Performance evaluation for production systems through queueing models

Shen, Yichi

2018

Shen, Y. (2018). Performance evaluation for production systems through queueing models.
Doctoral thesis, Nanyang Technological University, Singapore.

<http://hdl.handle.net/10356/75882>

<https://doi.org/10.32657/10356/75882>

**NANYANG
TECHNOLOGICAL
UNIVERSITY**

SINGAPORE

**PERFORMANCE EVALUATION FOR PRODUCTION
SYSTEMS THROUGH QUEUEING MODELS**

SHEN YICHI

SCHOOL OF MECHANICAL AND AEROSPACE ENGINEERING

2018

PERFORMANCE EVALUATION FOR PRODCUTION SYSTEMS THROUGH QUEUEING MODELS

SHEN YICHI

SCHOOL OF MECHANICAL AND AEROSPACE ENGINEERING

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2018

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Professor Wu Kan, for his support and guidance during the past few years. His patience, motivation, and immense knowledge have helped and influenced me a lot. From him, I have learnt how to be a researcher with responsibility, integrity and passion. It would be impossible for me to finish the Ph.D. study without his assistance.

Many thanks to Professor Yuan Xueming, Professor Chen Chun-Hsien and Professor Cai Yiyu for their valuable comments on my research.

I would like to express my gratitude to Dr. Sandeep Srivathsan. We cooperated in the study on the three-moment approximation for the mean queue time of a single-server queue.

Much appreciation to the reviewers of this thesis. Their comments have made this thesis more complete and rigorous.

I would also like to thank my wife for her continuous support. Her encouragement and company make me stronger.

Content

List of Figures.....	VII
List of Tables	IX
Summary	XI
Chapter 1 Introduction.....	1
1.1 Background and Motivation.....	1
1.1.1 The Role of Queueing Models.....	2
1.1.2 Stability of Queueing Systems	3
1.1.3 Mean Queue Time Approximations for Queueing Systems.....	5
1.2 Research Objectives and Contributions	7
1.3 Organization of the Thesis.....	11
Chapter 2 Literature Review	13
2.1 Stability	13
2.1.1 Stability of Single-Server Queues.....	13
2.1.2 Stability of Multiclass Queue Networks	15
2.2 Mean Queue Time Approximations	18
2.2.1 Approximations for Single-Server Queues	18
2.2.2 Approximation for General Queueing Systems.....	22
2.3 Dependence in Queueing Systems and the Research Gap	25
Chapter 3 Fundamentals of Queueing Models.....	27
3.1 Different Types of Queueing Models.....	27

3.1.1	Single-Server Queues.....	27
3.1.2	Multiclass Queueing Systems.....	28
3.2	Underlying Processes of Queueing Models	30
3.2.1	Queue Time Process $\{W_n, n \geq 1\}$	30
3.2.2	Queue Length Process $\{Q(t), t \geq 0\}$	31
3.2.3	Associated General Process $\{X(t) = (Q(t), U(t), V(t), t \geq 0)\}$	32
3.3	The Significance of Stability	33
3.4	Performance Measures and Their Relations.....	34
3.4.1	Mean Queue Time.....	34
3.4.2	Mean Queue Length.....	36
3.5	System Architecture of the Thesis	36
Chapter 4	Stability of GI/G/1 Queues	38
4.1	Brief	38
4.2	Different Types of Stability	39
4.2.1	Proper Limiting Distribution of Queue Time $\{W_n, n \geq 1\}$	39
4.2.2	Pathwise Stability of Queue Length $\{Q(t), t \geq 0\}$	42
4.2.3	Strong and Weak Stability of Queue Length $\{Q(t), t \geq 0\}$	43
4.2.4	Positive Harris Recurrence of Associated General Process $\{X(t), t \geq 0\}$	45
4.3	Relations among Different Types of Stability	46
4.4	The Growth Rate of the Queue Time W_n when $\rho = 1$	48
4.5	Summary	50
Chapter 5	Stability of Multiclass Queueing Networks.....	52

5.1	Brief	52
5.2	Problem Formulation	55
5.3	General Servers.....	56
5.4	Case Studies.....	62
5.4.1	The Lu-Kumar Network with Constant Interarrival and Service Times.....	62
5.4.2	The Lu-Kumar Network with Stochastic Arrivals and Constant Service Times	69
5.5	Stability Conditions.....	71
5.6	The Applications	72
5.6.1	The Feedforward Networks and FBFS Reentrant Lines	73
5.6.2	Stability of Queueing Networks in Practice	73
5.7	Comparison to the Fluid Model Approach.....	74
5.8	Summary	75
Chapter 6	Mean Queue Time Approximation for GI/G/1 Queues	78
6.1	Brief	79
6.2	Previous Analytical Results and Approximations.....	81
6.3	Impact of the Third Moments of Interarrival and Service Times	83
6.3.1	Simulation Analysis	84
6.3.2	The Dependence on the Interarrival Time and Service Time Distributions	86
6.4	The Three-Moment Approximation	87
6.5	Simulation Validation.....	90
6.5.1	Results for Segment 3	92

6.5.2	Result for Segment 4.....	97
6.6	Proposed Approximation for the Mean Queue Time	103
6.7	Summary	105
Chapter 7	Mean Queue Time of General Queueing Systems	107
7.1	Brief	107
7.2	Notations and Models	108
7.3	The Properties of Affine Structure.....	110
7.3.1	GI/G/1 and GI/G/n Queueing Systems	111
7.3.2	Single-Server Tandem Queues	115
7.3.3	Homogeneous-Server Tandem Queues	119
7.3.4	Reentrant Lines.....	120
7.4	The Approximation Algorithm	124
7.5	Validation of the Approximation Algorithm.....	126
7.5.1	Error Analysis for Single-Server Tandem Queues.....	127
7.5.2	Case Studies for Homogeneous-Server Tandem Queues.....	129
7.6	Summary	131
Chapter 8	Conclusions	134
8.1	Contributions of This Study	134
8.2	Limitations and Future Research	135
8.2.1	Identification of General Servers.....	135
8.2.2	Analytical Results for Approximation Errors.....	136
8.2.3	The Relationship between the Traffic Intensity and Mean Queue Time....	136

8.2.4 Capacity Planning	137
References	138
Appendix A1	151
Appendix A2	153
Appendix A3	162
Appendix A4	169
The Author's Publications Resulting from the Thesis	174

List of Figures

Figure 2-1 ASIA (left) and BSIA (right) systems Wu and McGinnis (2013).....	24
Figure 3-1 A GI/G/1 queueing model.....	28
Figure 3-2 Part of the procedure in a semiconductor fab (Bramson 2008).....	29
Figure 3-3 System architecture of the studies.	37
Figure 4-1 The relations among different types of stability.....	48
Figure 5-1 An M/M/1 queue.	57
Figure 5-2 A two-stage tandem queue.	58
Figure 5-3 A six-class queueing network.....	59
Figure 5-4 A tandem queue with four single-server stations.	62
Figure 5-5 The Lu-Kumar network.	63
Figure 5-6 WIP at job classes 2 and 4 when $\alpha = 1$	64
Figure 5-7 WIP at job classes 2 and 4 when $\alpha = 0.9$	66
Figure 5-8 WIP at job classes 2 and 4 when $\alpha \sim \exp(1)$	69
Figure 5-9 WIP at job classes 2 and 4 when $\alpha \sim \exp(1/0.85)$	70
Figure 5-10 Job mean queue time at compact server $\{2, 4\}$	71
Figure 6-1 Absolute error plot for the Gamma/Gamma/1 queueing system.	99
Figure 6-2 Absolute error plot for the Gamma/Lognormal/1 queueing system.	99
Figure 6-3 Absolute error plot for the Lognormal/Lognormal/1 queueing system.	99
Figure 6-4 Absolute error plot for the Weibull/Weibull/1 queueing system.	100
Figure 6-5 Recommended approximations.	105
Figure 7-1 A J -stage tandem queue with multiple servers at each stage.	109
Figure 7-2 Values of $\alpha 1\rho$ for a Gamma/Gamma/1 queue with different SCVs.	113

Figure 7-3 Values of $F(\rho)$ in different approximations when $c_a^2 = c_s^2 = 0.8$	114
Figure 7-4 Values of $F(\rho)$ in different approximations when $c_a^2 = c_s^2 = 1.25$	115
Figure 7-5 A single-server tandem queue.	116
Figure 7-6 Values of $IR_2(\rho)$ for systems with low SCVs.	120
Figure 7-7 Values of $IR_2(\rho)$ for systems with high SCVs.	120
Figure 7-8 The original reentrant line $\mathcal{S}$	121
Figure 7-9 The auxiliary systems $\mathcal{S}1$ (left) and $\mathcal{S}2$ (right).	121
Figure 7-10 The first coordinate of $\alpha^1(\rho)$	124
Figure 7-11 The second coordinate of $\alpha^1(\rho)$	124

List of Tables

Table 4-1 Relations between the traffic intensity and different types of stability of a GI/G/1 queue.....	47
Table 4-2 Growth rate of $E[W_n]$	49
Table 5-1 The differences between the general server and fluid model approaches.....	75
Table 6-1 Parameters and variables in a GI/G/1 queue.	81
Table 6-2 Results in literature (exact and approximate).....	83
Table 6-3 Relative effect of the interarrival and processing time distributions on the mean queue time in a Gamma/Gamma/1 queueing system.....	85
Table 6-4 Summary of results for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$	93
Table 6-5 Results of paired t-test for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$	94
Table 6-6 Summary of results for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ at low utilizations.....	94
Table 6-7 Summary of results for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ at medium utilizations.	95
Table 6-8 Summary of results for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ at high utilizations.....	95
Table 6-9 Results of paired t-test for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ at low utilizations. ..	96
Table 6-10 Results of paired t-test for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ at medium utilizations.	96
Table 6-11 Results of paired t-test for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ at high utilizations....	96
Table 6-12 Summary of results for $c_a^2 > 1$, $c_s^2 > 1$	98
Table 6-13 Results of paired t-test for $c_a^2 > 1$, $c_s^2 > 1$	98
Table 6-14 Summary of results for $c_a^2 > 1$, $c_s^2 > 1$ at low utilizations.	101
Table 6-15 Summary of Results for $c_a^2 > 1$, $c_s^2 > 1$ at medium utilizations.	101
Table 6-16 Summary of results for $c_a^2 > 1$, $c_s^2 > 1$ at high utilizations.....	102
Table 6-17 Results of paired t-test for $c_a^2 > 1$, $c_s^2 > 1$ at low utilizations.	102

Table 6-18 Results of paired t-test for $c_a^2 > 1, c_s^2 > 1$ at medium utilizations.....	102
Table 6-19 Results of paired t-test for $c_a^2 > 1, c_s^2 > 1$ at high utilizations.	103
Table 7-1 Distributions of the service times in each case.....	123
Table 7-2 Approximation errors when $c_{s_1}^2 = c_{s_2}^2 = 0.5$	130
Table 7-3 Approximation errors when $c_{s_1}^2 = c_{s_2}^2 = 5$	130
Table 7-4 Performance of the affine structure algorithm and the decomposition method employed by Hopp et al. (2002) when $c_{s_1}^2 = c_{s_2}^2 = 0.5$	131
Table 7-5 Performance of the affine structure algorithm and the decomposition method employed by Hopp et al. (2002) when $c_{s_1}^2 = c_{s_2}^2 = 5$	131

Summary

This thesis focuses on proposing a unified framework for the performance evaluation of production systems to support the effective and efficient system management (e.g., capacity planning and productivity improvement). Queueing systems are employed to model production systems and an important performance measure, the mean queue time, is analyzed explicitly. Furthermore, we will firstly clarify the stability of queueing systems in this thesis since it is a prerequisite for the study of other performance measures (e.g., mean queue time and mean queue length). Since single-server queues are elementary building blocks of complex queueing systems, for each topic, we start from the single-server queue and then move on to the queueing networks.

Stability of Queueing Systems

Stability of queueing systems is of fundamental importance in the application of queueing models. To determine the capacity or service rate of a queueing system, the stability must first be established. Although the stability of single-server queues has been a classic topic and there are various types of stability in literature, the relationships among them are not clear. To provide an overall picture of the single-server queue stability and make it easy to apply these different types of stability to practical situations, we survey the related research, classify them based on the underlying stochastic processes and clarify the relationships among them.

Different from the stability of single-server queues, the queueing network stability is more complicated due to the dependence among stations and “if and only if” conditions for

stability are still not available. In this thesis, we generalize the concept of servers in the context of queueing networks. We show that the general servers have similar impacts on the system stability as physical stations and a queueing network is pathwise stable if and only if the effective traffic intensity of each general server does not exceed one. Although the identification of general servers may be intricate, the concept can be employed indirectly to show the stability of queueing networks operating under the Work-in-Progress-Dependent dispatching policies.

Mean Queue Time Approximations for Queueing Systems

For single-server queues, the conventional two-moment approximations for mean queue times can produce large deviations in some situations, especially when the variability of the interarrival times is high. To improve the accuracy, we develop a unified three-moment approximation in this thesis. The new approximation has been shown to outperform the two-moment ones through extensive numerical experiments.

For queueing networks, the trade-off between the mean queue time and throughput is harder to establish due to the dependence within the system. To simplify the analysis, the prior approximations usually made the independence or heavy traffic assumptions, which may not hold in practical situations. To provide better estimations for system performance, we propose a unified approximation algorithm based on the properties of mean queue times in queueing systems. Compared to prior approximations, the new approximation makes no independence or heavy traffic assumptions and can be easily applied to different types of queueing systems. The accuracy has been verified through various queueing models.

Chapter 1 Introduction

The improvement of human being's life highly depends on the development of industry. Besides the “hardware” (e.g., plants and machines) in a production system, the “software” (e.g., system management) is also vital to the productivity improvement. To provide meaningful managerial insights and support the decision making, we will focus on the performance evaluation of production systems in this thesis. An overall introduction is given in this chapter.

1.1 Background and Motivation

System management plays a crucial role in the production industry. Since the last century, much effort has been made to improve production systems (e.g., maximize the throughput, reduce the mean cycle time and minimize the operation cost) and some well-known production policies and tools (e.g., JIT (Just-in-Time), TOC (Theory of Constraints), and OQNet (Optimized Queueing Network)) were developed (Goldratt and Cox 2016, Hopp and Spearman 2011, Hopp et al. 2002). Although the goals or methods of them may be different, all of them rely on the performance evaluation of production systems. For example, as in Hopp et al. (2002), capacity planning in the semiconductor industry commonly involves minimizing the overall operation cost through determining how much and what type of capacity (or machines) to install, subject to the following two conditions:

- (a) The mean cycle time of each product does not exceed the maximum cycle time, and
- (b) The throughput of each product equals the required throughput.

Since the cost of machines in a semiconductor production line can exceed one billion dollars, capacity planning is essential to the survival of a firm in the competitive environment (Hopp et al. 2002). To solve this optimization problem, we should first be able to evaluate the throughput and mean cycle time of the system under a given configuration to verify the above two conditions. Furthermore, the accuracy of performance evaluation will directly influence the final decision for capacity planning. Therefore, to support better system management, this study is devoted to thoroughly analyzing, understanding and evaluating the performance of production systems.

1.1.1 The Role of Queueing Models

Queueing theory was originated in Erlang (1909), in which a telegraphic system was studied. Since then, queueing systems have been widely utilized to evaluate the performance of service systems and production systems. With queueing systems, the variability of production systems (e.g., machine failures, setups and random process times) can be well modeled and analyzed. The process flows of products can also be well described via the job routings in queueing networks.

A single-stage workstation or service counter can be modeled by a simple GI/G/1 queue (Kendall's notation, see Kendall (1953)). If there are multiple servers, GI/G/n queues can be employed. Other queueing models, e.g., tandem queues and multiclass queueing networks, are also commonly employed in practice (Hopp et al. 2002). These flexible queueing models have provided us a theoretical framework to evaluate the performance of

production systems. In practice applications, the throughput and mean cycle time constraints correspond to the stability and mean queue time performance of queueing systems respectively.

1.1.2 Stability of Queueing Systems

The stability of queueing systems has been a classic topic since the last century. For GI/G/1 queues, the stability is easy to analyze and the stability condition can commonly be given in terms of utilizations. In literature, various types of stability have been studied based on different stochastic models, e.g., Lindley (1952) defined stability based on the queue time process and El-Taha and Stidham (2012) studied the pathwise stability based on the input-output process. However, the differences and connections among them are not clear. Hence, to provide an overall picture and a comprehensive understanding for these different types of stability, we systematically study the different underlying processes of a GI/G/1 queue, classify the various types of stability and study the relations among them.

Although the stability of GI/G/1 queues can be well characterized by the utilizations of servers, the stability of multiclass queueing networks depends on not only the utilization of each physical station. In queueing networks, the service disciplines and job routings can also significantly impact the system stability (Bramson 2008). Due to the difficulty in modeling the dependence within a queueing network, the stability of general queueing networks is much harder to analyze. For a long period, the study of stability has been focusing on some specific networks such as Jackson networks or Kelly type networks

(Baskett et al. 1975, Bramson 2008, Jackson 1957, Kelly 2011) and in these situations, each station can work like an independent queue in the steady state. However, for general multiclass queueing networks, it has been shown that the system can be unstable even if the utilization at each station is less than one. For example, Lu and Kumar (1991) showed the instability of a deterministic reentrant line under a SBP (static buffer priority) policy. Even a queueing network operating under the FIFO (First-in-First-Out) disciplines can be unstable even if the usual traffic condition is satisfied (Bramson 1994, Seidman 1994). Readers can refer to Bramson (2008) for more examples. All these phenomena imply that the theory of a GI/G/1 is not applicable to general queueing networks and stability depends on not only the physical stations but also the service disciplines and some other factors (e.g., service time distributions). The interdependence among different stations or job classes has a significant influence on the system performance.

Dai (1995) showed that a queueing network is positive Harris recurrent if the corresponding fluid limit model is stable. The fluid model, as a continuous analogue of the original queueing system, provides a new approach to consider the stability of queueing networks. However, the stability of a fluid model is itself hard to verify. Furthermore, there is a gap between the performance of a queueing network and its corresponding fluid model and the fluid model approach is not capable to get “if and only if” conditions (Bramson 2008, Bramson 1999, Dai et al. 1999). The connections between queueing networks and fluid models are not evident and the insights from the approach are limited.

To overcome the difficulties, we analyze the stability problem from the perspective of

servers. Inspired by the theory of a GI/G/1 queue, where the stability fully depends on the single server, we consider the mutual blocking effect among different job classes and generalize the concept of servers in the context of queueing networks. The negative dependence within a queueing network, which can significantly influence the system performance, can be well incorporated by the general servers. Through the new approach, we show that a queueing network is pathwise stable if and only if the effective traffic intensity of every general server does not exceed one.

1.1.3 Mean Queue Time Approximations for Queueing Systems

The exact analysis for the mean queue time in a single-server queue highly relies on the memoryless property of exponential distributions. For an M/G/1 queue, the mean queue time can be given by the famous Pollaczek–Khinchine formula (Khinchin 1932, Pollaczek 1930a, Pollaczek 1930b). However, in GI/G/1 queues, the exact analysis becomes difficult without the memoryless property and no exact closed-form solution for the mean queue time is available. Hence, much effort has been made to seek for a reliable approximation for the mean queue time of a GI/G/1 queue, which is also the first step towards the approximations for general queueing systems. Most of the approximations in literature are based on the first two moments of the interarrival and service times (e.g., see Kingman (1962), Heyman (1975) and Kraemer and Langenbach-Belz (1976)) and ignore the impact of the higher moments. However, in some cases, the third moment can also significantly influence the mean queue time and the conventional two-moment approximations will

produce large estimation errors (Myskja 1990). Hence, to provide more reliable estimations, we study the impact of the third moments of the interarrival and service times in this thesis and propose a three-moment approximation.

Similar to stability, the mean queue time of a complex network is much harder to analyze than a single-server queue. While the arrival process in a GI/G/1 queue is renewal, servers in a queueing network usually face a non-renewal arrival process. The non-renewal nature within a general queueing network makes the well-developed theory of GI/G/1 queues inapplicable. The dependence within a queueing network, which can significantly influence the queue time performance, is difficult to depict (see Hopp et al. (2002) and Kleinrock (1976)). Whitt (1983) employed the independence assumption to simplify the analysis and developed the an approximation tool called QNA (Queueing Network Analyzer). Harrison and Nguyen (1990) and (Dai and Harrison (1992) developed approximations based on the heavy traffic assumption. However, sometimes these additional assumptions are invalid and can lead to large deviations as large as 20% for practical manufacturing systems (e.g., Hopp et al. (2002)). Simulation models can also be employed to evaluate the performance of production systems. While they are suitable for the “what if” analysis, it is time-consuming if we need the performance under various configurations.

In this thesis, we don’t analyze the mean queue time of queueing systems under the independence or heavy traffic assumptions. Instead, we approximate the mean queue time via utilizing the auxiliary systems inspired by the research of Wu and McGinnis (2013). It

turns out that an “affine” relationship generally exists between the mean queue time of a queueing system and its corresponding auxiliary systems. Based on the affine structure and its properties, we develop an approximation algorithm for the mean queue time of a general queueing system. Since no additional assumption is made, the algorithm tends to provide reliable estimations for various queueing systems.

1.2 Research Objectives and Contributions

Although queueing systems can effectively model production systems with high fidelity, the analysis of queueing models is generally not easy. This research focuses on the performance evaluation of queueing systems. Since a complex queueing network is always comprised of the simpler GI/G/1 queues, the behavior of GI/G/1 queues should first be well understood before we can move to the more complex queueing networks. In this thesis, we will discuss two topics, stability and mean queue time.

Research Objective 1: Survey the stability of GI/G/1 queues.

Various types of stability have been defined for GI/G/1 queues based on different stochastic processes (e.g., the queue time process and the queue length process). Although the previous research on each type of stability appears to be adequate, the relations among them are not clear. Therefore, we will survey the research on the stability of GI/G/1 queues to get a comprehensive understanding and an overall picture, which will also provide

insights and foundations for the study of queueing network stability. In this research, we will

- (i) study different underlying processes of and classify the different types of stability,
- (ii) clarify the differences and connections among them, and
- (iii) propose the stability condition in terms of the utilization of a GI/G/1 queue.

Research Objective 2: Investigate the stability of multiclass queueing networks.

In a multiclass queueing network, due to the negative correlations, job classes at different physical stations can mutually block each other and behave like a virtual station (Dai and Vande Vate 1996). Dai and Vande Vate (1996) showed that “if and only if” conditions for the global stability of two-station queueing networks can be given based on the concept of virtual stations. Hasenbein (1997) generalized the concept to pseudostations and proposed the necessary conditions for stability. Both studies imply that, except for the physical stations, virtual servers, which also influence the system stability, may be induced due to the dependence within a queueing network. This inspires us to rethink the meaning of servers in the context of queueing networks. To propose a unified framework for the stability of queueing systems (i.e., single-server queues and queueing networks), we are going to

- (i) illustrate the negative correlations, i.e., mutual blocking, among job classes in multiclass queueing networks,

- (ii) generalize the concept of servers in the context of queueing networks,
- (iii) study the properties of queueing network stability and propose the stability conditions, and
- (iv) discuss the stability of queueing systems operating under specific service policies.

Research Objective 3: Propose a three-moment approximation for the mean queue time of a GI/G/1 queues.

While the prevalent two-moment approximations for the mean queue time of GI/G/1 queues ignored the impact of the third moments of interarrival and service times, there are situations where the third moment can make a big difference (Myskja 1990). Myskja (1991) developed a heuristic three-moment approximation based on the exact result of an $H_2/M/1$ queue. However, the applications of the heuristic approximation are limited (see Chapter 6 for details) and the approximation needs to be improved. In this thesis, we will

- (i) analyze the impact of the third moments of interarrival and service times on the mean queue time,
- (ii) propose an approximation formula for the mean queue time, and
- (iii) validate the approximation through extensive simulation experiments.

Research Objective 4: Investigate the properties of the mean queue times in general queueing systems and propose a unified approximation algorithm.

Dependence is prevalent in queueing systems and makes the analysis much intricate. Except for several special cases, there's no exact closed-form solution for the mean queue time of a general queueing system. The previous approximation tools in literature such as QNA can produce unsatisfactory estimates due to their additional assumptions. Wu and McGinnis (2013) proposed an interpolation method to approximate the mean queue time of single-server tandem queues through the concept of Intrinsic Ratio (IR), where the mean queue time of a tandem queue has an “affine” relationship with two well-defined virtual systems. It was shown that the IR of a tandem queue is nearly linear with respect to the system utilization. Based on the nice structure, an interpolation approximation for the mean queue time of single-server tandem queues was developed. The new approach, which can incorporate the impact of the dependence within the tandem queue, has been shown to outperform the conventional approximations through extensive case studies. Inspired by this phenomenon, we generalize the concept of Wu and McGinnis (2013) to more general settings in this research. In the following, we will

- (i) study the properties of mean queue time in general queueing systems,
- (ii) investigate the affine structure of mean queue times in general queueing systems,
- (iii) develop a unified approximation algorithm for mean queue times, and
- (iv) validate the algorithm through different queueing models and conduct the error analysis.

1.3 Organization of the Thesis

In Chapter 1, we have given an overall introduction for the performance evaluation of queueing systems. In Chapter 2, we will review the previous studies in this area and clarify the research gap of this study. Chapter 3 presents the fundamentals of queueing models and provides the overall system architecture for this thesis.

In Chapter 4, we first survey the different underlying stochastic models of GI/G/1 queues, which can describe the system evolution from different perspectives. Different types of stability are then classified, and the stability conditions are given in terms of utilizations. We have also summarized the relationships among these different types of stability. The study will provide insights for the performance analysis of a single-machine system and serve as a foundation for the study of complex queueing networks.

The stability of queueing networks is studied in Chapter 5. We propose the concepts of mutual blocking and general servers in the context of queueing networks. We show that the stability of queueing networks fully depends on the utilizations of these general servers. Although the identification of general servers for general queueing networks may be hard, we can employ the concept to prove the stability of queueing networks under some specific service disciplines.

Chapter 6 presents the three-moment approximation for the mean queue time of GI/G/1 queues. We show that the new approximation outperforms the conventional two-moment ones. Since the approximations for queueing networks generally require the results of single-server queues, the three-moment approximation will also support the performance

evaluation of complex queueing networks.

Chapter 7 focuses on the mean queue time of general queueing systems. We show that the mean queue time of general queueing systems enjoys a nice affine structure. Properties of the affine structure are investigated and we develop a unified approximation algorithm for the mean queue time of a general queueing system. The approximation algorithm has been validated through various queueing models.

Chapter 8 concludes the study and indicates directions for future research.

Since each chapter focuses on different topics, the notations in different chapters may be slightly different. However, some conventional notations (e.g., ρ (utilization or traffic intensity), λ (arrival rate), and μ (service rate)) are consistent throughout the whole thesis. Note that the terms “utilization” and “traffic intensity” have the same meaning. We use both terms in this thesis.

Chapter 2 Literature Review

This research focuses on studying the properties of queueing systems to provide fundamentals and insights for the production management. The studies on the stability of queueing systems are firstly reviewed. Different approaches to approximate the mean queue time are then introduced. Furthermore, from the perspective of dependence, we discuss the main issue of the previous studies and the technical difficulty in studying queueing systems. This provides an overview for the literature and also clarifies the research gap of this thesis.

2.1 Stability

2.1.1 Stability of Single-Server Queues

Although there are various types of stability in probability theory (e.g. recurrence, positive Harris recurrence and pathwise stability), they cannot be applied to a queueing system directly. Before discussing stability, one should first specify the underlying process of the queue. The queue time process, queue length process, and associated general process have received the most attention in the prior research.

The queue time process $\{W_n, n \geq 1\}$ of a $GI/G/1$ queue was first thoroughly studied by Lindley (1952), where W_n is the n th customer's queue time. A queue is called stable if the queue time sequence has a proper limiting distribution. It was shown that the proper limiting distribution exists if and only if the traffic intensity is less than 1. Although the queue time process is unstable when the traffic intensity equals 1, the limiting distribution of $W_n/\sqrt{n}$ exists and is that of the absolute value of a normal random variable (Asmussen

2008, Billingsley 2013, Chung 2001).

The queue length of a $GI/G/1$ queue is generally modeled as an input-output process $\{Q(t), t \geq 0\}$, where $Q(t) = Q(0) + A(t) - D(t)$, $Q(0)$ is the number of jobs in queue at time 0, $A(t)$ is the total arrivals at time t , and $D(t)$ is the total departures before time t . The process $\{Q(t), t \geq 0\}$ is said to be stable if its growth rate is $o(t)$ almost surely (Altman et al. 1998, El-Taha and Stidham 1992, El-Taha and Stidham 1993, Mazumdar et al. 1992). Besides pathwise stability, weak stability and strong stability can also be defined on the queue length process (Courcoubetis et al. 1989, Courcoubetis and Weber 1994).

Exploiting the model used by Dai (1995) to describe queueing networks, we can also define an associated Markov process $\{X(t) = (Q(t), U(t), V(t)), t \geq 0\}$ to depict the dynamics of a $GI/G/1$ queue, where $Q(t)$ is the number of jobs in the queue at time t , $U(t)$ is the remaining interarrival time, and $V(t)$ is the residual service time for the job in service ($V(t) := 0$ if $Q(t) = 0$). In this regime, a queue is said to be stable if the Markov process defined above is positive Harris recurrent. Positive Harris recurrence guarantees the existence of a finite invariant measure (Meyn and Tweedie 2012).

Since the 1960s, queues with more general settings have been studied. Loynes (1962) studied queues with non-independent but stationary arrival processes. Queues with non-stationary input streams were discussed in Rolski (1981). Szczotka (1986a, 1986b) studied queues from the viewpoint of stationary representations. Sigman (1988) studied queues with inputs governed by a Harris recurrent Markov chain. Furthermore, through the sample-path analysis, El-Taha and Stidham (2012) only assumed the arrival processes

satisfy $\lim_{t \rightarrow \infty} A(t)/t = \alpha$. Morozov and Delgado (2009) studied the regenerative queueing systems. The stability of an M/G/1 retrial queue with unreliable servers was studied by Sherman et al. (2009).

2.1.2 Stability of Multiclass Queue Networks

In production lines, congestion at a station usually leads to congestion at its downstream stations in a later period. Dependence among stations is commonly seen in practical manufacturing systems. However, due to the nice properties of Brownian motions, people tend to evaluate system performance in the heavy traffic regime, where the dependence is weakened by randomness. Without resorting to heavy traffic, dependence among stations can be weak in some special cases. In a Jackson network, although the queue times among servers are dependent (Burke 1956), each server behaves like an independent M/M/1 queue in the steady state, and the entire network enjoys the nice product-form solution (Jackson 1957).

Sometimes the dependence among stations can be strong: a station's state can affect the states of other stations, so that the job classes from different stations cannot be processed simultaneously. This phenomenon has been observed by Kumar and Seidman (1990), Lu and Kumar (1991), Rybko and Stolyar (1992), Bramson (1994) and Seidman (1994). In these situations, people observed that the system can be unstable, even if the utilization of every station is less than 1 (Dai and Vande Vate 1996, Down and Meyn 1994).

Fluid model, which gives a deterministic and continuous analogue of a stochastic

system, has been used to determine the stability conditions for a wide range of queueing networks. Rybko and Stolyar (1992) studied the stability of multiclass networks with discrete state space via fluid models. Meyn (1995) treated transience for queueing network models through the associated fluid limit models. Dai (1995) proved that a queueing network is positive Harris recurrent if the corresponding fluid limit model eventually reaches zero regardless of the initial system configuration. This sufficient condition was simplified from the stability of a piecewise-linear fluid limit model to the stability of a linear fluid model by Chen (1995). However, Bramson (1999) showed the converse is not true, i.e., a queueing network with an unstable fluid model may be stable. Specifically, Dai et al. (2004) proved that the stability region of a two-station, five-class reentrant queueing network, operating under a non-preemptive static buffer priority service, depends on the distributions of interarrival and service times. As pointed by Dai (2009), “For a queueing network operating under some service policies, its stability region can depend on its distributions, the preemption mechanism, the way that simultaneous events are handled. Practical fluid models cannot capture these fine factors, and hence cannot be used to sharply determine stability of the corresponding queueing networks.” This can be considered as a thoughtful concluding remark to the applications of fluid models to the queueing network stability.

The notion of stability mentioned above centers around the positive Harris recurrence of the underlying Markov process (Dai 1995), which guarantees the existence of a finite invariant probability measure. Another interesting stability in literature is the pathwise

stability (Altman et al. 1998, El-Taha and Stidham 2012). It requires that the long-run arrival rate equals the departure rate. Chen (1995) first showed that a queueing network is pathwise stable if the corresponding fluid model is weakly stable.

Determining the stability region of a multiclass network can be difficult (Dai and Weiss 1996). Hence, some research has focused on global stability of queueing networks (Chen 1995, Dai et al. 1999, Dai and Vande Vate 2000), where global stability means that a queueing network with specific topology is stable under any work conserving policy. Dai and Vande Vate (1996) studied the global stability of a two-station network and obtained its sufficient and necessary conditions under pathwise stability based on the concept of virtual stations. Hasenbein (1997) extended the virtual stations to pseudostations for multi-station networks and obtained the necessary condition for global stability. Dai et al. (1999) searched for the monotone global stability region of a three-station fluid network. However, for the purpose of productivity improvement, knowing the stability region under a specific service policy is more important than finding the global stability region. Maglaras (1999) and Dai and Lin (2005) studied the stability of multiclass queueing networks under specific scheduling policies. Bramson (2011) analyzed the queueing networks under the JSQ (joint the shortest queue) policy. Mukhopadhyay and Mazumdar (2016) investigated the stability region of queueing networks under the randomized JSQ policy. However, none of them is able to present the general “if and only if” conditions for stability.

2.2 Mean Queue Time Approximations

Queueing theory is an analytical tool that can be used in the performance evaluation of a wide range of real life systems such as manufacturing and communication systems. As a key performance measure, the mean queue time of queueing systems has received plenty of attention since the origination of queueing theory.

2.2.1 Approximations for Single-Server Queues

Single-server queues, which are the fundamental building blocks of the practical systems, have been studied since the early 20th century. The early models concentrated on the special cases of the GI/G/1 queueing system, where either the interarrival or service times followed exponential distributions, or a combination of both. Erlang (1909) studied the queue time distribution of an M/D/1 queue. The mean queue time of an M/M/1 queue can be obtained by solving a Markov chain with the state space representing the number of customers in the system (e.g., Chen and Yao (2013)). Pollaczek (1930a, 1930b) and Khinchin (1932) developed the exact solution for the M/G/1 queue (i.e., the Pollaczek–Khinchine formula). The exact results for the GI/M/1 queueing system can be obtained by solving a Markov chain with the state space (L_k^a) denoting the number of customers in the system just before the k th customer arrives (see Adan and Resing (2002)). However, the Laplace-Stieltjes transform for the arrival process has to be obtained to arrive at a closed-form expression for the mean queue time of a GI/M/1 queue.

