論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Studies on Batch Arrival Infinite-Server Queues and Related Models
著者(和文)	矢島萌子
Author(English)	Moeko Yajima
出典(和文)	学位:博士(理学), 学位授与機関:東京工業大学, 報告番号:甲第11392号, 授与年月日:2020年3月26日, 学位の種別:課程博士, 審査員:三好 直人,樺島 祥介,渡邊 澄夫,福田 光浩,中野 張,増山 博之
Citation(English)	Degree:Doctor (Science), Conferring organization: Tokyo Institute of Technology, Report number:甲第11392号, Conferred date:2020/3/26, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Studies on
Batch Arrival Infinite-Server Queues and Related Models

Moeko Yajima

Thesis submitted for
the Degree of Doctor of Science

Department of Mathematical and Computing Science
School of Computing
Tokyo Institute of Technology

2020

Preface

Queueing theory is a mathematical approach to analyses of congestion in waiting lines. Queueing models are used to imitate waiting lines in queueing theory. In general, queueing theory is considered to be a division of operations research, because analysis results of queueing models are often used to make decisions about resources providing service.

This thesis studies batch arrival infinite-server queues and related models. Infinite-server queues have infinitely many servers, and thus all arriving customers can begin to receive service immediately upon arrivals without waiting. Infinite-server queues have many applications in various areas, such as inventory systems, road traffic systems, and telecommunication systems. In addition, infinite-server queues help us to understand the dynamics of customers in large-scale service systems (facilities), such as theme parks, large commercial complexes, and large parking lots.

Stability conditions for batch arrival infinite-server queues are paid little attention in previous studies. This thesis presents stability conditions for general infinite-server queues with batch arrivals. We first consider the stability for BMAP/M/∞ queues, which are infinite-server queues with a batch Markovian arrival process and an exponential service time distribution. We show that the stability condition for BMAP/M/∞ queues is that the logarithmic moment of batch sizes is finite. Furthermore, we extend this result to the multiclass case.

Next, we investigate the stability for GI^X/GI/∞ queues, which are batch arrival infinite-server queues such that batches arrive according to a renewal process and service times are independent and identically distributed with a general distribution. We show the stability condition for GI^X/GI/∞ queues. We also present a tractable sufficient condition for the stability under a moderate condition on the tail of the service time distribution. Furthermore, in the case that the service time distribution has an exponential tail, we show that the stability condition for the GI^X/GI/∞ queue is that the logarithmic moment of the batch size distribution is finite.

Markov-modulated queues change their parameters depending on a Markov chain. This thesis analyzes a Markov-modulated batch arrival infinite-server queue with catastrophe mechanism. Catastrophes can imitate situations such that customers may or may not leave the system without completing their service due to accidents.

In general, it is very difficult to exactly analyze Markov-modulated queues, except for some simple models. Thus, we consider the scaling model in a heavy traffic regime. We then establish a central limit theorem for the stationary queue length of our queueing model; that is, the centered and normalized stationary queue length distribution converges in distribution to a normal distribution. Furthermore, we derive an approximation of the stationary queue length distribution using the central limit theorem, and then confirm the accuracy of this approximation through numerical experiments.

In today's information society, it is a serious issue that energy consumption and transmission delay in data centers increase. In recent year, variable-speed CPUs have become popular because they can reduce energy consumption while maintaining acceptable transmission delay for jobs. Furthermore, a simple idea for saving energy is to keep servers powered off while the system is empty because idle CPUs still consume approximately 60% of their peak consumption while processing a job. We refer to such an idea for saving energy as the on-off policy.

In order to grasp the dynamics of data centers with a variable-speed CPU and the on-off policy, we study batch arrival single-server queues with variable service speed and the on-off policy. In this thesis, the service speed is assumed to change in proportion to the queue length. The queue length process of this single-server queue is identical to that of an infinite-server queue. We derive the probability generating function of the stationary queue length and the Laplace-Stieltjes transform of the stationary sojourn time distribution. In addition, we present some numerical results to show the energy-performance of the queueing model analyzed herein.

Moeko Yajima
February 2020

Acknowledgments

First of all, I would like to express my gratitude to my supervisor, Professor Naoto Miyoshi, who instructs me in bachelor, master and doctor course. He has taught me the ABC of research, and given me a lot of useful advice. He also actively assisted me to present at many international conferences. In addition, I was able to learn from him what a researcher should be, which is a great asset in my student life. Without his support, I could not complete this thesis. I would also like to express my gratitude to the laboratory members. Thanks to them, I had a good time in the laboratory.

I am grateful to Associate Professor Tuan Phung-Duc of University of Tsukuba. He was Assistant Professor of Tokyo Institute of Technology until 4 years ago, and gave a lot of valuable advice for my research even after transferring to University of Tsukuba. He gave me deep knowledge, which allowed me to proceed with my research.

I am thankful to Associate Professor Hiroyuki Masuyama of Kyoto University, who is my co-author of many papers and has participated in many international conferences. Thanks to his precise comments, I was able to proceed with my research. He also has given me a lot of benefit advice on how to write a paper by English.

Last but not least, I would like to thank my parents for giving me the opportunity to go on to the doctoral course and for keeping to watch over me in every situation. In addition to the people mentioned above, I was able to complete this paper thanks to the support of many people. I would like to express my deep appreciation here.

Contents

Notations and Abbreviations

Notations

$\mathbb{N}$	set of natural numbers; $\{1, 2, \dots\}$
$\mathbb{Z}$	set of integer numbers; $\{0, \pm 1, \pm 2, \pm 3, \dots\}$
$\mathbb{Z}_+$	set of non-negative integer numbers; $\{0, 1, 2, \dots\}$
$\mathbb{R}$	set of real numbers
$\mathbb{R}_+$	set of non-negative real numbers
$n!$	n factorial; $n \cdot (n-1) \cdot (n-2) \cdots 2 \cdot 1$ for $n \in \mathbb{N}$
$\binom{n}{k}$	binomial coefficient; $\frac{n!}{k!(n-k)!}$ for $0 \leq k \leq n$
f^{-1}	inverse function of function f
$f \circ g$	composite function; $f \circ g(x) = f(g(x))$
$f(x) = o(g(x))$	little-o symbol; $\lim_{x \downarrow 0} \lvert f(x) \rvert / \lvert g(x) \rvert = 0$
$f(N) = O(g(N))$	big-O symbol; $\lim_{N \to \infty} \lvert f(N)/g(N) \rvert < \infty$
e	Napier's constant
e^x, $\exp(x)$	exponential function of x
$\log x$	natural logarithm of x
i	imaginary unit; $\mathrm{i} = \sqrt{-1}$
$\mathrm{Re}(z)$	real part of complex number z
$\mathrm{Im}(z)$	imaginary part of complex number z
$\lvert z \rvert$	modulus of complex number z; $\lvert z \rvert = \sqrt{x^2 + y^2}$ for $z = x + \mathrm{i}y$

$\boldsymbol{X}$ (bold upper-case letter)	matrix
$\boldsymbol{I}$	identity matrix
$\boldsymbol{O}$	matrix whose all elements are 0
$\boldsymbol{X}^{-1}$	inverse of matrix $\boldsymbol{X}$

$\boldsymbol{x}$ (bold lower-case letter)	column vector
$\boldsymbol{x}^\top$	transposed vector of $\boldsymbol{x}$
$\|\|\boldsymbol{x}\|\|_2$	Euclidean norm of vector $\boldsymbol{x}$
$\mathrm{diag}(x_1, x_2, \ldots, x_n)$	diagonal matrix whose ith diagonal element is x_i
$\boldsymbol{e}$	column vector whose all elements are 1
$\boldsymbol{0}$	column vector whose all elements are 0

$\mathsf{E}[\cdot]$	expectation
$\mathsf{V}[\cdot]$	variance
$I(\chi)$	indicator function of event χ
Unif$\{a, b\}$	discrete uniform distribution with parameters $a \leq b$
Binom(n, p)	binomial distribution with parameters n and p
Geo(p)	geometric distribution with parameter p

Abbreviations

BMAP	batch Markovian arrival process
CF	characteristic function
i.i.d.	independent and identically distributed
LST	Laplace–Stieltjes transform
MGF	moment generating function
PGF	probability generating function
w.p.	with probability

List of Figures

List of Tables

Chapter 1

Introduction

Infinite-server queues help us to understand the dynamics of customers in large-scale service systems (facilities), such as theme parks, large commercial complexes, and large parking lots. In this thesis, we investigate batch arrival infinite-server queues and related models. In this chapter, Section 1.1 introduces the basics of queueing theory. Sections 1.2–1.6 present the preliminary knowledge of our work. Finally, Section 1.7 shows the organization of this thesis.

1.1 Basics of queueing theory

Queueing theory is a mathematical approach to analyses of congestion in waiting lines such as supermarkets, hospitals, road traffic systems, and computer systems. Queueing models are used to imitate waiting lines in queueing theory. In general, queueing theory is considered to be a division of operations research, because analysis results of queueing models are often used to make decisions about resources to provide service.

Queueing theory originated from Erlang's research at the beginning of the 20th century, which treated a telephone traffic [24, 25]. Beside the telephone exchange, queueing theory can deal waiting lines at various situation, e.g., a bank ATM, a supermarket cash register, an airport, a hospital, a road traffic system, an information network system, and a production system. Queueing theory is used to design and evaluate real-world systems which can be imitated by a queueing model. Thus, analysis results of queueing models can provide optimal design parameters for real-world systems.

The key elements of queueing models are customers, servers, and a waiting room (see Figure 1.1). Customers arrive at the system and request service. If there are available servers in the system, then an arriving customer occupies one of them and begins receiving service, otherwise the customer waits for service at the waiting room. Customers leave the system after service completion.

Kendall's notation is the standard way to describe a queueing model by using the four factors: A/B/n/m, where A refers to the arrival process, B refers to the service time distri-

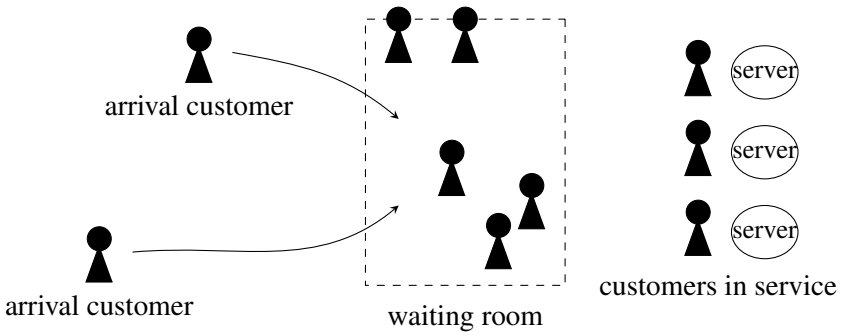

Figure 1.1: Standard queueing model

bution, n is the number of servers, and m is the capacity of the waiting room. If there is no limit for the capacity of the waiting room, i.e., $m = \infty$, then Kendall's notation A/B/n/∞ is usually shortened to A/B/n. Tables 1.1 and 1.2 show the symbols to describe the arrival processes and the service time distributions, respectively, in Kendall's notation. For example, an M/GI/1 queue refers to a single-server queue such that customers arrive according to a homogeneous Poisson process and service times of customers are independent and identically distributed (i.i.d.) with a general distribution.

1.2 Infinite-server queues

Infinite-server queues have infinitely many servers. All customers who arrives at infinite-server queues can begin to receive service immediately upon arrival without waiting. Many researchers have studied stationary and/or time-dependent infinite-server queues (see, e.g., [14, 29, 34, 58, 60, 63], and the references therein).

Applications of infinite-server queues are difficult to be imagined because it is impossible to prepare infinitely many servers in the real world. However, infinite-server queues

symbol	description
M	Poisson arrival process
M^X	batch Poisson arrival process
MAP	Markovian arrival process
BMAP	batch Markovian arrival process
GI	renewal process

Table 1.1: Symbols to describe arrival processes in Kendall's notation

symbol	description
M	exponential distribution
D	deterministic distribution
GI	general distribution
PH	phase-type distribution

Table 1.2: Symbols to describe service time distributions in Kendall's notation

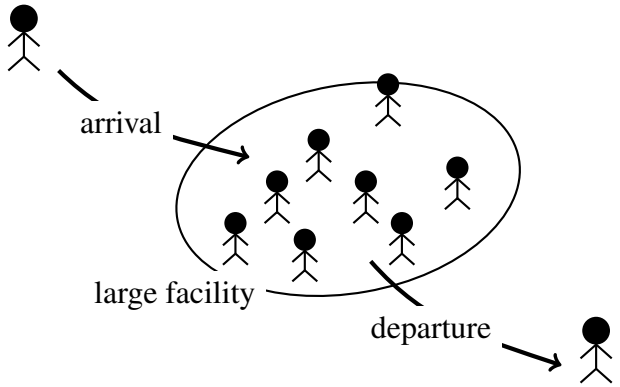

Figure 1.2: Application of an infinite-server queue to a large facility

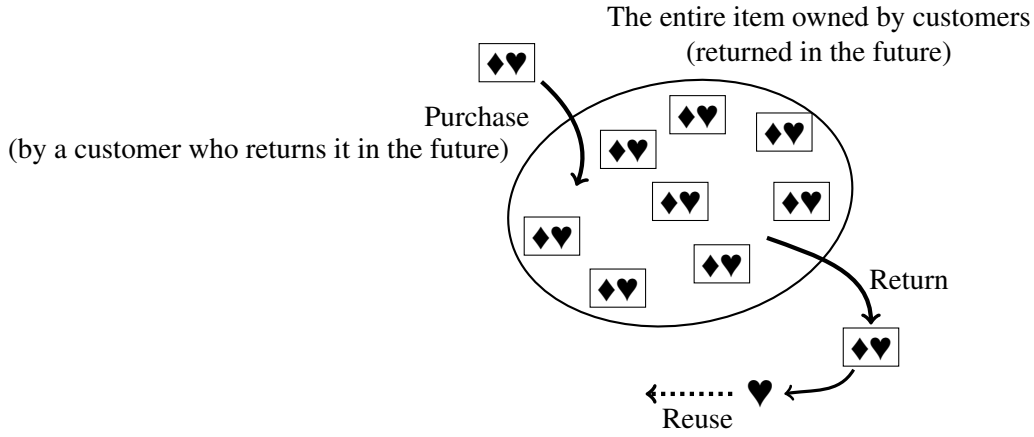

Figure 1.3: Application of an infinite-server queue to a production system

help us to understand the dynamics of customers in large-scale service systems (facilities), such as theme parks, large commercial complexes, and large parking lots. Furthermore, infinite-server queues have many applications in various areas, such as inventory systems [7], road traffic systems [68], and telecommunication systems [50]. We now present two application examples of infinite-server queues.

