$, defined just as in (2.27), at the last queue of ten queues in series as a function of the mean service time and traffic intensity ρ there with the corresponding values in the $E_2/M/1$ queue (left) and with the RQ approximation $c_{W*}^2(\rho)$, defined just as in (2.36). Figure 6.8 and 6.6 look similar, except that there is a significant difference for small values of ρ . In general, we do not expect RQ to be effective for extremely low ρ , because (i) the CLT is not appropriate for only a few summands and (ii) the mean waiting time is known to depend on other factors when ρ is small. The mean waiting time and mean workload actually are quite different in light traffic; see Section IV.A of [60]. As explained there, the mean workload tends to be more robust to model detail.

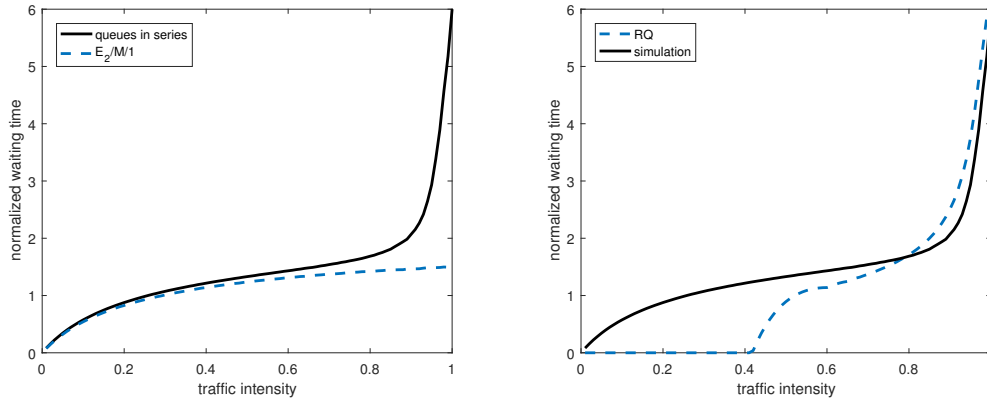


Figure 6.8: Contrasting the discrete-time and continuous-time views: the analog of Figure 6.6 for the waiting time.

6.2.4 A MAP/MAP/1 Example

We now include the dependence among service times by considering a MAP/MAP/1 model. With the dependence in service times, we can no longer enjoy the convenient decomposition in (2.26). Hence, we apply the RQ-IDW formulation in (2.34). For simplicity, we use simulation estimation here to obtain the IDW function.

We specify the generating matrix of the arrival MAP in (6.1). To consider the full range of $\rho \in (0, 1)$, we define the service MAP in the ρ -th model by the following matrices

$$\mathbf{D}_0^{s,\rho} = \mathbf{D}_0/\rho, \quad \mathbf{D}_1^{s,\rho} = \mathbf{D}_1/\rho \quad (6.5)$$

for $(\mathbf{D}_0, \mathbf{D}_1)$ defined in (6.1), where the division is entry-wise.

Figure 6.9(top) show the simulation estimation of the IDW for this model, whereas the bottom plot show the simulation estimation and RQ approximation of the mean steady-state workload as functions of the traffic intensity ρ . Note that the light traffic limit is $c_Z^*(0) = 1 + c_a^2 = 5.1$ and the heavy-traffic limit is $c_Z^*(0) = c_A^2 + c_S^2 = 18.14$, where $c_a^2 = 4.1$ is the scv of the interarrival time, $c_A^2 = c_S^2 = 9.07$ are the limiting variability parameters for the arrival and service MAP, see (6.2).

The somewhat poor performance in the range $\rho \in (0, 0.3)$ is expected. Note that, for small ρ , the optimal solution of (2.17) is obtained at small s^* , as shown in Theorem 2.2. However, CLT does not provide a good approximation for small s in uncertainty set (2.16).

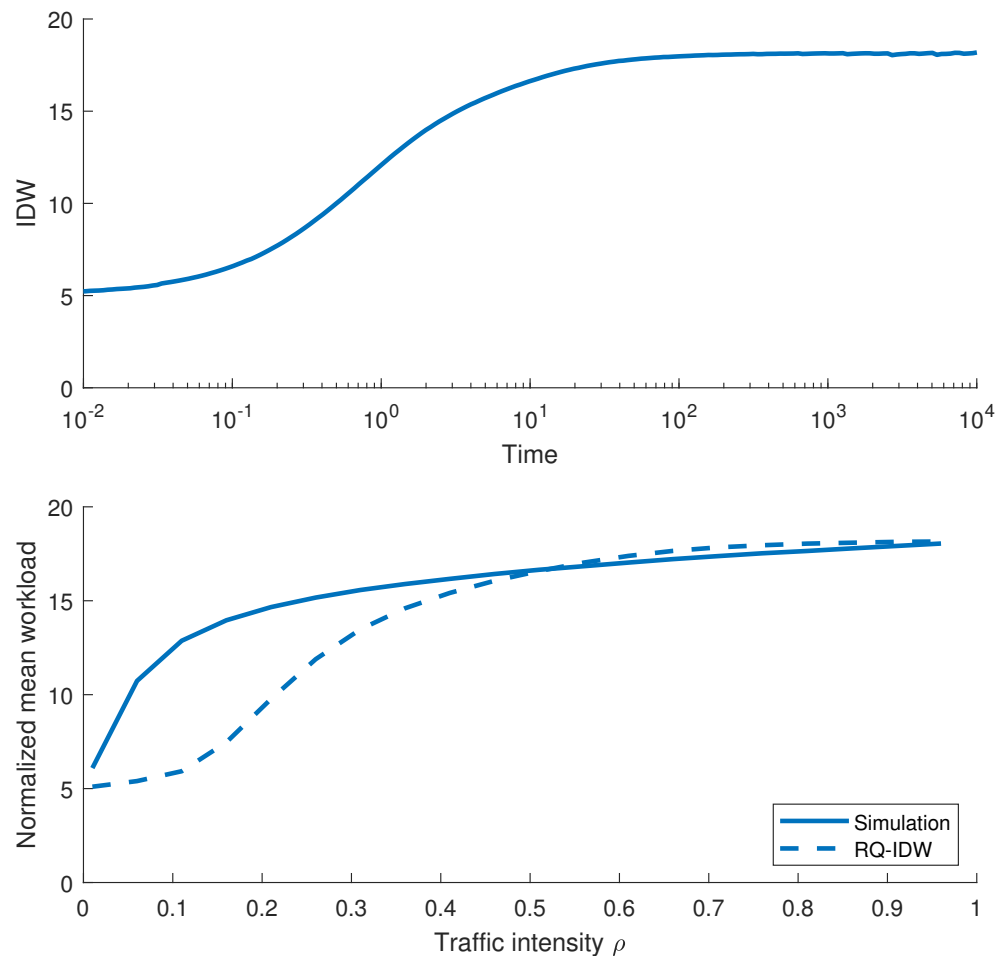


Figure 6.9: The IDW and RQ approximation for a MAP/MAP/1 model.

ρ	0.01	0.06	0.11	0.16	0.21	0.26	0.31	0.36	0.41
Abs. err.	0.0001	0.0102	0.0472	0.0990	0.1365	0.1498	0.1457	0.1306	0.1068

Table 6.9: Absolute error of the RQ-IDW approximation in the MAP/MAP/1 model.

Nevertheless, the absolute error *before normalization* in this case will be small as shown in Table 6.9. For example, if the mean interarrival time is 1 minute, then for $\rho = 0.26$ the absolute error will be 0.1498 minutes ≈ 9 seconds.

6.2.5 The Queues in Series Models

We now present a comprehensive simulation study of the performance of our RQ approximation in the two queues in series model. We denote such a system by $GI/GI/1 \rightarrow \cdot/GI/1$.

As in Section 6.2.1, we consider the same 5 cases for the renewal external arrival processes, i.e. $E_4, LN(0.25), H_2(4), LN(4), G(4)$; and the same 6 cases for the service time distributions at both stations, i.e. $E_4, LN(0.25), M, H_2(4), LN(4), G(4)$. The external arrival processes have rate-1. For the service rate at station 1, we consider two cases $\mu_1 = 1/\rho_1$, where the traffic intensity $\rho_1 = 0.7$ or 0.9 . So we have a total of $5 \times 6 \times 6 \times 2 = 360$ cases. To show the performance impact of dependence under the full range of traffic intensity levels, we assume the service-rate at station 2 to be $\mu_2 = 1/\rho_2$ with $\rho_2 \in \{0.01, 0.06, 0.11, \dots, 0.96\}$. We assess the performance of RQ based on the ability to predict the mean workload at the second station, for different traffic intensities.

Details of the simulation are discussed in Section 6.1. Since the simulation time scales roughly linearly in system size, for two queues in series, it takes on average 60 minutes to obtain simulation estimations of all 20 ρ 's for each case. Again, about 70% of the simulation time is devoted to the case $\rho = 0.96$. On the other hand, the RQ algorithm now requires the IDC of the total input process, which is not readily available. In this section, we resort to simulation estimation of the IDC's, as discussed in Section 2.3.4. In Section 6.3, we will look at the RQNA alternative, where we rely on our RQNA algorithm instead of simulation estimation. For the simulation methodology for the estimation of the IDC is described in Section 6.1. Again, we use a single simulation run of $1.1 \cdot 10^9$ time units and discard the first 10^8 time units for the arrival process to approach stationarity. The simulation time

required for each case is on average 8 minutes on a 4.8GHz CPU (implemented in C++), each case uses 8GB of RAM memory since we need to store the entire sample path in order to speed up the estimation, see Section 2.3.4.

Table 6.10 contains some of the hardest cases among the 360 models, the $GI/E_4/1 \rightarrow \cdot/E_4/1$ models with $\rho_1 = 0.9$ and $GI = H_2(4), LN(4)$ or $G(4)$. We discuss it in the following observations.

Table 6.11-6.16 summarizes the absolute relative error of the RQ approximation with simulated IDC at station 2, in the case of $\rho_1 = 0.9$. For each table, we fix the service-time distribution at station 2 (specified in the upper left entry), and display 30 cases, corresponding the combination of 5 types of external arrival processes (row), and 6 choices of service-time distributions at station 1 (column). In the top half of each table, we show the average relative error, defined as

$$\frac{1}{20} \sum_{k=1}^{20} |RE(\rho_k)|, \text{ for } \rho_k = -0.04 + 0.05k;$$

in the bottom half of each table, we show the maximum relative error, defined in (6.4). We make the following observations.

1. The RQ performance increases as the service scv at station 2 increases. In Table 6.11, we see that in the case of E_4 service distribution at station 2, the absolute relative error can be as large as 86%, whereas the absolute relative error stays below 13% in Table 6.16. This happens for the same reason as Observation 2 in Section 6.2.1.
2. The hardest cases for RQ appear to be the ones with high variability external arrival process ($LN(4)$ and H_4), and low variability service-time distributions (E_4 , $LN(0.25)$ and M) at both stations, e.g. in the right three columns of Table 6.10 and in the lower left corner of Table 6.11-6.13. In intuitively, these are the hardest cases for queueing approximations, because the highly variable arrival processes are smoothed out by the low-variable service times at station 1. As a result, when the traffic intensity at station 2 is low to moderate, the queue sees a arrival process (the departure from station 1) with low variability; whereas for higher traffic intensity, the arrival process sees the highly variable external arrival prcess. This resonates with the heavy-traffic bottleneck phenomenon studied in [125]. The sharp transition between the

aforementioned two cases is the major cause of the unsatisfactory performance of RQ. This effect becomes more pronounced ρ_1 increases, e.g. compare Table 6.11 and Table 6.17, where the later one is an analog of the former one with $\rho_1 = 0.7$ instead of 0.9. Figure 6.10 displays some of the worst cases in terms of RQ approximations, i.e. the $LN(4)/E_4/1 \rightarrow \cdot/GI/1$ models with $GI = E_4, M$ and $H_2(4)$.

3. In Table 6.18 (top), we show the maximum absolute relative error (6.4) in the case of $\rho_1 = 0.9$, M service at station 2 and across the full range of $\rho_2 \in (0, 1)$; where as in Table 6.18 (bottom), we restrict to $\rho_2 \in (0.5, 1)$. As in Observation 3 in Section 6.2.1, we observe significant improvement of the RQ performance in most cases, except for the ones with high variability external arrival process and low variability service-time distributions, which are demonstrated in Figure 6.10.
4. The RQ performance under $\rho_1 = 0.7$ is roughly the same as that of $\rho_1 = 0.9$. In Table 6.19, we show the comparison under $LN(4)$ service at station 2. We omit most of the $\rho = 0.7$ cases.

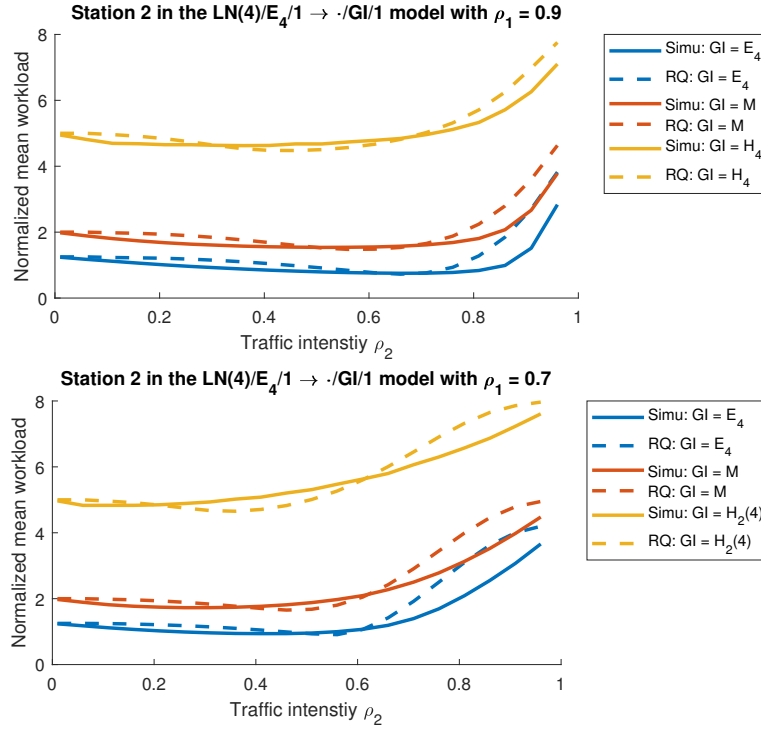
6.2.6 The Limitation of IDC

Recall from (1.1) that the IDC is a continuous-time function defined only through the mean and variance functions. We have shown above that the IDC-based RQ algorithm can capture the essential impact of dependence on queueing performance in models with non-renewal input flows. We have also shown in Theorem 3.1 that the IDC can recover the full distribution of a renewal process; however, such statement does not extend to non-renewal processes. In this section, we demonstrate that there can still be serious limitation in the approximation of non-renewal processes using the IDC.

Consider a hyperexponential renewal process with rate-1, whose IDC I_{H_2} of this process can be found in Example 2.2. Now, we show that there are infinitely many MMPP(2) that have rate-1 and the same IDC's as I_{H_2} . Towards this end, we equate (2.62) to (2.60) and

$\rho_1 = 0.9$		E_4 arrival		$LN(0.25)$ arrival		$H_2(4)$ arrival		$LN(4)$ arrival		G_4 arrival	
ρ_2		Sim	RQ	Sim	RQ	Sim	RQ	Sim	RQ	Sim	RQ
Mean steady-state workload	0.01	0.00	0.00(1%)	0.00	0.00(1%)	0.00	0.00(1%)	0.00	0.00(1%)	0.00	0.00(1%)
	0.06	0.00	0.00(6%)	0.00	0.00(6%)	0.00	0.00(6%)	0.00	0.00(6%)	0.00	0.00(6%)
	0.11	0.01	0.01(11%)	0.01	0.01(11%)	0.01	0.01(10%)	0.01	0.01(10%)	0.01	0.01(10%)
	0.16	0.02	0.02(16%)	0.02	0.02(16%)	0.02	0.02(13%)	0.02	0.02(13%)	0.02	0.02(13%)
	0.21	0.03	0.03(19%)	0.03	0.03(19%)	0.03	0.03(13%)	0.03	0.03(13%)	0.03	0.03(12%)
	0.26	0.04	0.05(22%)	0.04	0.05(22%)	0.05	0.05(11%)	0.05	0.05(12%)	0.05	0.05(9%)
	0.31	0.06	0.08(26%)	0.06	0.08(26%)	0.06	0.08(25%)	0.06	0.08(25%)	0.06	0.08(25%)
	0.36	0.09	0.11(27%)	0.09	0.11(27%)	0.09	0.11(25%)	0.09	0.11(25%)	0.09	0.11(25%)
	0.41	0.12	0.15(26%)	0.12	0.15(26%)	0.12	0.15(23%)	0.12	0.15(23%)	0.12	0.15(23%)
	0.46	0.15	0.19(24%)	0.16	0.19(24%)	0.16	0.19(20%)	0.16	0.19(20%)	0.16	0.19(20%)
	0.51	0.20	0.24(20%)	0.20	0.24(20%)	0.21	0.24(15%)	0.21	0.24(15%)	0.21	0.24(15%)
	0.56	0.26	0.30(15%)	0.26	0.30(15%)	0.28	0.30(9%)	0.28	0.30(9%)	0.28	0.30(9%)
	0.61	0.33	0.36(9%)	0.33	0.36(9%)	0.36	0.37(1%)	0.36	0.37(1%)	0.36	0.37(1%)
	0.66	0.43	0.44(3%)	0.43	0.44(3%)	0.48	0.47(-4%)	0.48	0.46(-4%)	0.48	0.46(-4%)
	0.71	0.56	0.54(-3%)	0.56	0.54(-2%)	0.66	0.68(2%)	0.66	0.67(1%)	0.66	0.65(-0%)
	0.76	0.74	0.70(-5%)	0.74	0.70(-5%)	0.94	1.13(20%)	0.93	1.12(20%)	0.93	1.05(12%)
	0.81	1.02	0.95(-7%)	1.02	0.95(-6%)	1.46	2.09(44%)	1.44	2.21(53%)	1.44	1.91(33%)
	0.86	1.49	1.39(-7%)	1.49	1.40(-6%)	2.65	4.43(67%)	2.61	4.88(87%)	2.62	4.11(57%)
	0.91	2.49	2.35(-5%)	2.49	2.37(-5%)	7.21	11.49(59%)	6.96	12.44(79%)	7.15	11.03(54%)
	0.96	5.97	5.79(-3%)	5.97	5.80(-3%)	33.74	43.19(28%)	32.66	44.03(35%)	33.70	42.58(26%)
Normalized mean steady-state workload	0.01	1.24	1.25(1%)	1.24	1.25(1%)	1.24	1.25(1%)	1.24	1.25(1%)	1.24	1.25(1%)
	0.06	1.18	1.25(6%)	1.18	1.25(6%)	1.18	1.25(6%)	1.18	1.25(6%)	1.18	1.25(6%)
	0.11	1.12	1.24(11%)	1.12	1.24(11%)	1.13	1.24(10%)	1.13	1.24(10%)	1.13	1.24(10%)
	0.16	1.06	1.23(16%)	1.06	1.23(16%)	1.09	1.23(13%)	1.09	1.23(13%)	1.09	1.23(13%)
	0.21	1.01	1.21(19%)	1.01	1.21(19%)	1.07	1.21(13%)	1.06	1.21(13%)	1.08	1.21(12%)
	0.26	0.97	1.18(22%)	0.96	1.18(22%)	1.06	1.18(11%)	1.06	1.18(12%)	1.08	1.18(9%)
	0.31	0.91	1.14(26%)	0.91	1.14(26%)	0.92	1.14(25%)	0.92	1.14(25%)	0.92	1.14(25%)
	0.36	0.87	1.10(27%)	0.87	1.10(27%)	0.88	1.10(25%)	0.88	1.10(25%)	0.88	1.10(25%)
	0.41	0.83	1.04(26%)	0.83	1.04(26%)	0.85	1.04(23%)	0.85	1.04(23%)	0.85	1.04(23%)
	0.46	0.79	0.98(24%)	0.79	0.98(24%)	0.82	0.98(20%)	0.82	0.98(20%)	0.82	0.98(20%)
	0.51	0.76	0.91(20%)	0.76	0.91(20%)	0.79	0.91(15%)	0.79	0.91(15%)	0.79	0.91(15%)
	0.56	0.72	0.84(15%)	0.73	0.84(15%)	0.77	0.84(9%)	0.77	0.84(9%)	0.77	0.84(9%)
	0.61	0.70	0.76(9%)	0.70	0.76(9%)	0.76	0.77(1%)	0.76	0.77(1%)	0.76	0.77(1%)
	0.66	0.67	0.68(3%)	0.67	0.68(3%)	0.75	0.73(-4%)	0.75	0.72(-4%)	0.75	0.72(-4%)
	0.71	0.64	0.62(-3%)	0.64	0.62(-2%)	0.76	0.78(2%)	0.76	0.77(1%)	0.75	0.75(-0%)
	0.76	0.61	0.58(-5%)	0.61	0.58(-5%)	0.78	0.94(20%)	0.78	0.93(20%)	0.78	0.87(12%)
	0.81	0.59	0.55(-7%)	0.59	0.55(-6%)	0.84	1.21(44%)	0.84	1.28(53%)	0.84	1.11(33%)
	0.86	0.56	0.53(-7%)	0.56	0.53(-6%)	1.01	1.68(67%)	0.99	1.85(87%)	0.99	1.56(57%)
	0.91	0.54	0.51(-5%)	0.54	0.52(-5%)	1.57	2.50(59%)	1.51	2.70(79%)	1.55	2.40(54%)
	0.96	0.52	0.50(-3%)	0.52	0.50(-3%)	2.93	3.75(28%)	2.83	3.82(35%)	2.93	3.70(26%)