Lindley’s integral equation (Lindley 1952) is the seminal work in developing analytical

results for the GI/G/1 queue. However, solving Lindley's equation involves techniques of complex variable theory. Obtaining closed-form solutions has proved challenging with the exception of some special cases. As a consequence, there was a spurt in research that developed bounds and/or approximate closed-form expressions for the mean queue time of the GI/G/1 queue. The performance of a GI/G/1 queue is mostly approximated by Brownian motions. In order to gain enough randomness (so that the queue can behave like Brownian motions), the queue must be resorted to heavy traffic, where Donsker's theorem (Donsker 1951, Donsker 1952) can be applied. Since the behavior of Brownian motions is dominated by its first two moments, the performance of a GI/G/1 queue can be approximated accurately by its first two moments in heavy traffic. The pioneering study by Kingman (1962) demonstrated that, as the utilization ρ approaches one, the limiting distribution of $(1 - \rho)w$ is negative exponential, where w is the equilibrium queue time. The heavy traffic mean queue time provides an upper bound for the mean queue time in a GI/G/1 queue. The bound is weak at small utilizations and becomes tighter as the utilization approaches unity. Before achieving heavy traffic, the behavior of a GI/G/1 queue will deviate from Brownian motions. Hence, the approximate error in general situations may come from the following two: (1) deviation from the Brownian motion itself (i.e., the errors caused by the first two moments), and (2) the higher moments, which are not captured by Brownian motions.

Heyman (1975) developed an approximation for the queue length and queue time distributions using a diffusion model. The model used the mean and variance of the

asymptotic rate of change of an appropriately chosen diffusion process for the queue length process. As ρ approaches one, the approximation converges to the heavy traffic approximation provided by Kingman (1962). It was heuristically extended by using a correction factor ' g ', which is a function of the utilization and the SCVs of the interarrival and service times, to enhance its accuracy in light and moderate traffic (Kraemer and Langenbach-Belz 1976). Gelenbe (1975) and Kobayashi (1974) have also proposed diffusion approximations for the mean queue time in a GI/G/1 queue. Shanthikumar and Buzacott (1980) reviewed these two-moment approximations, and performed extensive numerical experiments for queues with small variability (i.e., SCVs of the arrival and service processes are not greater than one). In this range, the heuristic approximations were found to outperform the diffusion approximations. Based on a study by Yu (1977), Shanthikumar and Buzacott (1980) commented that anomalies might occur when the variability of the arrival process is large ($SCV \geq 1$) using an example of the $H_2/M/1$ queueing system (the mean queue length can vary between 624.62 to 23.48 while the utilization and arrival SCV were fixed at 0.95 and 64, respectively). Based on the example, they conclude that *this example obviously excludes the possibility of obtaining a good approximation for the more general case GI/G/1 with the SCV of interarrival times > 1 .* This inference has been attributed to the effect of the third moment of the arrival process (Myskja 1990, Myskja 1991).

Kimura (1985) first developed a diffusion approximation that applies the Laplace-Stieltjes transform for the arrival process, and then applied heuristic modifications

to further enhance its performance. This approximation serves as a hybrid between a two-moment approximation and the full distribution approximation for the arrival process (Myskja 1990). The drawback of this approach is the absence of an explicit formula for the mean queue time in a GI/G/1 queue.

Myskja (1990, 1991) used the exact results for the $H_2/M/1$ queueing system to develop heuristic approximations for the mean queue time in a GI/G/1 queue. Myskja (1990) applied an interpolation between two extreme cases of the exact solution of the $H_2/M/1$ queueing system to approximate the mean queue time. The first extreme case represented a batch Poisson arrival process with geometric batch size, where the solution of the $M^B/GI/1$ queueing system yields the mean queue time. The second extreme case represented a Poisson arrival process with long empty arrivals, for which the solution of an $M/GI/1$ queueing system yields the mean queue time. A heuristic adjustment was made to interpolate between the two extremes, namely, $q_a = q_0$ and $q_a = \infty$. Note that q_a represents the relative third moment of the arrival process and is given by the ratio of the third moment to six times the cube of the first moment (See Table 6-1 for more details), and q_0 represents the lowest value of the relative third moment for a given mean and SCV. The approximation was found to overestimate the mean queue time in moderate and heavy traffic. Myskja (1991) heuristically extended the exact results for the $H_2/M/1$ queueing system to obtain a 3/2-approximation for the mean queue time, which uses the first three moments of the arrival process, and the first two moments of the service time distribution. The approximation was found to outperform other existing ones when the interarrival times

followed H_2 distributions and the service time distribution followed either H_2 or Erlang distributions. Since Myskja (1991) only presented the expression without a proof, a complete proof is given in the Appendix A3.

Hence, we should consider the impact of the third moments of interarrival and service time properly to provide more accurate estimates for GI/G/1 queues.

2.2.2 Approximation for General Queueing Systems

For queues with multiple servers, there is also exact analytical results for the M/M/n queues based on the Markov chain analysis (e.g., see Adan and Resing (2002) and Chen and Yao (2013)). However, the evaluation of general cases is complicated and Gupta et al. (2010) even rigorously explained why the approximation of M/G/n queues is hard. Till now, the exact analysis of an M/G/n queue remains an open problem. Inspired by the case of GI/G/1 queues, Cosmetatos (1982), Kimura (1986), and Page (1982) tried to approximate the mean queue time of GI/G/n queues through interpolating the results of M/M/n, M/D/n and D/M/n queues, of which the performance is easier to evaluate. Whitt (1993) further developed a unified approximation for the mean queue time of a GI/G/n queue through comparing it with the mean queue time of the corresponding M/M/n queue. It turns out that the mean queue time of a GI/G/n queue ($W^{GI/G/n}(\rho)$) can also be approximated by $F(\rho)W^{M/M/n}(\rho)$, where $W^{M/M/n}(\rho)$ is the mean queue time of the corresponding M/M/n queue with the same utilization and $F(\rho)$ is the coefficient term.

The above phenomenon implies that, although the exact solution for the mean queue

time of a GI/G/1 or GI/G/n queue is unavailable, it is closely related to the mean queue time of its corresponding M/M/1 or M/M/n queue. There is an affine structure among the mean queue times of these systems and based on the affine structure, the mean queue times of GI/G/1 and GI/G/n queues can be well approximated.

Approximating queueing networks relies on some specific assumptions. For example, in a Jackson network, where the external arrival processes are Poisson and the service times are exponentially distributed, the performance can be evaluated directly (Jackson 1957). When the service times at all servers are constant in a tandem queue, the mean queue time can also be established through the reduction method (Friedman 1965). Hulett and Damodaran (2011) also studied the performance of Markovian type manufacturing systems. However, for more general systems, the system performance is commonly analytically intractable due to the dependence among stations. In a tandem queue, although the long term arrival rate at each station is the same, the downstream stations commonly face a non-renewal process fed by the departures from upstream stations. To approximate the system performance, Whitt (1983) proposed QNA based on Kingman's heavy traffic approximation and the independence assumption (Kleinrock 1976), i.e., regarding each station as an independent queue, and estimating the variability of arrivals based on the variability equation (see equation (10) in Marshall (1968)). However, Whitt (1985) pointed out that QNA may produce bad estimations especially when the system variability of arrival processes is high or the variability of service times is low. Boxma (1979a, 1979b) also provided the situations in which the independence assumption is not valid. Another well-known approximation for the mean queue time is

QNET, which was developed by Harrison and Nguyen (1990) and Dai and Harrison (1992) based on the heavy traffic assumption. However, although the approach tends to provide reliable approximations when the system utilization is high with high variability, the algorithm can be time-consuming due to the slow convergence speed when the variability is low.

Figure 2-1 ASIA (left) and BSIA (right) systems (Wu and McGinnis 2013).

Inspired by the Jackson networks and the reduction method (Avi-Itzhak 1965, Friedman 1965, Jackson 1957), Wu and McGinnis (2013) found that the mean queue time of a tandem queue can be characterized by its corresponding All-See-the-Initial-Arrivals (ASIA) and Bottleneck-See-the-Initial-Arrivals (BSIA) systems as shown in Figure 2-1. In ASIA system, all stations are assumed to face the initial arrival process. This means that all stations face a renewal arrival process and hence the mean queue time can be approximated by the results of GI/G/1 queues. In BSIA system, only the bottleneck (i.e., the station with the largest utilization) sees the initial arrival process, and it is motivated by the fact that the mean queue

time of a tandem queue is dominated by the bottleneck in heavy traffic.

The Intrinsic Ratio (IR) is defined through comparing the actual queue time in the original tandem queue with the mean queue times in BSIA and ASIA systems. Based on the nice properties of IR (e.g., nearly linear), an interpolation method was proposed to approximate the mean queue time in single-server tandem queues. For example, the mean queue time at the second station (W_2) of a Simple Tandem Queue with Backend bottleneck (STQB) can be given by $y_2 W_2^{ASIA} - (1 - y_2) IG_2$, where y_2 is the IR of station 2, W_2^{ASIA} is the mean queue time of station 2 in ASIA system, W_2^{BSIA} is the mean queue time of station 2 in the BSIA system and $IG_2 = W_2^{ASIA} - W_2^{BSIA}$ is the Intrinsic Gap of station 2 (Wu and McGinnis 2013). Note that the mean queue times in ASIA and BSIA systems can be established efficiently based on the results of single-server queues since it's not necessary to consider the dependence among different stations in the two systems. Wu and McGinnis (2013) showed that the interpolation method can produce accurate estimates and outperforms the conventional approximation approaches.

2.3 Dependence in Queueing Systems and the Research Gap

As shown by Sections 2.1 and 2.2, the key difference between a single-server queue and a queueing system is the dependence. For example, while a GI/G/1 queue only faces an external arrival process, the arrival processes at the stations in multiclass queueing systems commonly depend on its upstream stations. The mechanism of dependence is complicated and there is still no unified approach to tackle the nut.

Although it is well known that the dependence has a significant impact, the previous studies commonly ignored it. For example, Dai (1995) used a set of fluid limit equations to describe the original queueing networks. However, the equations can only capture the job routings and service disciplines. Dependence among different stations cannot be captured by the fluid models and thus the approach is only capable to give the sufficient conditions for stability.

In the queue time approximations, the researchers commonly decomposed a complex queueing system into simpler ones. To utilize the approximation formulas for GI/G/1 queues, the arrival process at each station was regarded as a renewal process (Kleinrock 1976, Whitt 1983, Hopp et al. 2002). However, the arrival process at a downstream station can be non-renewal due to the dependence in a multiclass queueing network and ignoring the effect will induce large approximation errors (Wu et al. 2017a, Whitt 1985).

Therefore, instead of ignoring the dependence, we will study its impact explicitly to provide a more reliable framework for the analysis of queueing networks, which is indeed the main research gap of this thesis. Furthermore, besides general queueing systems, we will discuss the single-server queues first to reinforce the foundations of queueing theory.

Chapter 3 Fundamentals of Queueing Models

Fundamentals of queueing models will be introduced in this chapter. Queueing models and their underlying processes are discussed and they are the foundations for the studies in this thesis. The significance of stability and mean queue time is also clarified.

3.1 Different Types of Queueing Models

3.1.1 Single-Server Queues

Single-server queues are the simplest and most basic queueing models and they are elementary building blocks of other general queueing models. In general, a single-server queue is comprised of three key elements, i.e., the arrival process of jobs, the queue, and the servers. Other features, e.g., the behavior of jobs and the service disciplines, will further characterize the evolution of a queueing model (Adan and Resing 2002).

Figure 3-1 shows a GI/G/1 queue. Customers arrive at time epochs $\{A_n, n \geq 1\}$ and S_n is the service time of the n th customer. More than one customer arriving at the same epoch is not allowed. T_n is the interarrival time between n th and $(n + 1)$ th customer. $\{T_n, n \geq 1\}$ is a sequence of iid random variables with mean $E[T] \in (0, +\infty)$. Similarly, $\{S_n, n \geq 1\}$ is a sequence of iid random variables with mean $E[S] \in (0, +\infty)$. Arrival rate λ and service rate μ are $1/E[T]$ and $1/E[S]$ respectively. Furthermore, the two families of random variables, $\{S_n, n \geq 1\}$ and $\{T_n, n \geq 1\}$, are mutually independent. Let $U_n = S_n - T_n$, then $\{U_n, n \geq 1\}$ is also a sequence of iid random variables with mean $E[U]$. Their variances are assumed to be positive, i.e., $0 < Var[U] = \sigma^2$. The traffic intensity ρ

$\rho := \lambda/\mu$.

Figure 3-1 A GI/G/1 queueing model.

Note that if both S_n and T_n are exponentially distributed, the system will become an M/M/1 queue, which was first introduced in the original work of Erlang (1909). Due to the memoryless property, an M/M/1 queue is easier to analyze through the Markov chain analysis.

3.1.2 Multiclass Queueing Systems

In practice, a single-server queue might be too simple to describe a real-life situation (e.g., an automobile assembly line and a semiconductor fab). In a semiconductor fab, a production system need to server multiple types of products and each type of product may need to visit a station multiple times. Figure 3-2 shows a part of the process flows in a semiconductor wafer fab. In this situation, jobs may need to visit each station multiple times.

Figure 3-2 Part of the procedure in a semiconductor fab (Bramson 2008).

We consider a network composed of J single-server stations indexed by $j = 1, 2, \dots, J$ and the J stations serve I types of customers indexed by $i = 1, 2, \dots, I$. Type i customers have an exogenous arrival process with interarrival times $\{\xi_i(n), n \geq 1\}$. The route for each type of customer is deterministic and the visits of a type i customer are numbered from 1 to c_i . Kelly (2011) referred to customers of type i during visit m as class (i, m) jobs. In this chapter, we simply denote these job classes, i.e., $\{(1, 1), (1, 2), \dots, (1, c_1), (2, 1), (2, 2), \dots, (2, c_2), \dots, (I, c_I)\}$, as $\{1, 2, 3, \dots, K\}$ and denote $\tau(k) = i$ if class k job is of type i . By convention, the station serving class k is denoted as $\sigma(k)$ and two stations cannot share the same class. Without loss of generality, we assume there are multiple classes at each station, i.e., the cardinality of the set $\{k | \sigma(k) = j\}$ ($1 \leq j \leq J$) is greater than one. The service times required by class k jobs at station $\sigma(k)$ are $\{\eta_k(n), n \geq 1\}$. We assume that $\{\xi_i(n), n \geq 1\}$ ($i = 1, 2, \dots, I$) and $\{\eta_k(n), n \geq 1\}$ ($k =$

$1, 2, \dots, K$) are both iid sequences and mutually independent. When there is only one type of customers (i.e., $I = 1$), the network formulated above becomes a reentrant line (Kumar 1993).

Let $\alpha_i = E[\xi_i(n)] < \infty$ ($i = 1, 2, \dots, I$) and $m_k = E[\eta_k(n)] < \infty$ ($k = 1, 2, \dots, K$) be the mean interarrival time of type i customer and mean service time of class k job respectively. $\lambda_i = 1/\alpha_i$ is the exogenous arrival rate of type i customers and $\mu_k = 1/m_k$ is the service rate of class k jobs. The traffic intensity (utilization or nominal workload) at each station is given by:

$$\rho_j = \sum_{k: \sigma(k)=j} \lambda_{\tau(k)} m_k, j = 1, 2, \dots, J.$$

In this thesis, we always consider queueing networks under a given but arbitrary work-conserving service discipline, i.e., a server cannot be idle if the number of job in the system is not zero.

3.2 Underlying Processes of Queueing Models

Before being able to analyze a queueing model, we should first employ a stochastic process to describe the dynamics of the queue model. In the following, we will introduce different underlying processes and they can provide different viewpoints (e.g., queue time or queue length) on a queueing model.

3.2.1 Queue Time Process $\{W_n, n \geq 1\}$

The dynamics of a single-server can be depicted by the queue times of the consecutive jobs. Let W_n denote the queue time of the n th customer in a GI/G/1 queue (Lindley 1952).

Since the queue is initially empty, we have $W_1 = 0$. By the definitions of S_n and T_n , the following relation can be established:

$$W_{n+1} = [W_n + S_n - T_n]^+ = [W_n + U_n]^+, \text{ where } [*]^+ = \max\{0, *\}.$$

According to the above Lindley equation, W_{n+1} only depends on W_n and U_n . Therefore, the queue time process $\{W_n, n \geq 1\}$ is a Markov process with state space R^+ and one can exploit the Markov analysis techniques to analyze the queueing model.

However, the above approach through queue time is only applicable to single-server queues. For multiclass queueing networks, there is no simple Lindley equation and the relationship between the queue times of consecutive jobs is much more complex.

3.2.2 Queue Length Process $\{Q(t), t \geq 0\}$

Queueing models can also be described from the perspective of queue length. However, different from queue time, queue length cannot simply be modeled as a discrete time process. The evolution of queue length is generally described by an input-output process $\{Q(t), t \geq 0\}$:

$$Q(t) = Q(0) + A(t) - D(t),$$

where $Q(t)$ is the number of jobs in the queue, $A(t)$ is the number of total arrivals during $[0, t]$, $D(t)$ is the number of total departures before t and $Q(0)$ is the initial queue length (see Appendix A1 for details). Although $\{Q(t), t \geq 0\}$ is a stochastic process, it can be analyzed through the sample-path analysis (El-Taha and Stidham 2012). For a multiclass queueing network, $Q(t)$ can be a multi-dimension vector, with each element representing

the queue before each station.

3.2.3 Associated General Process $\{X(t) = (Q(t), U(t), V(t)), t \geq 0\}$

Markov chain has been showed to be a powerful tool for analyzing stochastic systems. For an $M/M/1$ queue, the queue length process $\{Q(t), t \geq 0\}$ is a Markov process with countable state space due to the memorylessness of exponential distribution. The distribution of queue length can be analyzed through a birth and death process (Ross 1996) and the process is positive recurrent if and only if $\rho < 1$. The invariant distribution of queue length can be calculated based on the equilibrium balance equations (Chen and Yao 2013).

However, for a $GI/G/1$ queue without the nice memoryless property or a multiclass class queueing networks, it is easy to verify that the queue length process is not a Markov process (see Appendix A1 for details). To achieve the Markov property, the states have to include two more coordinates. Following Dai (1995) or Bramson (2008), the state of a queueing system at time t can be described as $X(t) = (Q(t), U(t), V(t))$, where $Q(t)$ captures how customers are lined up at each station, $U(t)$ captures the remaining time before the next arrival from outside, $V(t)$ is the residual service time at each class and $X(0) = (Q(0), U(0), V(0))$ is the initial state at time zero. When a customer arrives or a job completes its service, the system will determine which job to serve next according to the service discipline (e.g. the FIFO discipline and FBFS (first-buffer-first-serve)). The structure of $X(t)$ may be different under different service disciplines as observed by Dai

(1995), but this has no influence on our further analysis.

Remark. Note that $Q(t)$, $U(t)$ and $V(t)$ can be multi-dimension vectors depending on the structure of the queueing system. For single server queues, $Q(t)$, $U(t)$ and $V(t)$ are scalars, i.e., $X(t) = (Q(t), U(t), V(t)) \in \mathbb{Z}_+ \times \mathbb{R}_+ \times \mathbb{R}_+$, where $Q(t)$ is the number of jobs in the queue at time t , $U(t)$ is the remaining interarrival time, and $V(t)$ is the residual service time for the job in service ($V(t) := 0$ if $Q(t) = 0$). $U(t)$ and $V(t)$ will keep decreasing as time goes on and if either of them hits zero, a jump occurs for $X(t)$.

The process defined above is known as a piecewise-deterministic Markov process (Dai 1995, Davis 1984). The associated general process $\{X(t) = (Q(t), U(t), V(t)), t \geq 0\}$ models the queues with higher fidelity than the queue time and queue length processes.

Theorem 3.1 The process $\{X(t) = (Q(t), U(t), V(t)), t \geq 0\}$ defined above is a Markov process.

In this section, we have introduced three stochastic processes which commonly exists in literature. Based on these underlying processes, queueing systems can be analyzed from differently perspectives.

3.3 The Significance of Stability

An important issue in the study of a queueing (or production) system is stability. Although there're various types of stability, they commonly mean that, in the long term, the system will converge to a steady equilibrium in some sense (El-Taha and Stidham 2012). In other words, the stability is a prerequisite for the study of some performance measures. It

guarantees that a queueing system has a stationary limiting behavior and hence makes the long-term performance indicators (e.g., mean queue time) meaningful. In practice, the stability also commonly links to the capacity and throughput rate of production systems. For example, Wu (2014) defined the capacity of production system as the maximum potential throughput rate that a system can provide given that the system is stable. In most occasions, the throughput rate of a stable production system equals the input rate (e.g., the job arrival rate) if there is no yield loss. Therefore, understanding the stability is important in both theory and practice. In next chapter, we will thoroughly study the stability of queueing systems, from single-server queues to multiclass queueing networks.

3.4 Performance Measures and Their Relations

As introduced in the above section, stability guarantees the existence of some performance measures.

3.4.1 Mean Queue Time

In production systems, the mean cycle (sojourn) time of jobs or products needed to be considered explicitly and there are often constraints on the mean cycle time (Hopp et al. 2002). In this thesis, we will focus on the mean queue time approximations. Note that studying the mean queue time, which is more common in queueing theory, is equivalent to the study of mean cycle time, since it is just the sum of the mean queue time and total processing time.

For a particular job, its queue time before a station is defined as $W_{job} = D_{job} - A_{job}$,

where A_{job} is the arrival time of the job, and D_{job} is the time it began to receive service. Then the mean queue time before this station will be $E[W] = \lim_{n \rightarrow \infty} W_n$, or equivalently $E[W] = \lim_{n \rightarrow \infty} \sum_{i=1}^n W_n / n$, where W_n is the queue time of the n th customer before the station. In other words, jobs arriving at the station need to wait for $E[W]$ in average before they can receive service. For a general queueing system, the total mean queue time will be the summation of the mean queue time before each station in the system.

From the perspective of stability, we may desire the throughput rate of a production system to be as high as possible to get faster return on the investment, especially for manufacturing systems with high equipment costs (e.g., a semiconductor wafer fab). Unfortunately, there's a trade-off between the throughput and cycle time. In a stochastic production system, a higher throughput rate (utilization) will lead to a longer cycle time (Braverman et al. 2017, Wu and McGinnis 2012). On one hand, the cycle time need to be controlled carefully to satisfy the customers' due dates. On the other hand, due to the perishability of products, a company in the high-tech industry generally needs to release its new product as early as possible to earn more market shares. Therefore, the mean cycle time of a production system is an important performance measure which needs to be understood clearly. Note that studying the mean queue time, which is a more common term in queueing theory, is equivalent to the study of mean cycle time in queueing systems, since it is just the sum of the mean queue time and total processing time.

3.4.2 Mean Queue Length

Besides the mean queue time, studying the mean queue length of a queueing system is also of great importance since it directly reflects the WIP (Work-in-Progress) levels. A system with large WIP bubbles is hard to control and most production policies (e.g., JIT) try to reduce the average WIP. Furthermore, a higher WIP level usually induces higher operations costs or inventory costs. However, we will not consider the mean queue length separately in this thesis since it can be evaluated directly based on the mean queue time and Little's law (Little 1961). According to Little's law, for any stable queueing system, there is

$$Q = \lambda W,$$

where Q is the mean queue length, λ is the arrival rate of jobs, and W is the mean queue time. Therefore, for a queueing system with a given arrival rate (i.e., λ is fixed), longer queue time also implies a longer queue and there's a linear relationship between them. In other words, the performance of mean queue time and mean queue length is the same in some sense.

3.5 System Architecture of the Thesis

The above sections have illustrated the basic elements of queueing models. In the following chapter, we will study the stability and mean queue time problems in detail.

As shown in Figure 3-3, stability is a prerequisite for the study of mean queue time. Hence, in Chapter 6 and Chapter 7, we always assume the queueing system is stable when discussing mean queue time. Before analyzing the general queueing systems, where the

dependence plays a crucial role, we first comprehensively study the single-servers to get a clearer picture for the stability and mean queue time problems. The insights and results of this thesis can be further developed to support the management of production systems.

Figure 3-3 System architecture of the studies.

Chapter 4 Stability of GI/G/1 Queues

Since the last century, queueing systems have been widely utilized to evaluate the performance of practical systems, e.g., production lines and communication networks. Stability, which directly impacts the system capacity and throughput (e.g., see Wu (2014)), plays a key role in the analyzing of queueing systems. For example, in the capacity planning for a semiconductor fab (Hopp et al. 2002), the stability condition must be satisfied to meet the throughput constraint. Furthermore, stability commonly implies that a system will have stationary limiting behavior, which means that some limiting performance measures (e.g., mean queue time) are meaningful only if they are defined on a stable system. Hence, understanding the stability of queueing systems is of crucial importance. In this chapter, we will focus on the GI/G/1 queues and aim at presenting an overall picture for the stability problem. The survey will support the applications of queueing models in practical situations (e.g., capacity planning) and also provide theoretical foundations for the stability of more complex queueing networks (Shen and Wu 2018).

4.1 Brief

Since GI/G/1 queues are basic building blocks of queueing networks and stability of queueing networks is vital to the capacity planning for manufacturing systems, thoroughly understanding the stability of GI/G/1 queues is the first step towards this goal. The types of stability of GI/G/1 queues are various according to their underlying processes. Unfortunately, although the stability of GI/G/1 queues has been a classic topic, there is no

summary to compare the various underlying processes and the corresponding stability defined on them. To provide an overall picture of this fundamental topic, we classify these different types of stability and clarify their relations in this chapter. Furthermore, the growth rate of the queue time W_n when the traffic intensity equals 1 is analyzed.

In this chapter, we study the single-server queueing model as introduced in Section 3.1.1. In the following, different types of stability and the stability conditions are studied in Section 4.2. The relations among these different types of stability are given in Section 4.3. Section 4.4 studies the limiting behavior of the queue time when the traffic intensity equals one. Section 4.5 summarizes this chapter.

4.2 Different Types of Stability

4.2.1 Proper Limiting Distribution of Queue Time $\{W_n, n \geq 1\}$

We will focus on the limiting behavior of $\{W_n, n \geq 1\}$ in this section. Based on Lindley (1952), since $W_{n+1} = [W_n + U_n]^+$, the evolution of queue times only depends on U_n . Because $\{U_n\}$ are i.i.d. random variables, the behavior of $\{W_n, n \geq 1\}$ is just that of a random walk with a barrier at the origin. The distribution function $F_n(x) = P(W_n \leq x)$ is the probability that W_n doesn't exceeds x and $F_n(x) = 0$ for any $x < 0$. Since the system is initially empty, we have $W_1 = 0$ and $F_1(x) \equiv 0$. By the relation $W_{n+1} = [W_n + U_n]^+$, for any $x \geq 0$,

$$\begin{aligned} F_{n+1}(x) &= P(W_{n+1} \leq x) = P(W_{n+1} = 0) + P(0 < W_{n+1} \leq x) \\ &= P(W_n + U_n \leq 0) + P(0 < W_n + U_n \leq x) = P(W_n + U_n \leq x) \end{aligned}$$

$$= \int_{W_n+U_n \leq x} dF_n(W_n) dG(U_n) = \int_{U_n \leq x} F_n(x - U_n) dG(U_n),$$

where $G(x)$ is the distribution function of any U_n . Although the expressions of $F_n(x)$ can be computed recursively based on the relation, the properties of $F_n(x)$ are not obvious.

Alternatively, because $W_1 = 0$, for $x \geq 0$, we have

$$F_2(x) = P(W_2 \leq x) = P([W_1 + U_1]^+ \leq x) = P(U_1 \leq x),$$

and

$$\begin{aligned} F_3(x) &= P(W_3 \leq x) = P([W_2 + U_2]^+ \leq x) = P([U_1]^+ + U_2 \leq x) \\ &= P(U_2 \leq x, U_1 + U_2 \leq x). \end{aligned}$$

Inductively, for any n ,

$$F_{n+1}(x) = P(U_n \leq x, U_n + U_{n-1} \leq x, \dots, U_n + U_{n-1} + \dots + U_2 + U_1 \leq x).$$

Since $\{U_n\}$ are i.i.d. random variables, after rearranging the order of U_n , we get:

$$F_{n+1}(x) = P(U_1 \leq x, U_1 + U_2 \leq x, \dots, U_1 + U_2 + \dots + U_{n-1} + U_n \leq x).$$

This implies that $F_n(x)$ is the probability of a series of decreasing events for any $x \geq 0$.

By the continuity of probability measure, the existence of $F(x) := \lim_{n \rightarrow \infty} F_n(x)$ is guaranteed.

Definition 4.1 A $GI/G/1$ queue is called stable if its queue time process $\{W_n, n \geq 1\}$ has a proper limiting distribution, i.e., $F(x)$ defined above is non-degenerate.

This stability means that the queue times of the in-coming customers converge to a steady distribution and finally attain equilibrium. Distributions of the queue times will remain consistent in the long run. As commented by Konheim (1975), Lindley (1952) and Loynes (1965), the solution of $F(x)$ is hard to get. However, investigating whether $F(x)$

is a proper distribution function or not is accessible. $F(x)$ can be represented as:

$$F(x) = P(\sum_{i=1}^n U_i \leq x, \forall n \geq 1).$$

By the strong law of large numbers (SLLN), $\sum_{i=1}^n U_i/n \rightarrow E[U]$ almost surely as $n \rightarrow \infty$. According to the above relation, properties of $F(x)$ are dominated by the expectation of the differences between service times and interarrival times.

Theorem 4.1 With respect to the queue time process $\{W_n, n \geq 1\}$, a GI/G/1 queue is stable if and only if $E[U] < 0$, i.e., $\rho < 1$. If $E[U] \geq 0$, $\lim_{n \rightarrow \infty} P(W_n < x) = 0$ for any x . Specifically, $W_n/n \rightarrow E[U]$ almost surely as n goes to infinity when $E[U] > 0$.

The above result is not surprising for $E[U] \neq 0$ by the SLLN. If $E[U] = 0$ and not all values assumed by U are integral multiples of a constant, Chung and Donsker (1951) showed that $P(|\sum_{i=1}^n U_i - x| < \varepsilon \text{ occurs infinitely often}) = 1$ for any x and $\varepsilon > 0$. Note that if U can only assume integral multiples of a constant, $P(|\sum_{i=1}^n U_i - x| < \varepsilon \text{ occurs infinitely often}) = 1$ still holds for x having those values (Lindley 1952). This implies that $F(x) \equiv 0$ for any x .

The queue time process $\{W_n, n \geq 1\}$ focuses on the evolution of individual customer's queue time and in this regime, a queue is called stable if there is a proper limiting distribution for W_n as n goes to infinity. The Lindley (queue time) process was also discussed in Borovkov (2012), Feller (2008) and Wolff (1989). Besides the individual customer's queue time, the queue length of a GI/G/1 queue is also of interest.

4.2.2 Pathwise Stability of Queue Length $\{Q(t), t \geq 0\}$

Definition 4.2 A GI/G/1 queue is pathwise stable if $Q(t)/t \rightarrow 0$ as t goes to infinity almost surely.

Remark. Pathwise stability is the stochastic version of rate stability (see Appendix A1), which is defined on a deterministic input-output process (El-Taha and Stidham 1992, El-Taha and Stidham 2012, El-Taha and Stidham 1993). Mazumdar et al. (1992) and Altman et al. (1998) first studied the pathwise behavior of queues. Chen (1995) and Dai and Vande Vate (1996) studied the pathwise stability of queueing networks.

This definition means that a queue is pathwise stable if the growth rate of $Q(t)$ is $o(t)$ almost surely. However, note that $\lim_{t \rightarrow \infty} Q(t)/t = 0$ only claims the growth of $Q(t)$ is slower than linear. In some cases, although the queue length process is pathwise stable, $Q(t)$ still diverges to infinity with a positive probability.

For GI/G/1 queues subjected to the settings in Section 3.1.1, we have the following theorem.

Theorem 4.2 A GI/G/1 queue is pathwise stable if and only if $\lim_{t \rightarrow \infty} D(t)/t = \alpha$, where α is the arrival rate.

Hence, a pathwise stable GI/G/1 queue reaches an equilibrium in the sense that the long-run arrival rate equals the long-run throughput rate. The following theorem clarifies the relation between the pathwise stability and traffic intensity.

Theorem 4.3 The queue length process $\{Q(t), t \geq 0\}$ of a GI/G/1 queue is pathwise stable if and only if $E[U] \leq 0$, i.e., $\rho \leq 1$. And if $E[U] > 0$, $\lim_{t \rightarrow \infty} Q(t)/t = \alpha - \beta$ almost

surely.

Proof of Theorem 4.3. When $E[U] > 0$, $\lim_{t \rightarrow \infty} Q(t)/t = (\alpha - \beta) > 0$ is a direct consequence of the SLLN. Hence, it suffices to show if $E[U] \leq 0$, the queue length process is pathwise stable. According to Corollary 3.2 in Dai and Vande Vate (1996), pathwise stability of the queue length process can be implied by the weak stability of the corresponding fluid model, which is defined as:

$$Q(t) = Q(0) + \alpha t - \beta T(t),$$

$$I(t) = t - T(t),$$

$$Q(t) \geq 0,$$

$$T(0) = 0,$$

$$I(0) = 0, I(\cdot) \text{ is non-decreasing, and}$$

$$\dot{I}(t) = 0, \text{ when } Q(t) > 0 \text{ and } I(t) \text{ is differentiable at } t.$$

The fluid model is said to be weakly stable if the solution to the above equations is unique with $Q(t) = 0$ for any $t \geq 0$ when $Q(0) = 0$. Based on Theorem 3.1 in Chen (1995), $\rho \leq 1$ implies the weak stability of the above fluid model. Q.E.D.

4.2.3 Strong and Weak Stability of Queue Length $\{Q(t), t \geq 0\}$

Besides pathwise stability, weak stability and strong stability of the queue length process are also interesting topics, and were introduced by Courcoubetis et al. (1989) and Courcoubetis and Weber (1994).

Definition 4.3 The queue length process $\{Q(t), t \geq 0\}$ is weakly stable if

$$P(Q(t) \rightarrow \infty \text{ as } t \rightarrow \infty) = 0;$$

and is strongly stable if

$$\sup_{t \geq 0} E[Q(t)] < \infty.$$

Carr and Hajek (1993) gave an equivalent definition for the weak stability of the queue length process.

Definition 4.4 The queue length process $\{Q(t), t \geq 0\}$ is weakly stable if the family of distribution functions of $Q(t)$ is tight:

$$\forall \epsilon > 0, \exists c > 0, \text{ s.t. } P(Q(t) > c) \leq \epsilon \text{ for any } t \geq 0.$$

However, a process can behave erratically under weak stability. For example, consider the following artificial process:

$$Q(t) = \lfloor t + 1 \rfloor t - \lfloor t \rfloor \lfloor t + 1 \rfloor, \quad t \geq 0,$$

where $\lfloor * \rfloor$ is the lower round function, i.e., $\lfloor t \rfloor = \max_{x \leq t, x \in \mathbb{Z}} x$. Obviously, $\lim_{t \rightarrow \infty} Q(t) \neq \infty$, since $Q(t) = 0$ for all integers. Hence, it is weakly stable. However, $Q(t)$ will oscillate with the amplitude of the oscillation goes to infinity.

Strong stability guarantees that the supremum of $E[Q(t)]$ is bounded and it is harder to achieve, even a classic M/G/1 queue can violate the condition with the traffic intensity less than one. It depends on not only the first moments of variables but also the variances.

Theorem 4.4 The queue length process $\{Q(t), t \geq 0\}$ is weakly stable if $\rho < 1$. If the process $\{Q(t), t \geq 0\}$ is strongly stable, then $\rho < 1$.