We first show the application example to a large facility, e.g., an amusement park or a shopping mall. Figure 1.2 illustrates that a customer arrives the large facility and then stays there until its request are satisfied. Customers of the large facility as illustrated in Figure 1.2 can be considered to customers of an infinite-server queue. The sojourn times of customers in the large facility are equivalent to the service times of customers in the infinite-server queue. Infinite-server queues help to understand the dynamics of customers in large facilities.

Next, we show the application example to a production system [65]. Figure 1.3 illustrates a product system of items $\boxed{\blacklozenge\heartsuit}$ such that purchased items will be returned in the future and parts $\heartsuit$ included returned items can be reused. In such a situation, items $\boxed{\blacklozenge\heartsuit}$

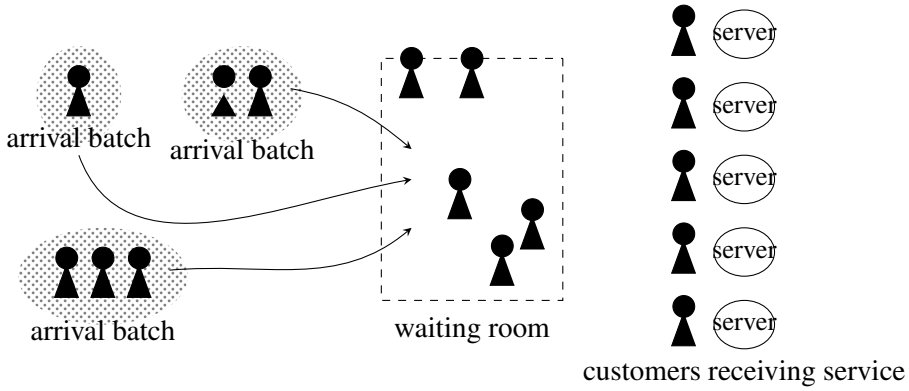

Figure 1.4: Batch arrival queue

which are purchased but not yet returned can be considered to customers of an infinite-server queue. The purchase of items in the production system is considered to the arrivals of customers in the infinite-server queue. The return of items in the production system are interpreted as the departures of customers in the infinite-server queue. Through the infinite-server queue, we can grasp the dynamics of the production system as illustrated in Figure 1.3. Furthermore, analyses of infinite-server queues help us to make the production plan parts ♦ and ♥.

1.3 Batch arrival queues

In batch arrival queues, multiple customers arrive the system in groups (i.e., batches). The number of customers belonging to a batch is referred to as the batch size. Batch sizes are generally assumed to have randomness. In queueing theory, the term bulk is sometimes used interchangeably with the term batch [20]. Bulk is often used in the application of transportation systems, whereas batch is often used in communication applications. This thesis uses the term batch.

Batch arrival processes are stochastic processes to imitate arrival times of batches and their sizes. Batch Poisson arrival processes are often seen in queueing theory. In the batch Poisson arrival process, batches arrive according to a Poisson process and batch sizes i.i.d. with a general distribution on $\mathbb{N}$. Batch Markovian arrival processes (BMAPs) [47] are also widely used ,where we introduce BMAPs in Section 2.2.1. BMAPs include various arrival processes as special cases, e.g., a batch Poisson arrival process, a phase-type (PH) renewal process [41], and a Markovian arrival process (MAP) [48]. Note that the MAP is a special case of BMAPs such that customers arrive one by one. Any simple point process is the weak limit of a sequence of MAPs [3].

1.4 Stability of queues

The stability of queues is defined as follows.

Definition 1.1 A queueing model is stable if its queue length process has a proper and non-degenerate limiting distribution.

In addition, the stability condition is defined as follows.

Definition 1.2 The stability condition is the necessary and sufficient condition that a queueing model is stable.

For simplicity, proper and non-degenerate limiting distributions are referred to as limiting distributions in this thesis. If the queueing model is stable, the queue length (i.e., the number of customers in the system) does not diverge infinitely under a long-time operation. On the other hand, if the queueing model is not stable, customers cannot finish receiving service in a finite time with a positive probability. Thus, the stability is an important property not only in theory but also in applications. We introduce the background of the study of the stability conditions in Sections 1.4.1 and 1.4.2.

1.4.1 Stability condition for finite multi-server queues

In this subsection, we introduce the background of the study of the stability condition for finite multi-server queues. Loynes [45] derived the stability condition for finite multi-server queues including batch arrival models. Loynes' stability criteria is the well known criteria in queueing theory.

We now consider a queueing model with $c \in \mathbb{N}$ servers and an infinite buffer space, where its queue length process is denoted by $\{L(t); t \in \mathbb{R}_+\}$. Let denote τ_n as the inter-arrival time between the nth and $(n-1)$st batches for $n \in \mathbb{Z}$. We define X_n as the size of the nth arriving batch for $n \in \mathbb{Z}$. We also define $\{S_{n,m}; 1 \leq m \leq X_n\}$ as the service times of customers belonging in the nth arriving batch for $n \in \mathbb{Z}$. Loynes [45] presented the stability condition for this finite multi-server queue as follows.

Theorem 1.1 *(Stability condition for finite multi-server queues [45]) Let assume that* $\{\tau_n; n \in \mathbb{Z}\}$, $\{X_n; n \in \mathbb{Z}\}$, *and* $\{S_n; n \in \mathbb{Z}\}$ *are independent of each other and stationary. Then, the queue length process* $\{L(t)\}$ *has a limiting distribution if and only if*

$$\frac{\mathsf{E}[X_1]}{\mathsf{E}[\tau_1]} \cdot \mathsf{E}[S_1] < c. \tag{1.1}$$

Theorem 1.1 gives a meaningful interpretation as follows. Finite multi-server queues are stable if and only if the average number of customers arriving per unit time is smaller than the average number of customers that the system can process per unit time.

1.4.2 Stability condition for infinite-server queues

In this subsection, we consider the stability condition for infinite-server queues. The stability of infinite-server queues means that the number of servers used simultaneously is finite with probability one. Assuming that multiple customers do not arrive at the same time, infinite-server queues are stable if and only if the mean inter-arrival time and the mean service time are finite [32]. On the other hand, in the case that multiple customers are allowed to arrive at the same time (i.e., batch arrival case), a batch-size distribution has significant effects on the stability condition.

Batch arrival infinite-server queues are not always stable even if the mean inter-arrival time and the mean service time are finite. However, few previous studies paid little attention to the stability condition for their models. Many researchers have studied batch arrival infinite-server queues at the steady state, assuming sufficient conditions for the stability (e.g., the first two moments of the batch-size distribution are finite) or the existence of a stationary queue length distribution. Some examples of such previous studies are presented in Section 2.1.

There slightly exist previous studies which derived the stability condition for batch arrival infinite-server queues. Actually, Pakes and Kaplan [56] obtained the stability condition of $\mathrm{GI}^X/\mathrm{GI}/\infty$ queues, which are infinite-server queues such that batches arrive according to a renewal process, batch sizes are i.i.d. with a general distribution on $\mathbb{N}$, and service times are i.i.d. with a general distribution on $\mathbb{R}_+$. In [56], the stability condition of $\mathrm{GI}^X/\mathrm{GI}/\infty$ queues can be obtained as a special case of a necessary and sufficient condition for the existence of a limiting distribution of the Bellman-Harris process. However, the stability condition derived in [56] does not appear to be well known, because most of the previous studies [10, 15, 28, 29, 35, 42, 44, 43, 51] do not cite the results in [56] and also do not mention the stability conditions for their own queueing models. Note that the general stability condition of $\mathrm{GI}^X/\mathrm{GI}/\infty$ queues is not explicitly presented in [56]. However, therein, the specific stability conditions of $\mathrm{GI}/\mathrm{GI}/\infty$ queues are explicitly presented for two cases: (i) the case in which the tail of the service time distribution is bounded (from above and below) by two Weibull-like tails and (ii) the case in which the tail of the service time distribution is regularly varying.

Cong [17] derived the stability condition for multiclass infinite-server queues with batch Poisson arrivals and class-dependent exponential service times, which is referred to as the $\mathrm{M}_K^X/\mathrm{M}_K/\infty$ queue. He showed that an $\mathrm{M}_K^X/\mathrm{M}_K/\infty$ queue is stable if and only if the logarithmic moment of the batch-size distribution is finite.

Remark 1.1 For the case in which the tail of the service time distribution is bounded (from above and below) by two Weibull-like tails, Pakes and Kaplan [56] presented an incorrect stability condition of $\mathrm{GI}^X/\mathrm{GI}/\infty$ queues. Theorem 3 of [56] states that a $\mathrm{GI}^X/\mathrm{GI}/\infty$ queue

is stable if and only if

$$\mathsf{E}[(\log X)^{\sigma}] < \infty, \qquad \text{for some } \sigma > 0, \tag{1.2}$$

provided that there exist some $x_0 > 0$ and $0 < b < a < \infty$ such that

$$\mathrm{e}^{-ax^{\sigma}} < 1 - H(x) < \mathrm{e}^{-bx^{\sigma}}, \qquad \text{for all } x > x_0.$$

However, (1.2) would not be correct and should be replaced by the following:

$$\mathsf{E}[(\log X)^{1/\sigma}] < \infty, \qquad \text{for some } \sigma > 0,$$

which can be easily proved using Lemma 1 of [56].

1.5 Markov-modulated queues

Markov-modulated queues change their parameters (e.g., arrival rates and service rates) depending on a Markov chain. Such a Markov chain is referred to as a background process. Markov-modulated queues have attracted a great deal of attention, in addition to the special cases of queueing models with constant parameters. Owing to dependence of the parameters on the background process, Markov-modulated queues can imitate more complex situation than queueing models with constant parameters. For example, in a transportation system, the background process may alternate between the accident state and the normal state. Under the accident state, the speed of cars is slower than that in the normal state [23]. In wireless communication, the transmission speed of a wireless channel may change between good and bad conditions [2].

In general, it is very difficult to exactly analyze Markov-modulated queues, except for some very simple models. Thus, some researchers have focused on their asymptotic model in some specific regimes.

1.6 Energy problem in data centers and queueing theory

In today's information society, it is a serious issue that energy consumption and transmission delay in data centers increase. In queueing theory, there exist many previous studies inspired by these problems. Analysis of queueing models can give today's data centers system parameters such as decreasing energy consumption and transmission delay. The important feature of today's data centers is that they are designed by a huge number of servers, which has become remarkable in recent years with the spread of cloud computing. It is also important that the arrivals of jobs are busty (see, e.g., [70]). Furthermore, the behavior of data centers tends to depend on the external environment.

It is desirable that data centers operate under high energy efficiency; that is, lower energy consumption and shorter transmission delay are realized simultaneously. However,

the higher the processing speed is, the higher the energy consumption per unit time is. Thus, it is difficult to achieve high energy efficiency.

1.6.1 Variable service speed

In recent years, variable-speed CPUs have become popular in order to reduce energy consumption while maintaining an acceptable transmission delay for jobs. Variable-speed CPUs can be automatically adjusted in terms of speed according to the workload or the number of jobs in the system by frequency scaling [59], dynamic voltage and frequency scaling techniques [52, 62], or other techniques. Working at high speed, the CPU reduces the transmission delay, but consumes more energy.