Table 6.10: The RQ approximation in various $GI/E_4/1 \rightarrow \cdot/E_4/1$ models.

Figure 6.10: The performance of RQ at station 2 of the $LN(4)/E_4/1 \rightarrow \cdot/GI/1$ models.

Service 2		Service 1, $\rho_1 = 0.9$					
E_4		E_4	$LN(0.25)$	M	$H_2(4)$	$LN(4)$	$G(4)$
Average RE	Arrival						
	E_4	13.95%	14.67%	2.26%	11.44%	10.31%	9.84%
	$LN(0.25)$	13.78%	14.47%	2.22%	11.43%	10.32%	9.82%
	$H_2(4)$	22.84%	24.32%	7.11%	12.84%	13.45%	10.66%
	$LN(4)$	25.83%	27.27%	8.76%	13.59%	14.38%	11.30%
	$G(4)$	20.79%	22.15%	5.79%	12.27%	12.81%	10.09%
Maximum RE	Arrival						
	E_4	26.81%	29.26%	10.78%	21.33%	18.58%	26.71%
	$LN(0.25)$	26.75%	29.29%	10.65%	21.23%	18.61%	26.36%
	$H_2(4)$	66.77%	68.02%	28.13%	23.50%	22.22%	28.86%
	$LN(4)$	86.79%	87.34%	39.50%	23.23%	22.07%	28.59%
	$G(4)$	56.67%	57.85%	24.23%	23.40%	22.04%	29.18%

Table 6.11: The absolute relative error of the RQ approximation at station 2 of two queues in series models with E_4 service times at station 2.

		Service 2	Service 1, $\rho_1 = 0.9$					
		$LN(0.25)$	E_4	$LN(0.25)$	M	$H_2(4)$	$LN(4)$	$G(4)$
Average RE	Arrival	E_4	13.78%	14.41%	2.25%	11.44%	10.31%	9.81%
		$LN(0.25)$	13.61%	14.20%	2.22%	11.42%	10.34%	9.83%
		$H_2(4)$	22.60%	24.00%	7.18%	12.82%	13.46%	10.54%
		$LN(4)$	25.59%	27.01%	8.76%	13.59%	14.37%	11.27%
		$G(4)$	20.58%	21.87%	5.84%	12.27%	12.77%	10.10%
Maximum RE	Arrival	E_4	26.20%	28.82%	10.88%	21.24%	18.60%	26.66%
		$LN(0.25)$	26.16%	28.76%	10.67%	21.20%	18.68%	26.32%
		$H_2(4)$	66.40%	68.26%	28.30%	23.44%	22.18%	28.85%
		$LN(4)$	86.47%	87.76%	39.15%	23.29%	22.32%	28.51%
		$G(4)$	56.43%	57.99%	24.29%	23.17%	21.91%	29.14%

Table 6.12: The absolute relative error of the RQ approximation at station 2 of two queues in series models with $LN(0.25)$ service times at station 2.

		Service 2	Service 1, $\rho_1 = 0.9$					
		M	E_4	$LN(0.25)$	M	$H_2(4)$	$LN(4)$	$G(4)$
Average RE	Arrival	E_4	7.34%	7.53%	1.71%	9.37%	8.17%	7.58%
		$LN(0.25)$	7.26%	7.44%	1.70%	9.36%	8.14%	7.57%
		$H_2(4)$	11.31%	11.80%	4.75%	10.37%	10.51%	8.12%
		$LN(4)$	13.01%	13.50%	6.02%	11.00%	11.17%	8.72%
		$G(4)$	10.35%	10.86%	3.97%	9.95%	9.99%	7.70%
Maximum RE	Arrival	E_4	17.11%	18.14%	7.05%	16.90%	14.18%	20.18%
		$LN(0.25)$	17.12%	18.14%	6.91%	17.17%	14.14%	20.17%
		$H_2(4)$	26.13%	26.21%	16.41%	19.73%	17.36%	22.88%
		$LN(4)$	35.20%	35.11%	24.01%	19.95%	17.32%	22.35%
		$G(4)$	23.82%	23.31%	14.71%	19.57%	17.40%	22.67%

Table 6.13: The absolute relative error of the RQ approximation at station 2 of two queues in series models with M service times at station 2.

Service 2		Service 1, $\rho_1 = 0.9$						
$H_2(4)$		E_4	$LN(0.25)$	M	$H_2(4)$	$LN(4)$	$G(4)$	
Average RE	Arrival	E_4	2.12%	2.02%	0.96%	5.62%	4.47%	3.88%
		$LN(0.25)$	2.10%	1.98%	0.93%	5.60%	4.43%	3.88%
		$H_2(4)$	3.68%	3.72%	2.21%	6.06%	5.60%	3.93%
		$LN(4)$	4.38%	4.37%	2.77%	6.44%	5.93%	4.37%
		$G(4)$	3.40%	3.46%	1.87%	5.77%	5.29%	3.73%
Maximum RE	Arrival	E_4	6.29%	6.10%	2.08%	9.63%	7.88%	9.81%
		$LN(0.25)$	6.33%	6.19%	2.14%	9.50%	7.94%	9.78%
		$H_2(4)$	8.67%	8.52%	6.70%	11.81%	9.95%	11.67%
		$LN(4)$	11.47%	11.20%	8.95%	11.65%	9.60%	11.38%
		$G(4)$	8.01%	7.80%	6.50%	11.92%	9.82%	11.59%

Table 6.14: The absolute relative error of the RQ approximation at station 2 of two queues in series models with $H_2(4)$ service times at station 2.

Service 2		Service 1, $\rho_1 = 0.9$						
$LN(4)$		E_4	$LN(0.25)$	M	$H_2(4)$	$LN(4)$	$G(4)$	
Average RE	Arrival	E_4	2.05%	1.97%	0.92%	5.33%	4.32%	3.89%
		$LN(0.25)$	2.02%	1.95%	0.91%	5.39%	4.37%	3.92%
		$H_2(4)$	3.84%	3.81%	2.28%	5.82%	5.55%	3.92%
		$LN(4)$	4.58%	4.51%	2.88%	6.20%	5.89%	4.37%
		$G(4)$	3.51%	3.54%	1.99%	5.52%	5.23%	3.75%
Maximum RE	Arrival	E_4	7.29%	6.65%	2.41%	9.50%	8.01%	9.52%
		$LN(0.25)$	7.21%	6.68%	2.35%	9.57%	8.18%	9.65%
		$H_2(4)$	9.67%	9.49%	7.29%	11.63%	10.11%	11.20%
		$LN(4)$	12.57%	12.26%	9.50%	11.43%	9.81%	11.18%
		$G(4)$	8.65%	8.72%	7.09%	11.63%	10.08%	11.11%

Table 6.15: The absolute relative error of the RQ approximation at station 2 of two queues in series models with $LN(4)$ service times at station 2.

Service 2		Service 1, $\rho_1 = 0.9$						
$G(4)$		E_4	$LN(0.25)$	M	$H_2(4)$	$LN(4)$	$G(4)$	
Average RE	Arrival	E_4	2.21%	2.28%	0.92%	5.71%	4.65%	4.11%
		$LN(0.25)$	2.18%	2.24%	0.93%	5.69%	4.70%	4.09%
		$H_2(4)$	3.68%	3.80%	2.19%	6.14%	5.75%	4.28%
		$LN(4)$	4.44%	4.54%	2.82%	6.50%	6.14%	4.65%
		$G(4)$	3.47%	3.59%	1.90%	5.74%	5.54%	4.11%
Maximum RE	Arrival	E_4	7.86%	8.06%	1.94%	10.51%	8.17%	10.92%
		$LN(0.25)$	7.82%	8.05%	1.88%	10.42%	8.32%	10.89%
		$H_2(4)$	8.07%	8.17%	6.72%	12.96%	10.70%	12.63%
		$LN(4)$	10.98%	10.81%	8.53%	12.77%	10.49%	12.36%
		$G(4)$	7.38%	7.50%	6.11%	12.44%	10.80%	12.72%

Table 6.16: The absolute relative error of the RQ approximation at station 2 of two queues in series models with $G(4)$ service times at station 2.

Service 2		Service 1, $\rho_1 = 0.7$						
E_4		E_4	$LN(0.25)$	M	$H_2(4)$	$LN(4)$	$G(4)$	
Average RE	External arrival	E_4	13.34%	13.63%	5.80%	10.84%	10.40%	8.89%
		$LN(0.25)$	13.21%	13.52%	5.81%	10.79%	10.34%	8.88%
		$H_2(4)$	18.57%	19.57%	10.77%	13.56%	13.35%	11.42%
		$LN(4)$	20.46%	21.33%	12.20%	14.45%	14.10%	12.16%
		$G(4)$	14.66%	15.71%	8.68%	12.63%	12.36%	10.14%
Maximum RE	External arrival	E_4	25.13%	26.83%	13.21%	27.51%	27.26%	27.12%
		$LN(0.25)$	25.31%	26.95%	12.97%	27.42%	27.11%	27.10%
		$H_2(4)$	41.71%	41.56%	24.82%	28.95%	26.08%	27.50%
		$LN(4)$	48.28%	48.09%	34.30%	27.88%	25.03%	26.93%
		$G(4)$	27.09%	27.00%	17.12%	29.13%	27.02%	27.40%

Table 6.17: The absolute relative error of the RQ approximation at station 2 of two queues in series models with $\rho_1 = 0.7$ and E_4 service times at station 2.

Service 2		Service 1, $\rho_1 = 0.9$					
M		E_4	$LN(0.25)$	M	$H_2(4)$	$LN(4)$	$G(4)$
$0 < \rho_2 < 1$	Arrival						
	E_4	17.11%	18.14%	7.05%	16.90%	14.18%	20.18%
	$LN(0.25)$	17.12%	18.14%	6.91%	17.17%	14.14%	20.17%
	$H_2(4)$	26.13%	26.21%	16.41%	19.73%	17.36%	22.88%
	$LN(4)$	35.20%	35.11%	24.01%	19.95%	17.32%	22.35%
	$G(4)$	23.82%	23.31%	14.71%	19.57%	17.40%	22.67%
$0.5 < \rho_2 < 1$	Arrival						
	E_4	4.43%	3.13%	7.05%	11.20%	9.13%	11.85%
	$LN(0.25)$	4.32%	3.04%	6.91%	11.30%	9.16%	12.01%
	$H_2(4)$	26.13%	26.21%	16.41%	10.63%	13.71%	3.86%
	$LN(4)$	35.20%	35.11%	24.01%	11.37%	16.01%	5.92%
	$G(4)$	23.82%	23.31%	14.71%	9.28%	12.11%	2.83%

Table 6.18: The maximum absolute relative error of the RQ approximation at station 2 of two queues in series models with $\rho_1 = 0.7$ and M service times at station 2.

Service 2		Service 1					
$LN(4)$		E_4	$LN(0.25)$	M	$H_2(4)$	$LN(4)$	$G(4)$
$\rho_1 = 0.7$	Arrival						
	E_4	6.19%	5.57%	3.35%	8.11%	7.76%	8.21%
	$LN(0.25)$	6.26%	5.60%	3.26%	8.08%	7.47%	8.24%
	$H_2(4)$	9.85%	10.03%	7.97%	13.24%	11.33%	11.07%
	$LN(4)$	12.80%	12.76%	10.55%	12.33%	10.47%	10.63%
	$G(4)$	8.43%	8.23%	6.54%	13.35%	10.76%	11.33%
$\rho_1 = 0.9$	Arrival						
	E_4	7.29%	6.65%	2.41%	9.50%	8.01%	9.52%
	$LN(0.25)$	7.21%	6.68%	2.35%	9.57%	8.18%	9.65%
	$H_2(4)$	9.67%	9.49%	7.29%	11.63%	10.11%	11.20%
	$LN(4)$	12.57%	12.26%	9.50%	11.43%	9.81%	11.18%
	$G(4)$	8.65%	8.72%	7.09%	11.63%	10.08%	11.11%

Table 6.19: Comparing the RQ performance for $\rho_1 = 0.7$ and 0.9.

stipulate that the rate $\lambda \equiv \frac{\lambda_1 r_2 + \lambda_2 r_1}{r_1 + r_2} = 1$, which yields the following solution

$$\begin{aligned}\lambda_2 &= \frac{2 - \gamma + c_a^2 \gamma - 2\lambda_1}{2(1 - \lambda_1)} \\ r_1 &= \frac{2\gamma(1 - \lambda_1)^2}{2 - \gamma + c_a^2 \gamma - 4\lambda_1 + 2\lambda_1^2} \\ r_2 &= \frac{(c_a^2 - 1)\gamma^2}{2 - \gamma + c_a^2 \gamma - 4\lambda_1 + 2\lambda_1^2},\end{aligned}\tag{6.6}$$

where $\gamma = (1 - p)\lambda_1 + p\lambda_2$ and λ_1 is a free variable such that $\lambda_1, \lambda_2, r_1, r_2 \geq 0$. Consider the special case of $c_a^2 = 4$, and $r = 0.5$. The range of feasible λ_1 is $[0, 1) \cup [1.6, \infty)$, which can be easily solved from a set of quadratic inequalities $\lambda_1, \lambda_2, r_1, r_2 \geq 0$. Figure 6.11 displays the mean steady-state workload of the MMPP(2) defined by (2.61) and (6.6) with $\lambda = 1$, $c_a^2 = 4$ and $r = 0.5$. We show multiple choices of $\lambda_1 \in \{1.6, 3.2, 6.4, 12.8, 25.6\}$ within the feasible set. We do not show examples with $\lambda_1 \in [0, 1)$ because we can always swap the two states and assume that $\lambda_1 \geq 1.6$ without loss of generality. In Corollary 3.1, we show that $GI/GI/1$ is completely determined by the IDC. Note that the mean steady-state workload in the case of $\lambda = 1.6$ coincides with that of the model with H_2 arrival because it corresponds to the interrupted Poisson process, which is equivalent to the two-phase hyperexponential distribution [112].

The examples here show that even if the IDC's remain the same across all cases, the possible range of the performance measure can be substantial. Since the current version of RQ depends only on the IDC in this example, we cannot expect it to work consistently well across these examples. However, we would like to point out that these models are quite abnormal in the sense that the generating matrix of the MAP have extremely small entries. For example, when $\lambda_1 = 25.6$, we have $r_2 = 0.0004$. This implies that the underlying state is rarely switched to state 2, but when it does, the arrival rate changes dramatically ($\lambda_1 = 25.6$ versus $\lambda_2 = 0.976$).

6.3 Robust Queueing Network Analyzer for Tandem Queues

In this section, we compare the RQNA approximation to the simulation estimation of the mean steady-state workload in various queues in series models. As discussed in Chapter 3.3, the only relevant network operation is the departure operation.