This theorem means that the usual traffic condition (i.e., $\rho < 1$) is sufficient for weak stability but only necessary for strong stability.

Although there are three types of stability with respect to the queue length process, all of them only focus on the bound of $Q(t)$ as time t goes to infinity. The approach through the input-output process is not capable to capture the property of $Q(t)$ with respect to the state space, e.g., the transitions between states and the recurrence of states.

4.2.4 Positive Harris Recurrence of Associated General Process $\{X(t), t \geq 0\}$

Although the process $\{X(t), t \geq 0\}$ becomes a Markov process after including two residual time processes, the state space becomes uncountable and the analysis of recurrence is harder. Stability of a Markov process with uncountable state space is commonly studied through the positive Harris recurrence. Similar to positive recurrence for a Markov process with countable state space, positive Harris recurrence guarantees that the states of the process will be in any proper set for infinitely many times during the evolution. All proper sets are accessible and the chain will finally converge to an invariant probability distribution and admit equilibrium (Meyn and Tweedie 2012). In many studies, a queueing network is called stable if its associated general process is positive Harris recurrent (Chen 1995, Dai 1995). As a direct corollary of Theorem 4.1 in Dai (1995) or Theorem 5.7 in Bramson (2008), the following theorem gives the stability condition of a GI/G/1 queue.

Theorem 4.5 A GI/G/1 queue is positive Harris recurrent if and only if $E[U] < 0$, i.e., $\rho < 1$.

Remark. As in Dai (1995), the above theorem holds only if the interarrival times are unbounded and spread out. It means that there exists some positive integer N and some

function $p(x) > 0$ on R_+ with $\int_0^\infty p(x) dx > 0$, such that $P(T \geq x)$ for any $x > 0$ and $P(a \leq \sum_{i=1}^N T_n \leq b) \geq \int_a^b p(x) dx$ for any $0 \leq a < b$.

Since $Q(t)$ will diverge to infinity almost surely when $\rho > 1$, there is $P_x(\tau(A) < \infty) < 1$ for any bounded set $A \in \mathcal{B}(X)$ and $x \notin A$. Therefore, the process $\{X(t) = (Q(t), U(t), V(t)), t \geq 0\}$ is not positive Harris recurrent if $\rho > 1$. When $\rho = 1$, although $P_x(\tau(A) < \infty) = 1$ for any $A \in \mathcal{B}(X)$ and the process is Harris recurrent, the invariant measure π is not finite.

4.3 Relations among Different Types of Stability

The above sections have introduced the stability of GI/G/1 queues from different perspectives. Although the definitions of these types stability are different, all of them imply some kinds of stationary limiting behavior. Furthermore, as shown by the above theorems, stability of GI/G/1 queues highly depends on the traffic intensity of the server. Table 4-1 gives the relations between them and the traffic intensity, where $\checkmark$ means the traffic intensity condition guarantees the stability, $\times$ means the traffic intensity condition leads to instability and $?$ means the relation depends on other factors.

As shown in Table 4-1, all types of stability are violated when $\rho > 1$, since the WIP will diverge to infinity almost surely according to the SLLN. When $\rho < 1$, all types of stability are guaranteed except for the strong stability. The case $\rho = 1$ is interesting and only pathwise stability is guaranteed. Although the service rate equals the arrival rate, the queue will tend to some sort of explosion because of the randomness. In this case, both the

mean queue length $E[Q]$ and mean queue time $E[W]$ are infinite. However, if both service times and interarrival times are constant and equal ($E[U] = 0, Var[U] = 0$), both $E[Q]$ and $E[W]$ will become finite and there exists a proper limiting distribution for $\{W_n, n \geq 1\}$. This trivial condition is sometimes overlooked (e.g., see Asmussen (2008)).

Table 4-1 Relations between the traffic intensity and different types of stability of a GI/G/1.

queue.

Different types of stability	$\rho < 1$	$\rho = 1$	$\rho > 1$
Proper limiting distribution of queue time $\{W_n, n \geq 1\}$	$\checkmark$	$\times$	$\times$
Positive Harris recurrence of $\{X(t), t \geq 0\}$	$\checkmark^*$	$\times$	$\times$
Pathwise stability of queue length $\{Q(t), t \geq 0\}$	$\checkmark$	$\checkmark$	$\times$
Weak stability of queue length $\{Q(t), t \geq 0\}$	$\checkmark$	$?$	$\times$
Strong stability of queue length $\{Q(t), t \geq 0\}$	$?$	$\times$	$\times$

*Positive Harris recurrence only holds if the interarrival times are unbounded and spread out.

Remark. When $\rho < 1$, a single-server queue may not be strongly stable. For example, if the variance of the service times of a M/G/1 queue is infinite (i.e., $Var[S] = \infty$), then the mean queue length will be infinite (Pollaczek 1930a, 1930b). This means $\lim_{t \rightarrow \infty} E[Q(t)] = \infty$ and the queue is not strongly stable. Similarly, when $\rho = 1$, the queue length process $\{Q(t), t \geq 0\}$ can diverge to infinity with a positive probability and the queue is not weakly stable.

Figure 4-1 The relations among different types of stability.

Figure 4-1 further summarizes the relations among these different types of stability, where “ $A \rightarrow B$ ” means A can imply B and “ $A \leftrightarrow B$ ” means A and B can imply each other.

4.4 The Growth Rate of the Queue Time W_n when $\rho = 1$

As shown in Table 4-1, $\rho = 1$ is a critical case which can lead to adverse results with respect to different types of stability. According to the analysis in Section 4.2.1, we know that when $\rho = 1$ the queue time W_n has no proper limiting distribution. The following theorem demonstrates the way and speed in which W_n explodes.

Theorem 4.6 If $\rho = 1$, $0 < \sigma^2 < \infty$, the process $\{W_n/\sqrt{n}, n \geq 1\}$ converges in distribution to the absolute value of a normal random variable with mean 0 and variance σ^2 :

$$W_n/\sqrt{n} \xrightarrow{d} |Z| \text{ as } n \rightarrow \infty, \text{ where } Z \sim N(0, \sigma^2).$$

The proof of the theorem can be found in Billingsley (2013) through applying Donsker’s

theorem. The convergence rate is $O(\sqrt{n \log \log n})$ (Chen and Yao 2013). Apparently, $E[W_n]$ diverges to infinity with rate $\sqrt{n}$ as n goes to infinity. This coincides with the fact that the process is not strongly stable if $\rho = 1$. The value of W_n grows much faster than $\sqrt{n}$ in the case of $\rho > 1$. By Asmussen (2008), $W_n/n \xrightarrow{a.s.} E[T](\rho - 1)$ as n goes to infinity, which states that W_n will explode linearly. When $\rho < 1$, the limit of $E[W_n]$ is bounded since W_n has a proper limiting distribution and $\sigma^2 < \infty$. These results are summarized in Table 4-2.

Table 4-2 Growth rate of $E[W_n]$.

Traffic intensity	$\rho < 1$	$\rho = 1$	$\rho > 1$
Growth rate of $E[W_n]$	$O(1)$	$O(\sqrt{n})$	$O(n)$

Theorem 4.6 only gives the growth rate of the expectation of W_n . The following theorem generalizes the prior results.

Theorem 4.7 If $\rho = 1$, $0 < \sigma^2 < \infty$, then for any $\epsilon > 0$,

$W_n/(n^\epsilon \sqrt{n})$ converges to 0 in probability as n goes to infinity.

Proof of Theorem 4.7. By the property of convergence in distribution, we have:

$$\lim_{n \rightarrow \infty} E[W_n/\sqrt{n}] = E[|Z|].$$

Since $E[|Z|]$ is a constant and is less than infinity, for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} E[W_n/(n^\epsilon \sqrt{n})] = \lim_{n \rightarrow \infty} E[|Z|]/n^\epsilon = 0.$$

By Markov inequality, for any $c > 0$,

$$P([W_n/(n^\epsilon \sqrt{n})] > c) \leq E[W_n/(n^\epsilon \sqrt{n})]/c,$$

which implies

$$\lim_{n \rightarrow \infty} P([W_n/(n^\epsilon \sqrt{n})] > c) = \lim_{n \rightarrow \infty} E[W_n/(n^\epsilon \sqrt{n})]/c = 0.$$

Since both c and ϵ are arbitrary positive numbers, this completes our proof. Q.E.D.

Theorem 4.7 gives an upper bound for the growth rate of queue times while Theorem 2.7 only gives the limiting distribution of $W_n/\sqrt{n}$ and the growth rate of $E[W_n]$. The result $W_n/\sqrt{n} \xrightarrow{d} |Z|$ implies that the growth rate of $E[W_n]$ is $O(\sqrt{n})$, which is only an expectation property but not a sample-path property. In this section, we have shown that the growth rate of W_n is less than $O(n^\epsilon \sqrt{n})$ for any $\epsilon > 0$, which is a sample-path property.

4.5 Summary

In this survey, we have summarized three underlying processes of a GI/G/1 queue, which describe the dynamics of queueing systems from different perspectives. The queue time process $\{W_n, n \geq 1\}$ and the queue length process $\{Q(t), t \geq 0\}$ focus on the two important measures, queue time and queue length, of a queueing system respectively. The associated general process contains more complete information and can be easily generalized to complex queueing networks. The different types of stability are essentially describing different types of limiting behavior of the underlying processes. For example, Section 4.2.1 considers the limiting distribution and Section 4.2.2 focuses on the growth rate. Although details of the different types of stability are different, all of them mean that, in the long term, the system can reach equilibrium in some sense.

In the case of $\rho = 1$, the limiting behavior of W_n is interesting. $W_n/\sqrt{n}$ converges

in distribution to the absolute value of a normal random variable. The author strengthens this result and shows that $W_n/(n^\epsilon\sqrt{n})$ converges to 0 in probability when n goes to infinity for any $\epsilon > 0$. This result asserts that the growth rate of a sample path is bounded by $O(n^\epsilon\sqrt{n})$ from above for an arbitrary ϵ . Whether the result cannot be strengthened to almost sure convergence is left for future research.

This survey has shown the differences among various types of stability and summarized the relations among different stabilities and the traffic intensity. It can help us understand the stability of GI/G/1 queues more thoroughly. Based on the relations among different types of stability, researchers and practitioners can also refer to this study to design and evaluate more complex queueing systems. Furthermore, this chapter has presented an overview for the stability of GI/G/1 queues, which is of fundamental importance for the study of general queueing networks.

Chapter 5 Stability of Multiclass Queueing Networks

In Chapter 4, we have studied the stability of GI/G/1 queues in detail and one may expect to extend the results to more general queueing networks. However, different from single-server queues, multiclass queueing networks have more complex state spaces and the transitions of states are subject to more factors (e.g., service disciplines). Furthermore, while the stability conditions of GI/G/1 queues can be given through the utilizations of servers, “if and only if” stability conditions for general queueing networks are still unknown. It has been shown that besides the utilizations of stations, other factors such as job routings and service disciplines also play a crucial role in queueing networks (Lu and Kumar 1991). In addition, the conventional approach through fluid models is only capable to provide sufficient conditions and there is a gap between fluid models and queueing networks (Bramson 1999). To tackle the nuts, in this chapter, we will study the stability of queueing network from the viewpoint of “servers” inspired by the results of GI/G/1 queues. We redefine the “servers” in the context of queueing networks, which can capture the dependence among stations. It turns out that, similar to GI/G/1 queues, the stability conditions can also be fully characterized by the utilizations of the general servers under this new setting.

5.1 Brief

Utilizing the installed capacity effectively is essential to the productivity improvement. However, if a system can be unstable when the utilization is less than one, capacity

decreases and the bottleneck is hard to locate. As a consequence, the system is difficult to manage and performance degrades. Furthermore, to determine the capacity or service rate of a queueing system, stability of the system must be established first. Since service rate is a basic input in a queueing model, stability plays a key role in analyzing queueing systems. Understanding the stability condition is essential in both practice and theory.

In this chapter, we propose the necessary and sufficient conditions for the pathwise stability of queueing networks inspired by the following observation. A station is said to have sufficient capacity if its service rate is as large as the arrival rate, or $\lim_{t \rightarrow \infty} Q_i(t)/t = 0$, where $Q_i(t)$ is the total number of jobs in the buffer of station i at time t . However, it is observed that a system may not have sufficient capacity, or $\sum_{i=1}^N Q_i(t)/t \not\rightarrow 0$ as $t \rightarrow \infty$, even if each station has (Banks and Dai 1997, Dai and Vande Vate 1996, Down and Meyn 1994). This means that the system capacity or service rate may decrease due to the dependence among stations. Specifically, Dai and Vande Vate (1996) showed the existence of virtual stations in a two-station queueing network, which influence the system stability as well as physical stations. Although the two classes from a virtual station may come from different physical stations, they cannot receive service simultaneously due to the negative correlation under static buffer disciplines. The observation forces us to rethink the meaning of server and capacity in the context of queueing networks: How should we define a server in a queueing network when there is dependence among stations? Is it enough to define stations simply based on their physical configurations? And, how to determine the capacity when there is dependence among stations? To study the stability of a queueing network, we

must clarify these fundamentals first.

In the following, we first introduce the *mutual blocking* phenomenon, which depicts the negative correlation among different classes in queueing networks. The concept of *general servers* is proposed based on the sample path properties. We find that general servers have similar impacts on system performance as physical stations and determining the stability of a multiclass queueing network under a given service discipline is equivalent to figuring out the traffic intensity of all general servers. However, one needs to verify numerous combinations of class sets to identify general servers, of which the complexity is NP. This implies that determining the stability of a queueing network in general settings could be essentially an NP problem, which is consistent with the case of pseudostations.

Although the identification of general servers is hard, the intuitive concept provides a new viewpoint and a unified framework to comprehend the properties of queueing network stability. The study implies that, in the context of queueing networks, we shouldn't overlook the dependence among classes and the structure of general servers should be considered explicitly. The demonstration for the properties of stability through the concept of general servers is straightforward and intuitive. Due to the high complexity of examining the sample paths, it is not practical to determine the stability of general queueing networks through identifying all general servers. However, employing the concept, it's capable to show that practical systems operating under *WIP-dependent* dispatching disciplines are always pathwise stable if every physical station has sufficient capacity.

In the following, Section 5.2 formulates the stability problem. Definitions and new

concepts are proposed in Section 5.3. Case studies are given in Section 5.4. Section 5.5 presents the stability conditions. Applications are given in Section 5.6. Section 5.7 compares the proposed new approach with the fluid model approach and Section 5.8 concludes this chapter. Proofs of the lemmas and theorems are given in the Appendix A2.

5.2 Problem Formulation

In this chapter, we study the multiclass queueing networks as introduced in Section 3.1.3 and the system dynamics are described by the associated general process as shown in Section 3.2.3.

The state of a multiclass queueing network at time t is $X(t) = (Q(t), U(t), V(t))$ and $|Q(t)|$ denotes the total number of jobs in the system. Let $R_k(t)$ denote the residual workload of class k jobs at time t and $\dot{R}_k(t)$ be its derivative.

The pathwise stability of queueing networks with infinite buffer size under work-conserving policies is defined as follows (Altman et al. 1998, Dai and Vande Vate 1996, El-Taha and Stidham 2012).

Definition 5.1 (Pathwise stability) A queueing system operating under a given service discipline is said to be pathwise stable (or stable), if $|Q(t)|/t \rightarrow 0$ as $t \rightarrow \infty$ almost surely.

Remark. This definition is the same as Definition 4.2 except that we use $|Q(t)|$ to denote the total number of jobs, i.e., the summation of all coordinates.

In the following, the term “stable” always means the pathwise stability unless

otherwise stated. Since for each class k , $Q_k(t) = Q_k(0) + A_k(t) - D_k(t)$ ($t \geq 0$), where $A_k(t)$ is the total number of arrivals in $[0, t]$ and $D_k(t)$ is the total number of departures in $[0, t]$, it is evident that pathwise stability implies the equivalence of the long run arrival (input) rate (i.e., $\lim_{t \rightarrow \infty} A_k(t)/t$) and departure (throughput) rate (i.e., $\lim_{t \rightarrow \infty} D_k(t)/t$).

5.3 General Servers

In the seminal work of Erlang (1909), he studied the performance of a telecom system through modeling it as a Markov chain based on the Poisson arrival process assumption. Due to the Poisson process assumption, the state of the queueing system is much simpler and it makes it possible to analytically study the system performance (e.g., mean queue length). Hence, Poisson process has been one of the most common assumptions in congestion theory.

The basic elements of an $M/M/1$ queue are the arrival rate λ and service rate (or capacity) μ as shown in Figure 5-1. The capacity μ is *the potential maximum throughput rate that a system could achieve in steady state* (Wu 2014), i.e., the system would be unstable if $\lambda > \mu$. And a server is defined as *a valuable resource with limited capacity*. Note that in a Markov chain, a server is defined based on the concept of capacity, but not vice versa. Furthermore, there is no implication that a server must be a physical station, although a physical station always has a limited capacity.

Figure 5-1 An M/M/1 queue.

In a queueing network, although the service rate of each class is known, the capacity of a class set is hard to establish when dependence exists among classes. When a class set $\mathcal{C}$ belongs to the same physical station, it is trivial that the classes in $\mathcal{C}$ cannot receive service simultaneously. When the classes belong to different stations, they should be able to receive service simultaneously. However, sometimes they cannot process jobs simultaneously due to the negative correlation among classes. An interesting example in literature is *mutual blocking* (Dai and Vande Vate 1996, Kumar and Seidman 1990, Lu and Kumar 1991), where between the two classes of a class set, when one class is working, the other tends to be idle. The two classes mutually block each other although they belong to different physical stations.

Definition 5.2 (Mutual blocking between two classes) Job classes k and l suffer mutual blocking if the following two conditions are satisfied almost surely:

- (a) $\lim_{t \rightarrow \infty} m(\{t | \dot{R}_k(t)\dot{R}_l(t) \neq 0\}) / m(\{t | \dot{R}_k(t) + \dot{R}_l(t) \neq 0\}) = 0$, and
- (b) If $(m_k/\alpha_{\tau(k)} + m_l/\alpha_{\tau(l)}) \leq 1$, $\lim_{t \rightarrow \infty} m(\{t | Q_k(t) + Q_l(t) \geq 2\})/t > 0$.

Remark. $m(\cdot)$ represents the Lebesgue measure on the measurable space $(\Omega, \mathcal{B})$, where Ω is the set of real numbers and $\mathcal{B}$ is the set of Borel sets. Note that the residual workload $R_k(t)$ is a decreasing function except for the time epoch at which a class k job arrives. Since such time epochs are countable, $\dot{R}_k(t)$ exists almost everywhere (e.g., see Royden and Fitzpatrick (1988)).

Figure 5-2 A two-stage tandem queue.

Mutual blocking reduces system capacity. Intuitively, mutual blocking means that classes k and l cannot receive service simultaneously except for a transient period. Condition (a) states that the busy periods of classes k and l are disjoint (except for a transient period). Condition (b) guarantees that condition (a) is not caused by the periodic shifting of the busy periods. In deterministic settings, when two classes have sufficient capacity, their busy periods may never intersect due to the periodic behaviour. For example, in the following tandem queue (see Figure 5-2) with $\lambda_1 = 1$, $\mu_1 = 1/0.9$, and $\mu_2 = 1/0.1$, if the interarrival and service times are constant, classes 1 and 2 cannot receive service simultaneously and condition (a) is satisfied. However, their busy periods alternate periodically and there is no mutual blocking between them in this case.

As long as mutual blocking exists between two classes, at most one class can receive service at any time except for a transient period. As observed by Hasenbein (1997), mutual blocking may also exist among multiple classes. In the six-class network shown in Figure 5-3, with classes 2, 4 and 6 having higher priority, at most two classes of class set $\{2, 4, 6\}$ may receive service at any time.

Figure 5-3 A six-class queueing network.

Definition 5.3 (Mutual blocking among multiple classes) A class set $\mathcal{C}$ of K classes suffers mutual blocking if the following two conditions are satisfied almost surely:

- (a) $\lim_{t \rightarrow \infty} m(\{t | \prod_{k \in \mathcal{C}} \dot{R}_k(t) \neq 0\}) / m(\{t | \sum_{k \in \mathcal{C}} \dot{R}_k(t) \neq 0\}) = 0$, and
- (b) If $\sum_{k \in \mathcal{C}} m_k / \alpha_{\tau(k)} \leq 1$, then $\lim_{t \rightarrow \infty} m(\{t | \sum_{k \in \mathcal{C}} Q_k(t) \geq |\mathcal{C}|\}) / t > 0$.

Remark. $|\mathcal{C}|$ is the cardinality of class set $\mathcal{C}$. Condition (b) makes the definition of mutual blocking complete. However, it has no impact on the stability condition since the classes will always have sufficient capacity in this case.

In this situation, the number of classes that can receive service simultaneously can be any number within the interval $[1, |\mathcal{C}| - 1]$. Similar to physical stations, any class set may have its specific capacity due to the mutual blocking and have an impact on the system stability.

Definition 5.4 (General server) For a queueing system operating under a given service discipline, a class set $\mathcal{C}$ is a general server if it suffers mutual blocking.

Corollary 5.1 If a class set $\mathcal{C}$ is a general server, then $\mathcal{C}'$ is also a general server if $\mathcal{C} \subseteq \mathcal{C}'$.

Physical stations, virtual stations (Dai and Vande Vate 1996) and pseudostations

(Hasenbein 1997) are special types of general servers. Due to the dependence among classes, any combination of multiple job classes in a queueing network may be a general server. When a queueing network consists of J physical stations, the number of general servers can increase exponentially with respect to $\sum_{j=1}^J K_j$, where K_j is the number of classes at physical station j . A general server is *intangible* if it is not a physical station.

Definition 5.5 (Effective number of classes) For a class set $\mathcal{C}$, its effective number of classes is M (denoted as $EF(S) = M$) if the following two conditions are satisfied almost surely:

- (a) If $F \subseteq \mathcal{C}$ and $|F| \geq M + 1$, then there is $\lim_{t \rightarrow \infty} m(\{t | \prod_{k \in F} \dot{R}_k(t) \neq 0\}) / m(\{t | \sum_{k \in \mathcal{C}} \dot{R}_k(t) \neq 0\}) = 0$, and
- (b) There exists $F \subseteq \mathcal{C}$ and $|F| = M$, s.t. $\lim_{t \rightarrow \infty} m(\{t | \prod_{k \in F} \dot{R}_k(t) \neq 0\}) / m(\{t | \sum_{k \in \mathcal{C}} \dot{R}_k(t) \neq 0\}) > 0$.

The effective number of classes measures the ability of a class set $\mathcal{C}$ to work as $|\mathcal{C}|$ standalone stations and is M if at most M classes of the class set can receive service simultaneously. For example, in the above six-class network, $EF(\{2, 4, 6\}) = 2$ and $EF(\{1, 4\}) = 1$.

According to Corollary 5.1, any class set containing a physical station is a general server. Hence, the number of general servers may be extremely large and one can find out many odd servers, such as $\{1, 2, 4\}$, $\{2, 5, 6\}$ and $\{3, 4, 6\}$ in the six-class network. The definition of general server is inefficient to capture the mutual blocking and studying these servers may be redundant since they are just subsets of a larger server with the same

effective number.

Definition 5.6 (Compact server) *A general server S with M effective classes is compact if the following two conditions are satisfied:*

- (a) For any $S' \supsetneq S$, $EF(S') > M$, and
- (b) For any decomposition $\{S_i\}_{i \in F}$ of S , i.e., $\bigcup_{i \in F} S_i = S$ and $S_i \cap S_j = \emptyset$ for any $i \neq j$, $\sum_{i \in F} EF(S_i) > M$.

For a compact server with M effective classes, one cannot find a larger set which still has M effective classes. Compact servers capture the mutual blocking within queueing systems with high fidelity. According to the definition, all physical stations are compact servers with $M = 1$.

Similar to physical stations, the *effective* load of each general server S is given by:

$$L_S = \sum_{k \in S} \lambda_{\tau(k)} m_k,$$

and the *effective* traffic intensity (utilization) of S is given by:

$$P_S = L_S / M,$$

where $M = EF(S)$ is the effective number of general server S . Obviously, for any physical station j , $P_j = \sum_{k: \sigma(k)=j} \lambda_{\tau(k)} m_k / 1 = \rho_j$, since the effective number of each physical station is one.

In queueing networks, due to the dependence among stations, job classes can suffer mutual blocking, which induces general servers and reduces system capacity. When addressing stability, we need to consider these general servers as well as physical stations.

While the capacity of a physical station is deterministic, the capacity of a general server is

sensitive to the system configurations. In the following, case studies are given to investigate the properties of general servers and stability.

5.4 Case Studies

5.4.1 The Lu-Kumar Network with Constant Interarrival and Service Times

To investigate the properties of general servers in queueing networks, we begin by studying a tandem queue consisting of four single-server stations (and four job classes) as shown in Figure 5-4.

We consider a production line in which jobs come at a constant rate with constant service times. The production line has four stations. Stations 2 and 3 are operated by the same worker. Stations 1 and 4 share one power supply. Due to the high energy consumption, stations 1 and 4 cannot be operated at the same time. When jobs are available at both stations 2 and 3, the worker gives priority to the jobs at station 2. When jobs are available at both stations 1 and 4, the power supply is allocated to station 4 after the current job (if any) is completed. The service policy is FIFO at each station with (non-preemptive) priority to station 2 over station 3 and to station 4 over station 1. Let $\alpha \in R^+$ denote the constant interarrival time at the first station. Let the constant service times m_k from the first to the last stations be 0.2, 0.6, 0.1 and 0.6, respectively.

Figure 5-4 A tandem queue with four single-server stations.

The first question is how many general servers exist in this system? In addition to the four physical stations, due to the above two constraints on capacity, at least two other servers (i.e., $\{1, 4\}$ and $\{2, 3\}$) need to be considered explicitly.

Indeed, in terms of system capacity and sojourn time, the tandem queue in Figure 5-4 is identical to the well-known Lu-Kumar network (Lu and Kumar 1991) as shown in Figure 3-5, where class 2 has priority over class 3 and class 4 has priority over class 1. The servers $\{1, 4\}$ and $\{2, 3\}$ in Figure 5-4 become physical stations A and B in Figure 5-5, and the four physical stations in Figure 5-4 correspond to the four different job classes in Figure 5-5.

Figure 5-5 The Lu-Kumar network.

Although there are four stations in Figure 5-4, there are only two stations in the Lu-Kumar network in Figure 5-5. When analyzing a queueing system, the focus should not be the physical stations in the system. Instead, we have to pay attention to all potential servers which play a role in determining system performance. Since servers $\{1, 4\}$ and $\{2, 3\}$ are not real stations in Figure 5-4 but suffer mutual blocking due to the dependence among job classes, they are general servers.

Since the capacity of station A in the Lu-Kumar network is $1/0.8$, the network is

unstable if the interarrival time α is smaller than 0.8. However, due to the static buffer priority discipline, after some simple bookkeeping, one can find that the Lu-Kumar network is also unstable when $\alpha = 1$. WIP in front of the class set $\{2, 4\}$ keeps increasing and oscillates between classes 2 and 4 as shown in Figure 5-6.

Figure 5-6 WIP at job classes 2 and 4 when $\alpha = 1$.

When $\alpha = 1$, WIP oscillates between job classes 2 and 4 because they will mutually block each other. Since class 4 has priority over class 1, when jobs are processed at class 4 new incoming jobs will be blocked at class 1 and class 2 will starve after its on-hand jobs are completed. Similarly, when jobs are processed by class 2, since class 2 has priority over class 3, jobs will be blocked at class 3 and class 4 starves. In this case, the class set $\{2, 4\}$ is a general server and its capacity is $1/1.2$. Therefore, the network is unstable.

In the Lu-Kumar network, small gaps between the busy periods may occur as shown in Figure 5-6. This is because after class 2 serves all jobs, class 4 can only start to serve jobs after a small delay of m_3 . Similarly, after class 4 serves all jobs, class 3 receives jobs after a small delay of m_1 . However, as shown in Figure 5-6, because the lengths of the busy

periods diverge to infinity while the gaps are at most 0.2, the ratio goes to zero with no impact on capacity. Note that these overlaps disappear under the corresponding preemptive-resume policy.

Due to mutual blocking, the network can be unstable even if $\alpha > 0.8$ (e.g. when $\alpha = 1$) as shown in Figure 5-6. Since classes 2 and 4 tend not to process jobs simultaneously, one would guess that this network is unstable if $\alpha < 1.2$. An interesting observation is that although the Lu-Kumar network is unstable when $\alpha = 1$, the network can be stable when the interarrival time becomes smaller than 1.

Lemma 5.1 (Stability region of the Lu-Kumar network) When $0.8 \leq \alpha < 0.9$, the Lu-Kumar network with constant arrivals in Figure 5 has the following properties:

- (a) It is stable, and at most two jobs are in the network.
- (b) Except for the first job entering the empty system, classes 2 and 4 always process jobs simultaneously.

The capacity of class set $\{2, 4\}$ is $1/1.2$ when $\alpha = 1$, but it becomes $1/0.6$ when $0.8 \leq \alpha < 0.9$.

Property 5.1 (Dependence among capacity and arrival rate) Even if the service rate μ_i of each class i is independent of arrival rate λ , capacity μ of a general server is dependent on λ .

When $0.8 \leq \alpha < 0.9$, classes 2 and 4 don't suffer mutual blocking because they are *synchronized*.

Definition 5.7 (Class synchronization) Denote the service start time and end time of job w at class i (or i') as a_i^w (or $a_{i'}^w$) and b_i^w (or $b_{i'}^w$) respectively. Job class i is synchronized with job class i' in a time interval (t_1, t_2) , if for any job k of class i with $[a_i^k, b_i^k] \subset (t_1, t_2)$, there exists a job k' of class i' with $[a_{i'}^{k'}, b_{i'}^{k'}] \subset (t_1, t_2)$, which satisfies either $[a_i^k, b_i^k] \supseteq [a_{i'}^{k'}, b_{i'}^{k'}]$ or $[a_i^k, b_i^k] \subseteq [a_{i'}^{k'}, b_{i'}^{k'}]$.

As shown in Lemma 5.1, the network can *synchronize* itself to avoid mutual blocking when $\alpha = 0.8$. As the interarrival time increases, the synchronization still holds before it reaches 0.9. When $\alpha = 0.9$, the clever synchronization breaks, mutual blocking occurs and the set $\{2, 4\}$ becomes a general server as shown in Figure 5-7. Similar phenomenon can also be observed when $1.1 \leq \alpha < 1.2$.

Figure 5-7 WIP at job classes 2 and 4 when $\alpha = 0.9$.

When $0.8 \leq \alpha < 0.9$ and $1.1 \leq \alpha < 1.2$, classes 2 and 4 are synchronized and the set $\{2, 4\}$ behaves like two standalone stations. In the following, we first show that the queue length between classes 2 and 4 (i.e., classes 2, 3 and 4) is at most two and then show that when $0.8 \leq \alpha \leq 1.2$, the classes 2 and 4 in Figure 5-5 can be synchronized only if $\alpha \in$

$[0.8, 0.9) \cup [1.1, 1.2)$.

Lemma 5.2 (Queue length under synchronization) If classes 2 and 4 in Figure 5-5 are synchronized, then there are at most two jobs between classes 2 and 4.

Lemma 5.3 (Synchronized region) When $0.8 \leq \alpha < 1.2$, classes 2 and 4 in Figure 5 can be synchronized if and only if $\alpha \in [0.8, 0.9) \cup [1.1, 1.2)$.

Remark. According to Lemma 5.2, when there initially are at least three jobs between classes 2 and 4, the two classes still cannot be synchronized even if $\alpha \in [0.8, 0.9) \cup [1.1, 1.2)$. This means that whether classes 2 and 4 suffer mutual blocking also depends on the initial WIP levels.

Besides the above lemmas, another trivial result is that the Lu-Kumar network is unstable when $\alpha < 0.8$ and is stable when $\alpha \geq 1.2$.

Corollary 5.2 (Stability region of the Lu-Kumar network) The Lu-Kumar network in Figure 3-5 with constant interarrival and service times is pathwise stable if and only if $\alpha \in [0.8, 0.9) \cup [1.1, +\infty)$.

Consistent with Dumas (1997) and Dai et al. (2004), Corollary 5.2 shows that the stability region of a queueing network may not be monotone.

Property 5.2 (Non-monotonicity of stability region) The stability region of a queueing network may not be monotone.

In terms of interarrival times, the original instability region is $\alpha \in (0, 0.8) \cup [0.9, 1.1)$. However, once mutual blocking occurs between classes 2 and 4, the system capacity decreases to $1/1.2$ and the new instability region becomes $\alpha \in (0, 1.2)$.

Property 5.3 (Asymmetry of instability region) For the Lu-Kumar network in Figure 5-5, the instability region is larger after mutual blocking occurs.

The above results show that the effective number of the set $\{2, 4\}$ is sensitive to the value of α . Therefore, the system capacity depends on the arrivals, which causes the discontinuity of the stability region. However, if we change the non-preemptive to preemptive-resume policy, the Lu-Kumar network will become a special case of the two-station network studied in Dai and Vande Vate (1996). Regardless of the value of α , the class set $\{2, 4\}$ is always a virtual station with capacity $1/1.2$. Therefore, the stability region of the network is $\alpha \in [1.2, +\infty)$. The effective number of the set $\{2, 4\}$ is always one under the preemptive-resume policy.

This case study shows that the capacity of general servers may change with arrival processes and initial WIP levels. One can show that the capacity of general servers also depends on service policies and service time distributions (Lu and Kumar 1991). Some researchers also studied the properties of stability through different examples and analyses (e.g., Dumas (1997) and Dai et al. (2004)). However, with insights from the fundamental queueing theory, the explanations via general servers are more straightforward and it provides a unified framework. Whether a single-server queue or a complex queueing network, the stability is always determined by the utilizations of servers.

5.4.2 The Lu-Kumar Network with Stochastic Arrivals and Constant Service Times

In this section, we consider stochastic arrivals and assume that $\alpha \sim \exp(1)$, $\alpha \sim U(0.8, 0.9)$, $\alpha \sim \exp(1/0.85)$ and $\alpha \sim \exp(1/1.3)$ respectively.

Figure 5-8 shows the WIP of classes 2 and 4 when the arrival process is Poisson with rate one. Classes 2 and 4 still suffer mutual blocking and only one class of jobs can receive service at a time (except during a transient period).

Figure 5-8 WIP at job classes 2 and 4 when $\alpha \sim \exp(1)$.

When $\alpha \sim U(0.8, 0.9)$, classes 2 and 4 can be synchronized to avoid mutual blocking. This observation coincides with Lemma 5.1. Although α is random, it always falls into the interval $(0.8, 0.9)$, in which the synchronization remains active. Therefore, $EF(\{2, 4\}) = 2$ and the system bottleneck is still the physical station A.

However, the system becomes unstable when $\alpha \sim \exp(1/0.85)$, although α has the same mean 0.85 with $U(0.8, 0.9)$. Figure 5-9 shows that classes 2 and 4 will suffer mutual

blocking when $\alpha \sim \exp(1/0.85)$. Since $\alpha \sim \exp(1/0.85)$, α can take values outside of the interval $[0.8, 0.9) \cup [1.1, 1.2)$ with a positive probability. Once α falls into the interval $(0.9, 1.1)$, the smart synchronization between classes 2 and 4 breaks and the set $\{2, 4\}$ becomes a compact server. According to Property 5.3, the instability region becomes $\alpha \in (0, 1.2)$. Hence, the system is unstable when $\alpha \sim \exp(1/0.85)$.