Motivated by the above, in queueing theory, many researchers have studied queueing models with variable service speed [4, 40, 54, 64]. For example, Lu et al. [46] considered a single-server queue such that customers arrive according to a Poisson process, service requirements of customers are i.i.d. with an exponential distribution, and the service speed changes in proportion to the queue length. They derived the stationary queue length distribution in the form of an infinite series. Adan and D'Auria [1] considered a single-server queue such that customers arrive according to a Poisson process, service requirements of customers are i.i.d. with an exponential distribution, and the service speed is controlled by thresholds. They derived the stationary queue length distribution and the Laplace–Stieltjes transform (LST) of the sojourn time distribution in an explicit form. Takine [64] studied a multiclass single-server queue in a Markovian random environment which govern the arrivals of customers and the service speed. He constructed a new queueing model with a constant service rate by means of time scale and then derived some quantities of the original model (variable service speed) using that of the new model (constant service speed).

1.6.2 On-off policy

CPUs still consume approximately 60% of their peak consumption during processing a job even while not processing jobs [5]. Thus, a simple idea for saving energy is to keep servers powered off while the system is empty. The server is turned off immediately after the system becomes empty, and the OFF server is reactivated immediately after a new customer arrives at the empty system. We refer to such an idea for saving energy as the on-off policy. However, a setup time is needed to reactivate the OFF server. Servers cannot process jobs during setup, but consume energy. The on-off policy is not always effective in decreasing energy consumption and transmission delay.

Some researchers have studied queueing models with the on-off policy (also referred to as queueing models with the setup time). For example, Baba [4] considered an M^X/M/1 queue with the exponential setup time. He derived the probability generating function (PGF) of the stationary queue length and the LST of the stationary sojourn time distribu-

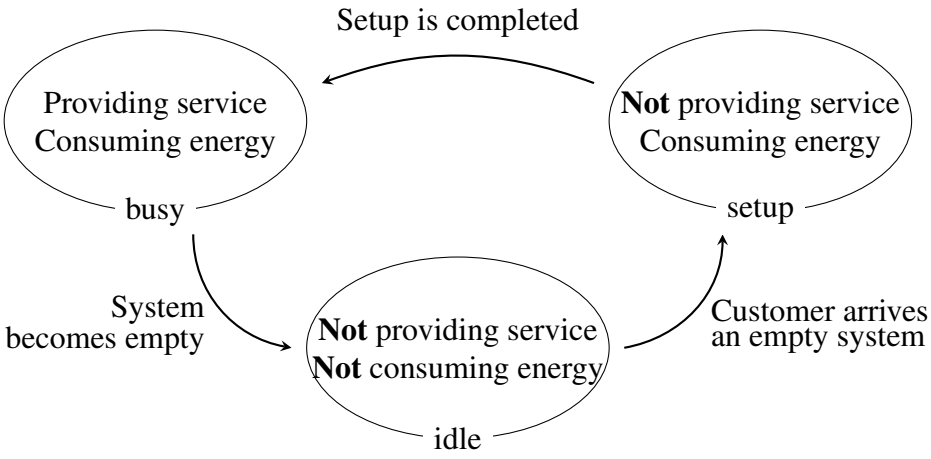

Figure 1.5: Queueing model with the on-off policy

tion. Phung-Duc [57] considered an M/M/c queue with the exponential setup time. He derived the closed form expression for the stationary queue length distribution by a generating function approach. Choudhury [16] studied an M^X/G/1 queue with the on-off policy in which the setup times follow a general distribution. He derived the PGF of the stationary queue length by using that of an ordinary M^X/G/1 queue.

Some researchers have also studied queueing models with the N policy, which is defined as follows. The server is turned off immediately after the system becomes empty. The OFF server is not turned on until N jobs are accumulated in the system even if a new customer arrives. The OFF server is turned on immediately after N customers are accumulated, but the setup time is needed to start providing service. The N policy can avoid frequent setups compared to the on-off policy. Hur and Paik [30] considered an M/G/1 queue with the N policy. They derived the stationary queue length distribution and the LST of the stationary waiting time. They investigated the optimal N reducing a cost function.

1.7 Organization of this thesis

This thesis studies infinite-server queues and related models. Chapter 2 investigates the stability for batch arrival infinite-server queues. Next, Chapter 3 considers a Markov-modulated batch arrival infinite-server queue such that customers may or may not leave the system without completing service due to accidents. Chapter 4 analyzes a batch arrival single-server queue such that the service speed changes in proportion to the queue length. Finally, Chapter 5 concludes this thesis and presents directions for future research. The contents of this thesis have been published as follows. Chapter 2 is based primarily on [71, 74], Chapter 3 is based primarily on [73], and Chapter 4 is based primarily on [72].

1.7.1 Chapter 2: Stability Condition for Batch Arrival Infinite-Server Queues

Chapter 2 studies the stability for batch arrival infinite-server queues. Our purpose is to obtain tractable stability conditions for batch arrival infinite-server queues. We first consider the stability for a BMAP/M/∞ queue, which is an infinite-server queue with a batch Markovian arrival process (BMAP) and an exponential service time distribution. We show that the stability condition of BMAP/M/∞ queues is that the logarithmic moment of the batch-size distribution is finite. In addition, using the stochastic ordering technique, we extend this result to the multiclass case.

Next, we study the stability for a GI^X/GI/∞ queue, which is an infinite-server queue such that batches arrive according to a renewal process and service times are i.i.d. with a general distribution. We show the stability condition of GI^X/GI/∞ queues using a different approach from [56]. Furthermore, in the case that the service time distribution has an exponential tail, we show that the GI^X/GI/∞ queue is stable if and only if the logarithmic moment of the batch-size distribution is finite.

1.7.2 Chapter 3: Central Limit Theorem for a Markov-Modulated Infinite-Server Queue with Binomial Catastrophes

Chapter 3 analyzes a Markov-modulated batch arrival infinite-server queue with catastrophe mechanism. Catastrophes can imitate situations such that customers may or may not leave the system without completing service due to accidents.

We establish a central limit theorem for the stationary queue length under a heavy traffic regime. Furthermore, we derive an approximation for the stationary queue length distribution using the central limit theorem, and then confirm the accuracy of this approximation through numerical experiments. In addition, we present the stability condition for this queueing model.

1.7.3 Chapter 4: Batch Arrival Single-Server Queue with Variable Service Speed

Chapter 4 studies a batch arrival single-server queue such that the service speed changes in proportion to the queue length. Note that the queue length process of this single-server queue is identical to that of an infinite-server queue because the service speed changes in proportion to the queue length.

We first present the stability condition of our queueing model. Next, we derive the probability generating function of the stationary queue length of our queueing model. We obtain the Laplace-Stieltjes transform of the stationary sojourn time distribution. Finally, we present numerical results to show the energy performance of our queueing model.

Chapter 2

Stability Condition for Batch Arrival Infinite-Server Queues

2.1 Introduction

Infinite-server queues have many applications in various areas, such as inventory systems [7], road traffic systems [68], and telecommunication systems [50]. Thus, many researchers have studied stationary and/or time-dependent infinite-server queues [14, 29, 34, 58, 60, 63]. However, almost all the previous works paid little attention to the *stability condition* for infinite-server queues. The stability condition is the necessary and sufficient condition that the queue length process has a proper and non-degenerate (i.e., the total probability on finite positive values is equal to one) limiting distribution. For simplicity, proper and non-degenerate limiting distributions are just called *limiting distributions* in this thesis.

Almost all the previous works have studied stationary infinite-server queues with batch arrivals, assuming sufficient conditions for stability (e.g., the first two moments of the batch-size distribution are finite) or the existence of the stationary queue length distribution. Holman et al. [29] derived some formulas for the mean and variance of the stationary queue length distribution in the $M^X/G/\infty$ queue, under the assumption that the first two moments of the batch-size distribution are finite. Keilson and Seidmann [34] assumed that the $M^X/G/\infty$ queue is stable and then proved that the stationary queue length distribution is a compound Poisson distribution under an additional condition. Breuer [12] derived the necessary and sufficient condition that the mean stationary queue length in the BMAP/G/∞ queue is finite.

As for the multiclass case, Liu and Templeton [44] considered an infinite-server queue (referred to as the $GR^{X_n}/G_n/\infty$ queue therein), where arrival times and types of customers are governed by a Markov renewal process and the batch sizes of customers depend on their types. For the $GR^{X_n}/G_n/\infty$ queue, they derived the probability generating function of the stationary queue length distribution under the assumption that all the moments of the

batch-size distribution are finite. Masuyama and Takine [51] derived explicit and numerically feasible formulas for the stationary joint queue length moments in an infinite-server queue with a multiclass batch Markovian arrival process and class-dependent phase-type service times, assuming that the stationary joint queue length distribution exists.

Actually, Pakes and Kaplan [56] obtained the stability condition of $GI^X/GI/\infty$ queues as a special case of a necessary and sufficient condition for the existence of a limiting distribution of the Bellman-Harris process. As far as we know, this fact does not seem well known because most of the previous studies [10, 15, 28, 29, 35, 42, 43, 44, 51] do not cite the results in [56] and, in the first place, they do not also mention the stability conditions of their own queueing models. We now note that the *general* stability condition of $GI^X/GI/\infty$ queues is not explicitly presented in [56] (see to Remark 1.1). However, therein, the specific stability conditions of $GI^X/GI/\infty$ queues are explicitly presented in the two cases: (i) the tail of the service time distribution is bounded (from above and below) by two Weibull-like tails; and (ii) is regularly varying.

Furthermore, Cong [17] derived the stability condition for multiclass infinite-server queues with batch Poisson arrivals and class-dependent exponential service times, which is referred to as the $M_K^X/M_K/\infty$ queue therein. He showed that an $M_K^X/M_K/\infty$ queues is stable if and only if the logarithmic moment of the batch-size distribution is finite. For convenience, we refer to this specific stability condition as the *the logarithmic batch size moment (LBSM) condition* in this chapter.

The main purpose of this chapter is to present a stability condition for general infinite-server queue with batch arrivals. We first consider the stability for BMAP/M/∞ queues, which is infinite-server queues with a batch Markovian arrival process (BMAP) and an exponential service time distribution. We show that the LBSM condition is the stability condition of BMAP/M/∞ queues. Using Foster's theorem (see, e.g., [11, Chapter 5, Theorem 1.1]), we prove that the LBSM condition is sufficient for the stability of the BMAP/M/∞ queue. We also show the necessity of the LBSM condition for stability in a similar way to Cong [17]. In addition, combining these results with the stochastic ordering technique, we prove that the LBSM condition is the stability condition of a multiclass BMAP/M/∞ queue, where customers arrive according to a multiclass batch Markovian arrival process (MBMAP) and service times of customers are independently distributed with class-dependent exponential distributions.

Next, we consider the stability for $GI^X/GI/\infty$ queues, which is infinite-server queues such that batches arrive according to a renewal process and service times of customers are independent and identically distributed (i.i.d.) with a general distribution. We present a stability condition of $GI^X/GI/\infty$ queues in the different way than [56]. We also show a tractable sufficient condition for the stability of $GI^X/GI/\infty$ queues under a moderate condition on the tail of the service time distribution. Furthermore, we prove that the LBSM condition is the stability condition of the $GI^X/GI/\infty$ queues whose service time distribu-

tions have exponential tails.

The reminder of this chapter is organized as follows. Section 2.2 derives the stability condition for the (single-class) BMAP/M/∞ queue. In addition, Section 2.3 extends the result in Section 2.2 to the multiclass case. Next, Section 2.4 studies the stability of GI^X/GI/∞ queues. Finally, Section 2.5 is devoted to concluding remarks and future work.

2.2 Stability condition for BMAP/M/∞ queues

This section considers a BMAP/M/∞ queue with a batch Markovian arrival process and an exponential service time distribution. We prove that the stability condition of BMAP/M/∞ queues are the LBSM condition; that is, the expectation of the logarithm of the batch-size distribution is finite.

2.2.1 Model description

We describe the BMAP/M/∞ queue. This queueing model has infinitely many servers. Customers arrive according to a batch Markovian arrival process (BMAP) [47]. The BMAP includes various arrival processes as special cases, e.g., a batch Poisson arrival process, a Phase-type (PH) renewal process [41], a Markovian arrival process (MAP) [48]. Note here that the MAP is a special case of BMAPs such that arrivals occur one by one. It is known [3] that any simple point process is the weak limit of a sequence of MAPs.