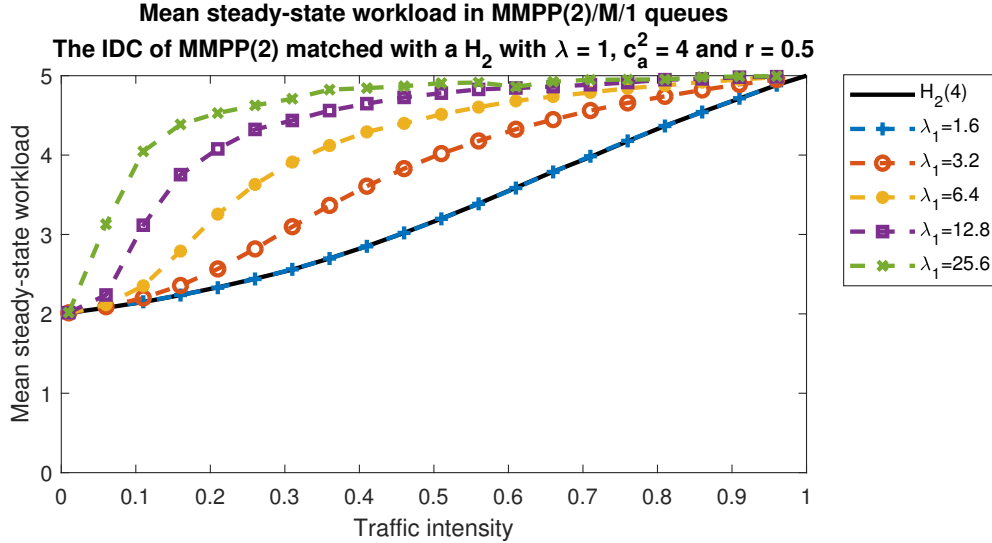


Figure 6.11: Comparing the mean steady-state workload in single-server queues with identical arrival IDC and service-time distribution.

In Section 6.3.1, we look at the performance of the RQNA approximation for the stationary departure processes. In Section 6.3.4, we investigate the performance of our RQNA algorithm in tandem queues.

6.3.1 Departure IDC Approximation in $G/G/1$ Models

We start with the RQNA approximation of the departure IDC, discussed in Section 5.1.

Figure 6.12 contrasts the simulation estimation of the departure IDC's with the RQNA approximations in (5.1) for $G(4)/GI/1$ and $E_4/GI/1$ models with various service time-distributions: E_4 , $LN(0.25)$, M , $H_2(4)$, $LN(4)$ and $G(4)$. For each model, two cases are displayed: $\rho_1 = 0.9$ and $\rho_1 = 0.7$. These plots validate Theorem 5.1 by stunning approximation performance. RQNA performance for $\rho_1 = 0.7$ is satisfactory but not as good as the cases with higher traffic intensity.

Recall that the theoretical support for our departure approximation is established in Theorem 5.1 only for generalized Jackson network. But we conjecture that this heavy-traffic limit theorem for the departure variance function holds in much more general settings in Conjecture 5.1. We now provide numerical support for it.

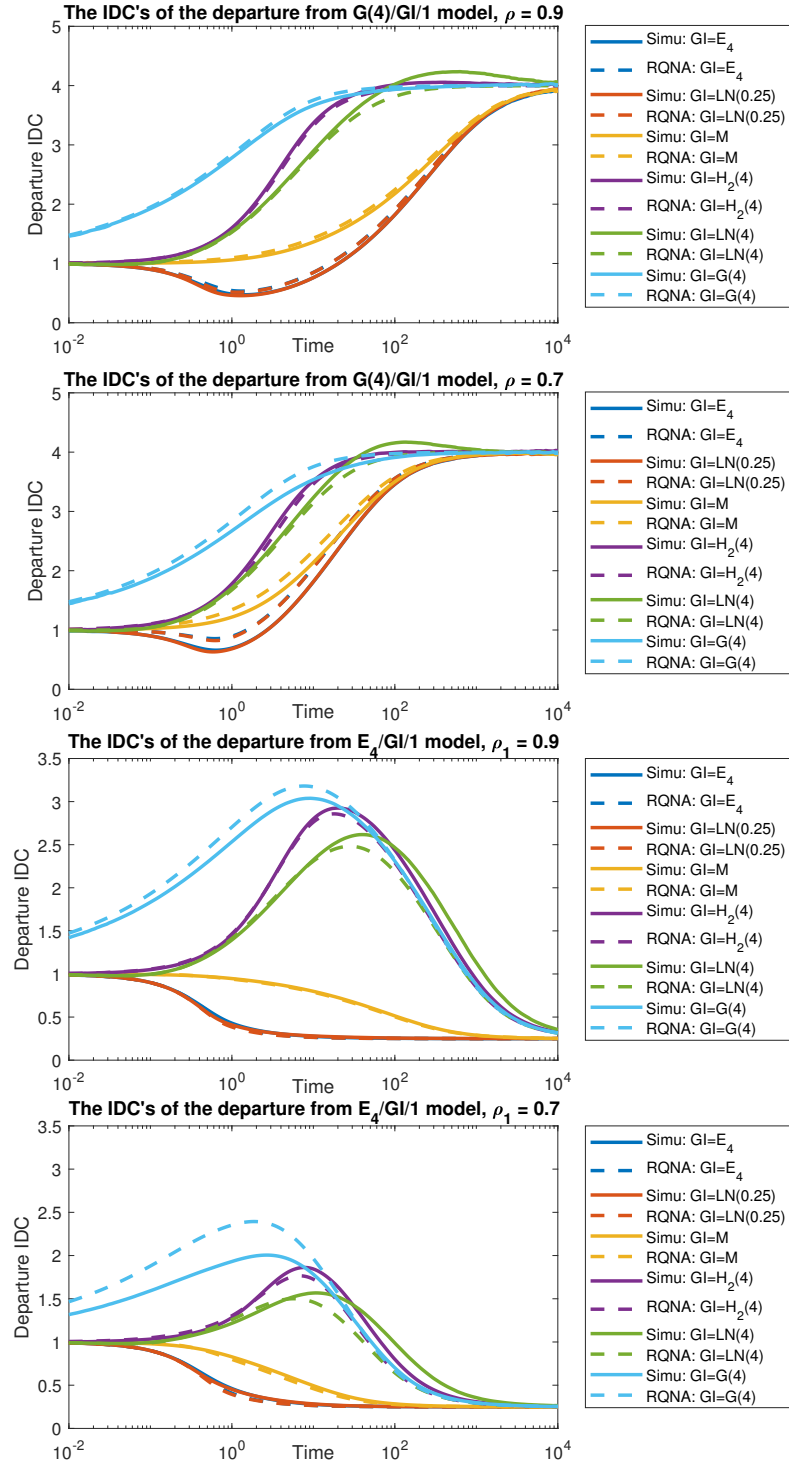


Figure 6.12: The RQNA approximation of the departure IDC of the $G(4)/GI/1$ and $E_4/GI/1$ models.

Consider the MAP defined by (6.1) for the arrival process, or by (6.5) for the service process. Figure 6.13 displays the RQNA approximation for the departure IDC in various $G/G/1$ models with MAP as arrival or service processes. Each plot focuses on the same $G/G/1$ model, but displays cases with $\rho_1 = 0.5, 0.7$ and 0.9 .

We observe that the RQNA approximation is asymptotically exact in HT limit for all cases, strongly supporting Conjecture 5.1. We also see that the approximation works well in $MAP/GI/1$ models but not as well in $GI/MAP/1$ models, leaving us some headroom for future refinements in light traffic.

6.3.2 An Illustrative Example

In this section, we consider a similar example as in Section 2.2.9, where the normalized workload as a function of ρ also has several modes, but the external arrival here has high variability.

In this example we use groups of queues in series with the same distribution and traffic intensity in order to better bring about an adjustment in the level of variability. Specifically, this example has 13 single-server queues in series. The external arrival process is a rate-1 renewal process with H_2 interarrival times with $c_a^2 = 10$. A group of three queues having E_{10} service times with mean 0.99 is then added to smooth the highly variable external arrivals. The next group of three queues has H_2 service times with mean 0.92 and squared coefficient of variation 5. These queues will bring up the variability of the departure process. Then, another group of three queues with mean 0.9 has E_{10} service times to smooth the departure process again. The variability is then raised by yet another group of three queues having H_2 service times with mean 0.3 and $c_s^2 = 10$. Finally, the last (13th) queue has exponential service times with mean and traffic intensity ρ . As before, we explore the impact of ρ on the performance of that last queue.

As explained in last example, for sufficiently low traffic intensities ρ at the last queue, the last queue should behave approximately the same as an $H_2/M/1$ queue, which has $c_a^2 = 10$, but as ρ increases, the arrival process at the last queue should inherit the variability of the previous service times and the external arrival process, and alter between $E_{10}/M/1$ and $H_2/M/1$ as the traffic intensity at the last queue increases. This implies that the normalized

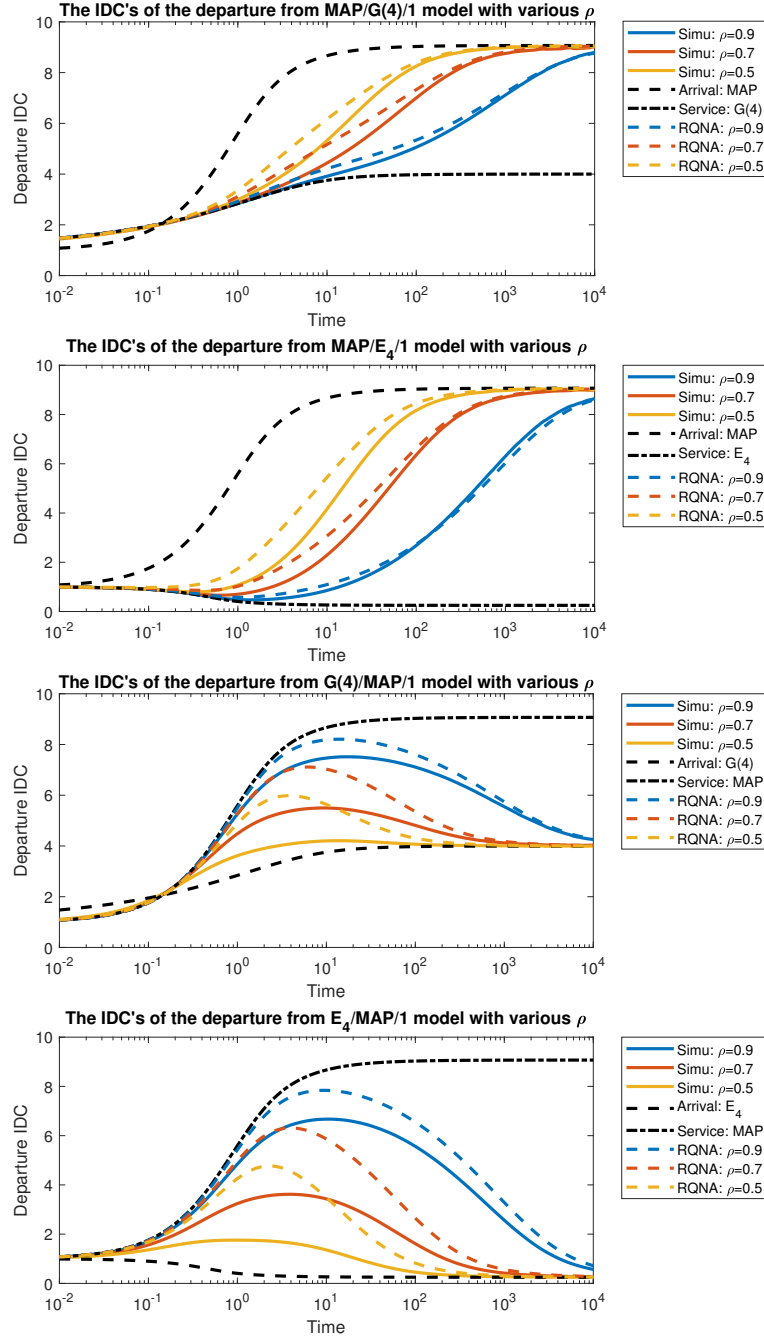


Figure 6.13: The RQNA approximation of the departure IDC of the models with MAP.

workload $c_Z^2(\rho)$ in (2.27) as a function of ρ should have several modes, corresponding to the variability of the external arrival process and the service processes at the first 4 groups of queues.

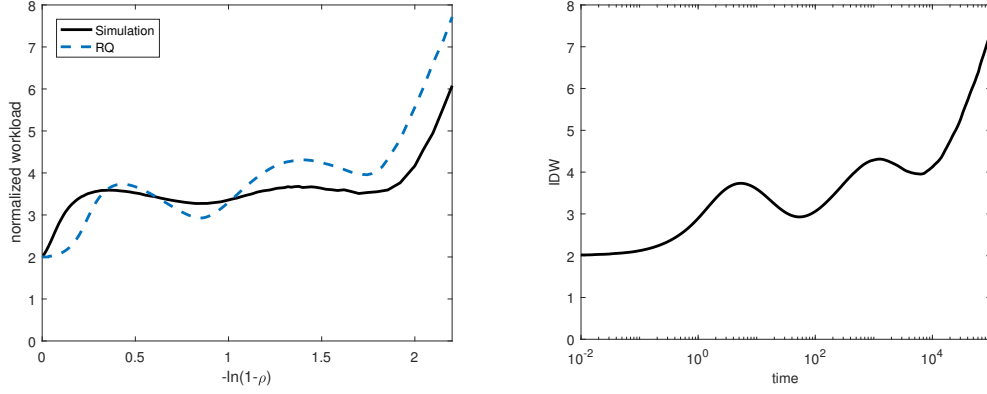


Figure 6.14: Comparing the simulation estimation to the RQ approximation for the workload, as a function of traffic intensity, at the last queue of a thirteen-queues-in-series model. The workload function have four internal modes.

We then have the similar plots in Figure 6.14, which compares simulation estimates of the normalized mean workload $c_Z^2(\rho)$ in (2.27) at the last queue with the RQ approximation $c_{Z^*}^2(\rho)$ in (2.36) (left) and shows the IDW for this example (right). Again, we are using the same scale as in Figure 2.2 (left), i.e., $-\ln(1-\rho)$, to stretch out the plot under heavy traffic.

Figure 6.14 (left) shows that the the normalized workload at the last queue again has four internal modes and that RQ successfully captures all modes and provides a reasonably accurate approximation for all ρ . Figure 6.14 (right) shows that the IDW has the same qualitative property as the RQ approximation, which is explained in (2.40). However, the fluctuations in the simulation values for $0 < \rho < 1$ in Figure 6.14 are much less than in Figure 2.2.

We conclude that (i) the IDW and RQ do capture the qualititative behavior and (ii) the RQ approximation based on the IDW is reasonably accurate in these difficult examples.

Queue	Sim	QNA	QNET	SBD	RQ	RQNA
1	0.290 (2.41%)	0.45 (55%)	0.45 (55%)	0.45 (55%)	0.30 (2.3%)	0.30(2.3%)
2	0.491 (1.43%)	0.61 (24%)	0.66 (35%)	0.66 (35%)	0.55 (13%)	0.53 (8.1%)
3	0.607 (1.32%)	0.72 (19%)	0.74 (22%)	0.74 (22%)	0.70 (15%)	0.66 (9.4%)
4	0.666 (1.20%)	0.78 (17%)	0.79 (18%)	0.79 (19%)	0.77 (16%)	0.74 (11%)
5	0.706 (1.42%)	0.83 (18%)	0.82 (16%)	0.82 (16%)	0.80 (14%)	0.79 (12%)
6	0.731 (1.78%)	0.85 (16%)	0.84 (14%)	0.84 (15%)	0.83 (13%)	0.82 (13%)
7	0.748 (1.34%)	0.87 (16%)	0.85 (14%)	0.85 (14%)	0.84 (12%)	0.85 (13%)
8	0.775 (1.68%)	0.88 (14%)	0.86 (11%)	0.86 (11%)	0.85 (9.2%)	0.86 (11%)
9	5.031 (4.31%)	7.99 (59%)	6.97 (39%)	4.05 (-20%)	4.95 (-2.0%)	4.50 (-11%)
Total	10.05	14.0 (39%)	13.0 (29%)	10.1 (0.09%)	10.6 (5.3%)	10.1 (0.13%)

Table 6.20: A comparison of four approximation methods to simulation for 9 exponential (M) queues in series fed by a deterministic arrival process with $c_a^2 = 0$.

6.3.3 Comparisons with Previous Algorithms for Queues in Series

In this section, we compare the performance of our RQNA algorithm to the performance of QNA from [134], QNET from [73], SBD from [44] and RQ from [145], for the example with 9 queues in series considered by [125]. This example was introduced by [125] to illustrate the heavy-traffic bottleneck phenomenon and to show the limitation of traditional decomposition methods, e.g. the QNA algorithm.

In particular, we consider an OQN with 9 stations in tandem, each with i.i.d. exponential service times. Station 1 has the only external arrival process, which is a rate-1 general renewal process. The traffic intensities at the first 8 queues are set to $\rho_i = 0.6$ for $1 \leq i \leq 8$, while the last queue has the significantly higher traffic intensity $\rho_9 = 0.9$. As in [125], two specific external renewal arrival processes are considered: (i) deterministic interarrival times with $c_{a_0}^2 = 0$; and (ii) highly variable $H_2(8)$ interarrival times with $c_{a_0}^2 = 8$ (and again balanced means).

Table 6.20 (for low variability) and Table 6.21 (for high variability) compare the various approximations of the mean steady-state waiting time at each station, as well as the total waiting time in the system, to simulation estimates.

In the parentheses, we include (i) the relative half-width of the 95% confidence interval

Queue	Sim	QNA	QNET	SBD	RQ	RQNA
1	3.284 (3.50%)	4.05 (23%)	4.05 (23%)	4.05 (23%)	3.95 (20%)	3.95 (20%)
2	2.321 (4.18%)	2.92 (26%)	1.81 (22%)	1.82 (-22%)	2.61 (12%)	1.95 (-15%)
3	1.914 (3.40%)	2.19 (14%)	1.47 (-23%)	1.49 (-22%)	2.04 (6.7%)	1.07 (-44%)
4	1.719 (4.07%)	1.73 (0.64%)	1.16 (-33%)	1.19 (-31%)	1.72 (0.31%)	0.94 (-41%)
5	1.598 (3.69%)	1.43 (-11%)	1.07 (-33%)	1.10 (-31%)	1.53 (-4.1%)	0.91 (-43%)
6	1.478 (4.13%)	1.24 (-16%)	1.03 (-31%)	1.06 (-28%)	1.41 (-4.6%)	0.90 (-39%)
7	1.423 (3.23%)	1.12 (-21%)	1.00 (-30%)	1.03 (-28%)	1.33 (-6.8%)	0.90 (-37%)
8	1.413 (4.67%)	1.04 (-26%)	0.98 (-30%)	1.01 (-29%)	1.27 (-10%)	0.90 (-36%)
9	30.12 (16.8%)	8.90 (-71%)	6.04 (-80%)	36.5 (21%)	36.9 (23%)	32.8 (9.0%)
Total	45.27	24.6 (-46%)	18.6 (-59%)	49.8 (10%)	52.8 (17%)	44.4 (-2.0%)

Table 6.21: A comparison of four approximation methods to simulation for 9 exponential (M) queues in series fed by a highly-variable H_2 renewal arrival process with $c_a^2 = 8$.

for simulation estimates (column Sim); and (ii) the relative error of the approximations compared to the simulation estimates. The first 5 columns in Table 6.20 and Table 6.21 are taken directly from Tables VIII and IX of [44], but the simulation and QNA approximations come from [125]. The last column is obtained from the RQNA algorithm. The RQNA approximations of the workload are transformed into the approximations of the waiting time by (2.50).