Figure 5-9 WIP at job classes 2 and 4 when $\alpha \sim \exp(1/0.85)$.

On the other hand, although α falls into the interval $(0.9, 1.1)$ with a positive probability when $\alpha \sim \exp(1/1.3)$, the system is still stable. This is trivially because that the arrival rate is always less than the capacity of the bottleneck, which is at least $1/1.2$, no matter whether the mutual blocking exists or not.

Assume ω is a constant and X is a random variable. For a queueing network with a given topology, service policy, initial state and service time distribution, let $D(\omega)$ and $D(X)$ be the systems induced by the constant interarrival times ω and random interarrival times X respectively. $D(\omega) \in \Theta$ means that $D(\omega)$ suffers mutual blocking.

Conjecture 5.1 If there exists some set A , s.t. $D(\omega) \in \Theta$ for any $\omega \in A$ and $Pr(X \in$

$A) > 0$, then $D(X) \in \theta$ almost surely, i.e., mutual blocking occurs almost surely in the system induced by the random interarrival times X .

Based on Conjecture 5.1, the set $\{2, 4\}$ is always a compact server under exponential interarrival times. Just as physical stations, compact servers also have their specific queue time distributions. Figure 5-10 depicts the mean queue times at compact server $\{2, 4\}$ when the utilizations are between 0.1 and 0.95 from simulation. The largest half-width 95% confidence interval is within 4.5% of the mean at the 95% utilization.

Figure 5-10 Job mean queue time at compact server $\{2, 4\}$.

Conjecture 5.2 (Queue time of a compact server) The jobs at a compact server have their specific queue time distributions.

Just like physical stations, compact servers not only influence the stability of queueing networks but also have their specific queue time distributions.

5.5 Stability Conditions

For a queueing network operating under a given service discipline, any general server, no

matter a physical or intangible one, has an impact on system stability according to the previous analysis. The usual traffic condition will be sufficient if all general servers are considered.

Theorem 5.1 A queueing network under a given service discipline is stable if and only if the effective traffic intensity of every general server does not exceed one, i.e., $P_S \leq 1$ for any general server S .

Remark. In this theorem, we don't consider global stability but consider the stability of queueing networks under a given configuration (including the service discipline). The structure of general servers corresponds to the one under this configuration.

Theorem 5.2 A queueing network under a given service discipline is stable if and only if the effective traffic intensity of every compact server does not exceed one.

Since the identifying of general servers is intricate in general, we will not focus on finding out the general servers and calculating the utilizations. Instead, the relationship stated in the theorems will be applied conceptually.

5.6 The Applications

In the following, we will utilize the concept of general servers to address the stability of multiclass queueing networks. Firstly, we will give alternative and intuitive proofs for two known results through the concept of general servers. Secondly, we will show the stability of a practical dispatching policy.

5.6.1 The Feedforward Networks and FBFS Reentrant Lines

The multiclass feedforward networks (i.e., if $\tau(k) = \tau(l)$ and $k < l$, then $\sigma(k) \leq \sigma(l)$) and FBFS reentrant lines (i.e., if $\sigma(k) = \sigma(l)$ and $k < l$, then class k customers will have a higher priority over class j) have been shown to be stable if the usual traffic condition is satisfied (Dai 1995). In this section, we will give alternative proofs through the concept of general servers.

Theorem 5.3 A feedforward network is stable if and only if the traffic intensity of every physical station does not exceed one.

Theorem 5.4 A reentrant line operating under FBFS discipline is stable if and only if the traffic intensity of every physical station does not exceed one.

Theorem 5.3 and 5.4 are well known results. We give alternative proofs here to show that, although the identification of all general servers seems to be impractical, the concept is still valuable and we can employ it to gain insights and even prove the stability of specific service disciplines.

5.6.2 Stability of Queueing Networks in Practice

Stabilizing a queueing network is crucial for production management. Maglaras (1999) and Dai and Lin (2005) showed the stability of multiclass queueing networks under discrete-review and maximum pressure policies. However, these policies may not always be implemented in practice.

In practical production systems, larger WIP bubbles lead to higher variations of

interarrival times, and the buffer spaces cannot be infinite due to the layout constraint. Dispatching policies commonly try to prevent large WIP bubbles by giving higher priority to the job class with a longer queue length. Therefore, practical dispatching policies are *WIP-dependent*, i.e., class k will be assigned the highest priority if $Q_k > \varphi_k$, where Q_k is number of jobs in buffer k and φ_k is a preset threshold. If more than one job class is assigned the highest priority at the same time at a station, the station will choose from the two job classes randomly. In this situation, job classes from different physical stations can receive service simultaneously as long as their queue lengths are long enough. Therefore, mutual blocking (if any) will break.

Theorem 5.5 A practical queueing network operating under the WIP-dependent dispatching policy is stable, as long as the traffic intensity of every physical station does not exceed one.

When the dispatching policy is WIP-dependent, mutual blocking (among classes from different stations) cannot occur and compact servers simply reduce to physical stations. Hence, in practice, if every physical station has sufficient capacity, the system will always be stable as long as the dispatching policy is WIP-dependent. This provides a general guide for the dispatching of jobs in production systems.

5.7 Comparison to the Fluid Model Approach

The concept of general server has provided a new approach to analyze the stability of queueing systems. As shown by the previous section, through general servers, we can also prove the stability of feedforward networks and FBFS reentrant lines. Compared to the

conventional proof through fluid models, the new approach is more intuitive and insightful.

Table 5-1 The differences between the general server and fluid model approaches.

	Service Discipline?	Distributions of Variables?	Dependence?	IIF Condition?
General Server	Yes	Yes	Yes	Yes
Fluid Model	Yes	No	No	No

Table 5-1 further summarizes the differences between the two approaches, where Yes or No represents whether the approach can capture the corresponding feature in the first row. The fluid model, as a continuous analogue of the original queueing system, only concerns the first moment (i.e., mean) of the service or interarrival times. Other information (e.g., the distribution pattern) of the distributions is neglected although they may influence the system stability (Dai 2009). Furthermore, while the fluid model cannot effectively capture the dependence among different stations, the general servers are defined through depicting the dependence (i.e., mutual blocking) among stations. Consistent with the explanations in Chapter 2, the proposed general server approach manages to capture the mechanism of dependence and thus it's able to provide the sufficient and necessary conditions for stability.

5.8 Summary

Identifying the capacity of a queueing network is of critical importance in practice, where system stability plays a key role in this aspect. In this chapter, we have clarified the concept of servers in the context of queueing networks. In addition, the concept of traffic intensity has also been generalized to general servers through the effective number of classes. From

the viewpoint of general servers, stability of queueing networks solely depends on the utilizations, which is the same as the case of GI/G/1 queues introduced in Chapter 4. The usual traffic condition becomes sufficient if we consider all general servers. This implies that focusing on physical configurations is not enough and we shouldn't ignore the impacts of intangible servers induced by the dependence among classes.

Identifying the general servers relies on the examination of all class sets and the complexity will increase exponentially with respect to the network size. This is consistent with the case of pseudostations as shown in Hasenbein (1997). Hence, we conjecture that determining the stability of a queueing network in general settings is essentially an NP problem. This may also explain why it is hard to develop a simple “iff” condition for the stability of queueing networks.

Although the identification of general servers is hard, the concept is insightful and can be applied to prove the stability of practical systems. The structure of general servers depends on various parameters of system configurations, e.g. the interarrival time distributions, the service time distributions, the service policies, and the initial states. Intuitively, decreasing the arrival rate will preserve stability since the utilizations of physical stations become lower. However, decreasing arrival rate may induce general servers, which may indeed reduce capacity. As we have seen in the Lu-Kumar network, the stability region of a network may not be continuous, monotone, or convex, which coincides with the study of Dumas (1997) and Dai et al. (2004). The evolution of a queueing network is similar to a chaotic system, which is also sensitive to the initial conditions (Hirsch et al.

2012). A small difference in the initial conditions (i.e., arrival processes, service processes, or service policies) may induce significantly different future path (i.e., structure of general servers, queue length, or stability).

Identifying general servers may be intricate due to the high complexity. However, the applications of general servers don't rely on the identification of them all. As shown in Section 5.6, through the concept of general servers, we can provide more intuitive proofs for the stability of feedforward networks and FBFS reentrant lines. Furthermore, we can also show that, if every physical station has sufficient capacity, practical queueing networks under WIP-dependent service disciplines are always pathwise stable.

Although this chapter has been focusing on the pathwise stability of a queueing network, we believe that the approach also applies to the strong stability (i.e., positive Harris recurrence) since the structure of general servers remains unchanged. Given that the interarrival times are unbounded and spread out, the effective traffic intensity of every general server being less than one is supposed to be sufficient for the strong stability.

Chapter 6 Mean Queue Time Approximation for GI/G/1 Queues

The previous two chapters have been focusing on a fundamental feature of queueing systems, i.e., stability, which commonly links to the throughput rate of a production system and serves as a prerequisite for the study of mean queue time. If only considering stability, managers will always prefer a production system with a higher throughput rate to keep higher productivity and gain more revenues from the investment. However, as the utilization increases, the mean queue time of a production system will increase exponentially (Kingman 1962). In practical situations, products or orders generally have specific due dates. On one hand, products and orders should be finished before the designated due dates to meet the requirements of customers and hence keep high service quality. On the other hand, to satisfy the company's operational strategy, managers hope to release the products to the markets as early as possible to earn more market shares. Therefore, in nowadays' competitive environment, the mean cycle time of production systems shall be controlled carefully to avoid the delay of products and orders. To support the system management, we will focus on characterizing the mean queue time performance for queueing systems.

In this chapter, we start from the GI/G/1 queues, which are elementary building blocks of complex queueing systems. Proposing an accurate approximation for the GI/G/1 queues is the first step towards a reliable approximation for the entire practical production systems.

Since the last century, most of the approximation formulas have been built on the first

two moments of the interarrival and service times. To improve the accuracy, the third moment will be considered (Wu et al. 2017b).

6.1 Brief

Myskja (1990) and Myskja (1991) assumed that the exact results for the $H_2/M/1$ queueing system can be modified to approximate the $H_2/G/1$ queueing system as well as a $GI/G/1$ queueing system. While the former assumption may hold, the latter need not be the case, especially in the practical setting. When the mean and variance of the hyper-exponential distribution are fixed, the third moment can be varied over a wide range. Hence, the approximations use a term q_0 (see Tables 6-1 and 6-2 for the details), representing the lowest value of the relative third moment (q_a). The value of q_0 was set to a value of r^2 , which is exact for the 2-phase hyper-exponential (H_2) distribution (r represents the relative second moment that is given as a ratio of the second moment to two times the square of the first moment). However, the “*lower limit on q_0 does not necessarily apply in general for other distributions* (Myskja 1991).” For instance, when empirical data from a practical setting is used, the fitted distribution (e.g., beta, gamma, lognormal and Weibull) will have a fixed mean, variance and third moment, i.e., $q_0 = q$. Another drawback of the models developed by Myskja (1990) and Myskja (1991) in a practical setting is the use of H_2 distribution for the interarrival time and H_2 or Erlang distribution for the service time distributions. These two distributions are special cases of the phase type distributions which can be used to approximate non-negative distributions, but it has the following shortcomings. According to

Neuts (1981), “Although the density plots of the PH-distributions, even with m as small as five or ten, can be highly versatile, there are interesting distributions, for example with steeply increasing or decreasing densities or with intervals of constancy, which are difficult to approximate. Foremost among these are “delayed distributions,” for which $F(x) = 0$, for $0 \leq x \leq a$, for some $a > 0$. Such distributions are of interest to many applications, but even the simple delayed exponential distribution is difficult to approximate by phase type distributions.” For these distributions, the time taken for the computational algorithms used to parameterize the equivalent phase-type distributions can be infinitely long. Hence, it is necessary to develop approximations with closed-form solutions that are computationally less intensive and can deliver results instantaneously for practical applications.

The above discussion has led us to a conclusion that the existing two-moment approximations may not always perform well when the variability of the arrival process is high. Also, the approximations developed by Myskja (1990) and Myskja (1991) may not perform well when the interarrival times do not follow H_2 distribution. Hence, it is necessary to develop a better approximation for the cases of a general arrival process with high variability.

Gathering insights from the existing approximations, we develop a three-moment approximation that can be used for more general interarrival or service time distributions. We perform a comprehensive numerical experimentation to study the performance of our approximation and further compare it with the other existing methods. The paired t-tests show that our approach provides better approximations when the interarrival time SCV is

greater than one.

The model and notations in this chapter are based on Section 3.1.1 and Section 3.4.1. For easier references, the notations of the parameters and variables in a GI/G/1 queue are summarized in Table 6-1.

Table 6-1 Parameters and variables in a GI/G/1 queue.

Model Parameters	
λ, c_a^2	Arrival rate and inter-arrival time SCV.
μ, c_s^2	Service rate and service time SCV.
r_a, r_s	Relative second moment of the inter-arrival and service time distributions, $r_a = (c_a^2 + 1)/2$ and $r_s = (c_s^2 + 1)/2$.
m_a^i, m_s^i	i^{th} moment of the inter-arrival and service time distributions.
q_a, q_s	Relative third moment of the inter-arrival and service time distributions, $q_a = m_a^3 / 6(m_a^1)^3$ and $q_s = m_s^3 / 6(m_s^1)^3$.
Model Variable	
ρ	Utilization of the server.
σ	Load at arrival instants, $0 < \sigma < 1$ (σ is the Laplace-Stieltjes transform of the arrival process)
W	Queue time at the GI/G/1 queue

In the following, we present the previous analytical results and approximations in Section 6.2. The impact of the third moments and its property are proposed in Section 6.3. The three-moment approximation for GI/G/1 queues is given in Section 6.4. In Section 6.5, we present simulation studies and error analyses to validate the proposed approximation. Summary of this chapter is given in Section 6.6.

6.2 Previous Analytical Results and Approximations

When either the interarrival and/or service times follow an exponential distribution (e.g., M/M/1 queue, M/G/1 queue and the G/M/1 queue), there are exact results available for the

mean queue time. However, in practice, both arrival and service processes could follow general distributions and a GI/G/1 queueing model has to be used.

For practical applications, since it is too cumbersome to obtain the exact mathematical results for these general queueing models (Kraemer and Langenbach-Belz 1976), research efforts have focused on simple two-moment bounds or approximations that have closed-form expressions (Gelenbe 1975, Heyman 1975, Kingman 1962, Kobayashi 1974, Kraemer and Langenbach-Belz 1976, Marchal 1976). However, when the interarrival time SCV (Squared Coefficient of Variation = variance / mean²) is greater than one, Shanthikumar and Buzacott (1980) concluded that it is difficult to obtain good approximations. The third moment of the arrival process has a significant influence on the mean queue time of a GI/G/1 queue in this range (Myskja 1990, Myskja 1991). Hence, (Myskja 1990, Myskja 1991) developed a heuristic three-moment approximation for the mean queue time using the exact results for the H₂/M/1 queueing system, where H₂ refers to a 2-phase hyper-exponential interarrival time distribution.

Table 6-2 has classified the exact formulas for some specific queueing systems and some approximation formulas for general GI/G/1 queues in literature.

Table 6-2 Results in literature (exact and approximate).

Exact Results for Special Cases of the GI/G/1 Queueing System	
M/M/1 queue	$E[W] = \left(\frac{\rho}{1-\rho} \right) \cdot \frac{1}{\mu}$
M/G/1 queue (P-K formula)	$E[W] = \left(\frac{\rho}{1-\rho} \right) \cdot \frac{1}{\mu} \cdot \left(\frac{1+c_s^2}{2} \right)$
G/M/1 queue	$E[W] = \left(\frac{\sigma}{1-\sigma} \right) \cdot \frac{1}{\mu}, 0 < \sigma < 1$
Approximate Results and Bounds for the GI/G/1 Queueing System	
Kingman upper bound (1962)	$E[W] \leq \frac{\rho(c_a^2 / \rho^2 + c_s^2)}{2\mu(1-\rho)}$
Kobayashi (1974)	$E[W] \approx \left(\frac{\hat{\rho}}{1-\hat{\rho}} \right) \cdot \frac{1}{\mu}, \quad \hat{\rho} = \exp \left[-\frac{2(1-\rho)}{\rho(c_a^2 + c_s^2 / \rho)} \right]$
Heyman (1975)	$E[W] \approx \left(\frac{\rho}{1-\rho} \right) \cdot \frac{1}{\mu} \cdot \left(\frac{c_a^2 + c_s^2}{2} \right)$
Gelenbe (1975)	$E[W] \approx \left[\frac{\rho c_a^2 + c_s^2}{2(1-\rho)} \right] \cdot \frac{1}{\mu}$
Marchal (1976)	$E[W] \approx \frac{\rho(1+c_s^2)}{2\mu(1-\rho)} \cdot \left(\frac{c_a^2 + \rho^2 c_s^2}{1 + \rho^2 c_s^2} \right)$
Kramer and Langenbach- Belz (1976)	$E[W] \approx \left(\frac{\rho}{1-\rho} \right) \cdot \frac{1}{\mu} \cdot \left(\frac{c_a^2 + c_s^2}{2} \right) \cdot g(\rho, c_a^2, c_s^2), \quad g(\rho, c_a^2, c_s^2) = \begin{cases} \exp \left[-\frac{2(1-\rho)(1-c_a^2)^2}{3\rho(c_a^2 + c_s^2)} \right] & ; c_a^2 \leq 1 \\ \exp \left[-\frac{(1-\rho)(c_a^2 - 1)}{(c_a^2 + 4c_s^2)} \right] & ; c_a^2 \geq 1 \end{cases}$
Kimura (1985)	$E[W] \approx \frac{\sigma(c_a^2 + c_s^2)}{\mu(1-\sigma)(c_a^2 + 1)}$
Myskja (1990)	$E[W] \approx \frac{\rho}{2\mu(1-\rho)} \left\{ (c_s^2 + 1) + \left(\frac{q_0}{q_a} \right)^{\left(\frac{1-\rho}{\rho} \right)} \cdot \frac{1}{\rho} \cdot (c_a^2 - 1) \right\}$
Myskja (1991)	$E[W] \approx \left(\frac{\rho}{1-\rho} \right) \cdot \frac{1}{\mu} \cdot \left\{ r_s + \frac{1}{\rho} \cdot \left[\sqrt{(r_s - \theta)^2 + (2r_s - 1 + d) \cdot (r_a - 1) - (r_s - \theta)} \right] \right\},$ $\theta = \frac{\rho(q_a - r_a) - (q_a - r_a^2)}{2\rho(r_a - 1)}, \quad d = \left(1 + \frac{1}{r_a} \right) \cdot (1 - r_s) \cdot \left[1 - \left(\frac{q_0}{q_a} \right)^3 \right] \cdot (1 - \rho^3)$

6.3 Impact of the Third Moments of Interarrival and Service Times

To include the third moments in the approximation formula, which can produce better estimates for the mean queue time of general single-server queues, we should first

apprehend the impact of the third moments and then reflect it in the final close-form approximation expression.

6.3.1 Simulation Analysis

In this section, we first studied the relative importance of the interarrival and service time distributions on the mean queue time in a GI/G/1 queue. The simulation results showed that the mean queue time is asymmetrically dependent on the interarrival time SCV and the service time SCV. Table 6-3 presents the simulation results of the mean queue time for the GI/G/1 queueing systems, where different combinations of Gamma, Weibull and Lognormal distributions are examined. The results are furnished for 10%, 50% and 95% utilizations and SCV combinations of (1.5, 1.5), (1.5, 3.0) and (3.0, 1.5). The SCV combinations correspond to third moment combinations of (1.67, 1.67), (1.67, 4.67), and (4.67, 1.67), respectively.

Table 6-3 Relative effect of the interarrival and processing time distributions on the mean queue time in a Gamma/Gamma/1 queueing system.

Queueing System	ρ	0.1			0.5			0.95		
	(c_a^2, c_s^2)	(1.5, 1.5)	(1.5, 3.0)	(3.0, 1.5)	(1.5, 1.5)	(1.5, 3.0)	(3.0, 1.5)	(1.5, 1.5)	(1.5, 3.0)	(3.0, 1.5)
Gamma / Gamma / 1	(q_a, q_s)	(1.67, 1.67)	(1.67, 4.67)	(4.67, 1.67)	(1.67, 1.67)	(1.67, 4.67)	(4.67, 1.67)	(1.67, 1.67)	(1.67, 4.67)	(4.67, 1.67)
	$E[W]$	0.25	0.35	0.64	1.65	2.42	2.85	28.62	42.89	43.36
Gamma / Lognormal / 1	(q_a, q_s)	(1.67, 2.60)	(1.67, 10.67)	(4.67, 2.60)	(1.67, 2.60)	(1.67, 10.67)	(4.67, 2.60)	(1.67, 2.60)	(1.67, 10.67)	(4.67, 2.60)
	$E[W]$	0.74	1.03	1.86	4.89	7.23	8.50	85.48	128.05	130.69
Gamma / Weibull / 1	(q_a, q_s)	(1.67, 1.74)	(1.67, 5.56)	(4.67, 1.74)	(1.67, 1.74)	(1.67, 5.56)	(4.67, 1.74)	(1.67, 1.74)	(1.67, 5.56)	(4.67, 1.74)
	$E[W]$	0.25	0.35	0.63	1.65	2.42	2.84	28.58	43.01	43.54
Lognormal / Gamma / 1	(q_a, q_s)	(2.60, 1.67)	(2.60, 4.67)	(10.67, 1.67)	(2.60, 1.67)	(2.60, 4.67)	(10.67, 1.67)	(2.60, 1.67)	(2.60, 4.67)	(10.67, 1.67)
	$E[W]$	0.08	0.17	0.16	6.53	10.62	9.42	267.73	404.01	390.59
Lognormal / Lognormal / 1	(q_a, q_s)	(2.60, 2.60)	(2.60, 10.67)	(10.67, 2.60)	(2.60, 2.60)	(2.60, 10.67)	(10.67, 2.60)	(2.60, 2.60)	(2.60, 10.67)	(10.67, 2.60)
	$E[W]$	0.26	0.52	0.47	3.85	6.30	5.54	83.87	127.01	123.59
Lognormal / Weibull / 1	(q_a, q_s)	(2.60, 1.74)	(2.60, 5.56)	(10.67, 1.74)	(2.60, 1.74)	(2.60, 5.56)	(10.67, 1.74)	(2.60, 1.74)	(2.60, 5.56)	(10.67, 1.74)
	$E[W]$	0.24	0.51	0.47	3.92	6.36	5.66	84.00	127.24	124.11
Weibull / Gamma / 1	(q_a, q_s)	(1.74, 1.67)	(1.74, 4.67)	(5.56, 1.67)	(1.74, 1.67)	(1.74, 4.67)	(5.56, 1.67)	(1.74, 1.67)	(1.74, 4.67)	(5.56, 1.67)
	$E[W]$	0.22	0.32	0.43	8.00	11.85	12.69	270.76	409.03	407.83
Weibull / Lognormal / 1	(q_a, q_s)	(1.74, 2.60)	(1.74, 10.67)	(5.56, 2.60)	(1.74, 2.60)	(1.74, 10.67)	(5.56, 2.60)	(1.74, 2.60)	(1.74, 10.67)	(5.56, 2.60)
	$E[W]$	0.64	0.92	1.24	4.75	6.99	7.51	85.78	128.33	128.90
Weibull / Weibull / 1	(q_a, q_s)	(1.74, 1.74)	(1.74, 5.56)	(5.56, 1.47)	(1.74, 1.74)	(1.74, 5.56)	(5.56, 1.47)	(1.74, 1.74)	(1.74, 5.56)	(5.56, 1.47)
	$E[W]$	0.22	0.32	0.43	1.60	2.37	2.53	28.68	42.93	43.01

Table 6-3 shows that the asymmetric dependence is prevalent at low utilizations, but diminishes as the utilization increases. Their impact becomes nearly the same (i.e., symmetric) in heavy traffic, which is consistent with Kingman's heavy traffic approximation, where the service time and interarrival time variations have the same weight or impact on the mean queue time.

The results presented in Table 6-3 also indicate that the third moment of the service time distribution can be important in developing a three-moment approximation. This means that Myskja (1991)'s conjecture on the effect of the third moment of the service time distribution being insignificant may not always hold.

6.3.2 The Dependence on the Interarrival Time and Service Time Distributions

This asymmetry in light and moderate traffic can be explained using the fundamental relationship in the Lindley equation (Lindley 1952) involving the queue time of the $(n + 1)$ th customer (W_{n+1}), queue time of the n th customer (W_n), the interarrival time between the n th and the $(n + 1)$ th customer (T_n), and the service time of the n th customer (S_n) as shown in

$$W_{n+1} = \max\{0, W_n + S_n - T_n\}. \quad (1)$$

(a) If $W_n + S_n - T_n \leq 0$, $W_{n+1} = 0$. This implies that the mean queue time depends on the entire distribution of the interarrival time as well as the service time.

(b) If $W_n + S_n - T_n > 0$ and $W_n - T_n < 0$, we have $S_n > T_n - W_n > 0$. It follows that $W_{n+1} = W_n + S_n - T_n$ and $W_{n+1} - W_n = S_n - T_n$. The conditional expectation given that $S_n > T_n - W_n > 0$ is

$$\begin{aligned} E[W_{n+1} - W_n | S_n > T_n - W_n > 0] &= E[S_n - T_n | S_n > T_n - W_n > 0] \\ &= E[S_n | S_n > T_n - W_n > 0] - E[T_n | S_n > T_n - W_n > 0]. \end{aligned}$$

In this situation, the mean queue time depends on the entire distribution of the interarrival time as well as the service time.

(c) If $W_n + S_n - T_n > 0$ and $W_n - T_n \geq 0$, we have $T_n \leq W_n$. The conditional expectation given that $T_n \leq W_n$ is

$$\begin{aligned} E[W_{n+1} - W_n | T_n \leq W_n] &= E[S_n - T_n | T_n \leq W_n] \\ &= E[S_n | T_n \leq W_n] - E[T_n | T_n \leq W_n] \\ &= E[S_n] - E[T_n | T_n \leq W_n]. \end{aligned}$$

Since $T_n \leq W_n$ is independent on S_n , the queue time of the $(n + 1)$ th customer is dependent only on the mean service time of the n th customer. As a result, the mean queue time is dependent on the entire interarrival time distribution, but only on the mean of the service time distribution.

Condition (c) shows that the dependence on the interarrival time and service time distributions is asymmetric. However, Kingman's heavy traffic approximation shows that the asymmetry diminishes as the queue goes to heavy traffic. Hence, we have the following property for the mean queue time in a GI/G/1 queueing system.

Property 6.1 The dependence of the mean queue time in a GI/G/1 queueing system on the interarrival time and service time distributions is asymmetric except in heavy traffic.

6.4 The Three-Moment Approximation

In order to conduct a detailed study, we partition the possible values of c_a^2 and c_s^2 into four segments: (i) $0 \leq c_a^2, c_s^2 \leq 1$, (ii) $0 \leq c_a^2 \leq 1, c_s^2 > 1$, (iii) $c_a^2 > 1, 0 \leq c_s^2 \leq 1$, and (iv) $c_a^2 > 1, c_s^2 > 1$. When $0 \leq c_a^2, c_s^2 \leq 1$, the third and higher moments have insignificant impact on system performance, since the shapes (i.e., probability density

functions) of all distributions are similar when their mean and variance are fixed. The approximate error in this region mainly comes from the deviation from Brownian motions (as discussed in Section 6.3), and considering the third or higher moments into the approximate model cannot help to improve the accuracy. As a result, most of the two-moment approximations perform well in this range. Shanthikumar and Buzacott (1980) have performed extensive numerical studies to recommend approximations in this segment. The study found that heuristic approximations outperformed the diffusion approximations. We choose the approximation developed by Kraemer and Langenbach-Belz (1976) for this range.

The second segment covers the region $0 \leq c_a^2 \leq 1$, $c_s^2 > 1$. In this region, we consider the approximations developed by Heyman (1975) and Kraemer and Langenbach-Belz (1976). Our numerical experimentation for this region confirmed that the approximation developed by Kraemer and Langenbach-Belz (1976) outperformed the approximations developed by (Heyman 1975). The numerical results can be found at the supplemental of Wu et al. (2017b). Henceforth, we will refer to Heyman (1975) as Hey, and Kraemer and Langenbach-Belz (1976) as KL.

For the region $c_a^2 > 1$, which encompasses the segments $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ and $c_a^2 > 1$, $c_s^2 > 1$. We use a variant of Myskja (1991)'s approximation for this region as the approximation was derived based on the exact result for an $H_2/M/1$ queuing system. The following rationale was used for the modifications. Myskja (1991)'s approximation works well for two-phase hyper-exponential distributions, which cannot adequately approximate

delayed distributions that are common in practical scenarios (Neuts 1981). Further, Myskja's approximation uses a term q_0 that was set to a value of r^2 , which may not be valid for distributions other than H_2 . But no alternate form for q_0 was suggested. In practice, the available empirical data has a fixed mean, variance and skewness, which means that the value of q_0 can be set at a value of relative third moment of the interarrival time distribution, i.e., $q_0 = q_a$. This value would hold for most distributions that do not have a degree of freedom to vary the third moment when the first two moments are fixed, e.g., beta, gamma, lognormal and Weibull distributions. Additionally, we use the observation from Table 6-3 to develop a correction factor that utilizes the third moment of the service time distribution [see equation (4)]. In order to develop the factor, we first subtracted the contribution of the ∇_a term [see equation (3)] to the queue time from the simulation results. This value (shown in the supplementary file of (Wu et al. 2017)) was found to be an exponential function with dependence on the third moments of the interarrival and service time distributions, and the utilization of the server. Hence, a variant of the (Kraemer and Langenbach-Belz 1976)'s expression was used for the correction factor. However, since Kraemer and Langenbach-Belz (1976)'s approximation for $c_a^2 > 1$ tends to overestimate the mean queue time when the skewness is large and underestimate it when the skewness is small (with fixed mean and variance), a correction factor q_s/q_a is added to the original expression to capture these trends. We also heuristically modify the expression for θ using the exponent of 0.96 by regression as shown in equation (5). The resultant expression for the mean queue time is given by

$$E[W] \approx \left(\frac{\nabla_a + \nabla_s}{2} \right) \cdot \left(\frac{\rho}{1-\rho} \right) \cdot \frac{1}{\mu}, \quad (2)$$

$$\nabla_a = 1 + \frac{1}{\rho} \left[\sqrt{(c_s^2 + 1 - 2\theta)^2 + 2c_s^2(c_a^2 - 1)} - (c_s^2 + 1 - 2\theta) \right], \quad (3)$$

$$\nabla_s = c_s^2 \exp \left[-\frac{(1-\rho)(c_a^2 - 1)}{(c_a^2 + 4c_s^2)} \cdot \left(\frac{q_s}{q_a} \right)^{\sqrt{0.25\rho}} \right], \quad (4)$$

$$\theta = \frac{(\rho-1)(4q_a - 2c_a^2)^{0.96} - (2\rho-1) + (c_a^2)^2}{4\rho(c_a^2 - 1)}, \quad (5)$$

where,

$$q_a = m_a^3 / 6(m_a^1)^3.$$

Henceforth, this approximation will be referred to as TMA (i.e., Three-Moment Approximation).

6.5 Simulation Validation

In this section, we perform numerical experiments to recommend suitable approximations for the mean queue time in a GI/G/1 queueing system in the four segments. As discussed in Section 6.4, the approximation developed by Kraemer and Langenbach-Belz (1976) is suggested when $0 \leq c_a^2$, $c_s^2 \leq 1$ and $0 \leq c_a^2 \leq 1$, $c_s^2 > 1$. When the interarrival time SCV exceeds one ($c_a^2 > 1$, $0 \leq c_s^2 \leq 1$, and $c_a^2 > 1$, $c_s^2 > 1$), we compare the performance of TMA with that of the approximations developed by Kraemer and Langenbach-Belz (1976) and Heyman (1975).

Since TMA is developed based on the exact solution of an $H_2/M/1$ queue, it will provide accurate approximations for queues with phase-type distributed arrival or service

times (see Appendix A3). In the following, we will focus on the performance of TMA for queues with gamma, lognormal and Weibull distributed interarrival and service times. The mean service time was fixed at 10 time units. The interarrival time and service time SCVs considered in the study are 0.1, 0.5, 0.9, 1.5, 3.0, 5.0 and 10.0. The utilization levels considered are 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90 and 0.95. Thirty replications are conducted at each utilization level.

We compare the approximations with the results from an equivalent simulation model built using SimEvents®. The warmup period was chosen to be two times the sample size and the sample size was chosen such that the confidence interval was less than 2% of the mean queue time. Most queueing approximations result in large errors at small utilization levels, and are accurate in heavy traffic. However, the large errors at the small utilizations do not really matter as the mean queue times are a very small proportion of the service time at the server ($1/\mu$). Hence, the more interesting range for evaluation is the mid-utilization range (0.6 – 0.9). We define absolute error (ε) as a ratio of the aggregate absolute difference between the simulation estimates and the analytical results over the 10 utilization levels to the sum of the simulation estimates over the 10 utilization levels (i) as shown in

$$\varepsilon = \frac{\sum_{i=1}^{10} (| \text{Analytical Result} - \text{Simulation Estimates} |)}{\sum_{i=1}^{10} \text{Simulation Estimates}}. \quad (6)$$

We also define cumulative error as a ratio of the aggregate absolute difference between the simulation estimates and the analytical results over all utilization levels and SCV combinations to the sum of the simulation estimates over all utilization levels (i) and SCV

combinations (j) as shown in

$$\mathcal{E}_{cum} = \frac{\sum_{i=1}^{10} \sum_{j=1}^k (| \text{Analytical Result} - \text{Simulation Estimates} |)}{\sum_{i=1}^{10} \sum_{j=1}^k \text{Simulation Estimates}}, \quad (7)$$

where k is 12 for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$, and 16 for $c_a^2 > 1$, $c_s^2 > 1$.

At each segment, we conduct paired t-tests at 95% confidence interval to check if there is a statistical difference in the performance of the approximations and to rank the approximations. We first conduct a two-way test (difference = 0 Vs difference $\neq$ 0) to see if there is a difference in performance. If the null hypothesis is rejected, we conduct a one-way test (difference = 0 Vs difference $>$ 0) to see which of the two approximations is superior. On the other hand, if we fail to reject the null hypothesis in the two-tailed test, we conclude that there is no difference among the approximations.

6.5.1 Results for Segment 3

Table 6-4 summarizes the cumulative errors for each distribution combination (queueing system) considered in this study when $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$.

Table 6-4 Summary of results for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$.

Distribution		Hey	KL	TMA
Inter-arrival	Service	$\varepsilon_{cum} (\%)$	$\varepsilon_{cum} (\%)$	$\varepsilon_{cum} (\%)$
Gamma	Gamma	9.31	16.35	3.10
	Lognormal	9.31	16.35	3.07
	Weibull	9.43	16.46	2.98
Lognormal	Gamma	21.76	12.31	18.71
	Lognormal	21.92	12.46	18.64
	Weibull	21.77	12.32	18.68
Weibull	Gamma	2.57	7.40	5.79
	Lognormal	2.55	7.30	5.75
	Weibull	2.61	7.37	5.82

TMA dominates other approximations when the interarrival time distribution is Gamma. The approximations developed by Heyman (1975) and Kraemer and Langenbach-Belz (1976) seem to be the best approximations when the interarrival times follow Weibull and lognormal distributions, respectively. However, as shown in Table 6-4, TMA remains the second best when the interarrival times follow lognormal and Weibull distributions. The results of the paired t-tests, presented in Table 6-5, show that our approximation outperforms the other ones considered in this segment.