The BMAP is defined as follows. The BMAP is controlled by an irreducible time-homogeneous Markov chain $\{J(t); t \in \mathbb{R}_+\}$ in continuous time with finite state space $\mathbb{D} := \{1, 2, \ldots, d\}$, which is called the background Markov chain. Let $N(t)$, $t \in \mathbb{R}_+$, denote the total number of customers arriving from the BMAP during the time interval $(0, t]$, where $N(0) = 0$. We assume that, for $k \in \mathbb{Z}_+$ and $i, j \in \mathbb{D}$,

$$\begin{aligned}&\mathsf{P}(N(t+\Delta t) - N(t) = k, J(t+\Delta t) = j \mid J(t) = i)\\ &\quad = \begin{cases} 1 + D_{i,i}(0)\Delta t + o(\Delta t), & k = 0,\ i = j \in \mathbb{D},\\ D_{i,j}(k)\Delta t + o(\Delta t), & otherwise. \end{cases}\end{aligned}$$

Note here that $\boldsymbol{D}(k) := (D_{i,j}(k))_{i,j\in\mathbb{D}}$, $k \in \mathbb{N}$, is a nonnegative matrix and that $\boldsymbol{D}(0) := (D_{i,j}(0))_{i,j\in\mathbb{D}}$ is a diagonally dominant matrix with negative diagonal elements and nonnegative off-diagonal elements because of the irreducibility of the background Markov chain $J(t)$. Note also that $\boldsymbol{D} := \sum_{k=0}^{\infty} \boldsymbol{D}(k)$ is the infinitesimal generator of the background Markov chain $\{J(t)\}$; that is,

$$\boldsymbol{D}\boldsymbol{e} = \sum_{k=0}^{\infty} \boldsymbol{D}(k)\boldsymbol{e} = \boldsymbol{0}. \tag{2.1}$$

To avoid triviality, we assume that

$$\sum_{k=1}^{\infty} \boldsymbol{D}(k)\boldsymbol{e} \neq \boldsymbol{0}. \tag{2.2}$$

It is obvious that the joint stochastic process $\{(N(t), J(t)); t \in \mathbb{R}_+\}$ is a continuous-time Markov chain with state space $\mathbb{Z}_+ \times \mathbb{D}$, whose infinitesimal generator is given by

$$\begin{array}{c} \\ \mathbb{L}(0) \\ \mathbb{L}(1) \\ \mathbb{L}(2) \\ \mathbb{L}(3) \\ \vdots \end{array} \begin{array}{c} \begin{array}{ccccc} \mathbb{L}(0) & \mathbb{L}(1) & \mathbb{L}(2) & \mathbb{L}(3) & \cdots \end{array} \\ \begin{pmatrix} \boldsymbol{D}(0) & \boldsymbol{D}(1) & \boldsymbol{D}(2) & \boldsymbol{D}(3) & \cdots \\ \boldsymbol{O} & \boldsymbol{D}(0) & \boldsymbol{D}(1) & \boldsymbol{D}(2) & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{D}(0) & \boldsymbol{D}(1) & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{D}(0) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \end{array},$$

where $\mathbb{L}(k) = \{k\} \times \mathbb{D}$ for $k \in \mathbb{Z}_+$. As a result, the BMAP is characterized by $\{\boldsymbol{D}(k); k \in \mathbb{Z}_+\}$ and thus is referred to as BMAP $\{\boldsymbol{D}(k); k \in \mathbb{Z}_+\}$.

Each arriving customer occupies one of the servers immediately after its arrival, and leaves the system immediately after its service completion. The service times of customers are i.i.d. with the exponential distribution having mean $1/\mu \in (0, \infty)$. Therefore, customers behave independently of each other once they enter the system.

Let $L(t)$, $t \in \mathbb{R}_+$, denote the number of customers in the system at time t. It then follows from the Markov property of the BMAP and exponential service times that the joint stochastic process $\{(L(t), J(t)); t \in \mathbb{R}_+\}$ is a continuous-time Markov chain with state space $\mathbb{F} := \mathbb{Z}_+ \times \mathbb{D}$. Let $\boldsymbol{Q} := (q(k, i; \ell, j))_{(k,i),(\ell,j)\in\mathbb{F}}$ denote the infinitesimal generator of the Markov chain $\{(L(t), J(t))\}$. We then have

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{D}(0) & \boldsymbol{D}(1) & \boldsymbol{D}(2) & \boldsymbol{D}(3) & \cdots \\ \mu\boldsymbol{I} & \boldsymbol{\Lambda}_1(0) & \boldsymbol{D}(1) & \boldsymbol{D}(2) & \cdots \\ \boldsymbol{O} & 2\mu\boldsymbol{I} & \boldsymbol{\Lambda}_2(0) & \boldsymbol{D}(1) & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & 3\mu\boldsymbol{I} & \boldsymbol{\Lambda}_3(0) & \cdots \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & 4\mu\boldsymbol{I} & \ddots \\ \vdots & \vdots & \vdots & \ddots & \ddots \end{pmatrix}, \tag{2.3}$$

where $\boldsymbol{\Lambda}_k(0) := -k\mu\boldsymbol{I} + \boldsymbol{D}(0)$ for $k \in \mathbb{N}$.

Remark 2.1 From (2.2), (2.3) and the irreducibility of the background Markov chain $\{J(t)\}$, the Markov chain $\{(L(t), J(t))\}$ is irreducible. Therefore, $\{(L(t), J(t))\}$ has a limiting distribution (i.e., has a stationary distribution) if and only if it is positive recurrent (i.e., ergodic) [11, Chapter 8]. Furthermore, if there exists a stationary distribution of a n irreducible Markov chain, then it is unique and positive.

2.2.2 Stability condition

The following theorem shows the stability condition of BMAP/M/∞ queues.

Theorem 2.1 (Stability condition of BMAP/M/∞ queues) *$\{(L(t), J(t)); t \in \mathbb{R}_+\}$ is ergodic if and only if there exists some finite constant $C > 0$ such that*

$$\sum_{k=1}^{\infty} \log(k + \mathrm{e}) \boldsymbol{D}(k) \boldsymbol{e} \leq C \boldsymbol{e}. \tag{2.4}$$

The inequality (2.4) implies that the time average of the logarithm of the number of customers arriving in a batch is finite; that is, Theorem 2.1 shows that the LBSM condition (2.4) is the stability condition of the BMAP/M/∞ queue. In Sections 2.2.3 and 2.2.4, we separately prove the sufficiency and necessity of the LBSM condition (2.4) for the stability of the BMAP/M/∞ queue.

2.2.3 Proof for the sufficiency

For the proof for the sufficiency of Theorem 2.1, we use the following Lemma 2.1, which is Foster's theorem for continuous-time Markov chains.

Lemma 2.1 *[26, Chapter 2, Statement 8] Let consider a time-homogeneous Markov chain $\{X(t)\}$ with countable state space $\mathbb{S}$ and infinitesimal generator $\boldsymbol{Q} := (q_{s,p})_{s,p \in \mathbb{S}}$ such that $\boldsymbol{Q}\boldsymbol{e} = \boldsymbol{0}$. If there exists some function $\varphi : \mathbb{S} \to \mathbb{R}_+$ such that:*

1. *$\psi(s) := \sum_{p \in \mathbb{S}} q_{s,p} \varphi(p) < \infty$ for all $s \in \mathbb{S}$;*
2. *for some $\varepsilon > 0$, $\psi(s) \leq -\varepsilon$ for all $s \in \mathbb{S}$ except perhaps a finite number of states;*

then the Markov chain $\{X(t)\}$ is regular and ergodic.

In order to apply Lemma 2.1 to the Markov chain $\{(L(t), J(t))\}$, we present the following Lemma 2.2. It is immediate from Lemmas 2.1 and 2.2 that the LBSM condition (2.4) is a sufficient condition for the ergodicity of the irreducible generator $\boldsymbol{Q}$.

Lemma 2.2 *For $(k, i), (\ell, j) \in \mathbb{F}$, let $\upsilon(k, i)$ and $1_K(\ell, j)$ denote*

$$\upsilon(k, i) = \log(k + \mathrm{e}), \quad k \in \mathbb{Z}_+,\ i \in \mathbb{D},$$
$$1_K(\ell, j) = \begin{cases} 1, & \ell = 0, 1, \ldots, K, \qquad j \in \mathbb{D}, \\ 0, & \ell = K+1, K+2, \ldots, j \in \mathbb{D}, \end{cases}$$

respectively. If (2.4) holds, then there exist some $\delta \in (0, \infty)$ and $K \in \mathbb{Z}_+$ such that

$$\boldsymbol{Q}\boldsymbol{v} \leq -\delta \boldsymbol{e} + (\delta + C)\boldsymbol{1}_K, \tag{2.5}$$

where $\boldsymbol{v} = (\upsilon(k, i))_{(k,i) \in \mathbb{F}}$ and $\boldsymbol{1}_K = (1_K(\ell, j))_{(\ell,i) \in \mathbb{F}}$.

Proof. We define $\boldsymbol{y}(k)$, $k \in \mathbb{Z}_+$, as

$$\boldsymbol{y}(k) = \sum_{\ell=0}^{\infty} \boldsymbol{Q}(k;\ell)\boldsymbol{v}(\ell), \qquad k \in \mathbb{Z}_+,$$

where $\boldsymbol{Q}(k;\ell) = (q(k,i;\ell,j))_{i,j\in\mathbb{D}}$ for $k, \ell \in \mathbb{Z}_+$ and

$$\boldsymbol{v}(k) = (\upsilon(k,i))_{i\in\mathbb{D}} = \log(k+\mathrm{e})\boldsymbol{e}, \qquad k \in \mathbb{Z}_+.$$

We then have

$$\begin{aligned}\boldsymbol{y}(0) &= \sum_{\ell=0}^{\infty} \log(\ell+\mathrm{e})\boldsymbol{D}(\ell)\boldsymbol{e} \\ &= \boldsymbol{D}(0)\boldsymbol{e} + \sum_{\ell=1}^{\infty} \log(\ell+\mathrm{e})\boldsymbol{D}(\ell)\boldsymbol{e} \le C\boldsymbol{e}, \end{aligned} \tag{2.6}$$

where the last inequality follows from (2.4) and $\boldsymbol{D}(0)\boldsymbol{e} \le \boldsymbol{0}$. We also have, for $k \in \mathbb{N}$,

$$\begin{aligned}\boldsymbol{y}(k) &= k\mu[\boldsymbol{v}(k-1) - \boldsymbol{v}(k)] + \sum_{\ell=0}^{\infty} \boldsymbol{D}(\ell)\boldsymbol{v}(\ell+k) \\ &= k\mu \log\Big(1 - \frac{1}{k+\mathrm{e}}\Big)\boldsymbol{e} + \sum_{\ell=0}^{\infty} \boldsymbol{D}(\ell)\log(\ell+k+\mathrm{e})\boldsymbol{e}, \qquad k \in \mathbb{N}. \end{aligned} \tag{2.7}$$

Note here that

$$\log(\ell+k+\mathrm{e}) = \log(k+\mathrm{e}) + \log\Big(1 + \frac{\ell}{k+\mathrm{e}}\Big), \qquad k, \ell \in \mathbb{Z}_+.$$

Using this equation and (2.1), we obtain

$$\begin{aligned}\sum_{\ell=0}^{\infty} \log(\ell+k+\mathrm{e})\boldsymbol{D}(\ell)\boldsymbol{e} &= \log(k+\mathrm{e}) \sum_{\ell=0}^{\infty} \boldsymbol{D}(\ell)\boldsymbol{e} + \sum_{\ell=0}^{\infty} \log\Big(1 + \frac{\ell}{k+\mathrm{e}}\Big)\boldsymbol{D}(\ell)\boldsymbol{e} \\ &= \sum_{\ell=1}^{\infty} \log\Big(1 + \frac{\ell}{k+\mathrm{e}}\Big)\boldsymbol{D}(\ell)\boldsymbol{e}, \qquad k \in \mathbb{N}. \end{aligned}$$

It follows from this equation and (2.7) that

$$\boldsymbol{y}(k) = k\mu \log\Big(1 - \frac{1}{k+\mathrm{e}}\Big)\boldsymbol{e} + \sum_{\ell=1}^{\infty} \log\Big(1 + \frac{\ell}{k+\mathrm{e}}\Big)\boldsymbol{D}(\ell)\boldsymbol{e}, \qquad k \in \mathbb{N}. \tag{2.8}$$

We estimate the two terms in the right hand side of (2.8). It is easy to see that

$$\lim_{k\to\infty} k \log\Big(1 - \frac{1}{k+\mathrm{e}}\Big) = -1,$$

which shows that there exists some $\delta > 0$ such that

$$k\mu \log\Big(1 - \frac{1}{k+\mathrm{e}}\Big) \le -2\delta, \qquad k \in \mathbb{N}. \tag{2.9}$$

It also follows from (2.4) that, for all $k \in \mathbb{N}$,

$$\sum_{\ell=1}^{\infty} \log\Big(1 + \frac{\ell}{k+\mathrm{e}}\Big)\boldsymbol{D}(\ell)\boldsymbol{e} \le \sum_{\ell=1}^{\infty} \log(\ell+\mathrm{e})\boldsymbol{D}(\ell)\boldsymbol{e} \le C\boldsymbol{e}. \tag{2.10}$$

Applying (2.9) and (2.10) to (2.8), we obtain

$$\boldsymbol{y}(k) \le -2\delta\boldsymbol{e} + C\boldsymbol{e}, \qquad k \in \mathbb{N}. \tag{2.11}$$

In addition, by dominated convergence theorem, we have

$$\lim_{k\to\infty} \sum_{\ell=1}^{\infty} \log\Big(1 + \frac{\ell}{k+\mathrm{e}}\Big)\boldsymbol{D}(\ell)\boldsymbol{e} = \boldsymbol{0},$$

and thus there exists some $K := K_\delta \in \mathbb{Z}_+$ such that, for all $k = K+1, K+2, \ldots,$

$$\sum_{\ell=1}^{\infty} \log\Big(1 + \frac{\ell}{k+\mathrm{e}}\Big)\boldsymbol{D}(\ell)\boldsymbol{e} \le \delta\boldsymbol{e}.$$

Combining this inequality, (2.8) and (2.9), we obtain

$$\boldsymbol{y}(k) \le -\delta\boldsymbol{e}, \qquad k = K+1, K+2, \ldots. \tag{2.12}$$

Consequently, (2.5) follows from (2.6), (2.11) and (2.12). □

2.2.4 Proof for the necessity

The following lemma shows that the LBSM condition (2.4) holds if $\boldsymbol{Q}$ is ergodic; that is, $\boldsymbol{Q}$ has the unique stationary probability vector.