To put these performance measures in perspective, note that in an $M/M/1$ queue with arrival rate 1 we would have $EW = \rho^2/(1 - \rho)$, which would be 0.90 at the first 8 queues, but 8.1 at the last queue. For the D arrival process in Table 6.20, we expect that EW will be smaller; for the the H_2 arrival process in Table 6.21, we expect EW to be higher, but we see a big impact at the last queue, more than might be expected.

We make the following observations from this experiment:

1. The new RQNA algorithm does better than the QNA and QNET methods on total time spent waiting in queue, and is comparable with the SBD method, even though RQNA does not require solving an RBM.
2. The RQNA algorithm does exceptionally well at the final bottleneck queue and is competitive with all other methods for approximating the mean waiting time. The

new RQNA method is based on heavy-traffic limits just as the previous methods, but focuses on the flows, and exploits RQ instead of analyzing an RBM.

3. The RQNA algorithm can benefit from further improvement for light-to-medium traffic intensities. As demonstrated in Table 6.21, the mean waiting times at queues 3-8 are pushed too much towards the $M/M/1$ values in the departure IDC approximation for light to medium traffic intensity. That remains to be a direction for future research.

6.3.4 RQNA Performance in Tandem Queueus

In this section, we systematically investigate the RQNA approximation in the queues in series models, described in Section 3.3.2. In particular, we compare the simulation estimation, the RQ approximation and the RQNA approximation for examples in Section 6.2.5.

In Table 6.22 - 6.25, we look at various $GI_1/GI_2/1 \rightarrow \cdot/GI_3/1$ models with $\rho_1 = 0.9$, $GI_1, GI_3 \in \{E_4, LN(4)\}$ and $GI_2 \in \{E_4, LN(0.25), M, H_2(4), LN(4), G(4)\}$. In Table 6.26 - 6.29, we look at analog tables but with $\rho_1 = 0.7$. We observe that RQNA approximation matches closely with the RQ approximation, even though the IDC is calculated from our IDC equations instead of simulation estimations. Both approximation provide effective approximation of the simulated values.

6.4 Robust Queueing Network Analyzer for Open Queueing Networks

We discuss examples of networks with significant near-immediate feedback from [44]. We show that the near-immediate feedback in these examples have a significant impact on the performance measures. As discussed in Section 5.5, the RQNA algorithm can benefit from the feedback elimination procedure when customer feedback is present. Hence our predictions with and without feedback elimination are very different. We find that our RQNA with near-immediate feedback elimination performs as well or better than the other algorithms. We remark that the SBD algorithm performed remarkably well in these examples.

External arrival $LN(4)$, Service 2 = E_4 , $\rho_1 = 0.9$

ρ_2	Service 1 = E_4			Service 1 = $LN(0.25)$			Service 1 = M		
	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	1.12	1.24(11%)	1.23(10%)	1.11	1.24(12%)	1.24(11%)	1.26	1.25(-1%)	1.26(0%)
0.16	1.06	1.23(16%)	1.22(15%)	1.05	1.23(17%)	1.23(16%)	1.26	1.25(-1%)	1.26(0%)
0.21	1.01	1.21(20%)	1.21(20%)	1.00	1.21(21%)	1.21(22%)	1.27	1.25(-2%)	1.25(-1%)
0.26	0.96	1.18(23%)	1.18(23%)	0.94	1.18(25%)	1.18(25%)	1.28	1.25(-2%)	1.26(-2%)
0.31	0.92	1.15(25%)	1.14(25%)	0.90	1.15(27%)	1.15(27%)	1.28	1.25(-3%)	1.26(-2%)
0.36	0.88	1.10(25%)	1.10(25%)	0.86	1.10(28%)	1.10(28%)	1.29	1.25(-3%)	1.26(-2%)
0.41	0.85	1.05(24%)	1.04(23%)	0.82	1.05(27%)	1.04(26%)	1.30	1.25(-4%)	1.27(-3%)
0.46	0.82	0.99(21%)	0.98(20%)	0.79	0.98(24%)	0.97(23%)	1.32	1.26(-4%)	1.28(-3%)
0.51	0.79	0.92(16%)	0.91(15%)	0.77	0.91(18%)	0.90(17%)	1.33	1.26(-5%)	1.29(-3%)
0.56	0.77	0.85(9%)	0.84(9%)	0.75	0.83(11%)	0.82(10%)	1.35	1.28(-5%)	1.31(-2%)
0.61	0.76	0.77(2%)	0.77(1%)	0.74	0.75(2%)	0.74(1%)	1.37	1.30(-5%)	1.35(-2%)
0.66	0.75	0.70(-6%)	0.72(-4%)	0.73	0.68(-7%)	0.71(-3%)	1.40	1.33(-5%)	1.40(1%)
0.71	0.76	0.70(-8%)	0.77(1%)	0.74	0.68(-7%)	0.76(3%)	1.44	1.38(-4%)	1.50(4%)
0.76	0.78	0.77(-1%)	0.93(20%)	0.76	0.77(0%)	0.93(22%)	1.49	1.48(-0%)	1.66(11%)
0.81	0.84	1.00(19%)	1.28(53%)	0.83	1.00(21%)	1.28(55%)	1.58	1.67(6%)	1.92(22%)
0.86	0.99	1.50(51%)	1.85(87%)	0.99	1.50(52%)	1.85(87%)	1.75	2.03(16%)	2.34(34%)
0.91	1.51	2.40(59%)	2.70(79%)	1.51	2.40(59%)	2.70(78%)	2.12	2.69(27%)	2.96(39%)
0.96	2.83	3.71(31%)	3.82(35%)	2.83	3.71(31%)	3.82(35%)	3.03	3.75(24%)	3.87(28%)

ρ	Service 1 = $H_2(4)$			Service 1 = $LN(4)$			Service 1 = $G(4)$		
	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	1.36	1.26(-8%)	1.25(-8%)	1.30	1.24(-4%)	1.23(-5%)	2.47	1.78(-28%)	1.76(-29%)
0.16	1.43	1.26(-11%)	1.26(-11%)	1.37	1.23(-10%)	1.23(-10%)	2.66	1.96(-26%)	1.93(-27%)
0.21	1.50	1.28(-15%)	1.28(-15%)	1.45	1.23(-15%)	1.23(-15%)	2.81	2.15(-24%)	2.11(-25%)
0.26	1.60	1.30(-19%)	1.31(-18%)	1.54	1.24(-20%)	1.24(-19%)	2.95	2.34(-21%)	2.30(-22%)
0.31	1.71	1.33(-22%)	1.35(-21%)	1.64	1.27(-23%)	1.28(-22%)	3.08	2.54(-18%)	2.50(-19%)
0.36	1.84	1.39(-24%)	1.41(-23%)	1.74	1.33(-24%)	1.36(-22%)	3.19	2.74(-14%)	2.71(-15%)
0.41	2.00	1.49(-25%)	1.53(-23%)	1.86	1.44(-22%)	1.49(-20%)	3.29	2.95(-10%)	2.92(-11%)
0.46	2.18	1.67(-23%)	1.75(-20%)	1.99	1.60(-19%)	1.67(-16%)	3.39	3.15(-7%)	3.13(-8%)
0.51	2.38	1.98(-17%)	2.12(-11%)	2.12	1.82(-14%)	1.91(-10%)	3.48	3.35(-4%)	3.33(-4%)
0.56	2.60	2.40(-8%)	2.59(-0%)	2.27	2.08(-9%)	2.21(-3%)	3.58	3.53(-1%)	3.53(-1%)
0.61	2.83	2.85(1%)	3.07(8%)	2.44	2.37(-3%)	2.54(4%)	3.66	3.68(1%)	3.70(1%)
0.66	3.07	3.26(6%)	3.48(13%)	2.63	2.69(2%)	2.91(11%)	3.73	3.81(2%)	3.86(3%)
0.71	3.30	3.57(8%)	3.81(16%)	2.84	3.03(7%)	3.31(17%)	3.81	3.92(3%)	3.99(5%)
0.76	3.52	3.80(8%)	4.07(16%)	3.06	3.35(9%)	3.70(21%)	3.89	4.00(3%)	4.11(6%)
0.81	3.74	3.96(6%)	4.26(14%)	3.34	3.66(10%)	4.07(22%)	3.95	4.07(3%)	4.22(7%)
0.86	3.91	4.08(4%)	4.40(13%)	3.64	3.92(8%)	4.40(21%)	4.02	4.13(3%)	4.31(7%)
0.91	4.07	4.17(2%)	4.47(10%)	3.98	4.12(3%)	4.61(16%)	4.07	4.18(3%)	4.37(7%)
0.96	4.17	4.22(1%)	4.38(5%)	4.28	4.21(-1%)	4.55(6%)	4.11	4.22(3%)	4.33(5%)

Table 6.22: The RQNA approximation in various $LN(4)/GI/1 \rightarrow \cdot/E_4/1$ models with $\rho_1 = 0.9$.

External arrival E_4 , Service 2 = $LN(4)$, $\rho_1 = 0.9$

ρ_2	Service 1 = E_4			Service 1 = $LN(0.25)$			Service 1 = M		
	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	4.70	4.96(6%)	4.96(6%)	4.65	4.96(7%)	4.96(7%)	4.96	5.00(1%)	5.00(1%)
0.16	4.58	4.91(7%)	4.91(7%)	4.62	4.91(6%)	4.91(6%)	5.01	5.00(-0%)	4.99(-0%)
0.21	4.53	4.84(7%)	4.84(7%)	4.69	4.83(3%)	4.83(3%)	4.95	4.99(1%)	4.99(1%)
0.26	4.63	4.73(2%)	4.74(2%)	4.55	4.73(4%)	4.73(4%)	4.91	4.98(2%)	4.98(2%)
0.31	4.53	4.61(2%)	4.62(2%)	4.52	4.59(2%)	4.60(2%)	4.90	4.97(2%)	4.97(2%)
0.36	4.50	4.49(-0%)	4.51(0%)	4.45	4.46(0%)	4.47(0%)	4.88	4.95(1%)	4.96(1%)
0.41	4.44	4.40(-1%)	4.42(-0%)	4.45	4.38(-2%)	4.40(-1%)	4.89	4.93(1%)	4.94(1%)
0.46	4.43	4.35(-2%)	4.37(-1%)	4.41	4.34(-2%)	4.36(-1%)	4.86	4.91(1%)	4.92(1%)
0.51	4.42	4.32(-2%)	4.34(-2%)	4.41	4.31(-2%)	4.33(-2%)	4.85	4.89(1%)	4.90(1%)
0.56	4.37	4.30(-2%)	4.31(-1%)	4.36	4.29(-2%)	4.31(-1%)	4.85	4.87(0%)	4.87(0%)
0.61	4.35	4.28(-2%)	4.29(-1%)	4.35	4.28(-2%)	4.30(-1%)	4.87	4.84(-1%)	4.84(-0%)
0.66	4.33	4.27(-1%)	4.28(-1%)	4.36	4.27(-2%)	4.29(-2%)	4.81	4.80(-0%)	4.81(-0%)
0.71	4.34	4.26(-2%)	4.27(-2%)	4.35	4.26(-2%)	4.28(-2%)	4.76	4.76(-0%)	4.77(0%)
0.76	4.30	4.26(-1%)	4.26(-1%)	4.34	4.26(-2%)	4.27(-2%)	4.72	4.70(-0%)	4.71(-0%)
0.81	4.31	4.25(-1%)	4.26(-1%)	4.29	4.25(-1%)	4.26(-1%)	4.65	4.63(-1%)	4.64(-0%)
0.86	4.31	4.25(-1%)	4.25(-1%)	4.28	4.25(-1%)	4.26(-0%)	4.58	4.53(-1%)	4.54(-1%)
0.91	4.27	4.25(-0%)	4.25(-0%)	4.27	4.25(-1%)	4.25(-0%)	4.52	4.40(-3%)	4.41(-2%)
0.96	4.25	4.25(-0%)	4.25(-0%)	4.26	4.25(-0%)	4.25(-0%)	4.38	4.28(-2%)	4.28(-2%)

ρ	Service 1 = $H_2(4)$			Service 1 = $LN(4)$			Service 1 = $G(4)$		
	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	5.26	5.02(-4%)	5.02(-5%)	5.19	4.98(-4%)	4.98(-4%)	6.21	5.74(-8%)	5.66(-9%)
0.16	5.32	5.05(-5%)	5.05(-5%)	5.23	4.99(-5%)	4.99(-4%)	6.47	5.97(-8%)	5.86(-10%)
0.21	5.46	5.10(-7%)	5.09(-7%)	5.33	5.04(-5%)	5.04(-5%)	6.51	6.18(-5%)	6.05(-7%)
0.26	5.70	5.17(-9%)	5.16(-10%)	5.55	5.12(-8%)	5.12(-8%)	6.60	6.38(-3%)	6.23(-6%)
0.31	5.78	5.28(-9%)	5.27(-9%)	5.59	5.24(-6%)	5.23(-6%)	6.78	6.57(-3%)	6.40(-6%)
0.36	5.89	5.43(-8%)	5.42(-8%)	5.67	5.38(-5%)	5.36(-5%)	6.74	6.75(0%)	6.57(-3%)
0.41	5.97	5.65(-5%)	5.64(-6%)	5.72	5.54(-3%)	5.52(-4%)	6.83	6.90(1%)	6.71(-2%)
0.46	6.11	5.92(-3%)	5.92(-3%)	5.81	5.72(-2%)	5.69(-2%)	6.80	7.03(3%)	6.84(1%)
0.51	6.22	6.22(-0%)	6.23(0%)	5.91	5.90(-0%)	5.87(-1%)	6.88	7.12(3%)	6.95(1%)
0.56	6.31	6.50(3%)	6.52(3%)	5.98	6.07(2%)	6.06(1%)	6.80	7.17(6%)	7.01(3%)
0.61	6.39	6.71(5%)	6.75(6%)	6.05	6.24(3%)	6.24(3%)	6.88	7.18(4%)	7.04(2%)
0.66	6.45	6.83(6%)	6.88(7%)	6.13	6.37(4%)	6.41(5%)	6.85	7.14(4%)	7.02(2%)
0.71	6.52	6.85(5%)	6.92(6%)	6.22	6.46(4%)	6.54(5%)	6.74	7.04(4%)	6.95(3%)
0.76	6.47	6.77(5%)	6.87(6%)	6.23	6.48(4%)	6.61(6%)	6.66	6.88(3%)	6.82(2%)
0.81	6.44	6.58(2%)	6.70(4%)	6.25	6.39(2%)	6.59(5%)	6.52	6.64(2%)	6.60(1%)
0.86	6.30	6.25(-1%)	6.38(1%)	6.19	6.15(-1%)	6.42(4%)	6.28	6.28(-0%)	6.27(-0%)
0.91	5.95	5.69(-4%)	5.82(-2%)	5.97	5.65(-5%)	5.98(0%)	5.92	5.70(-4%)	5.71(-4%)
0.96	5.29	4.75(-10%)	4.82(-9%)	5.42	4.74(-12%)	4.98(-8%)	5.25	4.75(-10%)	4.77(-9%)

Table 6.23: The RQNA approximation in various $E_4/GI/1 \rightarrow \cdot/LN(4)/1$ models with $\rho_1 = 0.9$.

External arrival E_4 , Service 2 = E_4 , $\rho_1 = 0.9$

	Service 1 = E_4			Service 1 = $LN(0.25)$, $\rho_1 = 0.9$			Service 1 = M		
ρ_2	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	1.12	1.24(11%)	1.24(11%)	1.11	1.24(12%)	1.24(11%)	1.24	1.25(1%)	1.25(1%)
0.16	1.06	1.23(16%)	1.23(16%)	1.05	1.23(17%)	1.23(17%)	1.24	1.25(1%)	1.25(1%)
0.21	1.00	1.21(20%)	1.21(20%)	0.99	1.21(22%)	1.21(22%)	1.23	1.25(2%)	1.25(1%)
0.26	0.95	1.18(24%)	1.18(24%)	0.94	1.18(25%)	1.18(25%)	1.22	1.25(2%)	1.25(2%)
0.31	0.91	1.14(26%)	1.14(26%)	0.89	1.14(28%)	1.14(28%)	1.22	1.25(2%)	1.24(2%)
0.36	0.87	1.10(27%)	1.10(27%)	0.85	1.10(29%)	1.10(29%)	1.21	1.24(3%)	1.24(2%)
0.41	0.83	1.04(26%)	1.04(26%)	0.81	1.04(29%)	1.04(29%)	1.20	1.23(3%)	1.23(2%)
0.46	0.79	0.98(24%)	0.98(24%)	0.77	0.98(26%)	0.97(26%)	1.20	1.22(2%)	1.22(2%)
0.51	0.76	0.91(20%)	0.91(20%)	0.74	0.90(22%)	0.90(22%)	1.18	1.21(2%)	1.21(2%)
0.56	0.72	0.83(15%)	0.84(15%)	0.71	0.82(16%)	0.82(16%)	1.17	1.19(2%)	1.20(2%)
0.61	0.70	0.75(9%)	0.76(9%)	0.68	0.74(9%)	0.74(9%)	1.16	1.18(1%)	1.18(2%)
0.66	0.67	0.67(1%)	0.68(3%)	0.65	0.65(0%)	0.66(1%)	1.15	1.15(1%)	1.16(1%)
0.71	0.64	0.60(-5%)	0.62(-3%)	0.63	0.59(-6%)	0.61(-3%)	1.13	1.13(0%)	1.14(1%)
0.76	0.61	0.56(-8%)	0.58(-5%)	0.61	0.55(-9%)	0.58(-5%)	1.10	1.10(-0%)	1.10(0%)
0.81	0.59	0.53(-9%)	0.55(-7%)	0.59	0.53(-9%)	0.55(-6%)	1.07	1.05(-1%)	1.06(-1%)
0.86	0.56	0.52(-8%)	0.53(-7%)	0.56	0.51(-9%)	0.53(-6%)	1.01	0.99(-3%)	1.00(-2%)
0.91	0.54	0.51(-6%)	0.51(-5%)	0.54	0.51(-7%)	0.52(-5%)	0.93	0.88(-5%)	0.89(-4%)
0.96	0.52	0.50(-3%)	0.50(-3%)	0.52	0.50(-4%)	0.50(-3%)	0.78	0.69(-12%)	0.69(-11%)