Table 6-5 Results of paired t-test for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$.

Test #	Null Hypothesis (H_0)	Alternate hypothesis (H_1)	t-value	p-value	Decision	Result
1	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} = 0$	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} > 0$	5.37	0.000	Reject H_0	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} > 0$.
2	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} = 0$	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} > 0$	4.11	0.000	Reject H_0	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} > 0$.
3	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} = 0$	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} \neq 0$	-1.35	0.181	Fail to reject H_0	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} = 0$.

In addition to testing the accuracy of the approximations among different SCV combinations, we also tested the approximations across three utilization ranges: (i) low (0.1 – 0.5), (ii) medium (0.6 – 0.8), and (iii) high (0.90 and 0.95). The performances at the low, medium and high utilization levels are summarized in Tables 6-6 to 6-8, and the corresponding results for the paired t-tests are presented in Tables 6-9 to 6-11.

Table 6-6 Summary of results for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ at low utilizations.

Distribution		Hey	KL	TMA
Inter-arrival	Service	$\varepsilon_{cum} (\%)$	$\varepsilon_{cum} (\%)$	$\varepsilon_{cum} (\%)$
Gamma	Gamma	41.52	59.06	14.13
	Lognormal	41.47	59.02	14.10
	Weibull	41.61	59.12	14.06
Lognormal	Gamma	75.40	22.80	49.06
	Lognormal	76.68	23.69	48.76
	Weibull	75.27	22.70	49.11
Weibull	Gamma	11.68	37.83	48.09
	Lognormal	11.37	37.58	47.93
	Weibull	11.75	37.82	48.08

Table 6-7 Summary of results for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ at medium utilizations.

Distribution		Hey	KL	TMA
Inter-arrival	Service	$\varepsilon_{cum} (\%)$	$\varepsilon_{cum} (\%)$	$\varepsilon_{cum} (\%)$
Gamma	Gamma	14.29	26.86	4.27
	Lognormal	14.14	26.74	4.45
	Weibull	14.35	26.91	4.22
Lognormal	Gamma	39.54	19.07	43.15
	Lognormal	39.93	19.41	43.07
	Weibull	39.30	18.87	43.22
Weibull	Gamma	3.40	13.55	6.99
	Lognormal	3.47	13.32	6.97
	Weibull	3.41	13.57	7.02

Table 6-8 Summary of results for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ at high utilizations.

Distribution		Hey	KL	TMA
Inter-arrival	Service	$\varepsilon_{cum} (\%)$	$\varepsilon_{cum} (\%)$	$\varepsilon_{cum} (\%)$
Gamma	Gamma	3.05	6.81	1.13
	Lognormal	3.11	6.87	1.04
	Weibull	3.19	6.95	0.97
Lognormal	Gamma	14.60	10.16	11.35
	Lognormal	14.68	10.23	11.32
	Weibull	14.66	10.21	11.28
Weibull	Gamma	1.42	2.63	1.22
	Lognormal	1.41	2.60	1.20
	Weibull	1.47	2.59	1.25

Table 6-9 Results of paired t-test for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ at low utilizations.

Test #	Null Hypothesis (H_0)	Alternate hypothesis (H_1)	t-value	p-value	Decision	Result
1	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} = 0$	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} > 0$	2.96	0.002	Reject H_0	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} > 0$.
2	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} = 0$	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} > 0$	2.18	0.016	Reject H_0	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} > 0$.
3	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} = 0$	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} \neq 0$	0.82	0.412	Fail to reject H_0	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} = 0$.

Table 6-10 Results of paired t-test for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ at medium utilizations.

Test #	Null Hypothesis (H_0)	Alternate hypothesis (H_1)	t-value	p-value	Decision	Result
1	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} = 0$	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} > 0$	2.36	0.010	Reject H_0	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} > 0$.
2	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} = 0$	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} \neq 0$	1.61	0.055	Fail to reject H_0	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} \neq 0$.
3	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} = 0$	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} \neq 0$	-0.73	0.467	Fail to reject H_0	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} = 0$.

Table 6-11 Results of paired t-test for $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$ at high utilizations.

Test #	Null Hypothesis (H_0)	Alternate hypothesis (H_1)	t-value	p-value	Decision	Result
1	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} = 0$	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} > 0$	7.48	0.000	Reject H_0	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} > 0$.
2	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} = 0$	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} > 0$	5.91	0.000	Reject H_0	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} > 0$.
3	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} = 0$	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} \neq 0$	-1.03	0.307	Fail to reject H_0	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} = 0$.

The results in Tables 6-6 to 6-11 are consistent with the results in Tables 6-4 and 6-5 for the cumulative case. We also infer from Tables 6-6 to 6-8 that the cumulative errors for

all three approximations are higher at the low utilization range, decrease at the medium utilization range, and have the least errors at the high utilization range. As discussed earlier, since the mean queue time of a GI/G/1 queue is shorter compared to the mean service time ($1/\mu$) in light traffic, the relatively higher errors at lower utilizations have little impacts on the performance of the approximations. Hence, TMA is recommended for this segment ($c_a^2 > 1$, $0 \leq c_s^2 \leq 1$).

6.5.2 Result for Segment 4

Table 6-12 summarizes the cumulative error for each distribution combination (queueing system) considered in this section. We perform a paired t-test at 95% confidence interval to show that TMA dominates the other approximations over our numerical experiments. The results of the paired t-tests are presented in Table 6-13. Figure 6-1 through Figure 6-4 present the absolute error plots at different SCV combinations (c_a^2 , c_s^2) for the Gamma/Gamma/1, Gamma/Lognormal/1, Lognormal/Lognormal/1, and Weibull/Weibull/1 queueing systems, respectively.

The results from Tables 6-12 and 6-13 indicate that TMA presented in equation (8) outperforms all other approximations considered in this range ($c_a^2 > 1$, $c_s^2 > 1$). Of the 144 different SCV combinations (over the different queueing systems) that were considered in our study, TMA outperforms the other approximations in about 81% of the cases. The diffusion approximation by Heyman (1975) is the best in about 11% of the cases, while the heuristic approximation developed by Kraemer and Langenbach-Belz (1976) is the best in

about 8% of the cases. The paired t-test results also show that the approximation developed by Heyman (1975) outperforms the one developed by Kraemer and Langenbach-Belz (1976) in this segment.

Table 6-12 Summary of results for $c_a^2 > 1$, $c_s^2 > 1$.

Distribution		Hey	KL	TMA
Inter-arrival	Service	$\varepsilon_{cum} (\%)$	$\varepsilon_{cum} (\%)$	$\varepsilon_{cum} (\%)$
	Gamma	6.31	8.82	1.57
Gamma	Lognormal	6.00	8.51	1.84
	Weibull	6.34	8.85	1.47
	Gamma	6.35	3.94	3.91
Lognormal	Lognormal	7.48	4.71	3.79
	Weibull	6.55	4.05	3.71
	Gamma	2.46	4.73	1.71
Weibull	Lognormal	1.92	3.90	1.88
	Weibull	2.29	4.47	1.75

Table 6-13 Results of paired t-test for $c_a^2 > 1$, $c_s^2 > 1$.

Test #	Null Hypothesis (H_0)	Alternate hypothesis (H_1)	t-value	p-value	Decision	Result
1	$\varepsilon_{Hey} - \varepsilon_{TMA} = 0$	$\varepsilon_{Hey} - \varepsilon_{TMA} > 0$	11.77	0.000	Reject H_0	$\varepsilon_{Hey} - \varepsilon_{TMA} > 0$.
2	$\varepsilon_{KL} - \varepsilon_{TMA} = 0$	$\varepsilon_{KL} - \varepsilon_{TMA} > 0$	10.71	0.000	Reject H_0	$\varepsilon_{KL} - \varepsilon_{TMA} > 0$.
3	$\varepsilon_{Hey} - \varepsilon_{KL} = 0$	$\varepsilon_{Hey} - \varepsilon_{KL} < 0$	-2.37	0.009	Reject H_0	$\varepsilon_{Hey} - \varepsilon_{KL} < 0$.

Figure 6-1 Absolute error plot for the Gamma/Gamma/1 queueing system.

Figure 6-2 Absolute error plot for the Gamma/Lognormal/1 queueing system.

Figure 6-3 Absolute error plot for the Lognormal/Lognormal/1 queueing system.

Figure 6-4 Absolute error plot for the Weibull/Weibull/1 queueing system.

The domination of TMA in cases of gamma arrival processes is depicted in Figures 6-1 and 6-2. When the interarrival time follows either a lognormal or Weibull distributions, the overall performance of TMA seems to fall, but TMA still performs reasonably well and falls among the top 2 approximations for most cases (consistent with the ε_{cum} values in Table 6-6 for TMA).

As in the case of segment 3, we also investigated the performance of the three approximations across the three utilization ranges. The summaries of results are presented in Tables 6-14 to 6-16, and the corresponding paired t-test results are presented in Tables 6-17 to 6-19.

Table 6-14 Summary of results for $c_a^2 > 1$, $c_s^2 > 1$ at low utilizations.

Distribution		Hey	KL	TMA
Inter-arrival	Service	ε_{cum} (%)	ε_{cum} (%)	ε_{cum} (%)
Gamma	Gamma	31.99	39.43	7.43
	Lognormal	30.61	38.19	8.47
	Weibull	31.66	39.13	7.78
Lognormal	Gamma	13.14	5.95	17.58
	Lognormal	16.34	5.46	16.01
	Weibull	13.93	5.41	16.94
Weibull	Gamma	15.80	25.01	9.04
	Lognormal	12.91	22.43	9.18
	Weibull	14.76	24.08	9.05

Table 6-15 Summary of Results for $c_a^2 > 1$, $c_s^2 > 1$ at medium utilizations.

Distribution		Hey	KL	TMA
Inter-arrival	Service	ε_{cum} (%)	ε_{cum} (%)	ε_{cum} (%)
Gamma	Gamma	9.55	14.13	1.98
	Lognormal	9.18	13.78	2.43
	Weibull	9.57	14.14	2.01
Lognormal	Gamma	10.30	5.26	8.17
	Lognormal	12.36	6.68	7.82
	Weibull	10.67	5.46	7.84
Weibull	Gamma	3.25	7.79	1.89
	Lognormal	2.34	6.68	2.37
	Weibull	3.22	7.70	2.05

Table 6-16 Summary of results for $c_a^2 > 1$, $c_s^2 > 1$ at high utilizations.

Distribution		Hey	KL	TMA
Inter-arrival	Service	ε_{cum} (%)	ε_{cum} (%)	ε_{cum} (%)
Gamma	Gamma	2.09	3.35	0.70
	Lognormal	1.99	3.25	0.84
	Weibull	2.19	3.44	0.52
Lognormal	Gamma	4.74	3.42	1.65
	Lognormal	5.50	4.14	1.76
	Weibull	4.86	3.56	1.54
Weibull	Gamma	0.85	1.75	0.90
	Lognormal	0.69	1.24	1.01
	Weibull	0.74	1.53	0.92

Table 6-17 Results of paired t-test for $c_a^2 > 1$, $c_s^2 > 1$ at low utilizations.

Test #	Null Hypothesis (H_0)	Alternate hypothesis (H_1)	t-value	p-value	Decision	Result
1	$\varepsilon_{Hey} - \varepsilon_{TMA} = 0$	$\varepsilon_{Hey} - \varepsilon_{TMA} > 0$	8.73	0.000	Reject H_0	$\varepsilon_{Hey} - \varepsilon_{TMA} > 0$.
2	$\varepsilon_{KL} - \varepsilon_{TMA} = 0$	$\varepsilon_{KL} - \varepsilon_{TMA} > 0$	6.57	0.000	Reject H_0	$\varepsilon_{KL} - \varepsilon_{TMA} > 0$.
3	$\varepsilon_{Hey} - \varepsilon_{KL} = 0$	$\varepsilon_{Hey} - \varepsilon_{KL} \neq 0$	-1.41	0.160	Fail to reject H_0	$\varepsilon_{Hey} - \varepsilon_{KL} = 0$.

Table 6-18 Results of paired t-test for $c_a^2 > 1$, $c_s^2 > 1$ at medium utilizations.

Test #	Null Hypothesis (H_0)	Alternate hypothesis (H_1)	t-value	p-value	Decision	Result
1	$\varepsilon_{Hey} - \varepsilon_{TMA} = 0$	$\varepsilon_{Hey} - \varepsilon_{TMA} > 0$	12.19	0.000	Reject H_0	$\varepsilon_{Hey} - \varepsilon_{TMA} > 0$.
2	$\varepsilon_{KL} - \varepsilon_{TMA} = 0$	$\varepsilon_{KL} - \varepsilon_{TMA} > 0$	7.57	0.000	Reject H_0	$\varepsilon_{KL} - \varepsilon_{TMA} > 0$.
3	$\varepsilon_{Hey} - \varepsilon_{KL} = 0$	$\varepsilon_{Hey} - \varepsilon_{KL} \neq 0$	-1.91	0.059	Fail to reject H_0	$\varepsilon_{Hey} - \varepsilon_{KL} = 0$.

Table 6-19 Results of paired t-test for $c_a^2 > 1$, $c_s^2 > 1$ at high utilizations.

Test #	Null Hypothesis (H_0)	Alternate hypothesis (H_1)	t-value	p-value	Decision	Result
1	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} = 0$	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} > 0$	7.23	0.000	Reject H_0	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{TMA}} > 0$.
2	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} = 0$	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} > 0$	10.57	0.000	Reject H_0	$\varepsilon_{\text{KL}} - \varepsilon_{\text{TMA}} > 0$.
3	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} = 0$	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} \neq 0$	-1.82	0.071	Fail to reject H_0	$\varepsilon_{\text{Hey}} - \varepsilon_{\text{KL}} = 0$.

The results in Tables 6-14 to 6-16 reveal that for all approximations, the cumulative errors are highest at low utilizations, and lowest at high utilizations (which is consistent with the results for segment 3). Also, they are consistent with the results across all utilizations in Table 6-12. However, there seems to be a difference in the results for the paired t-tests. While the results for the paired t-test across all utilizations (see Table 6-13) allude to the fact that the approximation developed by Heyman (1975) has a lower cumulative error percentage compared to the approximation developed by Kraemer and Langenbach-Belz (1976). The results presented in Tables 6-17 to 6-19 for the low, medium and high utilization ranges seem to indicate no statistical difference between the two approximations. The paired t-test results (see Tables 6-17 to 6-19) indicate that the overall performance of the TMA approximation over the three utilization ranges is better than the other two approximations. Therefore, we recommend TMA when $c_a^2 > 1$, $c_s^2 > 1$.

6.6 Proposed Approximation for the Mean Queue Time

Based on the numerical experiments conducted, we recommend the approximation developed by Kraemer and Langenbach-Belz (1976) for $c_a^2 \leq 1$ and our approximation for

$c_a^2 > 1$. The general expression proposed for the mean queue time in a GI/G/1 queueing system is

$$E[W] \approx \begin{cases} \left(\frac{c_a^2 + c_s^2}{2} \right) \cdot \left(\frac{\rho}{1-\rho} \right) \cdot \frac{1}{\mu} \cdot g(\rho, c_a^2, c_s^2) & ; \quad c_a^2 \leq 1 \\ \left(\frac{\nabla_a + \nabla_s}{2} \right) \cdot \left(\frac{\rho}{1-\rho} \right) \cdot \frac{1}{\mu} & ; \quad c_a^2 > 1 \end{cases} \quad (8)$$

where

$$g(\rho, c_a^2, c_s^2) = \exp \left[-\frac{2(1-\rho)(1-c_a^2)^2}{3\rho(c_a^2 + c_s^2)} \right],$$

$$\nabla_a = 1 + \frac{1}{\rho} \left[\sqrt{(c_s^2 + 1 - 2\theta)^2 + 2c_s^2(c_a^2 - 1)} - (c_s^2 + 1 - 2\theta) \right],$$

$$\nabla_s = c_s^2 \exp \left[-\frac{(1-\rho)(c_a^2 - 1)}{(c_a^2 + 4c_s^2)} \cdot \left(\frac{q_s}{q_a} \right)^{\sqrt{0.25\rho}} \right],$$

$$\theta = \frac{(\rho - 1)(4q_a - 2c_a^2)^{0.96} - (2\rho - 1) + (c_a^2)^2}{4\rho(c_a^2 - 1)},$$

and $q_a = m_a^3/6(m_a^1)^3$.

When the utilization approaches one, the g -factor provided by Kraemer and Langenbach-Belz (1976) converges to 1, and the expression for $c_a^2 \leq 1$ would reduce to the approximation provided by Heyman (1975). In the expression for $c_a^2 > 1$, we have

$$\theta \rightarrow \frac{c_a^2 + 1}{4}, \quad \nabla_a \rightarrow c_a^2 \quad \text{and} \quad \nabla_s \rightarrow c_s^2 \quad \text{as the utilization approaches one. Hence, expression}$$

(8) is accurate in heavy traffic.

For an M/G/1 queueing system, the g -factor reduces to 1 and the expression for $c_a^2 \leq 1$ reduces to the Pollaczek–Khinchine formula. Hence, expression (8) is exact for the M/G/1 queueing system.

6.7 Summary

In this chapter, we have developed a general approximation for the mean queue time in a GI/G/1 queueing system. We first partition the possible values of c_a^2 and c_s^2 into four segments: (i) $0 \leq c_a^2 \leq 1$, $0 \leq c_s^2 \leq 1$, (ii) $0 \leq c_a^2 \leq 1$, $c_s^2 > 1$, (iii) $c_a^2 > 1$, $0 \leq c_s^2 \leq 1$, and (iv) $c_a^2 > 1$, $c_s^2 > 1$. In the first segment where both the interarrival time and service time SCV are less than 1, we inferred that the existing two-moment approximations should work well. We recommend the heuristic approximation by Kraemer and Langenbach-Belz (1976) for this segment ($0 \leq c_a^2 \leq 1$, $0 \leq c_s^2 \leq 1$). In the second segment, ($0 \leq c_a^2 \leq 1$, $c_s^2 > 1$), our numerical experiments confirmed that the approximation developed by Kraemer and Langenbach-Belz (1976) outperformed Heyman (1975)'s approximation. We have incorporated the insights from the existing approximations to develop a more general approximation for $c_a^2 > 1$. In the third segment ($c_a^2 > 1$, $0 \leq c_s^2 \leq 1$) and the fourth segment ($c_a^2 > 1$, $c_s^2 > 1$), we found that TMA outperforms the other existing approximations. Fig 6-5 presents our conclusions on the recommended approximations in the different segments.

Figure 6-5 Recommended approximations.

TMA yields accurate results in heavy traffic. It reduces the approximate errors when

the arrival process has large variability. While the famous KL approximation has only been checked through several simple distributions (D, E_2, E_4, M, H_2) , TMA has been verified through more arrival and service processes with a wider range of first and second moments as presented in this chapter and the supplementary file of Wu et al. (2017b). Hence, TMA gives reliable approximations for queues with large interarrival time SCVs. However, one should note that all the approximations presented in this chapter (including TMA) are only applicable with a renewal arrival process. When the interarrival intervals are not independent, e.g. the departure process in a general queueing network, large errors may appear especially when the service time variability is small (Wu and McGinnis 2013).

Chapter 7 Mean Queue Time of General Queueing Systems

The analysis in the previous chapter has provided us a closed-form formula to estimate the mean queue of single-server queues. However, in practice, what we face can be complex queueing networks, where the dependence among stations plays a critical role. In these cases, since stations can be fed by non-renewal processes, the conventional approximation formula for a single-server queue will no longer be applicable. In this chapter, we will investigate the properties of mean queue time of general queueing systems and propose a unified approximation algorithm.

7.1 Brief

As shown in the previous section, most of the prior approximations for queueing systems were developed based on additional assumptions (e.g., independence assumption in QNA and heavy traffic assumption in QNET), which can lead to unfavorable results in practical applications. Therefore, in this chapter, we hope to develop an approximation tool without any additional assumption to support the performance evaluation of general queueing systems.

Wu and McGinnis (2013) showed that although the mean queue time of a single-server tandem queue may have no exact closed-form solution, it can be well characterized by the mean queue times of its corresponding ASIA and BSIA systems. An affine structure exists among them and can be exploited to develop reliable approximations. Since no additional assumptions are made, the approximation tends to provide more accurate and reliable

estimates than previous approximations. Inspired by the above observations, we believe that the concept of affine structure applies to more general queueing systems.

In this chapter, we systematically present the affine structure in queueing systems and show the existence of affine structure in various queueing models. Based on the theoretical framework, an approximation algorithm for the mean queue time of general queueing systems is proposed. For single-server tandem queues, we show that the relative errors produced by the approximation algorithm will converge to zero as the system utilization goes to zero or one. This may also explain why the interpolation method proposed by Wu and McGinnis (2013) can produce accurate estimates and outperforms the prior approximation methods. Through numerical experiments, we show that the approximation algorithm also performs well for tandem queues with homogeneous servers.

In the following, Section 7.2 proposes the basic models and notations. Section 7.3 introduces the affine structure of mean queue time and its properties in general queueing systems. Based on the affine structure, Section 7.4 proposes an approximation algorithm for the evaluation of mean queue times in general queueing systems. The approximation algorithm is validated by both theoretical and numerical analyses in Section 7.5. Section 7.6 summarizes this chapter.

7.2 Notations and Models

In this chapter, we consider a multiclass queueing network $\mathcal{S}$ with a single type of job and J nodes, where there are m_j homogeneous stations at node j ($j = 1, \dots, J$). The external

arrival rate of customers to the system is λ . We use a random variable A to denote the interarrival time and $E[A] = 1/\lambda$. jobs leave the system after receiving K steps of service and we refer to them as K job classes. The service time of class k ($k = 1, \dots, K$) is denoted by S_k and $E[S_k] = 1/\mu_k$. The squared coefficient variations (SCVs) of the interarrival times and services time of each class are denoted as c_a^2 and $c_{s_k}^2$ ($k = 1, \dots, K$) respectively. The utilization (or traffic intensity) at node j is given by $\rho_j = \sum_{k: \sigma(k)=j} \lambda / (n_j \mu_k)$, where $\sigma(k)$ denotes the station that server class k . For stability, we assume $\rho_j < 1$ for $j = 1, 2, \dots, J$. ρ_{BN} is the utilization of the bottleneck, where $BN = \arg \max_{j=1,2,\dots,J} \{\rho_j\}$ is the bottleneck index (choose the smallest index if multiple nodes have the highest utilization). For simplicity, we denote the utilization of a tandem queue as ρ , which refers to the utilization of the bottleneck, i.e., $\rho = \rho_{BN}$. Note that the model can incorporate various systems such as tandem queues and reentrant line. For example, when $K = J$, the system will reduce to the following tandem queue (Figure 7-1) with homogeneous servers at each stage.

Figure 7-1 A J -stage tandem queue with multiple servers at each stage.

We denote the mean queue time of the jobs in a queueing system $\mathcal{S}$ as $\mathbf{W}^{\mathcal{S}}(\rho)$, where $\mathbf{W}^{\mathcal{S}}(\rho)$ is a vector with each coordinate representing the mean queue time at each station, i.e., $\mathbf{W}^{\mathcal{S}}(\rho) = (W_1^{\mathcal{S}}(\rho), W_2^{\mathcal{S}}(\rho), \dots, W_J^{\mathcal{S}}(\rho))$. Throughout this chapter, bold fonts are employed to

denote vectors. according to the above notations, the mean queue time of a two-stage single-server tandem queue $\mathcal{S}$ with Poisson arrivals and exponentially distributed service times can be written as

$$\mathbf{W}^{\mathcal{S}}(\rho) = \left(\frac{\rho_1}{1 - \rho_1} \frac{1}{\mu_1}, \frac{\rho_2}{1 - \rho_2} \frac{1}{\mu_2} \right). \quad (1)$$

In this chapter, we will study the performance of $\mathbf{W}^{\mathcal{S}}(\rho)$ as the utilization ρ changes. The other parameters (e.g., SCVs, services disciplines and capacity) of $\mathcal{S}$ will be kept fixed since we focus on the trade-off between the mean queue time and utilization (or throughput).

In queueing systems, the *affine structure* means that there exists a family of *coefficient vectors* $\{\boldsymbol{\alpha}^l(\rho), l = 1, 2, \dots, L\}$ such that

$$\mathbf{W}^{\mathcal{S}}(\rho) = \sum_{l=1}^L \boldsymbol{\alpha}^l(\rho) \circ \mathbf{W}^{\mathcal{S}_l}(\rho), \quad (2)$$

where $\boldsymbol{\alpha}^l(\rho) \circ \mathbf{W}^{\mathcal{S}_l}(\rho)$ is the Hadamard product (i.e., entrywise product), $\{\mathcal{S}_l, l = 1, 2, \dots, L\}$ are the family of *auxiliary systems* and $\mathbf{W}^{\mathcal{S}_l}(\rho)$ ($l = 1, 2, \dots, L$) is the mean queue times of system $\mathcal{S}_l$. The affine structure means that the performance of a queueing system can be characterized by the performance of some auxiliary queueing systems. Hence, if the performance of the auxiliary systems is known, we only need to approximate the coefficient vectors to estimate the performance of the original system.

7.3 The Properties of Affine Structure

In this section, we will verify the affine structure and investigate its properties in some common queueing systems.

7.3.1 GI/G/1 and GI/G/n Queueing Systems

GI/G/1 and GI/G/n queues are sing-node queueing systems, where there's no interdependence among different stages.

Although the mean queue time of an M/M/1 can be evaluated through Markov chain analysis and $W^{M/M/1}(\rho)$ can be given by

$$W^{M/M/1}(\rho) = \frac{\rho}{1 - \rho} \frac{1}{\mu}, \quad (3)$$

the analysis becomes intricate when the arrival process is not Poisson. For GI/G/1 queues, Lindley (1952) first thoroughly studied the evolution of the consecutive jobs' queue times. Although the evolution can be described by the concise Lindley equation, solving the equation to get a closed-form solution for the mean queue time in steady state is hard (Konheim 1975, Lindley 1952, Loynes 1965).

Kingman (1962) first developed an upper bound

$$W^{GI/G/1}(\rho) \leq \frac{c_a^2/\rho^2 + c_s^2}{2} \frac{\rho}{1 - \rho} \frac{1}{\mu} \quad (4)$$

for the mean queue time of GI/G/1 queues, which becomes tight when the system goes to heavy traffic. The heavy traffic approximation of GI/G/1 is given by

$$W^{GI/G/1}(\rho) \cong \frac{c_a^2 + c_s^2}{2} \frac{\rho}{1 - \rho} \frac{1}{\mu}. \quad (5)$$

Note that $(c_a^2 + c_s^2)/2$ in formula (5) is called the *variability factor* by Hopp and Spearman (2011). The approximation is exact for queues with Poisson arrivals, i.e., an M/G/1 queue. The approximation formula can also be developed through the diffusion process analysis by Heyman (1975). Compared to the result of an M/M/1 queue, the only difference is the coefficient term $(c_a^2 + c_s^2)/2$. To increase the accuracy of the

approximation, some researchers adjusted the formula through modifying the coefficient term (Kraemer and Langenbach-Belz 1976, Marchal 1976, Myskja 1991, Myskja 1990, Wu et al. 2017b). Although the final approximation formulas are different in these approximation, all of them are of the form $F(\rho)W^{M/M/1}(\rho)$. For example, $F(\rho)$ refers to $g(\rho, c_a^2, c_s^2)$ in Kraemer and Langenbach-Belz (1976) and refers to $(\nabla_a^2 + \nabla_s^2)/2$ in Wu et al. (2017b). From the viewpoint of affine structure introduced in Section 7.2, the term $F(\rho)$ in these approximations essentially tries to estimate the actual coefficient $\alpha^1(\rho)$ with the corresponding M/M/1 queue being the auxiliary system $\mathcal{S}_1$.

The analysis of GI/G/n queues is much harder. Even with a Poisson arrival process, an M/G/n has no exact closed-form solution for the mean queue time. Inspired by the results of GI/G/1 queues, a nature way to approximate the mean queue time of a GI/G/n queue is comparing it with its corresponding systems, e.g., an M/M/n queue, of which the performance is easier to evaluate. In literature, the mean queue time ($W^{GI/G/n}(\rho)$) of a GI/G/n queue were also approximated by formulas of the form $F(\rho)W^{M/M/n}(\rho)$ (e.g., see Iglehart and Whitt (1970) and Whitt (1993)). For example, in Whitt (1993), $F(\rho)$ refers to the term $\phi(\rho, c_a^2, c_s^2, k)(c_a^2 + c_s^2)/2$, which is an estimate for the actual coefficient $\alpha^1(\rho)$ with the corresponding M/M/n queue being the auxiliary system $\mathcal{S}_1$.

In the following, we will introduce some properties of the coefficient vector $\alpha^1(\rho)$ in the affine structure of GI/G/1 and GI/G/n queues.

Property 7.1 If $c_a^2 = c_s^2 = 1$, then $\alpha^1(\rho) \equiv 1$ for $\rho \in (0,1)$. If $c_a^2 = c_s^2 = 0$, then $\alpha^1(\rho) \equiv 0$ for $\rho \in (0,1)$.

Property 7.1 is a direct corollary of the definition of $\alpha^1(\rho)$. When the system is deterministic, i.e., both the interarrival and service times are constant, the mean queue time will be zero as long as the system is stable.

Property 7.2 $\lim_{\rho \rightarrow 1} \alpha^1(\rho) = (c_a^2 + c_s^2)/2$.

Property 7.2 is the heavy traffic property as shown by Kingman (1962). Kingman's heavy traffic approximation converges to the exact mean queue when the utilization goes to one.

Figure 7-2 Values of $\alpha^1(\rho)$ for a Gamma/Gamma/1 queue with different SCVs.

Besides properties 1 and 2, we also conduct cases studies to investigate the behavior of $\alpha^1(\rho)$. A Gamma/Gamma/1 queueing model under different combinations of SCVs is studied. There are 30 replications of simulation experiments for each combination of SCVs. Each replication consists of 2,000,000 customers after disregarding the first 4,000,000 customers for a warm-up. The 90% half-width confidence interval is within 2% of the mean queue time. Figure 7-2 presents the values of $\alpha^1(\rho)$ for a Gamma/Gamma/1 queue with

different SCVs. Figure 7-2 shows that although $\alpha^1(\rho)$ is different under different settings, it is always a smooth function with respect to the utilization ρ and converges to the constant $(c_a^2 + c_s^2)/2$ in heavy traffic.

Intuitively, to provide a reliable approximation for the coefficient $\alpha^1(\rho)$, $F(\rho)$ should at least satisfy the above properties and conjecture. This guarantees that the approximation error converges to zero when the system goes to heavy traffic and is zero for queues with exact solutions (e.g. M/M/1, D/D/1 and M/M/n queues).

Figure 7-3 Values of $F(\rho)$ in different approximations when $c_a^2 = c_s^2 = 0.8$.

Figure 7-4 Values of $F(\rho)$ in different approximations when $c_a^2 = c_s^2 = 1.25$.

Remark. $F(\rho)$ depends on the number of servers in the approximation of (Whitt 1993).

Figures 7-3 and 7-4 are drawn in the case of $n = 2$.

We examine the properties of $F(\rho)$ for the approximations proposed by Heyman (1975), Marchal (1976), Kraemer and Langenbach-Belz (1976) and Whitt (1993) for the cases (i) $c_a^2 = c_s^2 = 0.8$, and (ii) $c_a^2 = c_s^2 = 1.25$. As shown by Figures 7-3 and 7-4, the behavior of $F(\rho)$ in these approximations is consistent with the above observations and properties. These approximations can produce reliable estimates and are widely used in practical performance evaluation (Hopp et al. 2002).

7.3.2 Single-Server Tandem Queues

Tandem queues are commonly used to model productions lines. In this section, we study the properties of a single-server tandem queue as shown in Figure 7-5. The queue time analysis is difficult in general due to the dependence among stations and the non-renewal arrival

processes of the downstream stations (e.g., Marshall (1968)). Although the dependence among stations has a significant influence on the system performance, most studies ignored it and employed the independence assumption to simplify the analysis (e.g., see Whitt (1983)).

Figure 7-5 A single-server tandem queue.

Inspired by the results of Burke (1956), Jackson (1957) and Friedman (1965), Wu and McGinnis (2013) proposed the concepts of ASIA (All-See-Initial-Arrivals) systems and BSIA (Bottleneck-See-Initial-Arrival) systems for single-server tandem queues. They show that the mean queue time of a tandem queue can be approximated through interpolating the results of its corresponding ASIA and BSIA systems.

Definition 7.1 In an ASIA system, the mean queue time at stage j is $W_j = W_j^{ASIA}$, where W_j^{ASIA} is the mean queue time at stage j when it sees the initial arrival process.

Definition 7.2 In a BSIA system, $\sum_{j=1}^J W_j = W^{BSIA}$, where W^{BSIA} is the mean queue time of the bottleneck when it sees the initial arrival process, i.e., $W^{BSIA} = W_{BN}^{ASIA}$.

For the BSIA system, each W_j can be calculated based on the Fully Coupled System (FCS) (see Wu and McGinnis (2013) and Friedman (1965)).

Definition 7.3 For a server in tandem queues, its Intrinsic Ratio (IR) is defined as

$$IR_i = \frac{\text{Actual } W_i - W_i \text{ in BSIA}}{W_i \text{ in ASIA} - W_i \text{ in BSIA}}.$$

Note that when $i = 1$, both the denominator and numerator in Definition 7.3 equal

zero. To unify the notations, we define $IR_1 = 1$.

Employing the notations presented in Section 7.2, the above definitions can be written as

$$\mathbf{W}^{\mathcal{S}}(\rho) = \mathbf{IR}(\rho) \circ \mathbf{W}^{ASIA\ of\ \mathcal{S}}(\rho) + (\mathbf{1} - \mathbf{IR}(\rho)) \circ \mathbf{W}^{BSIA\ of\ \mathcal{S}}(\rho), \quad (6)$$

where $\mathbf{IR}(\rho) = (IR_1(\rho), IR_2(\rho), \dots, IR_N(\rho))$ and $\mathbf{1}$ is a N -dimension vector with each element being one. Recall that ρ is the utilization of the system bottleneck as stated in Section 7.2. In this situation, $\mathbf{IR}(\rho)$ and $(\mathbf{1} - \mathbf{IR}(\rho))$ correspond to the coefficient vectors $\alpha^1(\rho)$ and $\alpha^2(\rho)$ respectively. Some properties of $\mathbf{IR}(\rho)$ have been observed by Wu and McGinnis (2013). In the following, we will rigorously analyze the properties of $\mathbf{IR}(\rho)$.

Property 7.3 $\mathbf{IR}(\rho) \equiv \mathbf{1}$ if the arrival process is Poisson and the service times at all stages are exponentially distributed.

Property 7.3 follows from the fact that the departure process of an M/M/1 queue is again a Poisson process (Burke 1956). In this case, we have $\mathbf{W}^{\mathcal{S}}(\rho) = \mathbf{W}^{ASIA\ of\ \mathcal{S}}(\rho)$, which implies that $\mathbf{IR}(\rho) \equiv \mathbf{1}$ according to the definition.