Lemma 2.3 *If $\boldsymbol{Q}$ has the unique stationary probability vector $\boldsymbol{\pi} = (\pi(k,i))_{(k,i)\in\mathbb{F}}$, then (2.4) holds for some finite constant $C > 0$.*

Proof. Let $\boldsymbol{\pi}(k) = (\pi(k,i))_{i\in\mathbb{D}}$ for $k \in \mathbb{Z}_+$, which is positive (see to Remark 2.1). It follows from the global balance equation $\boldsymbol{\pi}\boldsymbol{Q} = \boldsymbol{0}$ that

$$k\mu\boldsymbol{\pi}(k) = (k+1)\mu\boldsymbol{\pi}(k+1) + \sum_{\ell=0}^{k} \boldsymbol{\pi}(k-\ell)\boldsymbol{D}(\ell), \quad k \in \mathbb{Z}_+.$$

Multiplying the above equation by z^k and taking the sum over $k \in \mathbb{Z}_+$, we obtain, for $|z| \le 1$,

$$\mu\sum_{k=1}^{\infty} kz^k\boldsymbol{\pi}(k) = \mu\sum_{k=0}^{\infty}(k+1)z^k\boldsymbol{\pi}(k+1) + \sum_{k=0}^{\infty}\sum_{\ell=0}^{k} z^k\boldsymbol{\pi}(k-\ell)\boldsymbol{D}(\ell),$$

which leads to

$$\mu z \frac{\mathrm{d}}{\mathrm{d}z}\widehat{\boldsymbol{\pi}}(z) = \mu \frac{\mathrm{d}}{\mathrm{d}z}\widehat{\boldsymbol{\pi}}(z) + \widehat{\boldsymbol{\pi}}(z)\sum_{k=0}^{\infty} z^k \boldsymbol{D}(k), \tag{2.13}$$

where $\widehat{\boldsymbol{\pi}}(z) = \sum_{k=0}^{\infty} z^k \boldsymbol{\pi}(k)$. Postmultiplying both sides of (2.13) by $\boldsymbol{e}$ and rearranging the terms of the resulting equation, we have, for $|z| \le 1$,

$$\begin{aligned}
\mu(1-z)\frac{\mathrm{d}}{\mathrm{d}z}\widehat{\boldsymbol{\pi}}(z)\boldsymbol{e} &= -\widehat{\boldsymbol{\pi}}(z)\sum_{k=0}^{\infty} z^k \boldsymbol{D}(k)\boldsymbol{e} \\
&= -\widehat{\boldsymbol{\pi}}(z)\sum_{k=0}^{\infty} \boldsymbol{D}(k)\boldsymbol{e} + \widehat{\boldsymbol{\pi}}(z)\sum_{k=1}^{\infty}(1-z^k)\boldsymbol{D}(k)\boldsymbol{e} \\
&= \widehat{\boldsymbol{\pi}}(z)\sum_{k=1}^{\infty}(1-z^k)\boldsymbol{D}(k)\boldsymbol{e},
\end{aligned} \tag{2.14}$$

where we use (2.1) in the third equality. Furthermore, it follows from (2.14) that

$$\mu \frac{\mathrm{d}}{\mathrm{d}z}\widehat{\boldsymbol{\pi}}(z)\boldsymbol{e} = \widehat{\boldsymbol{\pi}}(z)\sum_{k=1}^{\infty}\frac{1-z^k}{1-z}\boldsymbol{D}(k)\boldsymbol{e}.$$

Integrating both sides of this equation over $z \in (0,1)$ and using $\widehat{\boldsymbol{\pi}}(z) \ge \widehat{\boldsymbol{\pi}}(0) = \boldsymbol{\pi}(0)$, we have

$$\begin{aligned}
\mu\{\widehat{\boldsymbol{\pi}}(1) - \boldsymbol{\pi}(0)\}\boldsymbol{e} &= \sum_{k=1}^{\infty}\int_0^1 \frac{1-z^k}{1-z}\widehat{\boldsymbol{\pi}}(z)\mathrm{d}z \cdot \boldsymbol{D}(k)\boldsymbol{e} \\
&\ge \boldsymbol{\pi}(0)\sum_{k=1}^{\infty}\boldsymbol{D}(k)\boldsymbol{e}\int_0^1 \frac{1-z^k}{1-z}\mathrm{d}z.
\end{aligned} \tag{2.15}$$

Note here that

$$\begin{aligned}
&\widehat{\boldsymbol{\pi}}(1)\boldsymbol{e} = 1, \\
&\int_0^1 \frac{1-z^k}{1-z}\mathrm{d}z = \sum_{\ell=1}^{k}\frac{1}{\ell} \ge \log(k+1) \ge \log(k+\mathrm{e})\frac{\log 2}{\log(1+\mathrm{e})}, \qquad k \in \mathbb{N}.
\end{aligned}$$

Substituting these into (2.15), we obtain

$$\boldsymbol{\pi}(0)\sum_{k=1}^{\infty}\log(k+\mathrm{e})\boldsymbol{D}(k)\boldsymbol{e} \le \frac{\mu\log(1+\mathrm{e})}{\log 2}\{1-\boldsymbol{\pi}(0)\boldsymbol{e}\}. \tag{2.16}$$

Since $\boldsymbol{\pi}(0) > \boldsymbol{0}$ and $0 < \boldsymbol{\pi}(0)\boldsymbol{e} < 1$, the inequality (2.16) completes the proof. □

2.3 Extension to the multiclass case of BMAP/M/∞ queues

In this section, we extend the result in Section 2.2 to the multiclass case; that is, we present the stability condition of an infinite-server queue with a multiclass batch Markovian arrival process and class-dependent exponential service times.

2.3.1 Model description

In this subsection, we describe the multiclass model of BMAP/M/∞ queues. This queueing model has infinitely many servers where customers arrive according to a multiclass batch Markovian arrival process (MBMAP). We assume that arriving customers are classified into K classes and the set of class indices is denoted by $\mathbb{K} := \{1, 2, \ldots, K\}$. For each $\nu \in \mathbb{K}$, the service times of class ν customers are i.i.d. with the exponential distribution having mean $1/\mu_\nu \in (0, \infty)$. We denote the multiclass infinite-server queue described above by MBMAP$_K$/M$_K$/∞, where the subscript "K" represents the number of classes.

The MBMAP is an extension of the BMAP described in Section 2.2.1. As in Section 2.2.1, the MBMAP has the background Markov chain $\{J(t); t \in \mathbb{R}_+\}$ with state space $\mathbb{D}$ and irreducible infinitesimal generator $\boldsymbol{D}$. For $\nu \in \mathbb{K}$, let $N_\nu(t)$, $t \in \mathbb{R}_+$, denote the total number of class ν customers who arrive from the MBMAP during the time interval $(0, t]$, where $N_\nu(0) = 0$. Let $N(t) := \sum_{\nu\in\mathbb{K}} N_\nu(t)$ for $t \in \mathbb{R}_+$. We then assume that, for $i, j \in \mathbb{D}$,

$$\begin{aligned}&\mathsf{P}(N(t+\Delta t) - N(t) = 0, J(t+\Delta t) = j \mid J(t) = i)\\ &\quad= \begin{cases} 1 + D_{i,i}(0)\Delta t + o(\Delta t), & i = j \in \mathbb{D},\\ D_{i,j}(0)\Delta t + o(\Delta t), & otherwise,\end{cases}\end{aligned}$$

where $\boldsymbol{D}(0) := (D_{i,j}(0))_{i,j\in\mathbb{D}}$ is a diagonally dominant matrix with negative diagonal and nonnegative off-diagonal elements. We also assume that, for $\nu \in \mathbb{K}$, $k \in \mathbb{N}$, and $i, j \in \mathbb{D}$,

$$\begin{aligned}&\mathsf{P}(N_\nu(t+\Delta t) - N_\nu(t) = k, J(t+\Delta t) = j \mid J(t) = i)\\ &\quad= D_{\nu,i,j}(k)\Delta t + o(\Delta t),\end{aligned} \tag{2.17}$$

where $\boldsymbol{D}_\nu(k) := (D_{\nu,i,j}(k))_{i,j\in\mathbb{D}}$, $\nu \in \mathbb{K}$, $k \in \mathbb{N}$, is a nonnegative matrix such that $\boldsymbol{D}(0) + \sum_{\nu\in\mathbb{K}}\sum_{k=1}^{\infty}\boldsymbol{D}_\nu(k)$ is equal to the infinitesimal generator of the background Markov chain $\{J(t)\}$; that is,

$$\boldsymbol{D}(0) + \sum_{\nu\in\mathbb{K}}\sum_{k=1}^{\infty}\boldsymbol{D}_\nu(k) = \boldsymbol{D}. \tag{2.18}$$

It follows from (2.17) and (2.18) that the classes of the customers in a batch are same and thus their service times are independently distributed with the same exponential distribution.

To avoid triviality, we assume that

$$\sum_{k=1}^{\infty}\boldsymbol{D}_\nu(k)\boldsymbol{e} \neq \boldsymbol{0}, \qquad \text{for all } \nu \in \mathbb{K}.$$

As a result, the MBMAP is characterized by $\{\boldsymbol{D}(0), \boldsymbol{D}_\nu(k); \nu \in \mathbb{K}, k \in \mathbb{N}\}$. We denote the MBMAP described above by MBMAP $\{\boldsymbol{D}(0), \boldsymbol{D}_\nu(k); \nu \in \mathbb{K}, k \in \mathbb{N}\}$.

2.3.2 Stability condition

Let $\boldsymbol{L}(t) = (L_1(t), L_2(t), \ldots, L_K(t))$ for $t \in \mathbb{R}_+$, where $L_\nu(t)$ denotes the number of class ν customers in the system at time t. It then follows that the joint stochastic process $\{(\boldsymbol{L}(t), J(t)); t \in \mathbb{R}_+\}$ is an irreducible Markov chain with state space $\mathbb{Z}_+^K \times \mathbb{D}$. Thus, the following theorem means that the stability condition of MBMAP$_K$/M$_K$/∞ queues is the LBSM condition, which is the same as the single-class model; that is, BMAP/M/∞ queues.

Theorem 2.2 (Stability condition of MBMAP$_K$/M$_K$/∞ queues) *The Markov chain* $\{(\boldsymbol{L}(t), J(t))\}$ *is ergodic if and only if there exists some finite constant* $C > 0$ *such that*

$$\sum_{k=1}^{\infty} \log(k + \mathrm{e}) \boldsymbol{D}_*(k) \boldsymbol{e} \le C \boldsymbol{e}, \tag{2.19}$$

where $\boldsymbol{D}_*(k) = \sum_{\nu \in \mathbb{K}} \boldsymbol{D}_\nu(k)$ *for* $k \in \mathbb{N}$.

Remark 2.2 Theorem 2.2 is a generalization of [17, Lemma 2], which presents a necessary and sufficient condition for the stability of a multiclass infinite-server queue with batch Poisson arrivals and class-dependent exponential service times.

Proof of Theorem 2.2. Besides the original MBMAP$_K$/M$_K$/∞ queue, we consider two MBMAP$_K$/M$_K$/∞ queues, denoted by Queues 1 and 2, which are fed by the same arrival process as that of the original queue; that is, fed by MBMAP $\{\boldsymbol{D}(0), \boldsymbol{D}_\nu(k); \nu \in \mathbb{K}, k \in \mathbb{N}\}$. In Queue 1 (resp. 2), all the service times are i.i.d. with the exponential distribution having mean $1/\mu_{\min}$ (resp. $1/\mu_{\max}$), where

$$\mu_{\min} = \min_{\nu \in \mathbb{K}} \mu_\nu, \qquad \mu_{\max} = \max_{\nu \in \mathbb{K}} \mu_\nu.$$

Clearly, Queues 1 and 2 can be considered single-class BMAP/M/∞ queues when the class of customers are ignored, where the arrival process is reduced to BMAP $\{\boldsymbol{D}(0), \boldsymbol{D}_*(k); k \in \mathbb{N}\}$.

Let $|\boldsymbol{L}(t)| = \sum_{\nu \in \mathbb{K}} L_\nu(t)$ for $t \in \mathbb{R}_+$, which denotes the total number of customers in the system of the original MBMAP$_K$/M$_K$/∞ queue at time t. Let $L^{(1)}(t)$ (resp. $L^{(2)}(t)$), $t \in \mathbb{R}_+$, denote the total number of customers in the system of Queue 1 (resp. 2) at time t. From the assumption of Queues 1 and 2, we can construct the three joint processes $\{(\boldsymbol{L}(t), J(t)); t \in \mathbb{R}_+\}$, $\{(L^{(1)}(t), J(t)); t \in \mathbb{R}_+\}$, $\{(L^{(2)}(t), J(t)); t \in \mathbb{R}_+\}$ in a common probability space such that the following pathwise ordered relation holds:

$$L^{(2)}(t) \le |\boldsymbol{L}(t)| \le L^{(1)}(t), \qquad \text{for all } t \in \mathbb{R}_+, \tag{2.20}$$

which is proved in Appendix A.