	Service 1 = $H_2(4)$			Service 1 = $LN(4)$			Service 1 = $G(4)$		
ρ	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	1.34	1.26(-6%)	1.26(-6%)	1.28	1.24(-3%)	1.24(-3%)	2.35	1.78(-24%)	1.72(-27%)
0.16	1.39	1.26(-9%)	1.26(-9%)	1.34	1.23(-8%)	1.23(-8%)	2.50	1.96(-22%)	1.87(-25%)
0.21	1.45	1.28(-12%)	1.27(-12%)	1.40	1.23(-12%)	1.23(-12%)	2.63	2.14(-19%)	2.03(-23%)
0.26	1.52	1.30(-15%)	1.29(-15%)	1.47	1.24(-16%)	1.24(-16%)	2.73	2.32(-15%)	2.19(-20%)
0.31	1.61	1.33(-17%)	1.32(-18%)	1.54	1.26(-18%)	1.27(-18%)	2.83	2.51(-11%)	2.35(-17%)
0.36	1.71	1.38(-19%)	1.37(-20%)	1.62	1.32(-19%)	1.32(-19%)	2.91	2.69(-8%)	2.51(-14%)
0.41	1.84	1.46(-21%)	1.44(-21%)	1.71	1.41(-17%)	1.40(-18%)	2.98	2.87(-4%)	2.68(-10%)
0.46	1.97	1.60(-19%)	1.58(-20%)	1.79	1.54(-14%)	1.52(-15%)	3.04	3.03(-0%)	2.84(-7%)
0.51	2.12	1.83(-14%)	1.82(-14%)	1.89	1.71(-10%)	1.68(-11%)	3.09	3.18(3%)	2.98(-3%)
0.56	2.28	2.16(-5%)	2.16(-5%)	1.99	1.90(-5%)	1.87(-6%)	3.13	3.30(5%)	3.11(-1%)
0.61	2.44	2.51(3%)	2.53(3%)	2.10	2.11(0%)	2.08(-1%)	3.17	3.38(7%)	3.21(1%)
0.66	2.59	2.81(8%)	2.84(10%)	2.21	2.32(5%)	2.31(4%)	3.17	3.43(8%)	3.27(3%)
0.71	2.72	3.01(11%)	3.06(12%)	2.33	2.51(8%)	2.52(9%)	3.16	3.42(8%)	3.29(4%)
0.76	2.82	3.10(10%)	3.17(12%)	2.43	2.66(9%)	2.71(12%)	3.13	3.35(7%)	3.25(4%)
0.81	2.87	3.07(7%)	3.15(10%)	2.53	2.73(8%)	2.84(12%)	3.05	3.21(5%)	3.13(3%)
0.86	2.84	2.90(2%)	3.00(6%)	2.60	2.68(3%)	2.85(10%)	2.90	2.97(2%)	2.92(1%)
0.91	2.64	2.52(-4%)	2.63(-0%)	2.54	2.42(-5%)	2.65(5%)	2.62	2.55(-3%)	2.53(-3%)
0.96	2.07	1.74(-16%)	1.83(-11%)	2.16	1.72(-21%)	1.97(-9%)	2.00	1.75(-13%)	1.76(-12%)

Table 6.24: The RQNA approximation in various $E_4/GI/1 \rightarrow \cdot/E_4/1$ models with $\rho_1 = 0.9$.

External arrival $LN(4)$, Service 2 = $LN(4)$, $\rho_1 = 0.9$

ρ_2	Service 1 = E_4			Service 1 = $LN(0.25)$, $\rho_1 = 0.9$			Service 1 = M		
	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	4.73	4.96(5%)	4.96(5%)	4.68	4.96(6%)	4.96(6%)	5.02	5.00(-0%)	5.00(-0%)
0.16	4.63	4.91(6%)	4.91(6%)	4.66	4.91(5%)	4.91(5%)	5.09	5.00(-2%)	5.00(-2%)
0.21	4.60	4.84(5%)	4.84(5%)	4.75	4.84(2%)	4.83(2%)	5.08	5.00(-1%)	5.01(-1%)
0.26	4.72	4.74(1%)	4.74(0%)	4.63	4.74(2%)	4.73(2%)	5.06	5.00(-1%)	5.02(-1%)
0.31	4.65	4.63(-0%)	4.63(-1%)	4.62	4.61(-0%)	4.60(-1%)	5.09	5.01(-2%)	5.03(-1%)
0.36	4.65	4.53(-3%)	4.53(-3%)	4.59	4.49(-2%)	4.50(-2%)	5.10	5.02(-2%)	5.05(-1%)
0.41	4.62	4.46(-3%)	4.48(-3%)	4.63	4.44(-4%)	4.46(-4%)	5.16	5.03(-2%)	5.07(-2%)
0.46	4.65	4.44(-5%)	4.48(-4%)	4.62	4.43(-4%)	4.46(-3%)	5.16	5.05(-2%)	5.10(-1%)
0.51	4.69	4.45(-5%)	4.50(-4%)	4.67	4.44(-5%)	4.50(-4%)	5.22	5.08(-3%)	5.15(-1%)
0.56	4.69	4.49(-4%)	4.57(-3%)	4.67	4.48(-4%)	4.56(-2%)	5.28	5.12(-3%)	5.22(-1%)
0.61	4.74	4.55(-4%)	4.67(-2%)	4.74	4.55(-4%)	4.66(-2%)	5.36	5.19(-3%)	5.31(-1%)
0.66	4.81	4.66(-3%)	4.81(0%)	4.83	4.65(-4%)	4.81(-0%)	5.42	5.27(-3%)	5.43(0%)
0.71	4.94	4.82(-2%)	5.03(2%)	4.94	4.81(-2%)	5.02(2%)	5.50	5.40(-2%)	5.60(2%)
0.76	5.05	5.06(0%)	5.31(5%)	5.08	5.06(-0%)	5.31(5%)	5.61	5.59(-0%)	5.83(4%)
0.81	5.29	5.42(2%)	5.71(8%)	5.27	5.42(3%)	5.71(8%)	5.77	5.87(2%)	6.14(6%)
0.86	5.67	5.96(5%)	6.25(10%)	5.60	5.96(6%)	6.25(12%)	6.01	6.29(5%)	6.55(9%)
0.91	6.20	6.77(9%)	6.98(13%)	6.21	6.77(9%)	6.97(12%)	6.50	6.92(6%)	7.12(10%)
0.96	7.07	7.69(9%)	7.76(10%)	7.08	7.69(9%)	7.76(10%)	7.18	7.71(7%)	7.79(8%)

ρ	Service 1 = $H_2(4)$			Service 1 = $LN(4)$			Service 1 = $G(4)$		
	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	5.34	5.02(-6%)	5.03(-6%)	5.26	4.98(-5%)	4.98(-5%)	6.39	5.74(-10%)	5.72(-11%)
0.16	5.44	5.05(-7%)	5.06(-7%)	5.34	4.99(-6%)	5.00(-6%)	6.69	5.97(-11%)	5.94(-11%)
0.21	5.59	5.11(-9%)	5.12(-8%)	5.48	5.05(-8%)	5.07(-7%)	6.78	6.20(-9%)	6.16(-9%)
0.26	5.89	5.19(-12%)	5.22(-11%)	5.74	5.14(-10%)	5.17(-10%)	6.88	6.42(-7%)	6.39(-7%)
0.31	6.00	5.31(-11%)	5.36(-11%)	5.82	5.27(-9%)	5.32(-9%)	7.13	6.64(-7%)	6.61(-7%)
0.36	6.17	5.50(-11%)	5.57(-10%)	5.94	5.44(-8%)	5.50(-7%)	7.14	6.85(-4%)	6.82(-4%)
0.41	6.30	5.76(-9%)	5.86(-7%)	6.03	5.64(-7%)	5.72(-5%)	7.28	7.04(-3%)	7.02(-4%)
0.46	6.50	6.10(-6%)	6.24(-4%)	6.20	5.86(-5%)	5.97(-4%)	7.26	7.22(-1%)	7.21(-1%)
0.51	6.69	6.48(-3%)	6.65(-1%)	6.35	6.11(-4%)	6.25(-2%)	7.43	7.38(-1%)	7.39(-1%)
0.56	6.84	6.85(0%)	7.04(3%)	6.49	6.38(-2%)	6.55(1%)	7.39	7.51(2%)	7.54(2%)
0.61	7.00	7.17(2%)	7.38(5%)	6.64	6.65(0%)	6.87(3%)	7.58	7.62(0%)	7.67(1%)
0.66	7.15	7.41(4%)	7.65(7%)	6.81	6.92(2%)	7.19(6%)	7.64	7.70(1%)	7.78(2%)
0.71	7.35	7.59(3%)	7.85(7%)	7.02	7.19(2%)	7.52(7%)	7.64	7.77(2%)	7.88(3%)
0.76	7.48	7.72(3%)	8.01(7%)	7.21	7.44(3%)	7.83(9%)	7.71	7.82(1%)	7.97(3%)
0.81	7.64	7.82(2%)	8.13(6%)	7.45	7.65(3%)	8.10(9%)	7.77	7.87(1%)	8.05(4%)
0.86	7.80	7.89(1%)	8.21(5%)	7.70	7.81(1%)	8.30(8%)	7.79	7.91(2%)	8.11(4%)
0.91	7.86	7.95(1%)	8.20(4%)	7.87	7.93(1%)	8.38(6%)	7.88	7.95(1%)	8.12(3%)
0.96	7.94	7.97(0%)	8.09(2%)	8.04	7.97(-1%)	8.21(2%)	7.91	7.98(1%)	8.06(2%)

Table 6.25: The RQNA approximation in various $LN(4)/GI/1 \rightarrow \cdot/LN(4)/1$ models with $\rho_1 = 0.9$.

External arrival $LN(4)$, Service 2 = E_4 , $\rho_1 = 0.7$

ρ_2	Service 1 = E_4			Service 1 = $LN(0.25)$			Service 1 = M		
	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	1.12	1.24(11%)	1.24(11%)	1.11	1.24(11%)	1.25(12%)	1.28	1.25(-2%)	1.25(-2%)
0.16	1.07	1.23(15%)	1.23(15%)	1.06	1.23(16%)	1.23(17%)	1.30	1.25(-4%)	1.25(-3%)
0.21	1.02	1.21(18%)	1.21(18%)	1.01	1.21(20%)	1.21(20%)	1.31	1.25(-5%)	1.26(-4%)
0.26	0.99	1.18(20%)	1.18(19%)	0.97	1.18(22%)	1.18(22%)	1.33	1.25(-6%)	1.26(-5%)
0.31	0.96	1.15(19%)	1.14(19%)	0.94	1.15(22%)	1.14(22%)	1.36	1.25(-8%)	1.27(-6%)
0.36	0.94	1.11(17%)	1.10(16%)	0.92	1.11(20%)	1.10(19%)	1.39	1.25(-9%)	1.29(-7%)
0.41	0.94	1.06(13%)	1.05(12%)	0.91	1.06(16%)	1.04(14%)	1.42	1.26(-11%)	1.31(-8%)
0.46	0.94	1.01(7%)	0.99(5%)	0.92	1.00(9%)	0.98(6%)	1.46	1.28(-12%)	1.34(-8%)
0.51	0.96	0.95(-1%)	0.93(-3%)	0.94	0.94(-0%)	0.91(-3%)	1.51	1.31(-13%)	1.40(-7%)
0.56	1.00	0.89(-11%)	0.91(-9%)	0.98	0.87(-11%)	0.89(-9%)	1.57	1.36(-13%)	1.49(-5%)
0.61	1.07	0.89(-17%)	1.07(-0%)	1.06	0.88(-17%)	1.07(1%)	1.66	1.46(-12%)	1.64(-1%)
0.66	1.19	1.11(-7%)	1.42(19%)	1.19	1.10(-7%)	1.42(20%)	1.77	1.64(-7%)	1.88(7%)
0.71	1.39	1.54(10%)	1.92(38%)	1.40	1.53(10%)	1.92(38%)	1.91	1.94(2%)	2.24(18%)
0.76	1.69	2.13(26%)	2.50(48%)	1.70	2.13(25%)	2.51(47%)	2.10	2.38(13%)	2.69(28%)
0.81	2.09	2.80(34%)	3.10(48%)	2.09	2.80(34%)	3.10(48%)	2.37	2.93(23%)	3.19(34%)
0.86	2.55	3.44(35%)	3.62(42%)	2.56	3.44(34%)	3.63(42%)	2.72	3.48(28%)	3.65(34%)
0.91	3.07	3.92(28%)	4.00(30%)	3.07	3.92(28%)	4.01(31%)	3.15	3.94(25%)	4.01(27%)
0.96	3.66	4.18(14%)	4.20(15%)	3.67	4.18(14%)	4.21(15%)	3.68	4.18(14%)	4.20(14%)

ρ	Service 1 = $H_2(4)$			Service 1 = $LN(4)$			Service 1 = $G(4)$		
	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	1.39	1.26(-10%)	1.26(-10%)	1.34	1.24(-7%)	1.23(-8%)	2.32	1.77(-23%)	1.69(-27%)
0.16	1.48	1.26(-14%)	1.27(-14%)	1.43	1.23(-14%)	1.23(-14%)	2.47	1.94(-21%)	1.84(-25%)
0.21	1.58	1.28(-19%)	1.29(-19%)	1.53	1.23(-20%)	1.24(-19%)	2.60	2.11(-19%)	1.99(-23%)
0.26	1.70	1.30(-24%)	1.32(-23%)	1.64	1.24(-25%)	1.27(-23%)	2.73	2.28(-17%)	2.15(-21%)
0.31	1.86	1.33(-28%)	1.37(-26%)	1.77	1.27(-28%)	1.32(-25%)	2.83	2.45(-14%)	2.31(-18%)
0.36	2.03	1.39(-32%)	1.46(-28%)	1.90	1.33(-30%)	1.43(-25%)	2.93	2.62(-11%)	2.49(-15%)
0.41	2.21	1.49(-33%)	1.62(-27%)	2.04	1.44(-29%)	1.58(-22%)	3.02	2.79(-8%)	2.67(-12%)
0.46	2.42	1.66(-31%)	1.91(-21%)	2.20	1.60(-27%)	1.80(-18%)	3.12	2.95(-5%)	2.85(-9%)
0.51	2.63	1.95(-26%)	2.31(-12%)	2.36	1.82(-23%)	2.07(-12%)	3.19	3.11(-2%)	3.03(-5%)
0.56	2.83	2.33(-18%)	2.75(-3%)	2.55	2.08(-19%)	2.39(-6%)	3.26	3.25(-0%)	3.20(-2%)
0.61	3.03	2.73(-10%)	3.15(4%)	2.73	2.37(-13%)	2.76(1%)	3.34	3.39(1%)	3.37(1%)
0.66	3.21	3.08(-4%)	3.49(9%)	2.94	2.69(-8%)	3.14(7%)	3.41	3.50(3%)	3.53(4%)
0.71	3.38	3.38(-0%)	3.76(11%)	3.16	3.03(-4%)	3.53(12%)	3.48	3.61(4%)	3.69(6%)
0.76	3.52	3.61(3%)	3.97(13%)	3.38	3.35(-1%)	3.90(15%)	3.55	3.72(5%)	3.85(8%)
0.81	3.66	3.80(4%)	4.13(13%)	3.59	3.66(2%)	4.20(17%)	3.60	3.85(7%)	4.00(11%)
0.86	3.75	3.98(6%)	4.23(13%)	3.81	3.92(3%)	4.39(15%)	3.67	3.99(9%)	4.13(12%)
0.91	3.87	4.13(7%)	4.26(10%)	3.99	4.12(3%)	4.41(10%)	3.78	4.13(9%)	4.21(11%)
0.96	4.02	4.21(5%)	4.24(6%)	4.14	4.21(2%)	4.28(3%)	3.97	4.21(6%)	4.26(7%)

Table 6.26: The RQNA approximation in various $LN(4)/GI/1 \rightarrow \cdot/E_4/1$ models with $\rho_1 = 0.7$.

External arrival E_4 , Service 2 = $LN(4)$, $\rho_1 = 0.7$

ρ_2	Service 1 = E_4			Service 1 = $LN(0.25)$			Service 1 = M		
	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	4.67	4.96(6%)	4.96(6%)	4.70	4.96(6%)	4.96(6%)	4.95	5.00(1%)	4.99(1%)
0.16	4.64	4.91(6%)	4.91(6%)	4.69	4.91(5%)	4.91(5%)	4.88	4.99(2%)	4.98(2%)
0.21	4.71	4.84(3%)	4.84(3%)	4.60	4.84(5%)	4.83(5%)	4.85	4.97(3%)	4.96(2%)
0.26	4.57	4.73(4%)	4.74(4%)	4.52	4.73(4%)	4.73(5%)	4.78	4.94(3%)	4.94(3%)
0.31	4.54	4.61(2%)	4.63(2%)	4.49	4.60(2%)	4.61(3%)	4.80	4.90(2%)	4.90(2%)
0.36	4.48	4.49(0%)	4.53(1%)	4.46	4.47(0%)	4.51(1%)	4.75	4.84(2%)	4.86(2%)
0.41	4.48	4.40(-2%)	4.46(-0%)	4.46	4.39(-2%)	4.45(-0%)	4.71	4.78(2%)	4.81(2%)
0.46	4.42	4.35(-2%)	4.40(-1%)	4.42	4.34(-2%)	4.40(-1%)	4.70	4.73(1%)	4.76(1%)
0.51	4.43	4.32(-3%)	4.36(-2%)	4.40	4.31(-2%)	4.36(-1%)	4.63	4.67(1%)	4.71(2%)
0.56	4.38	4.30(-2%)	4.33(-1%)	4.40	4.29(-2%)	4.33(-2%)	4.64	4.61(-1%)	4.65(0%)
0.61	4.36	4.28(-2%)	4.30(-1%)	4.42	4.28(-3%)	4.31(-2%)	4.60	4.55(-1%)	4.59(-0%)
0.66	4.37	4.27(-2%)	4.29(-2%)	4.37	4.27(-2%)	4.29(-2%)	4.58	4.49(-2%)	4.52(-1%)
0.71	4.36	4.26(-2%)	4.27(-2%)	4.34	4.26(-2%)	4.28(-2%)	4.55	4.43(-3%)	4.46(-2%)
0.76	4.34	4.26(-2%)	4.26(-2%)	4.32	4.26(-1%)	4.27(-1%)	4.49	4.37(-3%)	4.39(-2%)
0.81	4.29	4.25(-1%)	4.26(-1%)	4.29	4.25(-1%)	4.26(-1%)	4.45	4.32(-3%)	4.33(-3%)
0.86	4.28	4.25(-1%)	4.25(-1%)	4.27	4.25(-0%)	4.25(-0%)	4.40	4.28(-3%)	4.29(-2%)
0.91	4.27	4.25(-1%)	4.25(-1%)	4.28	4.25(-1%)	4.25(-1%)	4.34	4.26(-2%)	4.26(-2%)
0.96	4.26	4.25(-0%)	4.25(-0%)	4.26	4.25(-0%)	4.25(-0%)	4.29	4.25(-1%)	4.25(-1%)