Property 7.4 $\mathbf{IR}(\rho) \equiv \mathbf{0}$ for $\rho \in (0,1)$ if $c_{sj} = 0$ for $j = 1, 2, \dots, J$, i.e., the service times at each stage are constant.

The total mean queue time in a tandem queue with constant service times doesn't depend on the arrangement of stages based on (Friedman 1965). In such systems, there is $\mathbf{W}^{\mathcal{S}}(\rho) = \mathbf{W}^{BSIA\ of\ \mathcal{S}}(\rho)$ and we have $\mathbf{IR}(\rho) \equiv \mathbf{0}$.

Property 7.5 For a single-server tandem queue with a Poisson arrival process, we have

Property 7.5 For a single-server tandem queue with a Poisson arrival process,

$\lim_{\rho \rightarrow 0} IR_j(\rho) = C_j$, where C_j is a constant. Specifically, when $\rho \rightarrow 0$, $IR_j(\rho) = C_j + o(\rho)$.

Proof of this property is given in Appendix A4. It means that the intrinsic ratio converges to a constant in light traffic.

Property 7.6 For a single-server tandem queue with a Poisson arrival process, we have

$$\lim_{\rho \rightarrow 1} |IR_{BN}(\rho)/W_{BN}^{ASIA\ of\ S}(\rho)| = 0.$$

Proof of Property 7.6 is given in the Appendix A4. This property means that the slope of $IR_{BN}(\rho)$ with respect to ρ is much smaller than the slope of the mean queue time as $\rho \rightarrow 1$. The properties proved above are consistent with the observations by Wu and McGinnis (2013). Furthermore, we also have the following conjecture based on the simulation results in Wu and McGinnis (2013).

Conjecture 7.1 $IR(\rho)$ is approximately linear with respect to ρ .

Although the exact value of $IR(\rho)$ is hard to get, similar to the case of GI/G/1 queues, we can estimate $IR(\rho)$ based on the above properties and conjecture. We denote $F(\rho)$ as the approximation for $IR(\rho)$. The approximation for the mean queue time $W^S(\rho)$ is given by

$$W^S(\rho) \cong F(\rho) \circ W^{ASIA\ of\ S}(\rho) + (1 - F(\rho)) \circ W^{BSIA\ of\ S}(\rho). \quad (7)$$

Note that formula (7) is actually the interpolation approximation method proposed by Wu and McGinnis (2013) with our notations in Section 7.2.

7.3.3 Homogeneous-Server Tandem Queues

In this section, we apply the affine structure to a tandem queue $\mathcal{S}$ with homogeneous servers as shown in Section 7.2, where $J = 2$, $n_1 \geq 2$ and $n_2 \geq 2$. The first step is to find the auxiliary systems $\{\mathcal{S}_l, l = 1, 2, \dots, L\}$. It turns out that we can define the ASIA and BSIA systems of $\mathcal{S}$ in the same way as a single-server tandem queue. $IR(\rho)$ can also be computed according to Definition 7.3. In the following, we will first show that the affine structure also exists in tandem queues with homogeneous server and then examine the performance of the approximation algorithm. Since $IR_1(\rho) \equiv 1$, our focus will be on the second stage of the tandem queue.

To show that the affine structure also exists in tandem queues with homogeneous servers, we conduct simulation experiments to explore the properties of $IR_2(\rho)$. Figures 7-6 and 7-7 present the value of $IR_2(\rho)$ under different combinations of SCVs of the service times when $n_1 = n_2 = 2$, $E[S_1] = 40$ and $E[S_2] = 60$. The arrival process is Poisson and the service times are assumed to be Gamma distributed. For each combination of SCVs, 30 replications of experiments are conducted and each replication consists of 2,000,000 jobs after disregarding the first 4,000,000 jobs for a warm-up. The 90% half-width confidence interval is within 4% of the mean queue times.

Figure 7-6 Values of $IR_2(\rho)$ for systems with low SCVs.

Figure 7-7 Values of $IR_2(\rho)$ for systems with high SCVs.

As shown in Figures 7-6 and 7-7, the behavior of $IR_2(\rho)$ is similar to the case of single-server tandem queues. The affine structure of tandem queues with homogeneous servers has the same properties as introduced in the last section.

7.3.4 Reentrant Lines

In this section, we investigate the affine structure in a reentrant line, where a product can visit the stations multiple times. Figure 7-8 shows the original reentrant line $\mathcal{S}$ with two

stations and four steps. The first and third steps are at station 1 and station 2 processes the second and fourth steps. We define two auxiliary systems, $\mathcal{S}_1$ and $\mathcal{S}_2$, as shown in Figure 7-9. In the auxiliary systems, each station is recognized as a single-server queue, and the service times are given by $S_{1,3} = (S_1 + S_3)/2$ and $S_{2,4} = (S_2 + S_4)/2$. In system $\mathcal{S}_1$, each station sees the initial arrivals and behaves as a single server queue. In system $\mathcal{S}_2$, only the bottleneck station sees the initial arrivals. Note that the mean queue times of stations 1 and 2 in the two auxiliary systems can be given in the similar way as the ASIA and BSIA systems of tandem queues.

Figure 7-8 The original reentrant

line $\mathcal{S}$.

Figure 7-9 The auxiliary systems $\mathcal{S}_1$ (left) and

$\mathcal{S}_2$ (right).

To examine the affine structure, we assume Poisson arrivals and perform simulations for different combinations of S_1 , S_2 , S_3 and S_4 . According to the approximation formula (8) in Chapter 6, the mean queue time of system $\mathcal{S}_1$ can be given by

$$W^{\mathcal{S}_1}(\rho) = \left(\frac{1 + c_{S_{1,3}}^2}{2} \frac{\rho_1}{1 - \rho_1} E[S_{1,3}], \frac{1 + c_{S_{2,4}}^2}{2} \frac{\rho_2}{1 - \rho_2} E[S_{2,4}] \right).$$

If the first station is the bottleneck, i.e., $E[S_{1,3}] \geq E[S_{2,4}]$, the mean queue time in system $\mathcal{S}_2$ is given by

$$\mathbf{W}^{\mathcal{S}_2}(\rho) = \left(\frac{1 + c_{\mathcal{S}_{1,3}}^2}{2} \frac{\rho_1}{1 - \rho_1} E[S_{1,3}], 0 \right).$$

If the second station is the bottleneck, i.e., $E[S_{1,3}] < E[S_{2,4}]$, the mean queue time in system $\mathcal{S}_2$ is given by

$$\mathbf{W}^{\mathcal{S}_2}(\rho) = \left(\frac{1 + c_{\mathcal{S}_{1,3}}^2}{2} \frac{\rho_1}{1 - \rho_1} E[S_{1,3}], \frac{1 + c_{\mathcal{S}_{2,4}}^2}{2} \frac{\rho_2}{1 - \rho_2} E[S_{2,4}] - \frac{1 + c_{\mathcal{S}_{1,3}}^2}{2} \frac{\rho_1}{1 - \rho_1} E[S_{1,3}] \right).$$

Because $\mathbf{W}^{\mathcal{S}}(\rho) = \boldsymbol{\alpha}^1(\rho) \circ \mathbf{W}^{\mathcal{S}_1}(\rho) + \boldsymbol{\alpha}^2(\rho) \circ \mathbf{W}^{\mathcal{S}_2}(\rho)$, the coefficient vectors can be given by

$$\boldsymbol{\alpha}^1(\rho) = \left(\frac{W_1^{\mathcal{S}}}{W_1^{\mathcal{S}_1}}, \frac{W_2^{\mathcal{S}} - W_2^{\mathcal{S}_2}}{W_2^{\mathcal{S}_1} - W_2^{\mathcal{S}_2}} \right),$$

and

$$\boldsymbol{\alpha}^2(\rho) = \left(0, 1 - \frac{W_2^{\mathcal{S}} - W_2^{\mathcal{S}_2}}{W_2^{\mathcal{S}_1} - W_2^{\mathcal{S}_2}} \right).$$

Since $\boldsymbol{\alpha}^2(\rho)$ can be calculated according to $\boldsymbol{\alpha}^1(\rho)$, we will focus on the properties of $\boldsymbol{\alpha}^1(\rho)$ in the following.

Table 7-1 Distributions of the service times in each case.

Distributions	S_1	S_2	S_3	S_4
Case 1	$\Gamma(2, 0.2)$	$\Gamma(2, 0.25)$	$\Gamma(2, 0.2)$	$\Gamma(2, 0.25)$
Case 2	$\Gamma(0.5, 0.8)$	$\Gamma(0.5, 1)$	$\Gamma(0.5, 0.8)$	$\Gamma(0.5, 1)$
Case 3	$\Gamma(2, 0.25)$	$\Gamma(2, 0.2)$	$\Gamma(2, 0.25)$	$\Gamma(2, 0.25)$
Case 4	$\Gamma(0.5, 1)$	$\Gamma(0.5, 0.8)$	$\Gamma(0.5, 1)$	$\Gamma(0.5, 0.8)$
Case 5	$\Gamma(2, 0.25)$	$\Gamma(2, 0.25)$	$\Gamma(2, 0.25)$	$\Gamma(2, 0.25)$
Case 6	$\Gamma(0.5, 1)$	$\Gamma(0.5, 1)$	$\Gamma(0.5, 1)$	$\Gamma(0.5, 1)$
Case 7	$\Gamma(2, 0.175)$	$\Gamma(2, 0.2)$	$\Gamma(2, 0.275)$	$\Gamma(2, 0.3)$
Case 8	$\Gamma(0.5, 0.7)$	$\Gamma(0.5, 0.8)$	$\Gamma(0.5, 1.1)$	$\Gamma(0.5, 1.2)$
Case 9	$\Gamma(2, 0.2)$	$\Gamma(2, 0.175)$	$\Gamma(2, 0.3)$	$\Gamma(2, 0.275)$
Case 10	$\Gamma(0.5, 0.8)$	$\Gamma(0.5, 0.7)$	$\Gamma(0.5, 1.2)$	$\Gamma(0.5, 1.1)$
Case 11	$\Gamma(2, 0.275)$	$\Gamma(2, 0.3)$	$\Gamma(2, 0.175)$	$\Gamma(2, 0.2)$
Case 12	$\Gamma(0.5, 1.1)$	$\Gamma(0.5, 1.2)$	$\Gamma(0.5, 0.7)$	$\Gamma(0.5, 0.8)$

The service time distributions in the 12 simulation cases are presented in Table 7-1. For the simulation, 30 replications are performed for each case. Each replication consists of 2,000,000 jobs after disregarding the first 4,000,000 jobs for a warm-up. The 90% half-width confidence interval is within 2% of the mean queue times.

We study the trend $\alpha^1(\rho)$ as ρ approaches one, where ρ always represents the utilization of the bottleneck. Note that $W^{s_1}(\rho)$ and $W^{s_2}(\rho)$ can be calculated exactly based on the Pollaczek-Khinchine formuland $W^s(\rho)$ is estimated from simulations. The trends of the first and second coordinates of $\alpha^1(\rho)$ are given in Figures 7-10 and 7-11.

Figure 7-10 The first coordinate of $\alpha^1(\rho)$.

Figure 7-11 The second coordinate of $\alpha^1(\rho)$.

The above figures show that although $\alpha^1(\rho)$ takes different values under different settings, it has regular behavior similar to the IR and also enjoys the properties introduced in Section 7.3.2. This implies that the concept of affine structure can also be applied to reentrant lines.

7.4 The Approximation Algorithm

According to the affine structure, the mean queue time of a general queueing system can be

approximated by the following algorithm.

Algorithm: The Approximation of Mean Queue Times

Step 1. Determine appropriate auxiliary systems $\{\mathcal{S}_l, l = 1, 2, \dots, L\}$.

Step 2. For a set of utilization levels $\{\rho_t, t = 1, 2, \dots, T\}$, calculate the value of $\{\alpha^l(\rho_t), t = 1, 2, \dots, T\}$ for each l based on $W^S(\rho) = \sum_{l=1}^L \alpha^l(\rho) \circ W^{S_l}(\rho)$.

Step 3. For each l , fit the values $\{\alpha^l(\rho_t), t = 1, 2, \dots, T\}$ with a function $F_l(\rho)$, i.e., $F_l(\rho)$ serves as an approximation for the coefficient $\alpha^l(\rho)$.

Step 4. The mean queue time of the system at different utilization levels is approximated by $W^S(\rho) \cong \sum_{l=1}^L F_l(\rho) \circ W^{S_l}(\rho)$.

At Step 3, we need to fit the coefficient $\alpha^l(\rho)$ with the function $F^l(\rho)$. To guarantee $F^l(\rho)$ is a reliable approximation, $F^l(\rho)$ should at least have the same properties as $\alpha^l(\rho)$. For instance, if $\alpha^l(\rho)$ is a monotone function, then $F^l(\rho)$ should also be monotone. On the other hand, it would be easier to find an accurate $F^l(\rho)$ if $\alpha^l(\rho)$ can enjoy some nice properties (e.g., nearly linear). For example, for single-server tandem queues, if we set the ASIA and BSIA systems as the auxiliary systems, the coefficient $\alpha^l(\rho)$ (i.e., the IR) will be a nearly linear function as observed by Wu and McGinnis (2013). In this case, the coefficient $\alpha^l(\rho)$ can be fitted accurately and the above algorithm will produce can produce better approximations for the mean queue time of a tandem queue

than the conventional approaches, e.g., QNA. Therefore, in practical applications, we will try to find auxiliary systems satisfying (i) the performance of the auxiliary systems themselves (i.e., $\{W^{\delta_l}(\rho), l = 1, 2, \dots, L\}$) are easy to evaluate, and (ii) the coefficient vectors $\{\alpha^l(\rho), l = 1, 2, \dots, L\}$ enjoy nice properties (e.g., Properties 7.1-7.6 as shown in the previous sections), which can facilitate the approximation.

The affine structure and approximation algorithm imply that, the overall performance of a queueing system for $\rho \in (0,1)$ can be estimated through the performance at a few specific utilizations levels. Performance of complex systems can be predicted without repetitive time-consuming simulation runs. This is similar to the meta-model based methods in simulation optimization (Xu et al. 2015) and can be employed in practical large-scale problems.

7.5 Validation of the Approximation Algorithm

In this section, we will validate the proposed approximation algorithm from two aspects. For GI/G/1 queues, the approximation formula with the form $F(\rho)W^{M/M/1}(\rho)$ has been well developed in literature and can produce reliable estimates. For single-server tandem queues, since the accuracy has been verified through simulation studies by Wu and McGinnis (2013), we will conduct a rigorous error analysis in this section. For tandem queues with homogeneous servers, the rigorous error analysis is hard at the moment and we will perform simulation experiments to examine the accuracy of the approximation algorithm. For reentrant lines, the approximation algorithm can also provide reliable

approximations according to the simulation experiments. Since the procedures are similar to the case of tandem queues, the details are omitted here.

7.5.1 Error Analysis for Single-Server Tandem Queues

Recall that the approximation for the mean queue time $W^S(\rho)$, where S is single-server tandem queue, is given by $F(\rho) \circ W^{ASIA\ of\ S}(\rho) + (1 - F(\rho)) \circ W^{ASIA\ of\ S}(\rho)$.

Proposition 7.1 For a single-server tandem queue with Poisson arrivals, if $\lim_{\rho \rightarrow 0} F_j(\rho) = C_j$,

then the relative error of the approximation $W^S(\rho) \cong F(\rho) \circ W^{ASIA\ of\ S}(\rho) + (1 - F(\rho)) \circ W^{ASIA\ of\ S}(\rho)$ converges to zero when $\rho \rightarrow 0$.

The proof is given in the Appendix A4. While the conventional approximations commonly tend to produce large deviations in light traffic, the proposed new algorithm tends to produce accurate estimates.

Proposition 7.2 For a single-server tandem queue with Poisson arrivals, if $\lim_{\rho \rightarrow 1} F_{BN}(\rho) (1 - \rho) = 0$,

then the relative error of the approximation at the bottleneck $W_{BN}^S(\rho) \cong F_{BN}(\rho) W_{BN}^{ASIA\ of\ S}(\rho) + (1 - F_{BN}(\rho)) W_{BN}^{ASIA\ of\ S}(\rho)$ converges to zero when $\rho \rightarrow 1$.

The proof is given in the Appendix A4. Since the mean queue time in heavy traffic is dominated by the bottleneck, Proposition 7.2 implies that the approximation based on the affine structure also has the nice heavy traffic property, i.e., it will converge to the exact value as the system goes to heavy traffic. The mean queue times in ASIA and BSIA systems can be calculated exactly when the arrival process is Poisson. For general settings, the performance of ASIA and BSIA system can be computed based on the approximations for

GI/G/1 queues (e.g., the three-moment approximation in Chapter 4).

Properties and propositions in this section provide a theoretical fundamental for the approach proposed by Wu and McGinnis (2013). The above analysis can also explain why the interpolation method based on the affine structure is capable to provide reliable approximations.

Note that in Wu and McGinnis (2013) and the above analysis, two auxiliary systems (i.e., ASIA and BSIA) are employed and the relative error of the approximation algorithm enjoys the light and heavy traffic properties. Actually, the selection of auxiliary systems is not unique. For example, if we employ ASIA system as the only auxiliary system $\mathcal{S}_1$, then the coefficient vector is given by

$$\alpha^1(\rho) = \left(\frac{W_1^{\mathcal{S}}(\rho)}{W_1^{ASIA \text{ of } \mathcal{S}}(\rho)}, \frac{W_2^{\mathcal{S}}(\rho)}{W_2^{ASIA \text{ of } \mathcal{S}}(\rho)} \right). \quad (8)$$

One can verify that $\alpha^1(\rho)$ also has similar properties as the intrinsic ratio, e.g., smooth, monotone and nearly linear. In this setting, the queue time of the original tandem queue can be approximated by

$$QT^{\mathcal{S}}(\rho) \cong F^1(\rho) \circ W^{ASIA \text{ of } \mathcal{S}}(\rho), \quad (9)$$

where $F^1(\rho)$ is an estimate for $\alpha^1(\rho)$. However, although this approach can also provide an approximation for tandem queues, the coefficient vector doesn't satisfy Properties 7.5 and 7.6. Hence the relative errors produced by formula (9) wouldn't have the light and heavy traffic properties as shown in Propositions 5.1 and 5.2. This implies that, the performance of the approximation algorithm is dependent on the selection of auxiliary

systems. In practical application, the auxiliary systems should be determined deliberately and insightfully to achieve high-quality approximations.

7.5.2 Case Studies for Homogeneous-Server Tandem Queues

In the following, we will try to approximate the mean queue time of a tandem queue with homogeneous servers, where $n_1 = n_2 = 2$, $E[S_1] = 50$ and $E[S_2] = 60$. The arrival process is Poisson and the service times are Gamma distributed. We consider two cases: (i) $c_{S_1}^2 = c_{S_2}^2 = 0.5$, and (ii) $c_{S_1}^2 = c_{S_2}^2 = 5$.

According to the properties of $IR_2(\rho)$, we employ a linear function $F_2(\rho)$ to fit $IR_2(\rho)$ based on the value of $IR_2(0.3)$ and $IR_2(0.8)$, i.e.,

$$F_2(\rho) = \frac{IR_2(0.8) - IR_2(0.3)}{0.8 - 0.3} \rho + \frac{0.8IR_2(0.3) - 0.3IR_2(0.8)}{0.8 - 0.3}.$$

In this approximation, $IR_2(0.3)$ and $IR_2(0.8)$ are calculated based on the simulation results. Based on the linear function $F_2(\rho)$, the Approximated Queue Time (AQT) at station 2 is given by $F_2(\rho)W_2^{ASIA\ of\ S}(\rho) + (1 - F_2(\rho))W_2^{BSIA\ of\ S}(\rho)$.

Performance of the approximations are presented in Tables 7-2 and 7-3, where the relative errors (Diff%) are given by $(AQT-SQT)/SQT$, and SQT is the Simulation Queue Time at station 2. For the simulation, 30 replications are performed for each case. Each replication consists of 2,000,000 jobs after disregarding the first 4,000,000 jobs for a warm-up. The 90% half-width confidence interval is within 2% of the mean queue times.

As shown in Tables 7-2 and 7-3, the approximation algorithm based on the affine structure produces accurate approximations for tandem queues with homogeneous servers.

Table 7-2 Approximation errors when $c_{s_1}^2 = c_{s_2}^2 = 0.5$.

Utilization	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%
SQT	0.45	1.82	4.23	8.00	13.79	23.01	39.15	72.57	177.79	395.54
$IR_2(\rho)$	0.89	0.86	0.85	0.83	0.81	0.79	0.77	0.75	0.70	0.64
$F_2(\rho)$	0.89	0.87	0.85	0.83	0.81	0.79	0.77	0.75	0.73	0.72
AQT	0.45	1.82	4.23	7.98	13.78	23.01	39.09	72.57	179.04	400.08
Diff%	-0.20%	0.08%	-----	-0.26%	-0.08%	0.02%	-0.15%	-----	0.70%	1.15%

Table 7-3 Approximation errors when $c_{s_1}^2 = c_{s_2}^2 = 5$.

Utilization	10%	20%	30%	40%	50%	60%	70%	80%	90%	95%
SQT	1.67	7.18	17.73	35.57	64.58	111.75	195.94	367.63	866.43	1823.36
$IR_2(\rho)$	1.05	1.09	1.13	1.19	1.23	1.28	1.35	1.45	1.55	1.63
$F_2(\rho)$	1.00	1.06	1.13	1.19	1.25	1.32	1.38	1.45	1.51	1.54
AQT	1.62	7.08	17.73	35.66	65.22	113.47	198.01	367.63	858.36	1801.39
Diff%	-2.87%	-1.41%	-----	0.25%	0.99%	1.54%	1.06%	-----	-0.93%	-1.20%

Remark. Since this case study is designed to examine the errors contributed by $F_2(\rho)$, we assume $W_2^{ASIA\ of\ S}(\rho)$ and $W_2^{BSIA\ of\ S}(\rho)$ are known and employ the results from simulations. In practice, $W_2^{ASIA\ of\ S}(\rho)$ and $W_2^{BSIA\ of\ S}(\rho)$ can be computed through the approximations for GI/G/n queues (Whitt 1993).

Tables 7-4 and 7-5 further compare the performance of the approximation based on affine structure and the decomposition method employed by Hopp et al. (2002). As shown in the tables, the proposed new algorithm outperforms the traditional decomposition approach.

Table 7-4 Performance of the affine structure algorithm and the decomposition method

employed by Hopp et al. (2002) when $c_{s_1}^2 = c_{s_2}^2 = 0.5$.

Utilization	SQT	Affine Structure		Hopp et al. (2002)	
		AQT	Diff%	AQT	Diff%
10.00%	0.45	0.45	-0.20%	0.52	15.50%
20.00%	1.82	1.82	0.08%	2.12	16.36%
30.00%	4.23	4.23	---	4.73	11.87%
40.00%	8.00	7.98	-0.26%	8.76	9.46%
50.00%	13.79	13.78	-0.08%	14.85	7.66%
60.00%	23.01	23.01	0.02%	24.30	5.63%
70.00%	39.15	39.09	-0.15%	40.25	2.80%
80.00%	72.57	72.57	---	72.08	-0.69%
90.00%	177.79	179.04	0.70%	166.81	-6.18%
95.00%	395.54	400.08	1.15%	355.37	-10.16%

Table 7-5 Performance of the affine structure algorithm and the decomposition method

employed by Hopp et al. (2002) when $c_{s_1}^2 = c_{s_2}^2 = 5$.

Utilization	SQT	Affine Structure		Hopp et al. (2002)	
		AQT	Diff%	AQT	Diff%
10.00%	1.67	1.62	-2.87%	1.22	-26.78%
20.00%	7.18	7.08	-1.41%	5.17	-28.01%
30.00%	17.73	17.73	---	13.24	-25.29%
40.00%	35.57	35.66	0.25%	28.16	-20.84%
50.00%	64.58	65.22	0.99%	54.41	-15.75%
60.00%	111.75	113.47	1.54%	100.63	-9.96%
70.00%	195.94	198.01	1.06%	186.73	-4.70%
80.00%	367.63	367.63	---	372.45	1.31%
90.00%	866.43	858.36	-0.93%	956.39	10.38%
95.00%	1823.36	1801.39	-1.20%	2144.33	17.60%

7.6 Summary

Except for several special cases, the performance evaluation of queueing systems is hard in general. In this chapter, we have introduced an interesting phenomenon about the mean

queue times of general queueing systems: the mean queue time of a general queueing system has an affine relationship with its corresponding systems. In summary, the mean queue time of a system $\mathcal{S}$ can be represented as $\sum_{l=1}^L \alpha^l(\rho) \circ W^{\mathcal{S}^l}(\rho)$, where $\{W^{\mathcal{S}^l}(\rho), l = 1, 2, \dots, L\}$ are the mean queue times of the corresponding systems and $\{\alpha^l(\rho), l = 1, 2, \dots, L\}$ are the coefficient vectors. Therefore, the approximation for mean queue times will reduce to using functions $\{F^l(\rho), l = 1, 2, \dots, L\}$ to estimate the coefficient vectors $\{\alpha^l(\rho), l = 1, 2, \dots, L\}$ according to the affine structure.

If the corresponding systems are selected deliberately, the coefficient vectors can enjoy some nice properties, e.g., the IR converges to a constant in light traffic for single-server tandem queues. In these situations, the coefficient vectors can be estimated easily and the approximation algorithm will produce accurate estimates. In tandem queues, simply using a constant to approximate IR will provide reliable approximations for the system mean queue time (Wu and McGinnis 2013). In practice, we can get the value of the coefficient vectors at some specific utilization levels based on historical data or simulations. In this case, we can estimate the coefficient vectors at other utilization levels through approximations (e.g., linear regression). Furthermore, we proposed a conjecture (see Section 7.3) regarding the affine structure based on the simulation results. If the conjecture can be proved, we may approximate the coefficient vectors more accurately or even exactly. We have also shown that the proposed concept can be applied to more complex queueing models.

Of course, the key step of the approximation algorithm is to define the appropriate auxiliary systems. On one hand, the mean queue time of the auxiliary systems should be

easy to estimate. On the other hand, the coefficient vectors with respect to the auxiliary systems should enjoy some nice properties, e.g., smooth, bounded or some limiting behaviors, to make them easy to approximate. As shown by the analysis in this chapter, the ASIA and BSIA systems are reliable auxiliary systems for tandem queues. How to choose the auxiliary systems for more general queueing systems, e.g., queueing systems with multiple products, is left for future research.

The affine structure proposed in this chapter implies that, to capture the trade-off between the throughput and mean queue time for a queueing system for $\rho \in (0,1)$, we only need the performance at several specific utilizations levels. Utilizing the concept, it will save much effort for the performance evaluation of complex systems. The approximation algorithm can support the cycle time estimation at different utilizations in the capacity planning for manufacturing systems.

Chapter 8 Conclusions

8.1 Contributions of This Study

This thesis studies the performance evaluation of queueing system to support the management and improvement of production systems. We consider two important performance measures, stability and mean queue time. For each measure, we start from the simplest GI/G/1 queues and then move on to the general queueing networks, where the dependence within the system has a significant impact as introduced in Chapter 2.

Chapter 4 has surveyed the stability of single-server queues. We summarized the underlying stochastic processes commonly employed to describe the dynamics of queueing systems. Furthermore, we clarified the relations among different types of stability, which usually exist in literature. The survey presents an overall picture for the stability of GI/G/1 queues. It also serves as a theoretical foundation for the study of queueing networks.

Chapter 5 studied the stability of queueing networks, where the dependence among stations significantly influences the system performance. A queueing network can be unstable even if the traffic intensity of each physical station is less than one. In this study, we have clarified the concept of servers in the context of queueing network. Under the new framework, the stability of a queueing network is also dominated by the traffic intensity of general servers, which is consistent with the case of GI/G/1 queues. With the concept, we have shown that, in practice, queueing systems operating under the WIP-dependent dispatching policies are always stable if all physical stations have sufficient capacity.

Chapter 6 studied another performance measure of queueing systems, i.e., mean queue time. Most of the conventional queue time approximations for GI/G/1 queues only considered the first two moments of interarrival and service times. To improve the accuracy, we have considered the third moment and proposed a unified approximation formula in the chapter. The three-moment approximation has been validated through extensive case studies and provides more accurate estimates for single-server systems.

Chapter 7 studied the performance of mean queue times in general queueing systems. Rather than making additional assumptions to simplify the analysis, we characterize the mean queue time of a queueing system through comparing it with some auxiliary systems. It has been shown that the affine structure, which may come with some nice properties, exists in general queueing systems. Based on the affine structure, we proposed a unified approximation algorithm for general queueing systems without independence or heavy traffic assumption. The approximation algorithm is easy to apply and can produce reliable estimations.

8.2 Limitations and Future Research

8.2.1 Identification of General Servers

In Chapter 5, we have shown that the stability of a multiclass queueing network is fully dominated by the utilizations of general servers. However, the structure of general servers is not elegant. It depends on various factors, e.g., service disciplines and job routings, and the identification of general servers is hard. In future research, we will further investigate the

properties of general servers and try to derive a general algorithm to support the identifying of general servers.

8.2.2 Analytical Results for Approximation Errors

The validation for queue time approximations highly relies on simulation studies in literature. In this thesis, besides the simulations, we have analytically proven the limiting properties of approximation errors in light and heavy traffic. However, the results are limited to tandem queues. Whether the results still hold for more general queueing systems remains to be investigated. Furthermore, the upper bounds of the approximation errors (including the ones in literature) are not available. Besides simulations, it is worthy to investigate the upper bounds analytically to reinforce the proposed framework.

Furthermore, in some practical cases, the service times can also be dependent on other factors (e.g., see Wolff (1982), Sandmann (2010), Sandmann (2012) and Pang and Whitt (2012)). Whether the algorithm in Section 7.4 can be generalized to these settings need to be further verified.

8.2.3 The Relationship between the Traffic Intensity and Mean Queue Time

In stochastic settings, the mean queue time of a system increases with respect to the traffic intensity. For an M/M/1 queue, the relationship can be exactly characterized by the term $\rho/(1 - \rho)$. For general single-server queues, the relationship can also be well characterized through modifying the parameters as shown by equation (8) in Chapter 6.

However, the study on general queueing systems in this thesis is limited to the

approximation algorithms. Although one can exploit the approach in Chapter 7 to evaluate the performance for general queueing systems, the overall relationship between the traffic intensity and mean queue time is not clearly presented. To make the study more complete, we will further investigate the relationship in future.

8.2.4 Capacity Planning

This thesis is limited to the performance evaluation for queueing systems. In future research, we will apply this study to the capacity planning for semiconductor fabs, which involves utilizing a minimum cost to achieve the required performance (i.e., the throughput and cycle time constraints, see Hopp et al. (2002)). The techniques in this study can be employed to evaluate the performance of the semiconductor fabs. Therefore, the constraints (i.e., Conditions (a) and (b) in Section 1.1) of the optimization problem can be established more accurately and it will help to figure out better decisions.

References

- Adan, I. and Resing, J. (2002) Queueing theory, Eindhoven University of Technology Eindhoven.
- Altman, E., Foss, S., Riehl, E. and Stidham Jr, S. (1998) Performance bounds and pathwise stability for generalized vacation and polling systems. *Operations research*, 46, 137-148.
- Asmussen, S. (2008) Applied probability and queues, Springer Science & Business Media.
- Avi-Itzhak, B. (1965) A sequence of service stations with arbitrary input and regular service times. *Management Science*, 11, 565-571.
- Banks, J. and Dai, J. (1997) Simulation studies of multiclass queueing networks. *IIE Transactions*, 29, 213-219.
- Baskett, F., Chandy, K.M., Muntz, R.R. and Palacios, F.G. (1975) Open, closed, and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22, 248-260.
- Bharucha-Reid, A.T. (2012) Elements of the theory of markov processes and their applications, Courier Corporation.
- Billingsley, P. (2013) Convergence of probability measures, John Wiley & Sons.
- Borovkov, A. (2012) Stochastic processes in queueing theory, Springer Science & Business Media.

References

- Boxma, O. (1979a) On a tandem queueing model with identical service times at both counters, I. *Advances in Applied Probability*, 11, 616-643.
- Boxma, O. (1979b) On a tandem queueing model with identical service times at both counters, II. *Advances in Applied Probability*, 11, 644-659.
- Bramson, M. (1994) Instability of FIFO queueing networks with quick service times. *The Annals of Applied Probability*, 693-718.
- Bramson, M. (2011) Stability of join the shortest queue networks. *The Annals of Applied Probability*, 21, 1568-1625.
- Bramson, M. (2008) Stability of queueing networks. *Probability Surveys*, 5, 169-345.
- Bramson, M. (1999) A stable queueing network with unstable fluid model. *The Annals of Applied Probability*, 818-853.
- Braverman, A., Dai, J. and Miyazawa, M. (2017) Heavy traffic approximation for the stationary distribution of a generalized jackson network: the BAR approach. *Stochastic Systems*, 7, 143-196.
- Burke, P.J. (1956) The output of a queuing system. *Operations Research*, 4, 699-704.
- Carr, M. and Hajek, B. (1993) Scheduling with asynchronous service opportunities with applications to multiple satellite systems. *IEEE Transactions on Automatic Control*, 38, 1820-1833.

References

- Chen, H. (1995) Fluid approximations and stability of multiclass queueing networks: work-conserving disciplines. *The Annals of Applied Probability*, 637-665.
- Chen, H. and Yao, D.D. (2013) *Fundamentals of queueing networks: Performance, asymptotics, and optimization*, Springer Science & Business Media.
- Chung, K.L. (2001) *A course in probability theory*, Academic Press.
- Chung, K.L. and Donsker, M.D. (1951) *Four papers on probability*, American Mathematical Society.
- Cosmetatos, G.P. (1982) On the implementation of Page's approximation for waiting times in general multi-server queues. *The Journal of the Operational Research Society*, 33, 1158-1159.
- Courcoubetis, C., Konstantopoulos, P., Walrand, J. and Weber, R.R. (1989) Stabilizing an uncertain production system. *Queueing Systems*, 5, 37-54.
- Courcoubetis, C. and Weber, R. (1994) Stability of flexible manufacturing systems. *Operations Research*, 42, 947-957.
- Dai, J. (2009) *Fluid models and Stability of Multiclass Queueing Networks*. Technical Report, Cornell University.
- Dai, J. and Harrison, J.M. (1992) Reflected Brownian motion in an orthant: numerical methods for steady-state analysis. *The Annals of Applied Probability*, 65-86.

References

- Dai, J., Hasenbein, J.J. and Vate, J.H.V. (2004) Stability and instability of a two-station queueing network. *The Annals of Applied Probability*, 326-377.
- Dai, J., Hasenbein, J.J. and Vate, J.V. (1999) Stability of a three - station fluid network. *Queueing Systems*, 33, 293-325.
- Dai, J. and Vande Vate, J. (2000) The stability of two-station multitype fluid networks. *Operations Research*, 48, 721-744.
- Dai, J. and Vande Vate, J.H. (1996) Global stability of two-station queueing networks, in *Stochastic Networks*, Springer, pp. 1-26.
- Dai, J.G. (1995) On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *The Annals of Applied Probability*, 49-77.
- Dai, J.G. and Lin, W. (2005) Maximum pressure policies in stochastic processing networks. *Operations Research*, 53, 197-218.
- Dai, J.G. and Weiss, G. (1996) Stability and instability of fluid models for reentrant lines. *Mathematics of Operations Research*, 21, 115-134.
- Davis, M.H. (1984) Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 353-388.
- Donsker, M.D. (1951) An invariance principle for certain probability limit theorems. *AMS*, 6, 1-12.