It should be noted that $\{(L^{(1)}(t), J(t))\}$ and $\{(L^{(2)}(t), J(t))\}$ are the Markov chains of the same type as $\{(L(t), J(t))\}$ discussed in the previous section. Thus, it follows from Theorem 2.1 that (2.19) holds if and only if $\{(L^{(1)}(t), J(t))\}$ and $\{(L^{(2)}(t), J(t))\}$ are ergodic.

We now suppose that $\{(L^{(1)}(t), J(t))\}$ is ergodic. It then follows from (2.20) that $\{L^{(1)}(t)\}$ and thus $\{|\boldsymbol{L}(t)|\}$ take the value of zero infinitely many times with probability one and the mean recurrence time to state 0 is finite (see, e.g., [11, Chapter 8, Definitions 5.1, 5.2 and 5.4]). Therefore, $\{(\boldsymbol{L}(t), J(t))\}$ is ergodic.

On the other hand, we suppose that $\{(L^{(2)}(t), J(t))\}$ is not ergodic, i.e., is transient or null-recurrent. Note that if $\{(L^{(2)}(t), J(t))\}$ is transient then $\{L^{(2)}(t)\}$ and thus $\{|\boldsymbol{L}(t)|\}$ take the value of zero, at most, finitely many times with some positive probability. Note also that if $\{(L^{(2)}(t), J(t))\}$ is null-recurrent then the mean recurrence times to state 0 of $\{L^{(2)}(t)\}$ and thus $\{|\boldsymbol{L}(t)|\}$ are infinite. Therefore, in both cases, $\{(\boldsymbol{L}(t), J(t))\}$ is not ergodic. As a result, the above argument shows that (2.19) holds if and only if $\{(\boldsymbol{L}(t), J(t))\}$ is ergodic. □

2.4 Stability analysis for GI^X/GI/∞ queues

This section studies the stability of GI^X/GI/∞ queue. As mention in Section 2.1, Pakes and Kaplan [56] derived the stability condition of GI^X/GI/∞ queues. However, the stability condition presented in [56] is not explicitly and not physically interpretable, and then do not be cited by most of previous studies. In this section, we present the stability condition of GI^X/GI/∞ queues in the different way than [56]. We also show a tractable sufficient condition for the stability under a moderate condition on the tail of the service time distribution. Furthermore, supposing that the service time distribution has an exponential tail, we prove that the LBSM condition is the stability condition of the GI^X/GI/∞ queues.

2.4.1 Model description

In this subsection, we describe the GI^X/GI/∞ queue and define some notations. Let T_n, $n \in \mathbb{N}$, denote the nth arrival time, where

$$T_0 := 0 < T_1 < T_2 < \cdots .$$

Let $\tau_n := T_n - T_{n-1}$ for $n \in \mathbb{N}$, which is the inter-arrival time between the nth and $(n-1)$st arrivals. We then assume that $\{\tau_n; n \in \mathbb{N}\}$ is i.i.d. random variables with distribution function G on $\mathbb{R}_+$; that is,

$$\mathsf{P}(\tau_n \le x) = G(x), \qquad x \in \mathbb{R}_+.$$

Moreover, let X_n, $n \in \mathbb{N}$, denote the number of customers arriving in a batch at time T_n, where the X_n are referred to as batch sizes. We assume that $\{X_n; n \in \mathbb{N}\}$ is independent of $\{\tau_n; n \in \mathbb{N}\}$ and i.i.d. with distribution $(b_k; k \in \mathbb{N})$; that is,

$$\mathsf{P}(X_n = k) = b_k, \qquad k \in \mathbb{N}.$$

We also assume that each arriving customer immediately occupies one server and its service starts, and that the service times of customers are i.i.d. with distribution function H on $\mathbb{R}_+$ independently of $\{\tau_n\}$ and $\{X_n\}$. For later convenience, let $S_{n,1}, S_{n,2}, \ldots, S_{n,X_n}$, $n \in \mathbb{N}$, denote the service times of the customers arriving in a batch at time T_n. By definition, for $k \in \{1, 2, \ldots, X_n\}$,

$$\mathsf{P}(S_{n,k} \le x) = H(x), \qquad x \in \mathbb{R}_+.$$

We define $\{L(t); t \in \mathbb{R}_+\}$ as the queue length process of the GI^X/GI/∞ queue described above. We then have

$$L(t) = \sum_{n=1}^{\infty} \sum_{m=1}^{X_n} I\left(0 \le t - T_n < S_{n,m}\right), \qquad t \in \mathbb{R}_+. \tag{2.21}$$

Let τ and X denote generic random variables for $\{\tau_n\}$ and $\{X_n\}$, respectively. Let $\{S, S_1, S_2, S_3, \ldots\}$ denote a sequence of i.i.d. random variables with distribution H. In Section 2.4, we assume the following.

Condition 1 *The inter-arrival time distribution G is non-lattice and has the finite mean; that is,* $\mathsf{E}[\tau] < \infty$.

2.4.2 Stability condition

We derive the stability condition of GI^X/GI/∞ queues as follows.

Theorem 2.3 (Stability condition of GI^X/GI/∞ queues) *Under Condition 1, the queue length process $\{L(t)\}$ has a limiting distribution if and only if*

$$\mathsf{E}\left[\max_{m=1,2,\ldots,X} S_m\right] < \infty. \tag{2.22}$$

Remark 2.3 Toyoizumi [66] showed that the stability condition of M^X/GI/∞ queues is (2.22).

Remark 2.4 Theorem 1 of [56] presented a necessary and sufficient condition for the existence of the limiting distribution of the Bellman-Harris process. By setting $p_0 = 1$

(which implies that each object produces no progeny), we can show (see the paragraph between Theorems 2 and 3 of [56]) that a GI^X/GI/∞ queue is stable if and only if the following inequality holds.

$$\int_0^\infty \left\{ 1 - \sum_{k=1}^{\infty} \mathsf{P}(X = k) \cdot H(x)^k \right\} \mathrm{d}x < \infty. \tag{2.23}$$

By Fubini's theorem, we can also show that (2.23) is equivalent to (2.22). In the proof of Theorem 2.3, we use the different way than [56].

Proof of Theorem 2.3. We provide a simple and intuitive proof of this theorem, independently of Theorem 1 of [56]. Let

$$L_1(t) = \sum_{n=1}^{\infty} I\left(0 \le t - T_n < \max_{m=1,2,\ldots,X_n} S_{n,m} \right), \qquad t \in \mathbb{R}_+, \tag{2.24}$$

$$L_2(t) = \sum_{n=1}^{\infty} X_n \cdot I\left(0 \le t - T_n < \max_{m=1,2,\ldots,X_n} S_{n,m} \right), \qquad t \in \mathbb{R}_+. \tag{2.25}$$

It then follows from (2.21), (2.24) and (2.25) that

$$L_1(t) \le L(t) \le L_2(t), \qquad t \in \mathbb{R}_+.$$

By definition, $\{L_1(t); t \in \mathbb{R}_+\}$ is equivalent to the queue length process of a GI/GI/∞ queue obtained by treating customers arriving in each batch of the original GI^X/GI/∞ queue as a single *super customer* whose service time is equal to the maximum service time in the batch. On the other hand, $\{L_2(t); t \in \mathbb{R}_+\}$ is equivalent to the queue length process of a GI^X/GI/∞ queue obtained by assuming that customers in each batch leave the system simultaneously when the longest service in the batch is completed.

We note that $\{L_1(t)\}$ and $\{L_2(t)\}$ visit state 0 simultaneously and leave there simultaneously. Thus, these two processes visit state 0 simultaneously and leave there simultaneously. In addition, if the two processes visit state 0 infinitely many times, then they are regenerative processes with common regeneration times at which they leave state 0. Therefore, (3) implies that the stability conditions of $\{L(t)\}$, $\{L_1(t)\}$, and $\{L_2(t)\}$ are equivalent.

We now recall that $\{L_1(t)\}$ is the queue length process of the GI/GI/∞ queue where inter-arrival times follow non-lattice distribution G with finite mean, and where service times follow the distribution of $\max_{m=1,2,\ldots,X} S_m$. Thus, (2.22) holds if and only if $\{L_1(t)\}$ is stable (see [32, Theorem 0]). As a result, (2.22) is the stability condition of the queue length process $\{L(t)\}$ of the original GI^X/GI/∞ queue. □

2.4.3 Tractable sufficient conditions for the stability

We show a tractable sufficient condition for the stability under a moderate condition on the tail of the service time distribution.

Corollary 2.1 *Suppose that Condition 1 holds, and that there exists some increasing and convex function $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that it follows that, for any $c > 0$ and some $K_c > 0$,*

$$\frac{f^{-1}(cx)}{f^{-1}(x)} < K_c, \qquad \textit{for any } x \geq 0. \tag{2.26}$$

Under these conditions, if the following inequalities hold

$$\mathsf{E}[f(S)] < \infty, \tag{2.27}$$

$$\mathsf{E}[f^{-1}(X)] < \infty, \tag{2.28}$$

then the queue length process $\{L(t)\}$ has a limiting distribution .

Remark 2.5 Condition (2.26) means that the inverse function of f is dominated variation [21]. Typical examples of function f satisfying (2.26) are as follows: (i) $f(x) = \exp\{x^\alpha\}$ with $\alpha > 0$ and (ii) $f(x) = x^\beta$ with $\beta > 0$. For the second example of f, Toyoizumi [66] presented a result similar to Corollary 2.1, though his queueing model is an M^X/GI/∞ queue.

Proof of Corollary 2.1. It suffices to show that (2.22) holds. Noting f is convex, and using Jensen's inequality [31], we have

$$\begin{aligned}
\mathsf{E}\left[\max_{k=1,2,\ldots,X} S_k\right] &= \mathsf{E}\left[\mathsf{E}\left[\max_{k=1,2,\ldots,X} S_k \middle| X\right]\right] \\
&= \mathsf{E}\left[f^{-1} \circ f\left(\mathsf{E}\left[\max_{k=1,2,\ldots,X} S_k \middle| X\right]\right)\right] \\
&\leq \mathsf{E}\left[f^{-1}\left(\mathsf{E}\left[f\left(\max_{k=1,2,\ldots,X} S_k\right) \middle| X\right]\right)\right] \\
&= \mathsf{E}\left[f^{-1}\left(\mathsf{E}\left[\max_{k=1,2,\ldots,X} f(S_k) \middle| X\right]\right)\right].
\end{aligned} \tag{2.29}$$

Since $\{S_k\}$ is i.i.d. random variables and independent of X, we obtain

$$\mathsf{E}\left[\max_{k=1,2,\ldots,X} f(S_k) \middle| X\right] \leq \mathsf{E}\left[\sum_{k=1}^{X} f(S_k) \middle| X\right] = \sum_{k=1}^{X} \mathsf{E}[f(S_k)] = X\mathsf{E}[f(S)]. \tag{2.30}$$

Substituting (2.30) into (2.29) yields

$$\mathsf{E}\left[\max_{k=1,2,\ldots,X} S_k\right] \leq \mathsf{E}\left[f^{-1}(X\mathsf{E}[f(S)])\right].$$

Applying (2.26)– (2.28) to the above inequality, we obtain

$$\mathsf{E}\left[\max_{k=1,2,\ldots,X} S_k\right] \le \mathsf{E}\left[f^{-1}(X\mathsf{E}[f(S)])\right] \le K\mathsf{E}[f(S)] \cdot \mathsf{E}\left[f^{-1}(X)\right] < \infty.$$

Consequently, we complete the proof of Corollary 2.1. □

2.4.4 Stability condition for the special case

We also prove that the LBSM condition is the stability condition of the GI^X/GI/∞ queues whose service time distributions have exponential tails.

Corollary 2.2 *Suppose that Condition 1 holds and that there exist some $\alpha, \beta > 0$ such that*

$$0 < \liminf_{x\to\infty} \frac{1-H(x)}{\mathrm{e}^{-\alpha x}} \le \limsup_{x\to\infty} \frac{1-H(x)}{\mathrm{e}^{-\beta x}} < \infty. \tag{2.31}$$

Then, the queue length process $\{L(t)\}$ has a limiting distribution if and only if

$$\mathsf{E}[\log X] < \infty. \tag{2.32}$$

Proof. Let $f(x) = \mathrm{e}^{\theta x}$ for all $x \in \mathbb{R}_+$, where $0 < \theta < \beta$. It then follows from (2.31) that $\mathsf{E}[f(S)] = \mathsf{E}[\mathrm{e}^{\theta S}] < \infty$. Furthermore, $f^{-1}(x) = \theta^{-1}\log x$ for $x \in \mathbb{R}_+$ and thus (2.32) yields $\mathsf{E}[f^{-1}(X)] = \theta^{-1}\mathsf{E}[\log X] < \infty$. Therefore, from Corollary 2.1, if (2.32) holds then $\{L(t)\}$ has a limiting distribution.