ρ	Service 1 = $H_2(4)$			Service 1 = $LN(4)$			Service 1 = $G(4)$		
	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	5.17	5.02(-3%)	5.01(-3%)	5.07	4.98(-2%)	4.98(-2%)	5.85	5.71(-2%)	5.47(-6%)
0.16	5.18	5.05(-3%)	5.03(-3%)	5.22	4.99(-4%)	4.99(-4%)	5.84	5.90(1%)	5.60(-4%)
0.21	5.27	5.08(-4%)	5.06(-4%)	5.19	5.02(-3%)	5.02(-3%)	5.89	6.07(3%)	5.70(-3%)
0.26	5.47	5.13(-6%)	5.10(-7%)	5.25	5.08(-3%)	5.07(-3%)	6.00	6.20(3%)	5.79(-3%)
0.31	5.48	5.19(-5%)	5.16(-6%)	5.40	5.15(-5%)	5.13(-5%)	5.91	6.29(6%)	5.86(-1%)
0.36	5.53	5.27(-5%)	5.25(-5%)	5.36	5.23(-2%)	5.19(-3%)	5.93	6.35(7%)	5.92(-0%)
0.41	5.55	5.38(-3%)	5.37(-3%)	5.44	5.31(-2%)	5.26(-3%)	5.95	6.39(7%)	5.97(0%)
0.46	5.60	5.52(-2%)	5.53(-1%)	5.40	5.39(-0%)	5.34(-1%)	5.87	6.39(9%)	6.00(2%)
0.51	5.63	5.64(0%)	5.69(1%)	5.49	5.45(-1%)	5.42(-1%)	5.85	6.35(8%)	6.00(3%)
0.56	5.62	5.73(2%)	5.81(3%)	5.41	5.49(2%)	5.49(1%)	5.83	6.26(7%)	5.97(2%)
0.61	5.61	5.77(3%)	5.86(4%)	5.51	5.50(-0%)	5.54(1%)	5.77	6.13(6%)	5.89(2%)
0.66	5.58	5.72(3%)	5.83(5%)	5.50	5.46(-1%)	5.57(1%)	5.67	5.95(5%)	5.77(2%)
0.71	5.55	5.58(1%)	5.72(3%)	5.44	5.37(-1%)	5.55(2%)	5.57	5.72(3%)	5.60(1%)
0.76	5.43	5.37(-1%)	5.52(2%)	5.42	5.22(-4%)	5.47(1%)	5.45	5.44(-0%)	5.38(-1%)
0.81	5.33	5.08(-5%)	5.25(-2%)	5.34	4.99(-7%)	5.31(-1%)	5.28	5.11(-3%)	5.11(-3%)
0.86	5.18	4.75(-8%)	4.89(-6%)	5.21	4.71(-10%)	5.04(-3%)	5.11	4.75(-7%)	4.78(-6%)
0.91	4.92	4.45(-10%)	4.52(-8%)	5.03	4.44(-12%)	4.67(-7%)	4.87	4.45(-9%)	4.47(-8%)
0.96	4.59	4.29(-7%)	4.30(-6%)	4.70	4.29(-9%)	4.33(-8%)	4.55	4.29(-6%)	4.29(-6%)

Table 6.27: The RQNA approximation in various $E_4/GI/1 \rightarrow \cdot/LN(4)/1$ models with $\rho_1 = 0.7$.

External arrival E_4 , Service 2 = E_4 , $\rho_1 = 0.7$

ρ_2	Service 1 = E_4			Service 1 = $LN(0.25)$, $\rho_1 = 0.9$			Service 1 = M		
	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	1.12	1.24(11%)	1.24(11%)	1.11	1.24(11%)	1.24(11%)	1.22	1.25(3%)	1.25(2%)
0.16	1.06	1.23(16%)	1.23(16%)	1.05	1.23(16%)	1.23(16%)	1.20	1.25(4%)	1.25(3%)
0.21	1.01	1.21(20%)	1.21(20%)	1.00	1.21(21%)	1.21(21%)	1.19	1.25(5%)	1.24(5%)
0.26	0.96	1.18(23%)	1.18(23%)	0.95	1.18(24%)	1.18(24%)	1.17	1.24(6%)	1.23(6%)
0.31	0.92	1.14(25%)	1.14(25%)	0.91	1.14(26%)	1.14(26%)	1.15	1.23(7%)	1.22(6%)
0.36	0.88	1.10(25%)	1.10(25%)	0.87	1.10(27%)	1.10(27%)	1.13	1.22(8%)	1.21(7%)
0.41	0.84	1.04(24%)	1.04(24%)	0.83	1.04(26%)	1.04(26%)	1.11	1.20(8%)	1.19(8%)
0.46	0.81	0.98(22%)	0.98(22%)	0.80	0.98(23%)	0.98(23%)	1.09	1.17(8%)	1.17(8%)
0.51	0.77	0.91(18%)	0.92(19%)	0.76	0.90(18%)	0.90(18%)	1.06	1.13(6%)	1.14(7%)
0.56	0.74	0.83(12%)	0.84(14%)	0.73	0.82(12%)	0.83(13%)	1.04	1.08(5%)	1.10(6%)
0.61	0.71	0.75(6%)	0.77(9%)	0.71	0.74(5%)	0.76(7%)	1.01	1.03(3%)	1.06(5%)
0.66	0.68	0.67(-1%)	0.70(3%)	0.68	0.66(-3%)	0.69(2%)	0.97	0.98(1%)	1.01(4%)
0.71	0.65	0.60(-8%)	0.65(-1%)	0.65	0.60(-9%)	0.64(-2%)	0.94	0.93(-1%)	0.96(2%)
0.76	0.63	0.56(-10%)	0.60(-5%)	0.63	0.56(-11%)	0.60(-5%)	0.89	0.86(-4%)	0.89(-0%)
0.81	0.60	0.53(-11%)	0.56(-7%)	0.60	0.53(-12%)	0.56(-6%)	0.84	0.79(-7%)	0.82(-3%)
0.86	0.57	0.52(-10%)	0.53(-7%)	0.58	0.52(-11%)	0.54(-7%)	0.78	0.70(-10%)	0.73(-7%)
0.91	0.55	0.51(-7%)	0.51(-6%)	0.55	0.51(-8%)	0.51(-7%)	0.70	0.60(-15%)	0.62(-12%)
0.96	0.52	0.50(-4%)	0.50(-4%)	0.52	0.50(-4%)	0.50(-4%)	0.60	0.52(-14%)	0.52(-13%)

ρ	Service 1 = $H_2(4)$			Service 1 = $LN(4)$			Service 1 = $G(4)$		
	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	1.31	1.26(-4%)	1.25(-5%)	1.27	1.24(-2%)	1.24(-2%)	2.00	1.77(-11%)	1.59(-20%)
0.16	1.35	1.26(-7%)	1.26(-7%)	1.31	1.23(-6%)	1.24(-6%)	2.08	1.93(-7%)	1.69(-19%)
0.21	1.40	1.27(-9%)	1.26(-10%)	1.36	1.23(-9%)	1.23(-9%)	2.14	2.08(-2%)	1.79(-16%)
0.26	1.46	1.29(-12%)	1.28(-13%)	1.41	1.23(-12%)	1.24(-12%)	2.19	2.23(2%)	1.88(-14%)
0.31	1.53	1.31(-14%)	1.29(-16%)	1.46	1.25(-14%)	1.26(-14%)	2.22	2.35(6%)	1.96(-12%)
0.36	1.61	1.35(-16%)	1.32(-18%)	1.51	1.29(-14%)	1.29(-15%)	2.25	2.46(9%)	2.03(-10%)
0.41	1.69	1.39(-18%)	1.36(-20%)	1.56	1.35(-14%)	1.33(-15%)	2.27	2.54(12%)	2.10(-7%)
0.46	1.78	1.45(-18%)	1.42(-20%)	1.62	1.41(-13%)	1.38(-15%)	2.27	2.60(14%)	2.16(-5%)
0.51	1.86	1.55(-17%)	1.53(-18%)	1.67	1.49(-11%)	1.44(-14%)	2.27	2.64(16%)	2.21(-3%)
0.56	1.93	1.69(-13%)	1.70(-12%)	1.73	1.58(-9%)	1.52(-12%)	2.26	2.64(17%)	2.24(-1%)
0.61	1.98	1.83(-8%)	1.88(-5%)	1.77	1.66(-6%)	1.61(-9%)	2.23	2.61(17%)	2.25(1%)
0.66	2.02	1.95(-3%)	2.03(0%)	1.83	1.72(-6%)	1.70(-7%)	2.20	2.54(16%)	2.24(2%)
0.71	2.03	2.01(-1%)	2.10(4%)	1.86	1.75(-6%)	1.77(-5%)	2.14	2.42(13%)	2.18(2%)
0.76	2.02	1.99(-1%)	2.10(4%)	1.88	1.73(-8%)	1.81(-4%)	2.05	2.26(10%)	2.07(1%)
0.81	1.96	1.87(-4%)	1.99(2%)	1.87	1.66(-12%)	1.80(-4%)	1.93	2.03(5%)	1.90(-2%)
0.86	1.83	1.65(-10%)	1.77(-3%)	1.82	1.50(-18%)	1.70(-7%)	1.76	1.72(-2%)	1.66(-6%)
0.91	1.60	1.29(-19%)	1.41(-12%)	1.67	1.22(-27%)	1.46(-13%)	1.51	1.32(-12%)	1.31(-13%)
0.96	1.19	0.78(-34%)	0.86(-28%)	1.34	0.77(-43%)	0.98(-27%)	1.10	0.78(-29%)	0.80(-27%)

Table 6.28: The RQNA approximation in various $E_4/GI/1 \rightarrow \cdot/E_4/1$ models with $\rho_1 = 0.7$.

External arrival $LN(4)$, Service 2 = $LN(4)$									
	Service 1 = E_4			Service 1 = $LN(0.25)$, $\rho_1 = 0.7$			Service 1 = M		
ρ_2	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	4.77	4.96(-4%)	4.96(-4%)	4.79	4.96(-4%)	4.96(-4%)	5.14	5.00(-3%)	5.01(-3%)
0.16	4.81	4.92(-2%)	4.91(-2%)	4.84	4.92(-2%)	4.91(-1%)	5.15	5.00(-3%)	5.01(-3%)
0.21	4.94	4.85(-2%)	4.84(-2%)	4.83	4.85(-0%)	4.83(-0%)	5.21	5.00(-4%)	5.03(-3%)
0.26	4.87	4.77(-2%)	4.75(-2%)	4.83	4.76(-1%)	4.73(-2%)	5.22	5.01(-4%)	5.05(-3%)
0.31	4.93	4.69(-5%)	4.67(-5%)	4.88	4.66(-4%)	4.64(-5%)	5.33	5.03(-6%)	5.09(-4%)
0.36	4.95	4.63(-7%)	4.65(-6%)	4.94	4.60(-7%)	4.63(-6%)	5.39	5.07(-6%)	5.15(-4%)
0.41	5.07	4.63(-9%)	4.71(-7%)	5.04	4.61(-8%)	4.70(-7%)	5.44	5.12(-6%)	5.22(-4%)
0.46	5.13	4.70(-8%)	4.83(-6%)	5.11	4.69(-8%)	4.82(-6%)	5.56	5.20(-6%)	5.33(-4%)
0.51	5.26	4.83(-8%)	5.01(-5%)	5.24	4.82(-8%)	5.01(-4%)	5.61	5.31(-5%)	5.48(-2%)
0.56	5.35	5.03(-6%)	5.27(-2%)	5.40	5.03(-7%)	5.27(-2%)	5.77	5.46(-5%)	5.68(-2%)
0.61	5.55	5.31(-4%)	5.60(-1%)	5.58	5.31(-5%)	5.61(-0%)	5.89	5.68(-3%)	5.93(-1%)
0.66	5.76	5.68(-1%)	6.01(-4%)	5.75	5.68(-1%)	6.01(-4%)	6.05	5.97(-1%)	6.25(-3%)
0.71	5.99	6.14(-2%)	6.45(-8%)	5.96	6.14(-3%)	6.45(-8%)	6.23	6.33(-2%)	6.61(-6%)
0.76	6.23	6.64(-7%)	6.91(11%)	6.22	6.64(-7%)	6.91(11%)	6.43	6.75(-5%)	6.99(-9%)
0.81	6.49	7.13(10%)	7.32(13%)	6.49	7.13(10%)	7.32(13%)	6.67	7.18(-8%)	7.36(10%)
0.86	6.79	7.54(11%)	7.65(13%)	6.79	7.54(11%)	7.65(13%)	6.93	7.56(-9%)	7.66(11%)
0.91	7.20	7.82(-9%)	7.86(-9%)	7.20	7.82(-9%)	7.87(-9%)	7.23	7.83(-8%)	7.87(-9%)
0.96	7.61	7.95(-5%)	7.97(-5%)	7.60	7.95(-5%)	7.99(-5%)	7.63	7.95(-4%)	7.98(-5%)

	Service 1 = $H_2(4)$			Service 1 = $LN(4)$			Service 1 = $G(4)$		
ρ	Sim	RQNA	RQ	Sim	RQNA	RQ	Sim	RQNA	RQ
0.11	5.40	5.02(-7%)	5.03(-7%)	5.30	4.98(-6%)	4.99(-6%)	6.29	5.72(-9%)	5.62(-11%)
0.16	5.53	5.05(-9%)	5.08(-8%)	5.54	4.99(-10%)	5.03(-9%)	6.39	5.93(-7%)	5.82(-9%)
0.21	5.69	5.10(-10%)	5.15(-10%)	5.63	5.05(-10%)	5.11(-9%)	6.53	6.13(-6%)	6.01(-8%)
0.26	6.01	5.19(-14%)	5.27(-12%)	5.77	5.14(-11%)	5.24(-9%)	6.75	6.32(-6%)	6.20(-8%)
0.31	6.14	5.31(-14%)	5.44(-11%)	6.04	5.27(-13%)	5.40(-10%)	6.74	6.50(-4%)	6.39(-5%)
0.36	6.30	5.49(-13%)	5.69(-10%)	6.09	5.44(-11%)	5.61(-8%)	6.85	6.67(-3%)	6.57(-4%)
0.41	6.44	5.74(-11%)	6.02(-7%)	6.29	5.64(-10%)	5.85(-7%)	6.96	6.83(-2%)	6.75(-3%)
0.46	6.62	6.05(-9%)	6.39(-3%)	6.35	5.86(-8%)	6.12(-4%)	7.00	6.98(-0%)	6.92(-1%)
0.51	6.79	6.39(-6%)	6.77(-0%)	6.60	6.11(-7%)	6.43(-3%)	7.08	7.10(-0%)	7.08(-0%)
0.56	6.91	6.72(-3%)	7.10(-3%)	6.64	6.38(-4%)	6.75(-2%)	7.20	7.22(-0%)	7.22(-0%)
0.61	7.04	7.00(-1%)	7.37(-5%)	6.93	6.65(-4%)	7.08(-2%)	7.25	7.32(-1%)	7.37(-2%)
0.66	7.17	7.23(-1%)	7.59(-6%)	7.05	6.92(-2%)	7.40(-5%)	7.29	7.41(-2%)	7.50(-3%)
0.71	7.32	7.41(-1%)	7.76(-6%)	7.20	7.19(-0%)	7.70(-7%)	7.35	7.51(-2%)	7.64(-4%)
0.76	7.39	7.57(-2%)	7.89(-7%)	7.36	7.44(-1%)	7.95(-8%)	7.43	7.61(-2%)	7.76(-4%)
0.81	7.51	7.71(-3%)	7.97(-6%)	7.52	7.65(-2%)	8.11(-8%)	7.47	7.72(-3%)	7.87(-5%)
0.86	7.63	7.83(-3%)	8.01(-5%)	7.66	7.81(-2%)	8.17(-7%)	7.55	7.84(-4%)	7.94(-5%)
0.91	7.69	7.93(-3%)	8.01(-4%)	7.82	7.93(-1%)	8.11(-4%)	7.66	7.93(-3%)	7.98(-4%)
0.96	7.83	7.97(-2%)	7.99(-2%)	7.93	7.97(-1%)	8.02(-1%)	7.80	7.97(-2%)	8.01(-3%)

Table 6.29: The RQNA approximation in various $LN(4)/GI/1 \rightarrow \cdot/LN(4)/1$ models with $\rho_1 = 0.7$.

6.4.1 Feedback Elimination: A Three-Station Example.

We look at the suite of three-station examples depicted in Figure 6.15, as in Section 3.1 of [44]. This example is designed to have three stations that are tightly coupled with each other, so that the dependence among the queues and the flows is fairly complicated.

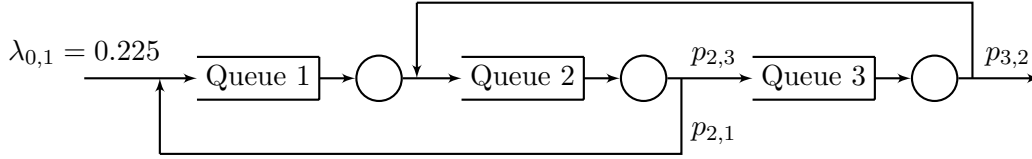


Figure 6.15: A three-station example.

In this example, we have three stations in tandem but also allow customer feedback from station 2 to station 1 and from station 3 to station 2, with probability $p_{2,1} = p_{2,3} = p_{3,2} = 0.5$. The only external arrival process is a Poisson process which arrives at station 1 with rate $\lambda_{0,1} = 0.225$, hence by (4.6) the effective arrival rate is $\lambda_1 = 0.675$, $\lambda_2 = 0.9$ and $\lambda_3 = 0.45$.

For the service distributions, we consider the same sets of parameters as in [44], summarized in Table 6.30 and 6.31. Note that Case 2 is relatively more challenging because there are two bottleneck stations; in contrast, all the other cases have only one.

Case	ρ_1	ρ_2	ρ_3	Case	$c_{s,1}^2$	$c_{s,2}^2$	$c_{s,3}^2$
1	0.675	0.900	0.450	A	0.00	0.00	0.00
2	0.900	0.675	0.900	B	2.25	0.00	0.25
3	0.900	0.675	0.450	C	0.25	0.25	2.25
4	0.900	0.675	0.675	D	0.00	2.25	2.25
				E	8.00	8.00	0.25

Table 6.30: Traffic intensity of the four cases in the three-station example.