References

- Donsker, M.D. (1952) Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *The Annals of Mathematical Statistics*, 277-281.
- Down, D. and Meyn, S. (1994) A survey of Markovian methods for stability of networks, in 11th International Conference on Analysis and Optimization of Systems Discrete Event Systems, Springer, pp. 490-504.
- Dumas, V. (1997) A multiclass network with non-linear, non-convex, non-monotonic stability conditions. *Queueing Systems*, 25, 1-43.
- El-Taha, M. and Stidham, S. (1992) Deterministic analysis of queueing systems with heterogeneous servers. *Theoretical Computer Science*, 106, 243-264.
- El-Taha, M. and Stidham, S. (2012) *Sample-path analysis of queueing systems*, Springer Science & Business Media.
- El-Taha, M. and Stidham, S. (1993) Sample-path analysis of stochastic discrete-event systems. *Discrete Event Dynamic Systems*, 3, 325-346.
- Erlang, A.K. (1909) The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik B*, 20, 87-98.
- Feller, W. (2008) *An introduction to probability theory and its applications*, John Wiley & Sons.
- Friedman, H.D. (1965) Reduction methods for tandem queueing systems. *Operations Research*, 13, 121-131.

References

- Gelenbe, E. (1975) On approximate computer system models. *Journal of the ACM*, 22, 261-269.
- Goldratt, E.M. and Cox, J. (2016) *The goal: a process of ongoing improvement*, Routledge.
- Gupta, V., Harchol-Balter, M., Dai, J. and Zwart, B. (2010) On the inapproximability of M/G/K: why two moments of job size distribution are not enough. *Queueing Systems*, 64, 5-48.
- Harrison, J.M. and Nguyen, V. (1990) The QNET method for two-moment analysis of open queueing networks. *Queueing Systems*, 6, 1-32.
- Hasenbein, J.J. (1997) Necessary conditions for global stability of multiclass queueing networks. *Operations Research Letters*, 21, 87-94.
- Heyman, D.P. (1975) A diffusion model approximation for the GI/G/1 queue in heavy traffic. *Bell Labs Technical Journal*, 54, 1637-1646.
- Hirsch, M.W., Smale, S. and Devaney, R.L. (2012) *Differential equations, dynamical systems, and an introduction to chaos*, Academic Press.
- Hopp, W.J. and Spearman, M.L. (2011) *Factory physics*, Waveland Press.
- Hopp, W.J., Spearman, M.L., Chayet, S., Donohue, K.L. and Gel, E.S. (2002) Using an optimized queueing network model to support wafer fab design. *IIE Transactions*, 34, 119-130.

References

- Hulett, M. and Damodaran, P. (2011) Analytical approximations to predict performance measures of markovian type manufacturing systems with job failures and parallel processing. *European Journal of Operational Research*, 212, 89-99.
- Iglehart, D.L. and Whitt, W. (1970) Multiple channel queues in heavy traffic. II: Sequences, networks, and batches. *Advances in Applied Probability*, 2, 355-369.
- Jackson, J.R. (1957) Networks of waiting lines. *Operations Research*, 5, 518-521.
- Kelly, F.P. (2011) *Reversibility and stochastic networks*, Cambridge University Press.
- Kendall, D.G. (1953) Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, 338-354.
- Khinchin, A. (1932) Mathematisches über die Erwartung vor einem öffentlichen Schalter. *Rec. Math*, 39, 72-84.
- Kimura, T. (1985) Refining diffusion approximations for GI/G/1 queues: a tight discretization method, Institute of Technology, Department of Information Sciences.
- Kimura, T. (1986) A two-moment approximation for the mean waiting time in the GI/G/s queue. *Management Science*, 32, 751-763.
- Kingman, J. (1962) On queues in heavy traffic. *Journal of the Royal Statistical Society. Series B (Methodological)*, 383-392.

References

- Kleinrock, L. (1976) Queueing systems, volume 2: Computer applications, Wiley New York.
- Kobayashi, H. (1974) Application of the diffusion approximation to queueing networks I: Equilibrium queue distributions. *Journal of the ACM*, 21, 316-328.
- Konheim, A.G. (1975) An elementary solution of the queueing system G/G/1. *SIAM Journal on Computing*, 4, 540-545.
- Kraemer, W. and Langenbach-Belz, M. (1976) Approximate formulae for the delay in the queueing system GI/G/1, in *Proceedings of the 8th International Teletraffic Congress*, pp. 235-1.
- Kumar, P. (1993) Re-entrant lines. *Queueing Systems*, 13, 87-110.
- Kumar, P. and Seidman, T.I. (1990) Dynamic instabilities and stabilization methods in distributed real-time scheduling of manufacturing systems. *IEEE Transactions on Automatic Control*, 35, 289-298.
- Lindley, D.V. (1952) The theory of queues with a single server, in *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, pp. 277-289.
- Little, J.D. (1961) A proof for the queueing formula: $L = \lambda W$. *Operations Research*, 9, 383-387.

References

- Loynes, R. (1965) On the waiting-time distribution for queues in series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 491-496.
- Loynes, R.M. (1962) The stability of a queue with non-independent inter-arrival and service times, in *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, pp. 497-520.
- Lu, S.H. and Kumar, P. (1991) Distributed scheduling based on due dates and buffer priorities. *IEEE Transactions on Automatic Control*, 36, 1406-1416.
- Maglaras, C. (1999) Dynamic scheduling in multiclass queueing networks: Stability under discrete-review policies. *Queueing Systems*, 31, 171-206.
- Marchal, W.G. (1976) An approximate formula for waiting time in single server queues. *AIIE Transactions*, 8, 473-474.
- Marshall, K.T. (1968) Some inequalities in queuing. *Operations Research*, 16, 651-668.
- Mazumdar, R.R., Guillemin, F., Badrinath, V. and Kannurpatti, R. (1992) On pathwise behavior of queues. *Operations Research Letters*, 12, 263-270.
- Meyn, S.P. (1995) Transience of multiclass queueing networks via fluid limit models. *The Annals of Applied Probability*, 946-957.
- Meyn, S.P. and Tweedie, R.L. (2012) *Markov chains and stochastic stability*, Springer Science & Business Media.

References

- Morozov, E. and Delgado, R. (2009) Stability analysis of regenerative queueing systems. Automation and Remote control, 70, 1977-1991.
- Mukhopadhyay, A. and Mazumdar, R.R. (2016) Analysis of randomized join-the-shortest-queue (JSQ) schemes in large heterogeneous processor-sharing systems. IEEE Transactions on Control of Network Systems, 3, 116-126.
- Myskja, A. (1991) An improved heuristic approximation for the GI/GI/1 queue with bursty arrivals, in Proceedings of the Teletraffic and Datatraffic in a Period of Change, ITC-13, Elsevier Science Publishers B.V. (North Holland), City, pp. 683-688.
- Myskja, A. (1990) On approximations for the GI/GI/1 queue. Computer Networks and ISDN Systems, 20, 285-295.
- Neuts, M.F. (1981) Matrix-geometric solutions in stochastic models: an algorithmic approach, Courier Corporation.
- Page, E. (1982) Tables of waiting times for M/M/n, M/D/n and D/M/n and their use to give approximate waiting times in more general queues. Journal of the Operational Research Society, 453-473.
- Pang, G. and Whitt, W. (2012) The impact of dependent service times on large-scale service systems. Manufacturing & Service Operations Management, 14, 262-278.
- Pollaczek, F. (1930a) Über eine aufgabe der wahrscheinlichkeitstheorie. I. Mathematische Zeitschrift, 32, 64-100.

References

- Pollaczek, F. (1930b) Über eine aufgabe der wahrscheinlichkeitstheorie. II. Mathematische Zeitschrift, 32, 729-750.
- Rolski, T. (1981) Queues with non-stationary input stream: Ross's conjecture. Advances in Applied Probability, 13, 603-618.
- Ross, S.M. (1996) Stochastic processes. 1996, Wiley, New York.
- Royden, H.L. and Fitzpatrick, P. (1988) Real analysis, Macmillan New York.
- Rybko, A.N. and Stolyar, A.L. (1992) Ergodicity of stochastic processes describing the operation of open queueing networks. Problemy Peredachi Informatsii, 28, 3-26.
- Sandmann, W. (2012) Delays in a series of queues with correlated service times. Journal of Network and Computer Applications, 35, 1415-1423.
- Sandmann, W. (2010) Delays in a series of queues: Independent versus identical service times, in 2010 IEEE Symposium on Computers and Communications (ISCC), IEEE, pp. 32-37.
- Seidman, T.I. (1994) " First come, first served" can be unstable! IEEE Transactions on Automatic Control, 39, 2166-2171.
- Shanthikumar, J. and Buzacott, J. (1980) On the approximations to the single server queue. International Journal of Production Research, 18, 761-773.
- Shen, Y. and Wu, K. (2018) Stability of a GI/G/1 Queue: A Survey. Asia-Pacific Journal of Operational Research, <https://doi.org/10.1142/S021759591850015X>.

References

- Sherman, N.P., Kharoufeh, J.P. and Abramson, M.A. (2009) An M/G/1 retrial queue with unreliable server for streaming multimedia applications. *Probability in the Engineering and Informational Sciences*, 23, 281-304.
- Sigman, K. (1988) Queues as Harris recurrent Markov chains. *Queueing Systems*, 3, 179-198.
- Szczotka, W. (1986a) Stationary representation of queues. I. *Advances in Applied Probability*, 18, 815-848.
- Szczotka, W. (1986b) Stationary representation of queues. II. *Advances in Applied Probability*, 18, 849-859.
- Whitt, W. (1993) Approximations for the GI/G/m queue. *Production and Operations Management*, 2, 114-161.
- Whitt, W. (1985) The best order for queues in series. *Management Science*, 31, 475-487.
- Whitt, W. (1983) The queueing network analyzer. *Bell Labs Technical Journal*, 62, 2779-2815.
- Wolff, R.W. (1989) *Stochastic modeling and the theory of queues*, Pearson College Division.
- Wolff, R.W. (1982) Tandem queues with dependent service times in light traffic. *Operations Research*, 30, 619-635.

References

- Wu, K. (2014) Classification of queueing models for a workstation with interruptions: a review. *International Journal of Production Research*, 52, 902-917.
- Wu, K. and McGinnis, L. (2013) Interpolation approximations for queues in series. *IIE Transactions*, 45, 273-290.
- Wu, K. and McGinnis, L. (2012) Performance evaluation for general queueing networks in manufacturing systems: Characterizing the trade-off between queue time and utilization. *European Journal of Operational Research*, 221, 328-339.
- Wu, K., Shen, Y. and Zhao, N. (2017a) Analysis of tandem queues with finite buffer capacity. *IIE Transactions*, 1-13.
- Wu, K., Srivathsan, S. and Shen, Y. (2017b) Three-moment approximation for the mean queue time of a GI/G/1 queue. *IIE Transactions*, 50 (2), 63-73.
- Xu, J., Huang, E., Chen, C.-H. and Lee, L.H. (2015) Simulation optimization: A review and exploration in the new era of cloud computing and big data. *Asia-Pacific Journal of Operational Research*, 32, 1-34.
- Yu, P.S. (1977) On Accuracy improvement and applicability conditions of diffusion approximation with applications to modelling of computer systems, Stanford University, California Stanford Electronics Labs.

Appendix A1

Markov Process and Positive Harris Recurrence. Notations and definitions follow (Bharucha-Reid 2012) and (Bramson 2008). Let $\{X(t), t \geq 0\}$ be a right-continuous process defined on $(X, \mathcal{B}(X))$, where X is the state space and $\mathcal{B}(X)$ is the Borel σ -algebra of X .

Definition A.1 $\{X(t), t \geq 0\}$ is a Markov process with transition probability kernel $P^t(x, A) = P_x(X(t) \in A)$, if for any $t > s$ and $A \in \mathcal{B}(X)$,

$$P(X(t) \in A | \mathcal{F}_s) = P(X(t) \in A | X(s)),$$

where $\mathcal{F}_s := \sigma(X(u), 0 \leq u \leq s)$ is the natural filtration.

Definition A.2 $\tau(A) := \inf \{t \geq 0, X(t) \in A\}$ is called the first hitting time of A .

Definition A.3 The Markov process $\{X(t), t \geq 0\}$ is Harris recurrent if for some non-trivial σ -finite measure φ , $\varphi(A) > 0$ implies $P_x(\tau(A) < \infty) = 1$ for every $x \in X$. Furthermore, it is called positive Harris recurrent if there exists an invariant probability measure π such that

$$\pi(A) = \int P^t(x, A) \pi(dx) \text{ for all } A \in \mathcal{B}(X).$$

Input-Output Process and Rate Stability. The notations and definitions are based on (El-Taha and Stidham 2012). Let $Q(t) = Q(0) + A(t) - D(t)$ denote the evolution of a general input-output system. $A(t)$ is the cumulative input to the system till time t , $D(t)$ is the cumulative output from the system till time t , and $Q(0)$ is the initial state of the system. $\{A(t), t \geq 0\}$, $\{D(t), t \geq 0\}$, and $\{Q(t), t \geq 0\}$ are all right-continuous processes

with state space R^+ (non-negative real numbers). Moreover, $\{A(t), t \geq 0\}$ and $\{D(t), t \geq 0\}$ are non-decreasing and $A(t) + Q(0) \geq D(t)$ holds for any $t \geq 0$.

Definition A.4 An input-output process $\{Q(t), t \geq 0\}$ is said to be (rate) stable if $\lim_{t \rightarrow \infty} Q(t)/t = 0$.

Rate stability is defined on a deterministic input-output system. In the context of stochastic setting, $\{Q(t), t \geq 0\}$ is a fixed sample-path of the stochastic process.

Appendix A2

Proof of Lemma 5.1. Consider any $\alpha \in [0.8, 0.9)$ and assume the 1st job arrives at class 1 at $t = 0$. Since the interarrival times are constant, the state $X(t)$ of the system at time t reduces to a two-tuple $X(t) = (Q(t), V(t))$. When $t = 0$, there is $X(0) = ((0, 0.2), (0, 0), (0, 0), (0, 0))$. At $t = \alpha$, the 1st job is still at class 3, and therefore $X(\alpha) = ((0, 0.2), (0, 0), (0, 0.9 - \alpha), (0, 0))$. At $t = 0.9$, the 1st job arrives at class 4 but needs to wait since the 2nd job is still receiving service at class 1. Hence, $X(0.9) = ((0, \alpha - 0.7), (0, 0), (0, 0), (1, 0))$. At $t = 2\alpha$, the 3rd job arrives, the 1st job has leaved the network, and the 2nd job is at class 3. Thus, $X(2\alpha) = ((0, 0.2), (0, 0), (0, 0.9 - \alpha), (0, 0))$. Note that the state of the network at time 2α is identical to the state at time α . Since the interarrival times and services times are both constant, $X(n\alpha) = ((0, 0.2), (0, 0), (0, 0.9 - \alpha), (0, 0))$ for any integer $n \geq 1$ and the evolutions in the interval $((n - 1)\alpha, n\alpha)$ are the same. In addition, the time epoch that the n th job begins to receive service at class 2 is $n\alpha + 0.2$, and so is the time epoch that the $(n - 1)$ th job begins to receive service at class 4. Therefore, except for the first job to initiate the system, classes 2 and 4 always process jobs simultaneously. Q.E.D.

Proof of Lemma 5.2. Let the arrival and departure times of the n th job at class i be A_n^i and D_n^i . Assume there are 3 jobs between classes 2 and 4, i.e., the n th, $(n + 1)$ th and $(n + 2)$ th jobs. There can be the following situations:

(a) The $(n + 1)$ th job at class 2 and the n th job at class 4 are processed simultaneously and the $(n + 2)$ th job is waiting at class 2. Since the $(n + 2)$ th job will block the $(n + 1)$ th job, there is no job receiving service at class 4 while the $(n + 2)$ th job is receiving service at class 2, i.e., class 4 is vacant during the interval (A_{n+2}^2, D_{n+2}^2) . Therefore, the synchronization breaks.

(b) The $(n + 2)$ th job at class 2 and the n th job at class 4 are processed simultaneously and the $(n + 1)$ th job is waiting at either class 3. After D_{n+2}^2 , the $(n + 1)$ th job will arrive at class 4 at time $D_{n+2}^2 + 0.1$ and A_{n+2}^4 will be $D_{n+2}^2 + 0.2$. Since at most one job (i.e., the $(n + 3)$ th job) can be at class 2 at time $D_{n+2}^2 + 0.2$, classes 2 and 4 cannot process jobs simultaneously at time $D_{n+2}^2 + 0.8$ and the synchronization breaks.

(c) The $(n + 2)$ th job at class 2 and the n th job at class 4 are processed simultaneously and the $(n + 1)$ th job waits at class 4. the $(n + 1)$ th job will block class 1 after the n th job completes its service, i.e., D_n^4 . Hence, classes 2 and 4 cannot process jobs simultaneously and the synchronization breaks.

Similar arguments can be applied when there are more than three jobs between classes 2 and 4. Q.E.D.

Proof of Lemma 5.3. Since the network is synchronized when $0.8 \leq \alpha < 0.9$ and $1.1 \leq \alpha < 1.2$, we will evaluate if the network is synchronized when $0.9 \leq \alpha < 1.1$.

By Lemma 5.2, there are at most two jobs between class 2 and class 4 in the Lu-Kumar network when $\{2, 4\}$ is synchronized. Classes 2 and 4 are synchronized if and only if both always are busy at the same time (except for the first job to initialize the system). Hence, if

the network is synchronized, the relations of the two jobs can be one of the following two:

(a) One job may wait at class 4: $A_n^4 \leq D_{n+1}^1$ (two consecutive jobs can be served at the same time), $A_{n+1}^1 < A_n^4$, $A_{n+1}^2 \geq D_n^3$ and $A_n^3 < A_{n+1}^2$ (no consecutive job can be blocked at classes 1, 2 and 3), or

(b) One job may wait at class 2: $A_{n+1}^2 \leq D_n^3$ (two consecutive jobs can be served at the same time), and $A_{n+1}^1 < A_n^4$, $A_n^3 < A_{n+1}^2$ and $A_{n+1}^4 \geq D_n^1$ (no consecutive job can be blocked at classes 1, 3 and 4).

For (a), since $A_{n+1}^1 < A_n^4 \leq D_{n+1}^1$, $A_{n+1}^1 < A_n^4 \leq A_{n+1}^1 + 0.2$, we have $0 < A_n^4 - A_{n+1}^1 \leq 0.2$. Because no consecutive job can be blocked at classes 1, 2 and 3, $A_n^4 - A_{n+1}^1 = 0.9$. Therefore, $0.7 \leq A_{n+1}^1 - A_n^1 < 0.9$, which is a contradiction to $0.9 \leq \alpha < 1.1$.

For (b), since $A_n^3 < A_{n+1}^2 \leq D_n^3$, $A_n^3 < A_{n+1}^2 \leq A_n^3 + 0.1$, we have $0 < A_{n+1}^2 - A_n^3 \leq 0.1$. Since $m_2 = 0.6$, $m_3 = 0.1$ and the n th and $(n-1)$ th jobs are served at classes 2 and 4 simultaneously, we have $A_n^2 + 0.6 \leq A_n^3 < A_n^2 + 0.7$ (if n th job is blocked by the $(n-1)$ th job at class 2). Hence, $0.6 < A_{n+1}^2 - A_n^2 < 0.8$. Because no consecutive job can be blocked at class 1, $0.6 < A_{n+1}^1 - A_n^1 < 0.8$, which is a contradiction to $0.9 \leq \alpha < 1.1$.

Together with Lemma 5.1, the set $\{2, 4\}$ (i.e., classes 2 and 4) can be synchronized if and only if $0.8 \leq \alpha < 0.9$ and $1.1 \leq \alpha < 1.2$. Q.E.D.

Proof of Theorem 5.1. (i) If a queueing network is stable, then effective traffic intensity of every general server does not exceed one. Proceeding by contradiction, we assume there exists some general server S with M effective classes such that $P_S > 1$. Since $\lim_{t \rightarrow \infty} D_k(t)/t = \lambda_{\tau(k)}$ for any $k \in S$ (due to the pathwise stability), we have $\lim_{t \rightarrow \infty} T_k(t)/$

$t = \lambda_{\tau(k)} m_k$ ($k \in S$), where $T_k(t)$ is the cumulative service time received by class k . Since $P_S > 1$, we have $\sum_{k \in S} \lim_{t \rightarrow \infty} T_k(t)/t = \sum_{k \in S} \lambda_{\tau(k)} m_k > M$. On the other hand, the definition of M implies that $\sum_{k \in S} \lim_{t \rightarrow \infty} T_k(t)/t \leq M$, which contradicts to our earlier conclusion. Hence, the effective traffic intensity of every general server does not exceed one if the network is stable. (ii) If the effective traffic intensity of every general server does not exceed one, then the queueing network is stable. It suffices to show that the corresponding fluid model is weakly stable if the effective traffic intensity of each general server does not exceed one ((Dai 1995), (Chen 1995)). For any $1 \leq j \leq J$ and $1 \leq k \leq K$, the basic fluid equations are given as follows:

$$Q_k(t) = Q_k(0) + A_k(t) - \mu_k T_k(t),$$

$$I_j(t) = t - \sum_{k: \sigma(k)=j} T_k(t),$$

$$Q_k(t) \geq 0,$$

$$T_k(0) = 0,$$

$$I_j(0) = 0, I_j(\cdot) \text{ is non-decreasing, and}$$

$$\dot{I}_j(t) = 0, \text{ when } \sum_{k: \sigma(k)=j} Q_k(t) > 0 \text{ and } I_j(t) \text{ is differentiable at } t,$$

where $A_k(t) = \lambda_{\tau(k)} t$ if k is the first class of type $\tau(k)$, otherwise $A_k(t) = \mu_{k-1} T_{k-1}(t)$. Similar to physical stations, for any general server S with effective number M , we have extra fluid equations: $I_S(t) = Mt - \sum_{k \in S} T_k(t)$ is non-decreasing and $P_S \leq 1$, where P_S is the effective traffic intensity of S . Proceeding by contradiction, we assume there is a fluid allocation which is not weakly stable, i.e., given $Q(0) = 0$, there exists a solution such that $Q(t_0) \neq 0$ for some t_0 . Then there exists a smallest k^* and a pair of t_1

and t_2 less than t_0 such that $Q(t) = 0$ for $t \in (0, t_1)$, $Q_{k^*}(t) > 0$ for $t \in (t_1, t_2]$.

Since the service policy is work-conserving, for at least one general server S^* which contains k^* , we have $\sum_{k \in S^*} (T_k(t_2) - T_k(t_1)) = EF(S^*)(t_2 - t_1)$. Furthermore, it follows that

$$\sum_{k \in S^*} \lambda_{\tau(k)}(t_2 - t_1)m_k - EF(S^*)(t_2 - t_1) = \sum_{k \in S^*} \lambda_{\tau(k)}(t_2 - t_1)m_k - \sum_{k \in S^*} (T_k(t_2) - T_k(t_1)) \geq \sum_{k \in S^*} Q_k(t_2)m_k > 0,$$

which implies that $P_{S^*} > 1$. This is a contradiction to the given condition. Hence, the fluid model is weakly stable and the network is pathwise stable. Q.E.D.

Remark. Since general servers are defined for a queueing network under a given service discipline and the structure of general servers also depends on the service discipline, general servers will include the information of the service discipline in some sense. Hence, in this proof, we don't need the extra fluid equation corresponding to the service discipline (e.g., FIFO policy). Considering the equations corresponding to the general servers under the given service discipline is sufficient.

Proof of Theorem 5.2. Based on Theorem 5.1, it suffices to show that if the effective traffic intensity of every compact server does not exceed one, it is also the case for general servers. Considering any general server S with M effective classes, we have the following four situations:

- (i) S itself is a compact server, then $P_S \leq 1$ trivially.
- (ii) S is a subset of a compact server, i.e., $S \subseteq S'$, where S' is a compact server with the same effective number M as S and we have $P_{S'} \leq 1$. For general server S , we have $P_S = L_S/EF(S) \leq L_{S'}/EF(S') = P_{S'} \leq 1$.

(iii) S is the union of compact servers, i.e., $S = \bigcup_{i \in F} S_i$, where each S_i is a compact server, F is an index set and $S_i \cap S_{i'} = \emptyset$ for any $i \neq i'$. We have $EF(S) \geq \sum_{i \in F} EF(S_i)$ because of the compactness of S_i . Then it follows that $P_S = L_S/EF(S) = \sum_{i \in F} L_{S_i}/EF(S) \leq \sum_{i \in F} L_{S_i}/\sum_{i \in F} EF(S_i) \leq 1$.

(iv) $S = (\bigcup_{i \in F} S_i) \cup S_j$, where each S_i is a compact server, F is an index set, $S_i \cap S_{i'} = \emptyset$ for any $i \neq i'$ and S_j is a subset of a compact server. We first consider $(\bigcup_{i \in F} S_i) \cup S_{j'}$, where $S_{j'}$ is a compact server containing S_j and $EF(S_{j'}) = EF(S_j)$. It follows from (iii) that $P_{(\bigcup_{i \in F} S_i) \cup S_{j'}} \leq 1$. Then we have $P_S = L_S/EF(S) \leq L_{(\bigcup_{i \in F} S_i) \cup S_{j'}}/EF(S) = L_{(\bigcup_{i \in F} S_i) \cup S_{j'}}/EF((\bigcup_{i \in F} S_i) \cup S_{j'}) \leq 1$.

Therefore, a queueing network is stable if and only if the effective traffic intensity of every compact server does not exceed one. Q.E.D.

Proof of Theorem 5.3. When there is only one station, i.e., $J = 1$, it is trivial that the physical station is the only compact server. Based on Theorem 5.2, the network is stable since the traffic intensity of the physical station does not exceed one. For a two-station feedforward network, it suffices to show that the effective traffic intensity of any class set $\mathcal{C} = \{c_1, c_2\}$, where $\sigma(c_j) = j$ for $j = 1, 2$, is no greater than one. It is trivial that $P_{\mathcal{C}} \leq 1$ if $m_{c_1} + m_{c_2} \leq 1/\lambda$. Next we will consider $m_{c_1} + m_{c_2} > 1/\lambda$. Since the first station is stable, we have $\lim_{t \rightarrow \infty} T_{c_1}(t)/t = \lambda m_{c_1}$ for station 1, where for any class k , $T_k(t)$ is the cumulative service time received by class k . Since traffic intensity of the second station is no larger than one, we also have $\lim_{t \rightarrow \infty} m\{t | R_{c_2}(t) > 0, \sum_{k \in \text{station } 2, k \neq c_2} R_k(t) = 0\}/t \geq \lambda m_{c_2}$. Since the service policy is work-conserving, we further have

$$\lim_{t \rightarrow \infty} m\{t | \dot{R}_{c_2}(t) > 0\} / t \geq \lambda m_{c_2}.$$

Since $m_{c_1} + m_{c_2} > 1/\lambda$ and $\lim_{t \rightarrow \infty} m\{t | \dot{R}_{c_1}(t) > 0\} / t = \lambda m_{c_1}$, we have

$$\lim_{t \rightarrow \infty} m\{t | \dot{R}_{c_1}(t) \dot{R}_{c_2}(t) \neq 0\} / m\{t | \dot{R}_{c_1}(t) + \dot{R}_{c_2}(t) \neq 0\} > 0.$$

Hence, $\mathcal{C} = \{c_1, c_2\}$ is not a general server and $P_{\mathcal{C}} = (\lambda m_{c_1} + \lambda m_{c_2})/2 \leq 1$. By induction, any feedforward network is stable if the traffic intensity at every physical station does not exceed one. Q.E.D.

Proof of Theorem 5.4. For a two-station reentrant line, it is sufficient to show that the effective traffic intensity of any class set $\mathcal{C} = \{c_1, c_2\}$, where $\sigma(c_j) = j$ for $j = 1, 2$, is no greater than one. It is trivial that $P_{\mathcal{C}} \leq 1$ if $m_{c_1} + m_{c_2} \leq 1/\lambda$. Next we will consider $m_{c_1} + m_{c_2} > 1/\lambda$. Without loss of generality, we assume $c_1 < c_2$. First, we consider $c_1 = 1$. Since class 1 has the highest priority at station 1, we have $\dot{R}_1(t) < 0$ if $R_1(t) > 0$. If $c_2 = 2$, then class 2 has the highest priority at station 2 and $\dot{R}_2(t) < 0$ when $R_2(t) > 0$. Since $m_{c_1} + m_{c_2} > 1/\lambda$, $\lim_{t \rightarrow \infty} m\{t | \dot{R}_{c_1}(t) \dot{R}_{c_2}(t) \neq 0\} / m\{t | \dot{R}_{c_1}(t) + \dot{R}_{c_2}(t) \neq 0\} > 0$. Hence, the set $\mathcal{C}$ is not a general server and $P_{\mathcal{C}} \leq 1$. If $c_2 = 3$ and class 2 belongs to the first station, we still have $\lim_{t \rightarrow \infty} m\{t | \dot{R}_{c_1}(t) \dot{R}_{c_2}(t) \neq 0\} / m\{t | \dot{R}_{c_1}(t) + \dot{R}_{c_2}(t) \neq 0\} > 0$. If $c_2 = 3$ and class 2 belongs to the second station, we have $\lim_{t \rightarrow \infty} m\{t | \dot{R}_2(t) < 0\} / t = \lambda m_2$ since class 2 has the highest priority. Hence, $\lim_{t \rightarrow \infty} m\{t | R_{c_2}(t) > 0, R_2(t) = 0\} / t = \min\{\lambda m_{c_2}, 1 - \lambda m_2\} = \lambda m_{c_2}$, which immediately implies $\lim_{t \rightarrow \infty} m\{t | \dot{R}_{c_1}(t) \dot{R}_{c_2}(t) \neq 0\} / m\{t | \dot{R}_{c_1}(t) + \dot{R}_{c_2}(t) \neq 0\} > 0$. Therefore, $P_{\mathcal{C}}$ is also no greater than one when $c_1 = 1$ and $c_2 = 3$. When $c_2 > 3$, the proof is similar. Hence, the set $\{c_1, c_2\}$ is not a general server and $P_{\mathcal{C}} \leq 1$ when $c_1 = 1$.

When $c_1 = 2$, the arguments are similar because class 1 has the highest priority and we can assume the external arrivals are at class 2. By induction, a two-station reentrant line is stable if the traffic intensity at every physical station does not exceed one. For an J -station reentrant line, the similar analysis applies. Therefore, any reentrant line operating under FBFS discipline is stable if the traffic intensity at every physical station does not exceed one. Q.E.D.

Proof of Theorem 5.5. It suffices to show that the mutual blocking (of classes belonging to different physical stations) will break if the dispatching policy is WIP-dependent, i.e., class k will be assigned the highest priority if $Q_k > \varphi_k$, where Q_k is the number of jobs in buffer k and φ_k is a preset threshold. Assume there exists a set S (not a physical station) such that $\lim_{t \rightarrow \infty} m(\{t | \prod_{k \in S} \dot{R}_k(t) \neq 0\}) / m(\{t | \sum_{k \in S} \dot{R}_k(t) \neq 0\}) = 0$ (S suffers mutual blocking), $P_S > 1$ and $\sum_{k \in S} Q_k(t)/t \nrightarrow 0$. If S only consists of two classes k and l , then $(Q_k(t) + Q_l(t))/t \nrightarrow 0$. Without loss of generality, we assume k is a class before l . For any large enough M , there exists a t_0 s.t. $Q_k(t_0) > M$. Therefore, we have $Q_k(t) \geq \varphi_k$ and $\dot{R}_k(t) < 0$ for any $t \in (t_0, t_0 + T)$, where T is a positive number and $T/(M - \varphi_k) \xrightarrow{a.s.} m_k$ as $M \rightarrow \infty$. If l is the next class of k (i.e., $l = k + 1$), Q_l will increase by $(M - \varphi_k)$ during the interval $(t_0, t_0 + T)$ since class l doesn't receive service. Hence, Q_l will exceed φ_l during the interval $(t_0, t_0 + T)$ and should have been assigned the highest priority. This contradicts to the previous assumption.

If $l = k + 2$, $(Q_{k+1} + Q_l)$ will increase by $(M - \varphi_k)$ during the interval $(t_0, t_0 + T)$. As long as Q_{k+1} exceeds φ_{k+1} , class $k + 1$ will receive service and Q_l increases.

Since M is large enough, Q_l will exceed its threshold φ_l during the interval $(t_0, t_0 + T)$, which is contradicted to the previous assumption.

By induction, Q_l will always exceed φ_l in the interval $(t_0, t_0 + T)$ since M is large enough and l is finite. Hence, classes k and l will both be assigned the highest priority during a period, which is a contradiction.

Similar arguments can be applied to the case that S consists of multiple classes. Therefore, a queueing network which satisfies the usual traffic condition will always be stable under the WIP-dependent policy. Q.E.D.

Appendix A3

Analysis of the Gamma/ H₂/1 and H₂/Gamma/1 Queues. We test the performance of TMA for Gamma/ H₂/1 and H₂/Gamma/1 queueing systems. As before, the utilization values chosen were 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90 and 0.95. The different SCV values chosen for the interarrival and processing time distributions were 1.5, 3.0 and 5.0.

Table A3-1 summarizes the results for the Gamma/ H₂/1 queueing system. It shows that TMA achieves the lowest cumulative error (ϵ_{cum}) and outperforms the others for each combination of the SCVs. The average cumulative error of TMA is about 1.08% while Hey and KL have higher average errors of 5.14% and 7.02%.

Table A3-1 The cumulative errors for the Gamma/ H₂/1 queue with different SCVs.

SCVs	(1.5, 1.5)	(1.5, 3.0)	(1.5, 5.0)	(3.0, 1.5)	(3.0, 3.0)	(3.0, 5.0)	(5.0, 1.5)	(5.0, 3.0)	(5.0, 5.0)	Average Error
Hey	2.61%	1.85%	1.45%	7.01%	5.39%	4.22%	8.79%	8.46%	6.51%	5.14%
KL	3.52%	2.18%	1.78%	9.86%	7.15%	5.39%	13.26%	11.42%	8.59%	7.02%
TMA	0.53%	0.57%	0.57%	1.42%	0.89%	0.83%	1.45%	2.08%	1.38%	1.08%

The results of the H₂/Gamma/1 queueing system are presented in Table A3-2. In this situation, the best approximation can be either Hey or TMA. The differences of the average cumulative errors between the two approximations are not significant.

Table A3-2 The cumulative errors for the H₂/Gamma/1 queue with different SCVs.

SCVs	(1.5, 1.5)	(1.5, 3.0)	(1.5, 5.0)	(3.0, 1.5)	(3.0, 3.0)	(3.0, 5.0)	(5.0, 1.5)	(5.0, 3.0)	(5.0, 5.0)	Average Error
Hey	0.18%	0.56%	0.74%	0.91%	1.19%	1.72%	2.02%	1.59%	1.53%	1.16%
KL	1.12%	1.08%	0.95%	2.87%	3.03%	2.92%	3.50%	4.24%	3.72%	2.60%
TMA	0.51%	0.40%	0.43%	1.77%	1.53%	1.14%	3.62%	3.40%	2.55%	1.71%

Since our TMA is developed based on the exact result of an H₂/M/1 queue, it should

perform well for queues with phase-type distributed interarrival or service times. The above simulation results are consistent with this observation.