In what follows, we prove the "only if" part of the statement that (2.22) implies (2.32). According to (2.31), there exist some $\theta_* \ge \alpha$ and $x_* > 0$ such that

$$\mathsf{P}(S > x) = 1 - H(x) \ge \mathrm{e}^{-\theta_* x}, \qquad \text{for all } x \ge x_*. \tag{2.33}$$

Using (2.33), we have

$$\begin{aligned}
\mathsf{E}\left[\max_{k=1,2,\ldots,X} S_k\right] &= \mathsf{E}\left[\int_0^\infty \left\{1-(H(t))^X\right\} \mathrm{d}t\right] \\
&\ge \mathsf{E}\left[\int_{x_*}^\infty \left\{1-(H(t))^X\right\} \mathrm{d}t\right] \\
&\ge \mathsf{E}\left[\int_{x_*}^\infty \left\{1-(1-\mathrm{e}^{-\theta_* x})^X\right\} \mathrm{d}t\right].
\end{aligned} \tag{2.34}$$

Furthermore, let $M_* = 1 - \mathrm{e}^{-\theta_* x_*} \in (0,1)$. We then have, for all $k \in \mathbb{N}$,

$$\begin{aligned}
\int_{x_*}^\infty \left\{1-(1-\mathrm{e}^{-\theta_* x})^k\right\} \mathrm{d}t &= \frac{1}{\theta_*}\int_{M_*}^1 \frac{1-z^k}{1-z}\mathrm{d}z = \frac{1}{\theta_*}\sum_{\ell=0}^{k-1}\int_{M_*}^1 z^\ell \mathrm{d}z \\
&= \frac{1}{\theta_*}\sum_{\ell=1}^{k}\frac{1}{\ell}(1-M_*^\ell) \ge \frac{1}{\theta_*}\sum_{\ell=1}^{k}\frac{1}{\ell}(1-M_*) \\
&\ge \frac{1-M_*}{\theta_*}\log k.
\end{aligned}$$

Applying this inequality to (2.34) yields

$$\mathsf{E}\left[\max_{k=1,2,\ldots,X} S_k\right] \geq \frac{1-M_*}{\theta_*}\mathsf{E}[\log X].$$

Therefore, (2.22) implies (2.32). □

2.5 Conclusion

In this chapter, we discussed the stability condition for batch arrival infinite-server queues. Section 2.2 showed that the LBSM condition is the stability condition of BMAP/M/∞ queues. Section 2.3 extended the result in Section 2.2 to the multiclass case; that is, we showed that the LBSM condition is also the stability condition of MBMAP$_K$/M$_K$/∞ queues. Section 2.4 considered the stability of GI^X/GI/∞ queues. We derived the stability condition of GI^X/GI/∞ queues in the different way than [56]. In addition, we presented a tractable sufficient condition for the stability for GI^X/GI/∞ queues under a modulate condition on the tail of the service time distribution. Furthermore, we proved that a GI^X/GI/∞ queue is stable if and only if the LBSM condition holds, provided that the service time distribution has an exponential tail.

In future work, we would like to derive a physically and interpretable stability condition of GI^X/GI/∞ queues without additional conditions. We predict that the condition of Corollary 2.1 is not only the sufficient condition but also the necessary condition. Furthermore, we would like to derive the stability condition for batch arrival infinite-server queues such that there exist correlations between inter-arrival times, batch sizes and service times.

Chapter 3

Central Limit Theorem for a Markov-Modulated Infinite-Server Queue with Binomial Catastrophes

3.1 Introduction

This chapter studies a Markov-modulated batch arrival infinite-server queue such that customers may or may not leave the system without completing service due to accidents. Markov-modulated queues are governed by a continuous-time Markov chain being independent the system, which is called *the background process*. In recent years, Markov-modulated queues have attracted much attention in addition to their special cases; that is, models with constant parameters [6, 8, 55]. Due to the dependence of parameters on the background process, Markov-modulated queues imitate more complex dynamics than queuing models with constant parameters. For example, in a transportation system, the background process may alternate between the accident state and the normal state. Under the accident state, the speed of cars is slower than that in the normal state [23].

In general, it is very difficult to exactly analyze Markov-modulated queues, except for some very simple models. Thus, researchers have usually considered their asymptotic model in some specific regimes. For example, Nazarov and Baymeeva [53] studied the Markov-modulated M/G/∞ queue and showed the asymptotic behavior of the characteristic function of the stationary queue length in a heavy traffic regime. Blom et al. [9] studied the Markov-modulated $\mathrm{M}/\mathrm{M}/\infty$ queue and established a central limit theorem for the stationary and transient queue length in a heavy traffic regime.

Catastrophe mechanism can imitate accidents inducing departure of customers. Queueing models with catastrophe mechanism are suitable for modeling computer systems subject to technical obstacles. For example, upon the occurrence of a virus, a job may or may not be affected. An affected job is removed, while the non-affected one is retained in the

system. Motivated by such applications, many researchers have studied queueing models with catastrophe mechanisms in recent years [19, 38, 39]. In particular, this chapter considers *the binomial catastrophe mechanism* [13, 33]. When a binomial catastrophe occurs, each customer is either retained with probability p or removed with probability $1-p$ without completing its service, independently of other customers. Thus, if a binomial catastrophe occurs when n customers are in the system, the number of retained customers follows the binomial distribution with parameters n and p.

In this chapter, we consider the Markov-modulated $\mathrm{M}^{\mathrm{X}}/\mathrm{M}/\infty$ queue with binomial catastrophes. Note here that an $\mathrm{M}^{\mathrm{X}}/\mathrm{M}/\infty$ queue is a batch arrival infinite-server queues with a batch Poisson arrival process and an exponential service time distribution. Binomial catastrophes are assumed to occur according to a homogeneous Poisson process. We assume that the arrival rate, the batch-size distribution, the service speed, and the occurrence rate and the retained probability of binomial catastrophes depend on the background process. When the batch sizes are one and the occurrence rates of binomial catastrophes are zero, our model coincides with the one in [9].

Our model may be used to estimate the distribution of the number of users in a service company. Let's consider a marketing situation, for example, in a mobile phone company. The company occasionally do some campaigns to increase its number of users. Thus, users join the company in batch upon such a campaign. The duration that the user stays with the company corresponds to the service time in our model. On the other hand, rival companies also do campaigns to increase their number of users. As a result, upon a campaign of a rival company, the user of the original company either opts for the rival with a probability or continues to stay with the original company with the commentary probability. The campaign of rival company can be interpreted as the catastrophe mechanism in our model. The random environment may reflect the satisfaction of users. In case users satisfy with the company, they will use the service for a longer time and more users will join. In case of dissatisfaction, users stay with the company for a shorter time and less users join the company. From these points of view, the environment corresponds to the satisfaction of the users with the company.

In this chapter, we establish a central limit theorem (CLT) for the stationary queue length of our model in a heavy traffic regime; that is, the centered and normalized stationary queue length distribution of the scaling model converges in distribution to a normal distribution. In our scaling regime, for scaling factor N and scaling coefficient α, the arrival rates are scaled by N, the transition rates of the background process are scaled by N^{α}, the occurrence rates of binomial catastrophes are scaled by N, and the removal probabilities are scaled by N^{-1}. In addition, we can easily obtain the approximation of the stationary queue length distribution by using the CLT, which is especially effective in heavy traffic situations. Furthermore, we show that the stability condition for our model is that the logarithmic moment of batch-size distribution is finite.

The reminder of this chapter is organized as follows. Section 3.2 describes the Markov-modulated $\mathrm{M}^{\mathrm{X}}/\mathrm{M}/\infty$ queue with binomial catastrophes and our scaling model. In Section 3.3, we establish the CLT for the stationary queue length, and derive an approximation of the stationary queue length distribution. In Section 3.4, we present the stability condition of our model. In Section 3.5, we show some numerical results to confirm the accuracy of the approximation of the stationary queue length distribution led by the CLT. Finally, Section 3.6 is devoted to concluding remarks and future work.

3.2 Model description

In this section, we describe our queueing models, i.e., the Markov-modulated $\mathrm{M}^{\mathrm{X}}/\mathrm{M}/\infty$ queue with binomial catastrophes. Section 3.2.1 presents the original model and, Section 3.2.2 presents the scaling model.

3.2.1 Original model

We consider a batch arrival infinite-server queue governed by the background process $\{J(t); t \in \mathbb{R}_+\}$. We assume that $\{J(t)\}$ is an irreducible continuous-time Markov chain with finite state space $\mathbb{D} = \{1, 2, \ldots, d\}$. Let $\boldsymbol{Q} = (q_{i,j})_{i,j\in\mathbb{D}}$ denote the infinitesimal generator of $\{J(t)\}$, and let $\boldsymbol{\tau} = (\tau_i)_{i\in\mathbb{D}}$ denote the stationary distribution of $\{J(t)\}$; that is, it follows that $\boldsymbol{\tau Q} = \boldsymbol{0}$.

For each $i \in \mathbb{D}$, when the background process is in state i, batches arrive according to the Poisson process with rate $\lambda_i \in (0, \infty)$ and batch sizes (i.e., the number of customers belonging to respective batches) are distributed with random variable X_i on $\mathbb{N}$. For $i \in \mathbb{D}$, we define $x_{i,k} = \mathsf{P}(X_i = k)$, $k \in \mathbb{N}$, and $\widehat{X}_i(z) = \sum_{k=1}^{\infty} x_{i,k} z^k$, $|z| \leq 1$. In order to show the CLT, we assume that the second moment of X_i is finite for any $i \in \mathbb{D}$.

Each arriving customer occupies one empty server, and leaves the system immediately after its service completion. Service requirements of customers are independently and identically distributed with the exponential distribution having mean 1. For each $i \in \mathbb{D}$, when the background process is in state i, the service rate is $\mu_i \in (0, \infty)$. For each $i \in \mathbb{D}$, when the background process is in state i, binomial catastrophes occur according to the Poisson process with rate $\gamma_i \in (0, \infty)$. Upon an occurrence of the binomial catastrophe, each customer in the system is either retained with probability $p_i \in (0, 1)$ or removed with probability $\overline{p_i} := 1 - p_i$, independently of other customers.

3.2.2 Scaling model

We consider the Markov-modulated $\mathrm{M}^{\mathrm{X}}/\mathrm{M}/\infty$ queue with binomial catastrophes under the heavy traffic regime. For the definition of the scaling model, we use the scaling factor N and scaling coefficient the $\alpha > 0$. We consider that N approaches to infinitely and α is

fixed. We define the scaling model as follows: for each $i \in \mathbb{D}$, the arrival rates are scaled as $\lambda_i \mapsto N\lambda_i$, the transition rates of the background process are scaled as $q_{i,j} \mapsto N^\alpha q_{i,j}$, the occurrence rates of the binomial catastrophes are scaled as $\gamma_i \mapsto N\gamma_i$, and the removal probabilities are scaled as $\overline{p_i} \mapsto N^{-1}\overline{p_i}$.

There are two points to notice in our scaling model. First, using the scaling coefficient α, we can express the relative relations between speed of arrivals and that of changing the background state. In particular, if α is strictly larger than one, the background process changes relatively faster than the arrival process. While, if α is strictly smaller than one, the speed of arrivals is relatively faster than that of changes of the background process. Second, the product of the occurrence rate of binomial catastrophes and the removal probability, $\sum_{i\in\mathbb{D}} \tau_i(\gamma_i\overline{p}_i + \mu_i)$, are always constant in the scaling model for any $i \in \mathbb{D}$. In addition, the rate that a customer in the system is removed, $\sum_{i\in\mathbb{D}} \tau_i(\gamma_i\overline{p}_i + \mu_i)$, does not depend on the scaling parameters N and α.

Let $L^{(N)}(t)$ and $J^{(N)}(t)$ denote the queue length in the scaled system and the state of the scaled background process, respectively, at time $t \in \mathbb{R}_+$. It is obvious that the joint stochastic process $\{(J^{(N)}(t), L^{(N)}(t)); t \in \mathbb{R}\}$ is an irreducible Markov chain in continuous time with state space $\mathbb{D} \times \mathbb{Z}_+$. We define $\boldsymbol{U} := (u(i,k;j,\ell))_{(i,k),(j,\ell)\in\mathbb{D}\times\mathbb{Z}_+}$ as the infinitesimal generator of $\{(J^{(N)}(t), L^{(N)}(t))\}$, which is given by

$$u(i,k;j,\ell) = \begin{cases} -\lambda_i - k\mu_i - \gamma_i(1-p_i^k) + q_{i,i}, & i=j,\ k=\ell, \\ \lambda_i x_{i,\ell-k}, & i=j,\ \ell>k, \\ q_{i,j}, & i\neq j,\ k=\ell, \\ k\mu_i + k\gamma_i p_i^{k-1}(1-p_i), & i=j,\ \ell=k-1, \\ \binom{k}{\ell}\gamma_i p_i^\ell (1-p_i)^{k-\ell}, & i=j,\ \ell<k-1, \\ 0, & otherwise. \end{cases} \tag{3.1}$$

Note that the Markov chain $\{(J^{(N)}(t), N^{(N)}(t))\}$ has a unique stationary distribution under the assumption that the second moment of X_i is finite for any $i \in \mathbb{D}$, which is proved in Theorem 3.2. We then define $L^{(N)}$ and $J^{(N)}$ as the queue length of the scaled system and the state of the scaled background process, respectively, at steady state. We also define $\pi_{i,n}^{(N)} = \mathsf{P}(J^{(N)} = i, L^{(N)} = n)$ for $(i,n) \in \mathbb{D} \times \mathbb{Z}_+$.