Table 6.31: Variability of the service distributions of the four cases in the three-station example.

We now compare the RQNA approximations and four previous algorithms as in Section 6.3.3, with the simulated mean sojourn times at each station, as well as the total sojourn

time of the network. The sojourn time for each station is defined as the waiting time plus the service time at that station, whereas the total sojourn time is the time that a customer stay in the system defined in (2.51). We consider two cases of the RQNA algorithm: (1) the plain RQNA algorithm without feedback elimination, as in Algorithm 1 and (2) the RQNA algorithm with feedback elimination, as discussed in Section 5.5.

For RQNA with feedback elimination, we apply feedback elimination to each station that has at least one feedback flow that only passes through stations with equal or lower traffic intensities. We eliminate all such flows in the feedback elimination procedure. Take Case 1 for example, we do not apply feedback elimination for Station 1 because all feedback customers go through Station 2, which has higher traffic intensity; we will, however, eliminate the flow from 2 to 1 as well as the flow from 3 to 2 for Station 2, since both Station 1 and 3 have lower traffic intensities. As another example, for both Station 2 and 3 in case 4, we eliminate the flow from 3 to 2, but we do not eliminate the flow from 2 to 1, since Station 2 and 3 share the same traffic intensity while Station 1 has higher traffic intensity.

Tables 6.32 and 6.33 expand Tables II and III in [44] by adding values for (1) the mean total sojourn time and (2) the RQ and RQNA approximations, with and without feedback elimination. For each table, we indicate by an asterisk in the last column the stations where elimination is applied.

We observed that the plain RQNA algorithm works well for stations with moderate to low traffic intensities, but not so satisfactory for congested stations. On the other hand, the accuracy of the RQNA algorithm with feedback elimination is on par with, if not better than the best previous algorithm.

6.4.2 A Ten-Station Example with Feedback.

We conclude with the 10-station OQN example with feedback in Figure 6.16, which was considered in Section 3.5 of [44].

The only exogenous arrival process is Poission with rate 1. For each station, if there are two routing destinations, the departing customer follows Markovian routing with equal probability, each being 0.5. The vector of mean service times is $(0.45, 0.30, 0.90, 0.30, 0.38571, 0.20, 0.1333, 0.20, 0.15, 0.20)$, so that the traffic intensity vector is $(0.6, 0.4, 0.6, 0.9, 0.9, 0.6,$

Case	Simulation	QNA	QNET	SBD	RQ	RQNA	RQNA (elim)	
A	1	40.39 (3.75%)	20.5 (-49%)	diverging	43.0 (6.4%)	73.9 (83%)	83.5 (107%)	44.8 (11.0%)
	2	59.58 (3.29%)	36.0 (-40%)	56.7 (-4.9%)	58.2 (-2.4%)	78.0 (31%)	94.3 (58%)	69.3 (16.4%)
	3	40.72 (4.78%)	24.0 (-41%)	38.7 (-5.0%)	40.2 (-1.3%)	57.2 (41%)	74.7 (83%)	43.3 (6.3%)
	4	42.12 (3.36%)	26.2 (-38%)	41.8 (-0.7%)	42.7 (1.3%)	59.3 (41%)	75.1 (78%)	41.2 (-2.2%)
B	1	52.40 (2.64%)	42.0 (-20%)	52.6 (0.4%)	50.2 (-4.2%)	72.4 (38%)	93.7 (79%)	53.1 (1.4%)
	2	91.52 (3.77%)	94.1 (2.8%)	83.7 (-8.5%)	95.3 (4.1%)	109 (20%)	169 (85%)	94.5 (3.2%)
	3	61.68 (3.44%)	72.2 (17%)	61.9 (0.4%)	60.9 (-1.3%)	79.4 (29%)	133 (115%)	60.5 (-1.9%)
	4	63.34 (2.83%)	75.8 (20%)	64.1 (1.3%)	64.7 (2.1%)	83.0 (31%)	135 (113%)	62.4 (-1.4%)
C	1	44.24 (1.96%)	31.3 (-29%)	37.0 (-16%)	47.1 (6.4%)	75.7 (71%)	91.4 (106%)	42.1 (-4.8%)
	2	92.42 (4.23%)	87.4 (-5.4%)	91.2 (-1.4%)	91.6 (-0.83%)	106 (15%)	156 (68%)	96.0 (3.8%)
	3	44.26 (4.69%)	33.2 (-25%)	44.0 (-0.7%)	45.0 (1.7%)	61.3 (38%)	84.2 (90%)	44.0 (-0.6%)
	4	50.20 (1.04%)	41.4 (-18%)	51.1 (1.7%)	52.2 (4.0%)	67.4 (34%)	91.2 (82%)	45.9 (-8.6%)
E	1	134.4 (4.77%)	265 (97%)	155 (15%)	116 (-14%)	158 (17%)	305 (127%)	120 (-11%)
	2	213.1 (3.47%)	308 (45%)	228 (7.1%)	206 (-3.3%)	234 (10%)	367 (72%)	173 (-19%)
	3	138.7 (3.97%)	244 (76%)	161 (16%)	135 (-2.5%)	163 (17%)	300 (116%)	136 (-2.0%)
	4	155.1 (4.37%)	252 (63%)	168 (8.2%)	147 (-5.0%)	178 (15%)	312 (101%)	148 (-4.8%)

Table 6.32: A comparison of six approximation methods for the total sojourn time in the three-station example in Figure 6.15 with parameters specified in Table 6.30 and 6.31.

Case	Station	Simulation	QNA	QNET	SBD	RQ	RQNA	RQNA (elim)
D1	1	2.476 (0.61%)	2.24 (-9.4%)	2.48 (0.3%)	2.47 (-0.1%)	2.47 (-0.28%)	2.68 (7.8%)	2.68 (7.8%)
	2	10.85 (3.21%)	14.9 (37%)	11.6 (6.5%)	11.4 (5.2%)	19.8 (83%)	28.4 (162%)	11.1* (2.7%)
	3	2.544 (0.63%)	2.53 (-0.8%)	2.54 (-0.0%)	2.59 (1.6%)	2.57 (1.2%)	2.53 (-0.7%)	2.53 (-0.7%)
	Total	55.81 (2.58%)	71.4 (28%)	58.8 (5.3%)	58.2 (4.3%)	91.8 (64%)	127 (127%)	57.6 (3.3%)
D2	1	11.35 (3.29%)	8.01 (-29%)	10.8 (-4.5%)	11.1 (-1.9%)	13.7 (20%)	16.6 (46%)	11.3* (0.1%)
	2	2.643 (1.25%)	2.96 (12%)	2.75 (4.0%)	2.82 (6.7%)	2.85 (7.8%)	3.06 (16%)	3.06 (16%)
	3	26.87 (2.04%)	32.9 (22%)	26.8 (-0.4%)	24.9 (-7.5%)	27.5 (2.2%)	36.4 (35%)	31.1* (16%)
	Total	98.36 (1.82%)	102 (3.4%)	97.2 (-1.2%)	94.4 (-4.0%)	104 (6.0%)	132 (34%)	105 (7.1%)
D3	1	11.39 (3.04%)	7.95 (-30%)	11.0 (-3.5%)	11.3 (-0.5%)	15.8 (39%)	16.5 (45%)	11.3* (-0.5%)
	2	2.290 (1.27%)	2.90 (27%)	2.53 (10%)	2.26 (-1.4%)	2.57 (12%)	3.04 (33%)	2.10* (-8.2%)
	3	2.220 (0.59%)	2.40 (7.9%)	2.38 (7.0%)	2.59 (16%)	2.39 (7.6%)	2.43 (9.6%)	2.43 (9.6%)
	Total	47.72 (2.51%)	40.2 (-16%)	47.8 (0.2%)	48.2 (1.0%)	62.6 (31%)	66.6 (39%)	47.5 (0.51%)
D4	1	11.30 (6.39%)	7.97 (-29%)	10.9 (-3.2%)	11.3 (0.3%)	14.2 (26%)	16.43 (45%)	11.3* (0.3%)
	2	2.414 (1.12%)	2.93 (21%)	2.64 (9.5%)	2.60 (7.7%)	2.65 (10%)	3.05 (26%)	2.10* (-13%)
	3	5.886 (1.05%)	6.83 (16%)	6.31 (7.3%)	6.17 (4.8%)	6.47 (10%)	6.85 (16%)	5.95* (1.1%)
	Total	55.24 (4.37%)	49.3 (-11%)	56.0 (1.4%)	56.7 (2.7%)	69.3 (25%)	75.5 (37%)	54.3 (-1.7%)
Average relative error			20.24%	4.72%	4.52%	21.61%	42.60%	5.51%

Table 6.33: A comparison of six approximation methods for the sojourn time at each station of the three-station example for Case D with parameters specified in Table 6.31.

0.4, 0.6, 0.6, 0.4). The scv's at these stations are (0.5, 2, 2, 0.25, 0.25, 2, 1, 2, 0.5, 0.5), where we assume a Erlang distribution if $c_s^2 < 1$, an exponential distribution if $c_s^2 = 1$ and a hyperexponential distribution if $c_s^2 > 1$.

Note that stations 4 and 5 are bottleneck queues, having equal traffic intensity, far greater than the traffic intensities at the other queues. Moreover, these two stations are quite closely coupled. Thus, at first glance, we expect that SBD with two-dimensional RBM should perform very well, which proves to be correct. Moreover, this example should be challenging for RQNA because it is based on heavy-traffic limits for OQN's with only a single bottleneck, thus involving only one-dimensional RBM.

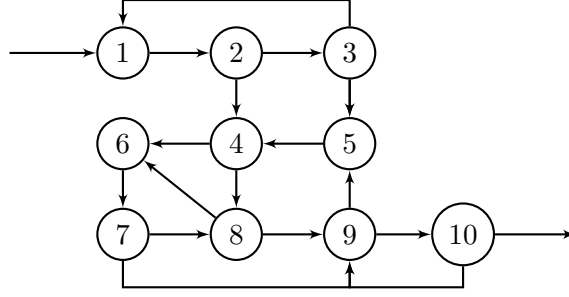


Figure 6.16: A ten-station with customer feedback example.

In Table 6.34, we report the simulation estimates and approximations for the steady-state mean sojourn time (waiting time plus service time) at each station, as well as the total sojourn time of the system, calculated as in (2.51). For the approximations, we compare QNA from [134], QNET from [73], SBD from [44], RQ from [145] (with estimated IDC), as well as the RQNA algorithms here. The simulation, QNA, QNET and SBD columns are taken from Table XIV of [44].

Again, we consider two versions of RQNA algorithm, the first one does not eliminate feedback, while the second one (marked by ‘elim’) applies the feedback elimination procedure. As before, in eliminating customer feedback, for each station, we identify the near-immediate feedback flows as the flows that come back to the station after completing service, without passing through any station with a higher traffic intensity. We then eliminate all near-immediate feedback flows, apply plain RQNA algorithm on the reduced network and use the new RQNA approximation as the approximation for that station.

Station	Simulation	QNA	QNET	SBD	RQ	RQNA	RQNA (elim)
1	0.99 (0.86%)	0.97 (-2.8%)	1.00 (0.2%)	1.00 (0.4%)	0.97 (-2.0%)	1.09 (9.2%)	1.00* (0.4%)
2	0.55 (0.69%)	0.58 (6.0%)	0.56 (2.6%)	0.55 (0.2%)	0.55 (-0.1%)	0.56 (1.3%)	0.56 (1.4%)
3	2.82 (1.93%)	2.93 (4.2%)	2.90 (3.2%)	2.76 (-2.0%)	2.96 (5.0%)	3.40 (21%)	2.75* (-2.5%)
4	1.79 (3.71%)	1.34 (-25%)	1.41 (-21%)	1.76 (-1.6%)	2.34 (31%)	3.51 (97%)	2.11* (18%)
5	2.92 (4.77%)	2.49 (-15%)	2.44 (-17%)	2.81 (-3.6%)	3.77 (29%)	9.07 (211%)	3.35* (15%)
6	0.58 (0.78%)	0.64 (10%)	0.62 (7.4%)	0.59 (2.2%)	0.60 (3.8%)	0.70 (20%)	0.49* (-16%)
7	0.24 (0.28%)	0.24 (-1.7%)	0.26 (7.1%)	0.27 (11%)	0.23 (-3.0%)	0.24 (-1.3%)	0.24 (-1.3%)
8	0.58 (0.67%)	0.64 (9.6%)	0.61 (4.6%)	0.60 (1.7%)	0.61 (3.9%)	0.70 (20%)	0.59* (0.6%)
9	0.34 (0.63%)	0.32 (-6.1%)	0.35 (2.0%)	0.43 (26%)	0.33 (-4.2%)	0.73 (111%)	0.42* (21%)
10	0.29 (0.19%)	0.30 (2.4%)	0.29 (1.4%)	0.28 (-1.7%)	0.28 (-1.5%)	0.26 (-8.7%)	0.26 (-8.7%)
Total	22.0 (2.45%)	20.3 (-7.9%)	20.4 (-7.3%)	22.4 (1.7%)	26.1 (18%)	44.5 (102%)	24.2* (9.9%)

Table 6.34: A comparison of six approximation methods to simulation for the mean steady-state sojourn times at each station of the open queueing network in Figure 6.16.

We make the following observations from this numerical example:

1. Particular attention should be given to the two bottleneck stations: 4 and 5. Note that QNA and QNET produce 15 – 25% error, which is satisfactory, but SBD does far better with only 1 – 4% error.
2. The RQNA algorithm without feedback elimination can perform very poorly with high traffic intensity and high feedback probability, presumably due to the break down of (2.26).
3. With feedback elimination, the RQNA algorithm performs significantly better and is competitive with previous algorithms in this complex setting, producing 15 – 18% error at stations 4 and 5. The performance of RQNA at the tightly coupled bottleneck queues evidently suffers because the current RQNA depends heavily on one-dimensional RBM.

Bibliography

- [1] J. Abate, G. L. Choudhury, and W. Whitt. Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Stochastic Models*, 10(1):99–143, 1994.
- [2] J. Abate and W. Whitt. Transient behavior of regulated Brownian motion I: Starting at the origin. *Advances in Applied Probability*, 19(3):560–598, 1987.
- [3] J. Abate and W. Whitt. The correlation functions of RBM and $M/M/1$. *Stochastic Models*, 4(2):315–359, 1988.
- [4] J. Abate and W. Whitt. The Fourier-series method for inverting transforms of probability distributions. *Queueing Systems*, 10:5–88, 1992.
- [5] J. Abate and W. Whitt. Numerical inversion of Laplace transforms of probability distributions. *ORSA Journal on Computing*, 7:36–43, 1995.
- [6] M. A. Ahmed and T. M. Alkhamis. Simulation optimization for an emergency department healthcare unit in Kuwait. *European journal of operational research*, 198(3):936–942, 2009.
- [7] S. L. Albin. On Poisson approximations for superposition arrival processes in queues. *Management Science*, 28(2):126–137, 1982.
- [8] S. L. Albin. Approximating a point process by a renewal process, II: Superposition arrival processes to queues. *Operations Research*, 32(5):1133–1162, 1984.
- [9] S. L. Albin, J. Barrett, D. Ito, and J. E. Mueller. A queueing network analysis of a health center. *Queueing Systems*, 7(1):51–61, 1990.

- [10] A. K. Aras, X. Chen, and Y. Liu. Many-server gaussian limits for overloaded non-markovian queues with customer abandonment. *Queueing Systems*, pages 1–45, 2018.
- [11] A. K. Aras, Y. Liu, and W. Whitt. Heavy-traffic limit for the initial content process. *Stochastic Systems*, 7(1):95–142, 2017.
- [12] S. Asmussen. *Applied Probability and Queues*. Springer, New York, second edition, 2003.
- [13] A. Azaron and S. M. T. F. Ghomi. Optimal control of service rates and arrivals in Jackson networks. *European Journal of Operational Research*, 147(1):17–31, 2003.
- [14] J. Azema, M. Kaplan-Duflo, and D. Revuz. Invariant measures for classes of Markov processes (in french). *Probability Theory and Related Fields*, 8(3):157–181, 1967.
- [15] F. Baccelli and P. Bremaud. *Elements of Queueing Theory: Palm Martingale Calculus and Stochastic Recurrences*. Springer, New York, second edition, 2003.
- [16] X. Bai, R. Gopal, M. Nunez, and D. Zhdanov. A decision methodology for managing operational efficiency and information disclosure risk in healthcare processes. *Decision support systems*, 57:406–416, 2014.
- [17] C. Bandi and D. Bertsimas. Tractable stochastic analysis in high dimensions via robust optimization. *Mathematical Programming*, 134:23–70, 2012.
- [18] C. Bandi, D. Bertsimas, and N. Youssef. Robust queueing theory. *Operations Research*, 63(3):676–700, 2015.
- [19] S. Banerjee, R. Johari, and C. Riquelme. Pricing in ride-sharing platforms: A queueing-theoretic approach. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, page 639. ACM, 2015.
- [20] D. Bertsimas, D. B. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501, 2011.
- [21] D. Bertsimas and D. Nakazato. The departure process from a $GI/G/1$ queue and its applications to the analysis of tandem queues, working paper, MIT. 1990.

- [22] J. Bi, Z. Zhu, R. Tian, and Q. Wang. Dynamic provisioning modeling for virtualized multi-tier applications in cloud data center. In *2010 IEEE 3rd International Conference on Cloud Computing*, pages 370–377. IEEE, 2010.
- [23] P. Billingsley. *Convergence of Probability Measures*. Wiley, New York, 1999.
- [24] A. A. Borovkov. Some limit theorems in the theory of mass service, II. *Theor. Probability Appl.*, 10:375–400, 1965.
- [25] A. A. Borovkov. Limit theorems for queueing networks, I. *Theory of Probability & Its Applications*, 31(3):413–427, 1986.
- [26] M. Brahim and D. Worthington. Queueing models for out-patient appointment systems-A case study. *Journal of the Operational Research Society*, 42(9):733–746, 1991.
- [27] A. Braverman, J. G. Dai, and M. Miyazawa. Heavy traffic approximation for the stationary distribution of a generalized Jackson network: The BAR approach. *Stochastic Systems*, 7(1):143–196, 2017.
- [28] L. Breiman. *Probability*. SIAM, Philadelphia, 1992.
- [29] R. Breunig. An almost unbiased estimator of the coefficient of variation. *Economics Letters*, 70(1):15–19, 2001.
- [30] S. Brumelle. On the relation between customer averages and time averages in queues. *J. Appl. Prob.*, 8(3):508–520, 1971.
- [31] A. Budhiraja and C. Lee. Stationary distribution convergence for generalized Jackson networks in heavy traffic. *Mathematics of Operations Research*, 34(1):45–56, 2009.
- [32] P. J. Burke. The output of a queueing system. *Operations Research*, 4(6):699–704, 1956.
- [33] C. W. Chan, J. Dong, and L. V. Green. Queues with time-varying arrivals and inspections with applications to hospital discharge policies. *Operations Research*, 65(2):469–495, 2016.