Proof for Myskja's Result. We first derive the exact results for the $H_2/M/1$ queueing system, and show that Myskja's general expression reduces to this exact result. We then conclude by presenting our expression.

The expected waiting time in a $GI/M/1$ queueing system is given by

$$E[W] = \left(\frac{\sigma}{1-\sigma} \right) \cdot \frac{1}{\mu}, \quad (1)$$

$$\sigma = \int_0^\infty e^{-\mu(1-\sigma)t} dF(t) = F^*(\mu(1-\sigma)). \quad (2)$$

where $F^*(s)$ is the Laplace-Stieltjes transform of the arrival interval distribution $F(t)$.

The $H_2/M/1$ queueing system is a special case of the $GI/M/1$ queueing system, where the distribution function of the arrival process is given as

$$F(t) = 1 - pe^{-\lambda_1 t} - (1-p)e^{-\lambda_2 t}, \quad 0 \leq p \leq 1. \quad (3)$$

The load at arrival instants, σ , for the $H_2/M/1$ queueing system is given as

$$\sigma = \frac{p \cdot \lambda_1}{[\lambda_1 + \mu(1-\sigma)]} + \frac{(1-p) \cdot \lambda_2}{[\lambda_2 + \mu(1-\sigma)]}. \quad (4)$$

After some algebraic manipulations, we have the following quadratic expression

$$\sigma^2 \mu^2 - \sigma \mu (\lambda_1 + \lambda_2 + \mu) + \lambda_1 \lambda_2 + \mu [p \lambda_1 + (1-p) \lambda_2] = 0. \quad (5)$$

One of the solutions of equation (4) is $\sigma = 1$. The other two solutions ($\sigma < 1$ and $\sigma > 1$) can be obtained from the roots of equation (5) as follows.

$$\sigma = \frac{\mu(\lambda_1 + \lambda_2 + \mu) \pm \sqrt{(\mu(\lambda_1 + \lambda_2 + \mu))^2 - 4\mu^2 [\lambda_1 \lambda_2 + p \lambda_1 \mu + (1-p) \lambda_2 \mu]}}{2\mu^2}.$$

After algebraic manipulations, we have the following solutions

$$\sigma = \frac{1}{2} \left\{ \frac{(\lambda_1 + \lambda_2 + \mu) \pm \sqrt{(\lambda_1 - \lambda_2 + \mu)^2 - 4p\mu(\lambda_1 - \lambda_2)}}{\mu} \right\}, \quad (6)$$

Since we have $0 \leq \sigma \leq 1$, the only feasible solution for equation (5) is

$$\begin{aligned} \sigma &= \frac{1}{2} \left\{ \frac{(\lambda_1 + \lambda_2 + \mu) - \sqrt{(\lambda_1 - \lambda_2 + \mu)^2 - 4p\mu(\lambda_1 - \lambda_2)}}{\mu} \right\} \\ &= \frac{1}{2} \left\{ 1 + \frac{\lambda_1 + \lambda_2}{\mu} - \frac{\sqrt{(\lambda_1 - \lambda_2 + \mu)^2 - 4p\mu(\lambda_1 - \lambda_2)}}{\mu} \right\}. \end{aligned} \quad (7)$$

(Myskja 1991) provided a general equation by using the first three independent moments of

$F(t)$ as

$$\sigma = \frac{1}{2} \left\{ 1 + \rho \left(\frac{q_a - r_a}{q_a - r_a^2} \right) - \sqrt{\left(1 - \rho \frac{q_a - 3r_a + 2}{q_a - r_a^2} \right)^2 + \rho^2 \frac{4(r_a - 1)^3}{(q_a - r_a^2)^2}} \right\}, \quad (8)$$

where $q_a = m_a^3/6(m_a^1)^3$ and $r_a = (c_a^2 + 1)/2$.

For the $H_2/M/1$ queueing system, we have the following

$$m_a^1 = \frac{p}{\lambda_1} + \frac{(1-p)}{\lambda_2} = \frac{\lambda_1 + p(\lambda_2 - \lambda_1)}{\lambda_1 \lambda_2}. \quad (9)$$

$$m_a^2 = \frac{2p}{\lambda_1^2} + \frac{2(1-p)}{\lambda_2^2} = 2 \left[\frac{\lambda_1^2 + p(\lambda_2^2 - \lambda_1^2)}{\lambda_1^2 \lambda_2^2} \right]. \quad (10)$$

$$m_a^3 = \frac{6p}{\lambda_1^3} + \frac{6(1-p)}{\lambda_2^3} = 6 \left[\frac{\lambda_1^3 + p(\lambda_2^3 - \lambda_1^3)}{\lambda_1^3 \lambda_2^3} \right]. \quad (11)$$

$$\rho = \frac{(m_a^1)^{-1}}{\mu} = \frac{\lambda_1 \lambda_2}{[\lambda_1 + p(\lambda_2 - \lambda_1)] \mu}. \quad (12)$$

$$r_a = \frac{(c_a^2 + 1)}{2} = \frac{\left(\frac{m_a^2 - (m_a^1)^2}{(m_a^1)^2} \right) + 1}{2} = \frac{m_a^2}{2(m_a^1)^2}.$$

Further algebraic manipulations would lead to the following equation

$$r_a = \frac{\lambda_1^2 + p(\lambda_2^2 - \lambda_1^2)}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^2}. \quad (13)$$

Similarly, the relative third moment, q_a , for a H₂/M/1 queueing system is given by

$$q_a = \frac{\lambda_1^3 + p(\lambda_2^3 - \lambda_1^3)}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^3}. \quad (14)$$

We first consider the term $\rho \left(\frac{q_a - r_a}{q_a - r_a^2} \right)$ from equation (8). Substituting equations (12),

(13) and (14) in the term, we obtain the numerator as follows.

$$\begin{aligned} \rho(q_a - r_a) &= \frac{\lambda_1 \lambda_2}{[\lambda_1 + p(\lambda_2 - \lambda_1)] \mu} \left\{ \frac{\lambda_1^3 + p(\lambda_2^3 - \lambda_1^3)}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^3} - \frac{\lambda_1^2 + p(\lambda_2^2 - \lambda_1^2)}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^2} \right\} \\ &= \frac{\lambda_1 \lambda_2}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^4 \mu} \left\{ \lambda_1^3 + p(\lambda_2^3 - \lambda_1^3) - [\lambda_1^2 + p(\lambda_2^2 - \lambda_1^2)] [\lambda_1 + p(\lambda_2 - \lambda_1)] \right\} \\ &= \frac{p(1-p)\lambda_1 \lambda_2 (\lambda_1 + \lambda_2)(\lambda_1 - \lambda_2)^2}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^4 \mu}. \end{aligned}$$

The denominator of the term is

$$\begin{aligned} q_a - r_a^2 &= \frac{\lambda_1^3 + p(\lambda_2^3 - \lambda_1^3)}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^3} - \left\{ \frac{\lambda_1^2 + p(\lambda_2^2 - \lambda_1^2)}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^2} \right\}^2 \\ &= \frac{1}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^4} \left\{ [\lambda_1^3 + p(\lambda_2^3 - \lambda_1^3)] [\lambda_1 + p(\lambda_2 - \lambda_1)] - [\lambda_1^2 + p(\lambda_2^2 - \lambda_1^2)]^2 \right\} \\ &= \frac{p(1-p)\lambda_1 \lambda_2 (\lambda_1 - \lambda_2)^2}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^4}. \end{aligned}$$

As a result, we have

$$\rho \left(\frac{q_a - r_a}{q_a - r_a^2} \right) = \frac{\frac{p(1-p)\lambda_1 \lambda_2 (\lambda_1 + \lambda_2)(\lambda_1 - \lambda_2)^2}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^4 \mu}}{\frac{p(1-p)\lambda_1 \lambda_2 (\lambda_1 - \lambda_2)^2}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^4}} = \frac{\lambda_1 + \lambda_2}{\mu}. \quad (15)$$

Next, we consider the first term in the square root $\left(1 - \rho \frac{q_a - 3r_a + 2}{q_a - r_a^2} \right)^2$. Substituting

equations (12), (13) and (14) in the term, we have

$$\begin{aligned}
 \rho \frac{q_a - 3r_a + 2}{q_a - r_a^2} &= \frac{\lambda_1 \lambda_2}{[\lambda_1 + p(\lambda_2 - \lambda_1)] \mu} \left\{ \frac{\lambda_1^3 + p(\lambda_2^3 - \lambda_1^3)}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^3} - 3 \frac{\lambda_1^2 + p(\lambda_2^2 - \lambda_1^2)}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^2} + 2 \right. \\
 &\quad \left. \frac{p(1-p)\lambda_1 \lambda_2 (\lambda_1 - \lambda_2)^2}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^4} \right\} \\
 &= \frac{\lambda_1 \lambda_2}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^4 \mu} \cdot \frac{[\lambda_1 + p(\lambda_2 - \lambda_1)]^4}{p(1-p)\lambda_1 \lambda_2 (\lambda_1 - \lambda_2)^2} \left\{ \lambda_1^3 + p(\lambda_2^3 - \lambda_1^3) + 2[\lambda_1 + p(\lambda_2 - \lambda_1)]^3 \right. \\
 &\quad \left. - 3[\lambda_1^2 + p(\lambda_2^2 - \lambda_1^2)][\lambda_1 + p(\lambda_2 - \lambda_1)] \right\} \\
 &= \frac{1}{\mu p(1-p)(\lambda_1 - \lambda_2)^2} \cdot (2p-1)(1-p)p(\lambda_1 - \lambda_2)^3 \\
 &= \frac{(2p-1)(\lambda_1 - \lambda_2)}{\mu}.
 \end{aligned}$$

As a result, we have

$$\begin{aligned}
 \left(1 - \rho \frac{q_a - 3r_a + 2}{q_a - r_a^2} \right)^2 &= \left(\frac{\mu - (2p-1)(\lambda_1 - \lambda_2)}{\mu} \right)^2 \\
 &= \frac{(\mu + \lambda_1 - \lambda_2)^2 - 4p(1-p)(\lambda_1 - \lambda_2)^2 + 4\mu p(\lambda_2 - \lambda_1)}{\mu^2}.
 \end{aligned}$$

Next we consider the second term in the square root $\rho^2 \frac{4(r_a - 1)^3}{(q_a - r_a^2)^2}$. Substituting

equations (12), (13) and (14) in the term, we have

$$\begin{aligned}
 \rho^2 \frac{4(r_a - 1)^3}{(q_a - r_a^2)^2} &= 4 \left\{ \frac{\lambda_1 \lambda_2}{[\lambda_1 + p(\lambda_2 - \lambda_1)] \mu} \right\}^2 \cdot \left\{ \frac{\left[\frac{\lambda_1^2 + p(\lambda_2^2 - \lambda_1^2)}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^2} - 1 \right]^3}{\left[\frac{p(1-p)\lambda_1 \lambda_2 (\lambda_1 - \lambda_2)^2}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^4} \right]^2} \right\} \\
 &= \frac{4\lambda_1^2 \lambda_2^2 [\lambda_1 + p(\lambda_2 - \lambda_1)]^8 \left\{ \lambda_1^2 + p(\lambda_2^2 - \lambda_1^2) - [\lambda_1 + p(\lambda_2 - \lambda_1)]^2 \right\}^3}{[\lambda_1 + p(\lambda_2 - \lambda_1)]^2 \mu^2 p^2 (1-p)^2 \lambda_1^2 \lambda_2^2 (\lambda_1 - \lambda_2)^4 [\lambda_1 + p(\lambda_2 - \lambda_1)]^6} \\
 &= \frac{4p(1-p)(\lambda_1 - \lambda_2)^2}{\mu^2}.
 \end{aligned}$$

Adding the two terms inside the square root, we have

$$\begin{aligned} \left(1 - \rho \frac{q_a - 3r_a + 2}{q_a - r_a^2}\right)^2 + \rho^2 \frac{4(r_a - 1)^3}{(q_a - r_a^2)^2} &= \frac{(\mu + \lambda_1 - \lambda_2)^2 + 4\mu p(\lambda_2 - \lambda_1)}{\mu^2} + \frac{4p(1-p)(\lambda_1 - \lambda_2)^2}{\mu^2} \\ &= \frac{(\mu + \lambda_1 - \lambda_2)^2 + 4\mu p(\lambda_2 - \lambda_1)}{\mu^2}. \end{aligned} \quad (16)$$

Substituting equations (15) and (16) in equation (8), we have

$$\begin{aligned} \sigma &= \frac{1}{2} \left\{ 1 + \rho \left(\frac{q_a - r_a}{q_a - r_a^2} \right) - \sqrt{\left(1 - \rho \frac{q_a - 3r_a + 2}{q_a - r_a^2} \right)^2 + \rho^2 \frac{4(r_a - 1)^3}{(q_a - r_a^2)^2}} \right\} \\ &= \frac{1}{2} \left\{ 1 + \frac{\lambda_1 + \lambda_2}{\mu} - \sqrt{\frac{(\mu + \lambda_1 - \lambda_2)^2 - 4\mu p(\lambda_1 - \lambda_2)}{\mu^2}} \right\} \\ \sigma &= \frac{1}{2} \left\{ 1 + \frac{\lambda_1 + \lambda_2}{\mu} - \frac{\sqrt{(\mu + \lambda_1 - \lambda_2)^2 - 4\mu p(\lambda_1 - \lambda_2)}}{\mu} \right\}. \end{aligned} \quad (17)$$

We have now shown that Myskja's (1991) result as shown in equation (8) reduces to equation (7) for the $H_2/M/1$ queueing system. Now, the expected waiting time in a GI/G/1 queue is given by

$$E[W] = \frac{\frac{1}{2} \left\{ 1 + \rho \left(\frac{q_a - r_a}{q_a - r_a^2} \right) - \sqrt{\left(1 - \rho \frac{q_a - 3r_a + 2}{q_a - r_a^2} \right)^2 + \rho^2 \frac{4(r_a - 1)^3}{(q_a - r_a^2)^2}} \right\}}{1 - \frac{1}{2} \left\{ 1 + \rho \left(\frac{q_a - r_a}{q_a - r_a^2} \right) - \sqrt{\left(1 - \rho \frac{q_a - 3r_a + 2}{q_a - r_a^2} \right)^2 + \rho^2 \frac{4(r_a - 1)^3}{(q_a - r_a^2)^2}} \right\}} \cdot \frac{1}{\mu}.$$

Let $x = q_a - r_a$, $y = q_a - r_a^2$, $z = r_a - 1$ and $\theta = \frac{\rho(q_a - r_a) - (q_a - r_a^2)}{2\rho(r_a - 1)} = \frac{\rho x - y}{2\rho z}$, then we

will have

$$\begin{aligned}
 \mu E[W] &= \frac{1 + \rho \frac{x}{y} - \sqrt{(1 - \rho \frac{x-2z}{y})^2 + \rho^2 \frac{4z^3}{y^2}}}{1 - \rho \frac{x}{y} + \sqrt{(1 - \rho \frac{x-2z}{y})^2 + \rho^2 \frac{4z^3}{y^2}}} = -1 + \frac{2}{1 - \rho \frac{x}{y} + \sqrt{(1 - \rho \frac{x-2z}{y})^2 + \rho^2 \frac{4z^3}{y^2}}} \\
 &= -1 + \frac{2 \frac{y}{2\rho z}}{\frac{y}{2z\rho} - \frac{x}{2z} + \sqrt{(\frac{y}{2\rho z} - \frac{x-2z}{2z})^2 + z}} = -1 + \frac{\frac{y}{\rho z}}{-\theta + \sqrt{(1-\theta)^2 + z}} \\
 &= -1 + \frac{\frac{y}{\rho z} (\sqrt{(1-\theta)^2 + z} + \theta)}{1 - 2\theta + z} = \frac{\rho}{1-\rho} (1 - \frac{1}{\rho} + \frac{y(\sqrt{(1-\theta)^2 + z} + \theta)}{\rho^2 z (1 - 2\theta + z)}) \\
 &= \frac{\rho}{1-\rho} (1 - \frac{1}{\rho} + \frac{y(\sqrt{(1-\theta)^2 + z} + \theta)}{\rho^2 z (1 - 2\frac{\rho x - y}{2\rho z} + z)}) = \frac{\rho}{1-\rho} (1 - \frac{1}{\rho} + \frac{(1-\rho)y(\sqrt{(1-\theta)^2 + z} + \theta)}{\rho^2 (z - \frac{\rho x - y}{\rho} + z^2)}) \\
 &= \frac{\rho}{1-\rho} (1 - \frac{1}{\rho} + \frac{(1-\rho)y(\sqrt{(1-\theta)^2 + z} + \theta)}{\rho^2 (\frac{y}{\rho} - y)}) = \frac{\rho}{1-\rho} (1 - \frac{1}{\rho} + \frac{\sqrt{(1-\theta)^2 + z} + \theta}{\rho}) \\
 &= \frac{\rho}{1-\rho} (1 + \frac{\sqrt{(1-\theta)^2 + z} + \theta - 1}{\rho}) = \left(\frac{\rho}{1-\rho} \right) \left\{ 1 + \frac{1}{\rho} \left[\sqrt{(1-\theta)^2 + r - 1} - (1-\theta) \right] \right\}
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 E[W] &= \left(\frac{\rho}{1-\rho} \right) \cdot \frac{1}{\mu} \cdot \left\{ 1 + \frac{1}{\rho} \cdot \left[\sqrt{(1-\theta)^2 + (r_a - 1)} - (1-\theta) \right] \right\}, \\
 \theta &= \frac{\rho(q_a - r_a) - (q_a - r_a^2)}{2\rho(r_a - 1)}.
 \end{aligned}$$

Myskja then heuristically modified the above expression by using the relative second moment of the service time distribution (r_s) as follows.

$$\begin{aligned}
 E[W] &\approx \left(\frac{\rho}{1-\rho} \right) \cdot \frac{1}{\mu} \cdot \left\{ r_s + \frac{1}{\rho} \cdot \left[\sqrt{(r_s - \theta)^2 + (2r_s - 1 + d) \cdot (r_a - 1)} - (r_s - \theta) \right] \right\}, \\
 \theta &= \frac{\rho(q_a - r_a) - (q_a - r_a^2)}{2\rho(r_a - 1)}, \quad d = \left(1 + \frac{1}{r_a} \right) \cdot (1 - r_s) \cdot \left[1 - \left(\frac{q_0}{q_a} \right)^3 \right] \cdot (1 - \rho^3).
 \end{aligned}$$

Appendix A4

Proof of Property 7.5. We consider a two-stage single-server tandem queue. It suffices to show the latter part of the property. For the first stage, it's trivial that $IR_1 \equiv 1$ and $C_1 = 1$.

For stage 2, there are two occasions: (i) $BN = 1$, i.e., the first station is the bottleneck, and (ii) $BN = 2$, i.e., the second station is the bottleneck.

(i) $BN = 1$. Note that $\rho = \rho_1 = \rho_2 \mu_2 / \mu_1$. In this case, we have

$$W_2^{ASIA \text{ of } S}(\rho) = \frac{1 + c_{s_2}^2}{2} \frac{\rho_2}{1 - \rho_2 \mu_2} \frac{1}{\mu_2},$$

and $W_2^{BSIA \text{ of } S}(\rho) = 0$. When $\rho \rightarrow 0$, $W_2^S(\rho)$ can be given by

$$W_2^S(\rho) = \rho_1 E[S_2 - S_1]^+ + \rho_2 E[S_2^e - S_1]^+ + o(\rho^2),$$

where S_2^e is the random variable denoting the remaining service time at station 2 (Wolff 1982).

Then it follows that, when $\rho \rightarrow 0$,

$$\begin{aligned} IR_2(\rho) &= \frac{\rho_1 E[S_2 - S_1]^+ + \rho_2 E[S_2^e - S_1]^+ + o(\rho^2)}{\frac{1 + c_{s_2}^2}{2} \frac{\rho_2}{1 - \rho_2 \mu_2} \frac{1}{\mu_2}} \\ &= \frac{2\mu_2}{1 + c_{s_2}^2} \left(\frac{\mu_2}{\mu_1} (1 - \rho_2) E[S_2 - S_1]^+ + (1 - \rho_2) E[S_2^e - S_1]^+ \right) + o(\rho) \\ &= \frac{2\mu_2}{1 + c_{s_2}^2} \left(\frac{\mu_2}{\mu_1} E[S_2 - S_1]^+ + E[S_2^e - S_1]^+ \right) + o(\rho) \\ &= C_2 + o(\rho). \end{aligned}$$

(ii) $BN = 2$. Note that $\rho = \rho_2 = \rho_1 \mu_1 / \mu_2$. In this case, we have

$$W_2^{ASIA \text{ of } S}(\rho) = \frac{1 + c_{s_2}^2}{2} \frac{\rho_2}{1 - \rho_2 \mu_2} \frac{1}{\mu_2},$$

and

$$W_2^{BSIA\ of\ \mathcal{S}}(\rho) = \frac{1 + c_{s_2}^2}{2} \frac{\rho_2}{1 - \rho_2 \mu_2} \frac{1}{\mu_2} - \frac{1 + c_{s_1}^2}{2} \frac{\rho_1}{1 - \rho_1 \mu_1} \frac{1}{\mu_1}.$$

When $\rho \rightarrow 0$, $W_2^{\mathcal{S}}(\rho)$ can be given by

$$W_2^{\mathcal{S}}(\rho) = \rho_1 E[S_2 - S_1]^+ + \rho_2 E[S_2^e - S_1]^+ + o(\rho^2).$$

Then it follows that, when $\rho \rightarrow 0$,

$$\begin{aligned} IR_2(\rho) &= \frac{\rho_1 E[S_2 - S_1]^+ + \rho_2 E[S_2^e - S_1]^+ + o(\rho^2)}{\frac{1 + c_{s_2}^2}{2} \frac{\rho_2}{1 - \rho_2 \mu_2} \frac{1}{\mu_2} - \left(\frac{1 + c_{s_2}^2}{2} \frac{\rho_2}{1 - \rho_2 \mu_2} \frac{1}{\mu_2} - \frac{1 + c_{s_1}^2}{2} \frac{\rho_1}{1 - \rho_1 \mu_1} \frac{1}{\mu_1} \right)} \\ &\quad + \frac{\frac{1 + c_{s_1}^2}{2} \frac{\rho_1}{1 - \rho_1 \mu_1} \frac{1}{\mu_1} - \frac{1 + c_{s_2}^2}{2} \frac{\rho_2}{1 - \rho_2 \mu_2} \frac{1}{\mu_2}}{\frac{1 + c_{s_2}^2}{2} \frac{\rho_2}{1 - \rho_2 \mu_2} \frac{1}{\mu_2} - \left(\frac{1 + c_{s_2}^2}{2} \frac{\rho_2}{1 - \rho_2 \mu_2} \frac{1}{\mu_2} - \frac{1 + c_{s_1}^2}{2} \frac{\rho_1}{1 - \rho_1 \mu_1} \frac{1}{\mu_1} \right)} \\ &= \frac{2\mu_1}{1 + c_{s_1}^2} \left((1 - \rho_1) E[S_2 - S_1]^+ + (1 - \rho_1) \frac{\mu_1}{\mu_2} E[S_2^e - S_1]^+ \right) + 1 \\ &\quad - \frac{1 + c_{s_2}^2}{1 + c_{s_1}^2} \frac{1 - \rho_1 \mu_1}{1 - \rho_2 \mu_2} + o(\rho) \\ &= \frac{2\mu_1}{1 + c_{s_1}^2} \left(E[S_2 - S_1]^+ + \frac{\mu_1}{\mu_2} E[S_2^e - S_1]^+ \right) + 1 - \frac{1 + c_{s_2}^2}{1 + c_{s_1}^2} \frac{1 - \rho_1 \mu_1}{1 - \rho_2 \mu_2} + o(\rho) \\ &= C_2 + o(\rho). \end{aligned}$$

This completes the proof of Property 7.5 for a two-stage tandem queue. For systems with multiple stages, $\mathbf{W}^{\mathcal{S}}(\rho)$, $\mathbf{W}^{ASIA\ of\ \mathcal{S}}(\rho)$ and $\mathbf{W}^{BSIA\ of\ \mathcal{S}}(\rho)$ can be figured out in the same way and the proof is similar. Q.E.D.

Proof of Property 7.6. We first consider the a two-stage tandem queue. There are two occasions: (i) $BN = 1$, i.e., the first station is the bottleneck, and (ii) $BN = 2$, i.e., the second station is the bottleneck.

(i) $BN = 1$. In this case, there is $IR_{BN}(\rho) \equiv 1$ and $\lim_{\rho \rightarrow 1} W_2^{ASIA\ of\ \mathcal{S}}(\rho) = \infty$. It follows

that $\left| \lim_{\rho \rightarrow 1} IR_{BN}(\rho) / W_{BN}^{ASIA\ of\ S}(\rho) \right| = 0$.

(ii) $BN = 2$. In this situation, we have

$$\begin{aligned} IR_{BN}(\rho) &= \frac{W_2^S(\rho) - W_2^{BSIA\ of\ S}(\rho)}{W_2^{ASIA\ of\ S}(\rho) - W_2^{BSIA\ of\ S}(\rho)} \\ &= \frac{W_2^S(\rho) - W_2^{BSIA\ of\ S}(\rho)}{W_1^{ASIA\ of\ S}(\rho)}. \end{aligned}$$

Based on Corollary 1 in (Wu et al. 2017), we have

$$W_2^S(\rho) \leq W_1^S(\rho) + W_2^S(\rho) \leq E[W^{BI}] + M,$$

where $E[W^{BI}]$ is the mean queue time of the BSIA system with virtual interruptions and

M is a finite positive number. On the other hand, we have

$$E[W^{BI}] = \frac{1 + c_{S_2'}^2}{2} \frac{\rho_2 + \rho_I}{1 - (\rho_2 + \rho_I)} E[S_2'],$$

where S_2' is the generalized service time at station 2 and ρ_I is the utilization of

interruptions. According to Corollary 3 in (Wu et al. 2017), we have $E[S_2'] \rightarrow E[S_2] =$

$1/\mu_2$, $c_{S_2'}^2 \rightarrow c_{S_2}^2$ and $\rho_I \rightarrow 0$ when $\rho_2 \rightarrow 1$.

Therefore,

$$\begin{aligned} \left| \frac{IR_{BN}(\rho)}{W_{BN}^{ASIA\ of\ S}(\rho)} \right| &\leq \left| \frac{E[W^{BI}] - W_2^{BSIA\ of\ S}(\rho) + M}{W_1^{ASIA\ of\ S}(\rho) W_2^{ASIA\ of\ S}(\rho)} \right| \\ &= \left| \frac{\frac{1 + c_{S_2'}^2}{2} \frac{\rho_2 + \rho_I}{1 - (\rho_2 + \rho_I)} E[S_2'] - \frac{1 + c_{S_2}^2}{2} \frac{\rho_2}{1 - \rho_2} \frac{1}{\mu_2} + W_1^{ASIA\ of\ S}(\rho) + M}{W_1^{ASIA\ of\ S}(\rho) \frac{1 + c_{S_2}^2}{2} \frac{\rho_2}{1 - \rho_2} \frac{1}{\mu_2}} \right|. \end{aligned}$$

Since $W_1^{ASIA\ of\ S}(\rho)$ and M are all finite, it follows that,

$$\lim_{\rho \rightarrow 1} \left| \frac{IR_{BN}(\rho)}{W_{BN}^{ASIA\ of\ S}(\rho)} \right| = \lim_{\rho \rightarrow 1} \left| \frac{1}{W_1^{ASIA\ of\ S}(\rho)} - \frac{1}{W_1^{ASIA\ of\ S}(\rho)} \right| = 0.$$

For systems with multiple stages, $\mathbf{W}^S(\rho)$, $\mathbf{W}^{ASIA\ of\ S}(\rho)$ and $\mathbf{W}^{BSIA\ of\ S}(\rho)$ can be figured out in the same way and the proof is similar. Q.E.D.

Proof of Proposition 7.1. We first consider a two-stage tandem queue. For stage 1, $IR_1(\rho) = 1$ and it's trivial that the approximation error equals zero at stage 1. For stage 2, there are two occasions: (i) $BN = 1$, i.e., the first station is the bottleneck, and (ii) $BN = 2$, i.e., the second station is the bottleneck.

(i) $BN = 1$. In this case, $W_2^{BSIA\ of\ S}(\rho) = 0$. Therefore, we have

$$\begin{aligned}\lim_{\rho \rightarrow 0} \varepsilon_2 &= \lim_{\rho \rightarrow 0} \frac{|(F_2(\rho) - IR_2(\rho))W_2^{ASIA\ of\ S}(\rho) - (F_2(\rho) - IR_2(\rho))W_2^{BSIA\ of\ S}(\rho)|}{IR_2(\rho)W_2^{ASIA\ of\ S}(\rho) + (1 - IR_2(\rho))W_2^{BSIA\ of\ S}(\rho)} \\ &= \lim_{\rho \rightarrow 0} \frac{|F_2(\rho) - IR_2(\rho)| |W_2^{ASIA\ of\ S}(\rho)|}{|IR_2(\rho)W_2^{ASIA\ of\ S}(\rho)|} \\ &= \lim_{\rho \rightarrow 0} \left| 1 - \frac{F_2(\rho)}{IR_2(\rho)} \right| = 0.\end{aligned}$$

(ii) $BN = 2$. In this case, we have

$$W_2^{BSIA\ of\ S}(\rho) = \frac{1 + c_{s_2}^2}{2} \frac{\rho_2}{1 - \rho_2 \mu_2} \frac{1}{\mu_2} - \frac{1 + c_{s_1}^2}{2} \frac{\rho_1}{1 - \rho_1 \mu_1} \frac{1}{\mu_1},$$

which implies that $\lim_{\rho \rightarrow 0} W_2^{BSIA\ of\ S}(\rho)/W_2^{ASIA\ of\ S}(\rho) < 1$ is a constant, and we denote the

limit as C . Therefore,

$$\begin{aligned}\lim_{\rho \rightarrow 0} \varepsilon_2 &= \lim_{\rho \rightarrow 0} \frac{|(F_2(\rho) - IR_2(\rho))W_2^{ASIA\ of\ S}(\rho) - (F_2(\rho) - IR_2(\rho))W_2^{BSIA\ of\ S}(\rho)|}{IR_2(\rho)W_2^{ASIA\ of\ S}(\rho) + (1 - IR_2(\rho))W_2^{BSIA\ of\ S}(\rho)} \\ &= \lim_{\rho \rightarrow 0} \frac{|F_2(\rho) - IR_2(\rho)| |1 - C|}{IR_2(\rho) + (1 - IR_2(\rho))C} \\ &= 0.\end{aligned}$$

The proof can be generalized to systems with more stages according to Section 6 in (Wu and McGinnis 2013). Q.E.D.

Proof of Proposition 7.2. For the bottleneck BN , we have

$$\begin{aligned}\varepsilon_{BN} &= \frac{|(F_{BN}(\rho) - IR_{BN}(\rho))W_{BN}^{ASIA\ of\ S}(\rho) - (F_{BN}(\rho) - IR_{BN}(\rho))W_{BN}^{BSIA\ of\ S}(\rho)|}{IR_{BN}(\rho)W_{BN}^{ASIA\ of\ S}(\rho) + (1 - IR_{BN}(\rho))W_{BN}^{BSIA\ of\ S}(\rho)} \\ &= \frac{|F_{BN}(\rho) - IR_{BN}(\rho)| |W_{BN}^{ASIA\ of\ S}(\rho) - W_{BN}^{BSIA\ of\ S}(\rho)|}{|IR_{BN}(\rho)W_{BN}^{ASIA\ of\ S}(\rho) + (1 - IR_{BN}(\rho))W_{BN}^{BSIA\ of\ S}(\rho)|} \\ &= \frac{|F_{BN}(\rho) - IR_{BN}(\rho)| |W_{BN}^{ASIA\ of\ S}(\rho) - W_{BN}^{BSIA\ of\ S}(\rho)|}{|IR_{BN}(\rho) (W_{BN}^{ASIA\ of\ S}(\rho) - W_{BN}^{BSIA\ of\ S}(\rho)) + W_{BN}^{BSIA\ of\ S}(\rho)|}.\end{aligned}$$

Note that $\lim_{\rho \rightarrow 1} |W_{BN}^{ASIA\ of\ S}(\rho) - W_{BN}^{BSIA\ of\ S}(\rho)|$ is a constant and we denote it as C . Hence,

we have

$$\begin{aligned}\lim_{\rho \rightarrow 1} \varepsilon_{BN} &= \lim_{\rho \rightarrow 1} \frac{|F_{BN}(\rho) - IR_{BN}(\rho)| C}{|IR_{BN}(\rho) C + W_{BN}^{BSIA\ of\ S}(\rho)|} \\ &\leq \lim_{\rho \rightarrow 1} \frac{\left(\left|\frac{F_{BN}(\rho)}{IR_{BN}(\rho)}\right| + 1\right) C}{\left|C + \frac{W_{BN}^{BSIA\ of\ S}(\rho)}{IR_{BN}(\rho)}\right|} \\ &= \lim_{\rho \rightarrow 1} \frac{C}{\left|C + \frac{W_{BN}^{BSIA\ of\ S}(\rho)}{IR_{BN}(\rho)}\right|} + \lim_{\rho \rightarrow 1} \frac{C \left|\frac{F_{BN}(\rho)}{IR_{BN}(\rho)}\right|}{\left|C + \frac{W_{BN}^{BSIA\ of\ S}(\rho)}{IR_{BN}(\rho)}\right|}.\end{aligned}$$

According to Property 7.6, the first term in the above formula equals zero. Furthermore,

because $\lim_{\rho \rightarrow 1} F_{BN}(\rho) (1 - \rho) = 0$, we have

$$\begin{aligned}\lim_{\rho \rightarrow 1} \frac{C \left|\frac{F_{BN}(\rho)}{IR_{BN}(\rho)}\right|}{\left|C + \frac{W_{BN}^{BSIA\ of\ S}(\rho)}{IR_{BN}(\rho)}\right|} &= \lim_{\rho \rightarrow 1} C \left|\frac{F_{BN}(\rho)}{W_{BN}^{BSIA\ of\ S}(\rho)}\right| \\ &= \lim_{\rho \rightarrow 1} C \left|\frac{F_{BN}(\rho)(1 - \rho)}{W_{BN}^{BSIA\ of\ S}(\rho)(1 - \rho)}\right| = 0.\end{aligned}$$

This completes the proof of Proposition 7.2. Q.E.D.

The Author's Publications Resulting from the Thesis

Wu, K., Srivathsan, S. and Shen, Y. (2017) Three-moment approximation for the mean queue time of a GI/G/1 queue. IIE Transactions, 50 (2), 63-73.

Shen, Y., Wu, K. (2018) Stability of a GI/G/1 queue: a survey. Asia-Pacific Journal of Operational Research, <https://doi.org/10.1142/S021759591850015X>.

Wu, K., Shen, Y. (2018) Stability of multiclass queueing networks. Under Review.

Shen, Y., Wu, K. (2018) The affine structure of mean queue time. Under review.