- [34] H. Chen and A. Mandelbaum. Discrete flow networks: bottleneck analysis and fluid approximations. *Math. Oper. Res.*, 16(2):408–446, 1991.
- [35] H. Chen and A. Mandelbaum. Stochastic discrete flow networks: diffusion approximations and bottlenecks. *The Annals of Probability*, 19(4):1463–1519, 1991.
- [36] H. Chen and D. D. Yao. *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer, New York, 2001.
- [37] S.-N. Chow and J. K. Hale. *Methods of Bifurcation Theory*, volume 251. Springer Science & Business Media, 2012.
- [38] J. W. Cohen. *The Single Server Queue*. North-Holland, Amsterdam, second edition, 1982.
- [39] D. R. Cox. *Renewal Theory*. Methuen, London, 1962.
- [40] D. R. Cox and P. A. W. Lewis. *The Statistical Analysis of Series of Events*. Methuen, London, 1966.
- [41] S. Creemers and M. Lambrecht. Healthcare queueing models. *Leuven Faculty of Business and Economics Report*, 2008.
- [42] J. Dai. On the positive Harris recurrence for multiclass queueing networks. *Ann Appl Probab*, 5:49–77, 1995.
- [43] J. Dai and S. P. Meyn. Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Transactions on Automatic Control*, 40(11):1889–1904, 1995.
- [44] J. Dai, V. Nguyen, and M. I. Reiman. Sequential bottleneck decomposition: an approximation method for generalized Jackson networks. *Operations Research*, 42(1):119–136, 1994.
- [45] J. Dai and P. Shi. Inpatient overflow: An approximate dynamic programming approach. *Manufacturing & Service Operations Management*, 2019.

- [46] J. G. Dai and J. M. Harrison. Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *The Annals of Applied Probability*, pages 65–86, 1992.
- [47] D. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: Elementary Theory and Methods*, volume I. Springer, Oxford, U. K., second edition, 2008.
- [48] D. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes: General Theory and Structure*, volume II. Springer, Oxford, U. K., second edition, 2008.
- [49] D. J. Daley. Weakly stationary point processes and random measures. *Journal of the Royal Statistical Society*, 33(3):406–428, 1971.
- [50] D. J. Daley. Further second-order properties of certain single-server queueing systems. *Stoch. Proc. Appl.*, 3(2):185–191, 1975.
- [51] D. J. Daley. Queueing output processes. *Adv. Appl. Prob.*, 8(2):395–415, 1976.
- [52] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes*. Springer, Oxford, U. K., second edition, 2008.
- [53] M. H. A. Davis. Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic processes. *J. Roy. Stat.Soc. B*, 46(3):353–388, 1984.
- [54] A. B. Dieker, S. Ghosh, and M. S. Squillante. Optimal resource capacity management for stochastic networks. *Operations Research*, 65(1):221–241, 2016.
- [55] R. L. Disney and D. Konig. Queueing networks: a survey of their random processes. *SIAM Review*, 27(3):335–403, 1985.
- [56] R. S. Dittus, R. W. Klein, D. J. DeBrot, M. A. Dame, and J. F. Fitzgerald. Medical resident work schedules: Design and evaluation by stimulation modeling. *Management Science*, 42(6):891–906, 1996.

- [57] W. Feller. *An Introduction to Probability Theory and its Applications*. John Wiley, New York, second edition edition, 1971.
- [58] K. W. Fendick, V. Saksena, and W. Whitt. Dependence in packet queues. *IEEE Trans Commun.*, 37:1173–1183, 1989.
- [59] K. W. Fendick, V. Saksena, and W. Whitt. Investigating dependence in packet queues with the index of dispersion for work. *IEEE Trans Commun.*, 39(8):1231–1244, 1991.
- [60] K. W. Fendick and W. Whitt. Measurements and approximations to describe the offered traffic and predict the average workload in a single-server queue. *Proceedings of the IEEE*, 71(1):171–194, 1989.
- [61] H. W. Ferng and J. F. Chang. Departure processes of $BMAP/G/1$ queues. *Queueing Systems*, 39(1):109–135, 2001.
- [62] D. Ferré, L. Hervé, and J. Ledoux. Limit theorems for stationary Markov processes with L^2 -spectral gap. In *Annales de l'IHP Probabilités et statistiques*, volume 48, pages 396–423, 2012.
- [63] S. Foss. Ergodicity of queuing networks. *Siberian Mathematical Journal*, 32(4):690–705, 1991.
- [64] D. Freund, S. G. Henderson, and D. B. Shmoys. Minimizing multimodular functions and allocating capacity in bike-sharing systems. In *International Conference on Integer Programming and Combinatorial Optimization*, pages 186–198. Springer, 2017.
- [65] D. Gamarnik and D. A. Goldberg. Steady-state $GI/GI/n$ queue in the Halfin-Whitt Regime. *Ann. Appl. Probability*, 23(6):2382–2419, 2013.
- [66] D. Gamarnik and A. Zeevi. Validity of heavy traffic steady-state approximations in generalized Jackson networks. *Advances in Applied Probability*, 16(1):56–90, 2006.
- [67] D. Gupta. Queueing models for healthcare operations. In *Handbook of Healthcare Operations Management*, pages 19–44. Springer, 2013.

- [68] A. A. Hanbali, M. Mandjes, Y. Nazarathy, and W. Whitt. The asymptotic variance of departures in critically loaded queues. *Adv. Appl. Prob.*, 43(1):243–263, 2011.
- [69] T. E. Harris. The existence of stationary measures for certain Markov processes. In *Proc. Third Berkeley Symp. Prob. and Stat.*, volume 2, pages 113–124. University of California, Berkely, CA, 1956.
- [70] J. M. Harrison. The heavy traffic approximation for single server queues in series. *Journal of Applied Probability*, 10(3):613–629, 1973.
- [71] J. M. Harrison. The diffusion approximation for tandem queues in heavy traffic. *Advances in Applied Probability*, 10(4):886–905, 1978.
- [72] J. M. Harrison. *Brownian Motion and Stochastic Flow Systems*. Wiley, New York, 1985.
- [73] J. M. Harrison and V. Nguyen. The QNET method for two-moment analysis of open queueing networks. *Queueing Systems*, 6(1):1–32, 1990.
- [74] J. M. Harrison and M. I. Reiman. Reflected Brownian motion on an orthant. *The Annals of Probability*, pages 302–308, 1981.
- [75] J. M. Harrison and R. J. Williams. Multidimensional reflected Brownian motions having exponential stationary distributions. *The Annals of Probability*, pages 115–137, 1987.
- [76] J. M. Harrison and R. J. Williams. On the quasireversibility of a multiclass Brownian service station. *Annals of Probability*, 18(3):1249–1268, 1990.
- [77] J. M. Harrison and R. J. Williams. Brownian models of feedforward queueing networks: quasireversibility and product-form solutions. *Ann. Appl. Prob.*, 2(2):263–293, 1992.
- [78] S. Hautphenne, Y. Kerner, Y. Nazarathy, and P. Taylor. The intercept term of the asymptotic variance curve for some queueing output processes. *European Journal of Operational Research*, 242:455–464, 2015.

- [79] A. Heindl, Q. Zhang, and E. Smirni. ETAQA truncation models for the $MAP/MAP/1$ departure process. In *First International Conference on the Quantitative Evaluation of Systems, 2004. QEST 2004. Proceedings.*, pages 100–109. IEEE, 2004.
- [80] H. Honnappa, R. Jain, and A. Ward. A queueing model with independent arrivals, and its fluid and diffusion limits. *Queueing Systems*, 80:71–103, 2015.
- [81] A. Horváth, G. Horváth, and M. Telek. A joint moments based analysis of networks of $MAP/MAP/1$ queues. *Performance Evaluation*, 67(9):759–778, 2010.
- [82] J. Q. Hu. The departure process of the $GI/G/1$ queue and its Maclaurin series. *Operations Research*, 44(5):810–815, 1996.
- [83] D. L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic, I. *Advances in Applied Probability*, 2(1):150–177, 1970.
- [84] D. L. Iglehart and W. Whitt. Multiple channel queues in heavy traffic, II: Sequences, networks and batches. *Advances in Applied Probability*, 2(2):355–369, 1970.
- [85] J. R. Jackson. Networks of waiting lines. *Operations Research*, 5(4):518–521, 1957.
- [86] D. L. Jagerman, B. Balcioglu, T. Altıok, and B. Melamed. Mean waiting time approximations in the $G/G/1$ queue. *Queueing Systems*, 46(3-4):481–506, 2004.
- [87] L. Jiang and R. E. Giachetti. A queueing network model to analyze the impact of parallelization of care on patient cycle time. *Health Care Management Science*, 11(3):248–261, 2008.
- [88] F. I. Karpelevich and A. Y. Kreinin. *Heavy-Traffic Limits for Multiphase Queues*, volume 137. American Mathematical Society, Providence, RI, 1994.
- [89] P. F. Kelly. *Reversibility and Stochastic Networks*. Cambridge University Press, revised edition, 2011.
- [90] J. G. Kemeny and J. L. Snell. *Finite Markov Chains*. Springer, New York, 1976.
- [91] S. Kim. Modeling cross correlation in three-moment four-parameter decomposition approximation of queueing networks. *Operations Research*, 59(2):480–497, 2011.

- [92] S. Kim. The two-moment three-parameter decomposition approximation of queueing networks with exponential residual renewal processes. *Queueing Systems*, 68:193–216, 2011.
- [93] S. Kim and W. Whitt. Are call center and hospital arrivals well modeled by non-homogeneous Poisson processes? *Manufacturing and Service Oper. Management*, 16(3):464–480, 2014.
- [94] S.-H. Kim, W. Whitt, and W. C. Cha. A data-driven model of an appointment-generated arrival process at an outpatient clinic. *INFORMS Journal on Computing*, 30(1):181–199, 2018.
- [95] J. F. C. Kingman. Inequalities for the queue $GI/G/1$. *Biometrika*, 49(3/4):315–324, 1962.
- [96] P. J. Kuehn. Approximate analysis of general queueing networks by decomposition. *IEEE Transactions on Communications*, 27(1):113–126, 1979.
- [97] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. SIAM, 1999.
- [98] D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.
- [99] S.-Q. Li and C.-L. Hwang. Queue response to input correlation functions: discrete spectral analysis. In *IEEE INFOCOM'92: The Conference on Computer Communications*, pages 382–394. IEEE, 1992.
- [100] S.-Q. Li and C.-L. Hwang. Queue response to input correlation functions: continuous spectral analysis. *IEEE/ACM transactions on networking*, 1(6):678–692, 1993.
- [101] D. V. Lindley. The theory of queues with a single server. *Math. Proceedings Cambridge Phil. Soc.*, 48:277–289, 1952.
- [102] R. M. Loynes. The stability of a queue with non-independent inter-arrival and service times. *Mathematical Proceedings of the Cambridge Philosophical Society*, 58(3):497–520, 1962.

- [103] Y. Lu, T. Abdelzaher, C. Lu, L. Sha, and X. Liu. Feedback control with queueing-theoretic prediction for relative delay guarantees in web servers. In *The 9th IEEE Real-Time and Embedded Technology and Applications Symposium, 2003. Proceedings.*, pages 208–217. IEEE, 2003.
- [104] D. M. Lucantoni, K. S. Meier-Hellstern, and M. F. Neuts. A single-server queue with server vacations and a class of non-renewal arrival processes. *Advances in Applied Probability*, 22(3):676–705, 1990.
- [105] A. Mandelbaum and M. I. Reiman. On pooling in queueing networks. *Management Science*, 44(7):971–981, 1998.
- [106] B. Melamed. On Poisson traffic processes in Discrete-state Markovian systems with applications to queueing theory. *Advances in Applied Probability*, 11(1):218–239, 1979.
- [107] M. F. Neuts. A versatile Markovian point process. *Journal of Applied Probability*, 16(4):764–779, 1979.
- [108] M. F. Neuts. *Structured Stochastic Matrices of M/G/1 Type and their Application*. Marcel Dekker, New York, 1989.
- [109] G. Nieuwenhuis. Equivalence of functional limit theorems for stationary point processes and their Palm distributions. *Probability Theory and Related Fields*, 81:593–608, 1989.
- [110] N. O’Connell and M. Yor. Brownian analogues of Burke’s theorem. *Stoch. Proc. Appl.*, 96:285–304, 2001.
- [111] E. Ozkan and A. Ward. Dynamic matching for real-time ridesharing, working paper, USC. 2017.
- [112] H. Perros and R. O. Onvural. On the superposition of arrival processes for voice and data. In *Data Communication Systems and their Performance*, pages 381–394. Elsevier, 1991.
- [113] M. I. Reiman. Open queueing networks in heavy traffic. *Math. Oper. Res.*, 9(3):441–458, 1984.

- [114] M. I. Reiman. Asymptotically exact decomposition approximations for open queueing networks. *Operations Research Letters*, 9(6):363–370, 1990.
- [115] K. W. Ross and D. Yao. Optimal dynamic scheduling in Jackson networks. *IEEE Transactions on Automatic Control*, 34(1):47–53, 1989.
- [116] S. M. Ross. *Stochastic Processes*. Wiley, New York, second edition, 1996.
- [117] C. H. Sauer and K. M. Chandy. Computer systems performance modeling. *SIAM Rev.*, pages 124—125, 1981.
- [118] J. Schuijbroek, R. C. Hampshire, and W.-J. V. Hoeve. Inventory rebalancing and vehicle routing in bike sharing systems. *European Journal of Operational Research*, 257(3):992–1004, 2017.
- [119] R. Serfozo. *Introduction to Stochastic Networks*, volume 44. Springer Science & Business Media, 2012.
- [120] K. Sigman. Queues as Harris recurrent Markov chains. *Queueing Systems*, 3(2):179–198, 1988.
- [121] K. Sigman. The stability of open queueing networks. *Stochastic Processes and their Applications*, 35(1):11–25, 1990.
- [122] K. Sigman. *Stationary Marked Point Processes: An Intuitive Approach*. Chapman and Hall/CRC, New York, 1995.
- [123] D. Sinreich and Y. Marmor. Emergency department operations: the basis for developing a simulation tool. *IIE transactions*, 37(3):233–245, 2005.
- [124] K. Sriram and W. Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications*, SAC-4(6):833–846, 1986.
- [125] S. Suresh and W. Whitt. The heavy-traffic bottleneck phenomenon in open queueing networks. *Operations Research Letters*, 9(6):355–362, 1990.

- [126] W. Szczotka. Exponential approximation of waiting time and queue size for queues in heavy traffic. *Advances in Applied Probability*, 22(1):230–240, 1990.
- [127] W. Szczotka. Tightness of the stationary waiting time in heavy traffic. *Advances in Applied Probability*, 31(3):788–794, 1999.
- [128] L. Takacs. *Introduction to the Theory of Queues*. Oxford University Press, 1962.
- [129] G. Tesauro, R. Das, W. E. Walsh, and J. O. Kephart. Utility-function-driven resource allocation in autonomic systems. In *Second International Conference on Autonomic Computing (ICAC'05)*, pages 342–343. IEEE, 2005.
- [130] P. Tuominen and R. L. Tweedie. Exponential ergodicity in Markovian queueing and dam models. *Journal of Applied Probability*, 16(4):867–880, 1979.
- [131] J. Walrand. Poisson flows in single-class open networks of quasireversible queues. *Soch. Proc. Appl.*, 13:292–303, 1982.
- [132] J. Walrand. *An Introduction to Queueing Networks*. Prentice-Hall, Englewood Cliffs, NJ, 1988.
- [133] W. Whitt. Approximating a point process by a renewal process: two basic methods. *Oper. Res.*, 30:125–147, 1982.
- [134] W. Whitt. The queueing network analyzer. *Bell Laboratories Technical Journal*, 62(9):2779–2815, 1983.
- [135] W. Whitt. Approximations for departure processes and queues in series. *Naval Research Logistics (NRL)*, 31(4):499–521, 1984.
- [136] W. Whitt. Departures from a queue with many busy servers. *Mathematics of Operations Research*, 9(4):534–544, 1984.
- [137] W. Whitt. On approximations for queues, III: Mixtures of exponential distributions. *AT&T Bell Laboratories Technical Journal*, 63(1):163–175, 1984.
- [138] W. Whitt. Queues with superposition arrival processes in heavy traffic. *Stochastic Processes and Their Applications*, 21:81–91, 1985.

- [139] W. Whitt. An interpolation approximation for the mean workload in a $GI/G/1$ queue. *Operations Research*, 37(6):936–952, 1989.
- [140] W. Whitt. Planning queueing simulations. *Management Science*, 35(11):1341–1366, 1989.
- [141] W. Whitt. A review of $L = \lambda W$. *Queueing Systems*, 9:235–268, 1991.
- [142] W. Whitt. Variability functions for parametric-decomposition approximations of queueing networks. *Management Science*, 41(10):1704–1715, 1995.
- [143] W. Whitt. *Stochastic-Process Limits*. Springer, New York, 2002.
- [144] W. Whitt and W. You. Heavy-traffic limit of the $GI/GI/1$ stationary departure process and its variance function. *Stochastic Systems*, 8(2):143–165, 2018.
- [145] W. Whitt and W. You. Using robust queueing to expose the impact of dependence in single-server queues. *Operations Research*, 66(1):184–199, 2018.
- [146] W. Whitt and W. You. The advantage of indices of dispersion in queueing approximations. *Operations Research Letters*, 47(2):99–104, 2019.
- [147] W. Whitt and W. You. Heavy-traffic limits for stationary network flows, working paper, Columbia university. 2019.
- [148] W. Whitt and W. You. Time-varying robust queueing, forthcoming, *Operaitons Research*. 2019.
- [149] K. Xiong and H. G. Perros. Service performance and analysis in cloud computing. In *2009 Congress on Services - I*, pages 693–700. Citeseer, 2009.
- [150] C. Zacharias and M. Armony. Joint panel sizing and appointment scheduling in outpatient care. *Management Science*, 63(11):3978–3997, 2016.
- [151] S. Zeltyn, Y. N. Marmor, A. Mandelbaum, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, A. Shtub, T. Lauterman, et al. Simulation-based models of emergency departments: Operational, tactical, and strategic staffing. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 21(4):24, 2011.

- [152] Q. Zhang, A. Heindle, and E. Smirni. Characterizing the *BMAP/MAP/1* departure process via the ETAQA truncation. *Stochastic Models*, 21(2-3):821–846, 2005.