Remark 3.1 Figures 3.1 and 3.2 show the sample paths of our scaling model with $N = 1000$. Figure 3.1 presents the case with $\alpha = 2$ (> 1), and Figure 3.2 with $\alpha = 0.5$ (< 1). In all cases, the other parameters except α are the same as follows.

$$d = 2, \quad \boldsymbol{Q} = \begin{pmatrix} -0.05, & 0.05 \\ 0.02, & -0.02 \end{pmatrix}, \quad (\lambda_1, \lambda_2) = (0.1, 1), \quad X_1, X_2 \sim U\{1, 10\},$$
$$(\mu_1, \mu_2) = (1, 1), \quad (\gamma_1, \gamma_2) = (0.0001, 0.0002), \quad (p_1, p_2) = (0.2, 0.1).$$

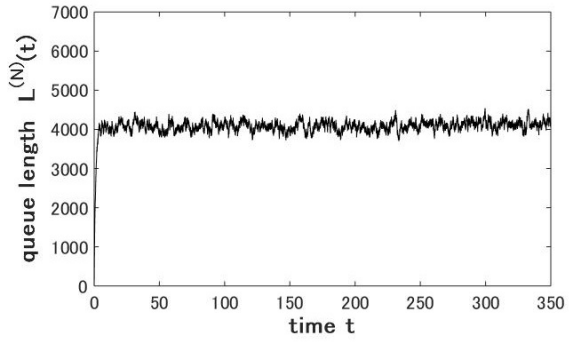

Figure 3.1: Sample path of the scaling model: $\alpha = 2$

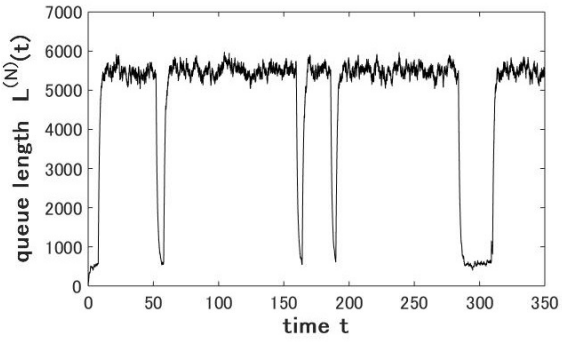

Figure 3.2: Sample path of the scaling model: $\alpha = 0.5$

Figure 3.1 looks like the sample path of a queue length process of an infinite-server queue with constant parameters. The reasons can be considered as follows. When α is strictly larger than one, the transition rates of background process are extremely larger than the arrival rates in the scaling model. It means that the inter-transition time of the background process is much shorter than the inter-arrival time of customers.

On the other hand, Figure 3.2 looks like that sample paths of queue length processes of infinite-server queues with two different parameters appear alternatively. The reasons cam be considered as follows. The queue length converges to a local equilibrium during each transition time because α is strictly smaller than one. In this case, the transition rates of the background process become extremely smaller than the arrival rate of the scaling model. Thus, there may be a large enough number of arrivals and departures before the state of the background process changes.

Note that Blom et al. gave a similar consideration for their model with Remark 3.1 in [9].

3.3 Central limit theorem

To avoid complicated expressions, we use the following diagonal matrices:

$$
\begin{aligned}
\boldsymbol{\Lambda} &= \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d), \\
\widehat{\boldsymbol{X}}(z) &= \mathrm{diag}(\widehat{X}_1(z), \widehat{X}_2(z), \ldots, \widehat{X}_d(z)), \\
\overline{\boldsymbol{X}} &= \mathrm{diag}(\mathsf{E}[X_1], \mathsf{E}[X_2], \ldots, \mathsf{E}[X_d]), \\
\underline{\boldsymbol{X}} &= \mathrm{diag}(\mathsf{E}[X_1^2], \mathsf{E}[X_2^2], \ldots, \mathsf{E}[X_d^2]), \\
\boldsymbol{M} &= \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_d), \\
\boldsymbol{\Gamma} &= \mathrm{diag}(\gamma_1, \gamma_2, \ldots, \gamma_d), \\
\boldsymbol{P} &= \mathrm{diag}(p_1, p_2, \ldots, p_d).
\end{aligned}
$$

In addition, we use the ergodic matrix $\boldsymbol{T} := \boldsymbol{e}\cdot\boldsymbol{\tau}$, the fundamental matrix $\boldsymbol{F} := (\boldsymbol{T}-\boldsymbol{Q})^{-1}$, and the deviation matrix $\boldsymbol{D} := \boldsymbol{F} - \boldsymbol{T}$ (see e.g. [18]). Note that it follows that

$$\boldsymbol{Q}\boldsymbol{F} = -\boldsymbol{I} + \boldsymbol{T}. \tag{3.2}$$

Using these notations, we present the CLT for the stationary queue length of the Markov-modulated $\mathrm{M}^{\mathrm{X}}/\mathrm{M}/\infty$ queue with binomial catastrophes. We show the proof of Theorem 3.1 in Sections 3.3.1–3.3.3.

Theorem 3.1 Central limit theorem (CLT) *Let $\beta = \min(\alpha, 1)$. The random variable*

$$N^{\beta/2}\left(\frac{L^{(N)}}{N} - \rho\right) \tag{3.3}$$

converges in distribution to the normal distribution with mean zero and variance σ^2, where ρ and σ^2 are given by

$$\begin{aligned}
\rho &= \frac{\boldsymbol{\tau}\boldsymbol{\Lambda}\overline{\boldsymbol{X}}\boldsymbol{e}}{\boldsymbol{\tau}\big[\boldsymbol{M} + \boldsymbol{\Gamma}(\boldsymbol{I}-\boldsymbol{P})\big]\boldsymbol{e}}, \\
\sigma^2 &= \sigma_1^2 I(\alpha \le 1) + \sigma_2^2 I(\alpha \ge 1), \\
\sigma_1^2 &= \frac{\boldsymbol{\tau}\big[\boldsymbol{\Lambda}\overline{\boldsymbol{X}} - \rho(\boldsymbol{M} + \boldsymbol{\Gamma}(\boldsymbol{I}-\boldsymbol{P}))\big]\boldsymbol{D}\big[\boldsymbol{\Lambda}\overline{\boldsymbol{X}} - \rho(\boldsymbol{M} + \boldsymbol{\Gamma}(\boldsymbol{I}-\boldsymbol{P}))\big]\boldsymbol{e}}{\boldsymbol{\tau}\big[\boldsymbol{M} + \boldsymbol{\Gamma}(\boldsymbol{I}-\boldsymbol{P})\big]\boldsymbol{e}}, \\
\sigma_2^2 &= \frac{\boldsymbol{\tau}\big[\boldsymbol{\Lambda}\underline{\boldsymbol{X}} + \rho(\boldsymbol{M} + \boldsymbol{\Gamma}(\boldsymbol{I}-\boldsymbol{P})) + \rho^2\boldsymbol{\Gamma}(\boldsymbol{I}-\boldsymbol{P})^2\big]\boldsymbol{e}}{2\boldsymbol{\tau}\big[\boldsymbol{M} + \boldsymbol{\Gamma}(\boldsymbol{I}-\boldsymbol{P})\big]\boldsymbol{e}}.
\end{aligned} \tag{3.4}$$

It should be noted that $L^{(N)}N^{-1}$ converges in probability to ρ, which is proven in Section 3.3.2. The constant ρ has a very simple form. The numerator of ρ is equal to the mean number of customers who arrive at the unscaled system per a unit time, and the denominator of ρ is equal to the mean departure (due to either service completion or binomial catastrophe) rate.

Furthermore, we can observe from Theorem 3.1 that the variance σ^2 strongly depends on scaling coefficient α. If α is strictly smaller than one, the variance is σ_1^2, which is characterized by the deviation matrix $\boldsymbol{D}$. If α is strictly larger than one, the variance is σ_2^2. If α is equal to one, the variance is the sum of σ_1^2 and σ_2^2.

Remark 3.2 Theorem 3.1 provides an approximation for the queue length distribution of our model with large arrival rates. Indeed, we have

$$\begin{aligned}
\mathsf{P}(L^{(N)} \le x) &= \mathsf{P}\left(N^{\beta/2}\left(\frac{L^{(N)}}{N} - \rho\right) \le N^{\beta/2}\left(\frac{x}{N} - \rho\right)\right) \\
&\simeq \Phi_{(0,\sigma^2)}\left(N^{\beta/2}\left(\frac{x}{N} - \rho\right)\right), \qquad \text{for } x \in \mathbb{R}_+,
\end{aligned} \tag{3.5}$$

where $\Phi_{(0,\sigma^2)}(\cdot)$ denotes the cumulative distribution function of the normal distribution with mean zero and variance σ^2.

3.3.1 Queue length distribution of the scaling model

We define the row vectors $\widehat{\boldsymbol{\pi}}^{(N)}(z) = (\widehat{\pi}_i^{(N)}(z))_{i\in\mathbb{D}}$ and $\widehat{\boldsymbol{\pi}}_p^{(N)}(z) = (\widehat{\pi}_{p,i}^{(N)}(z))_{i\in\mathbb{D}}$ as follows, respectively, for $|z| \le 1$ and $i \in \mathbb{D}$.

$$\widehat{\pi}_i^{(N)}(z) = \mathsf{E}[z^{L^{(N)}} I(J^{(N)} = i)], \tag{3.6}$$

$$\widehat{\pi}_{p,i}^{(N)}(z) = \widehat{\pi}_i^{(N)}(z + (1-z)N^{-1}\overline{p_i}). \tag{3.7}$$

For the proof of the CLT, we first derive the differential equations for the probability generating function (PGF) of the stationary queue length distribution.

Lemma 3.1 *The following differential equations hold for any* $|z| \le 1$.

$$\begin{aligned}(1-z)\frac{\mathrm{d}}{\mathrm{d}z}\widehat{\boldsymbol{\pi}}^{(N)}(z)\boldsymbol{M} = N\widehat{\boldsymbol{\pi}}^{(N)}(z)\boldsymbol{\Lambda}\big[\boldsymbol{I} - \widehat{\boldsymbol{X}}(z)\big] - N^{\alpha}\widehat{\boldsymbol{\pi}}^{(N)}(z)\boldsymbol{Q} \\ - N\{\widehat{\boldsymbol{\pi}}_p^{(N)}(z) - \widehat{\boldsymbol{\pi}}^{(N)}(z)\}\boldsymbol{\Gamma}.\end{aligned} \tag{3.8}$$

Proof. We first show (3.8) with $N = 1$. It follows from (3.1) that, for $i \in \mathbb{D}$ and $n \in \mathbb{Z}_+$,

$$\begin{aligned}(\lambda_i + \gamma_i + n\mu_i)\pi_{i,n}^{(1)} = \lambda_i \sum_{k=1}^{n} x_{i,k}\pi_{i,n-k}^{(1)} + (n+1)\mu_i\pi_{i,n+1}^{(1)} \\ + \gamma_i \sum_{k=0}^{\infty}\binom{n+k}{k} p_i^n \overline{p}_i^k \pi_{i,n+k}^{(1)} + \sum_{j\in\mathbb{D}} q_{j,i}\pi_{j,n}^{(1)},\end{aligned} \tag{3.9}$$

where the empty sum (i.e., summation from 1 to 0) is defined as 0. Multiplying (3.9) by z^n and taking the sum over $n \in \mathbb{Z}_+$, we obtain, for $i \in \mathbb{D}$,

$$\begin{aligned}(\lambda_i + \gamma_i)\widehat{\pi}_i^{(1)}(z) + z\mu_i\frac{\mathrm{d}}{\mathrm{d}z}\widehat{\pi}_i^{(1)}(z) = \lambda_i \sum_{n=1}^{\infty}\sum_{k=1}^{n} x_{i,k}\pi_{i,n-k}^{(1)} z^n + \mu_i\frac{\mathrm{d}}{\mathrm{d}z}\widehat{\pi}_i^{(1)}(z) \\ + \gamma_i \sum_{n=0}^{\infty}\sum_{k=0}^{\infty}\binom{n+k}{k} p_i^n \overline{p}_i^k \pi_{i,n+k}^{(1)} z^n + \sum_{j\in\mathbb{D}} q_{j,i}\widehat{\pi}_j^{(1)}(z).\end{aligned} \tag{3.10}$$

It is easy to see that

$$\sum_{n=1}^{\infty}\sum_{k=1}^{n} x_{i,k}\pi_{i,n-k}^{(1)} z^n = \widehat{X}_i(z)\widehat{\pi}_i^{(1)}(z),$$

and

$$\sum_{n=0}^{\infty}\sum_{k=0}^{\infty}\binom{n+k}{k} p_i^n \overline{p}_i^k \pi_{i,n+k}^{(1)} z^n = \widehat{\pi}_{p,i}^{(1)}(z).$$

Substituting these equations into (3.10) yields

$$\begin{aligned}(\lambda_i + \gamma_i)\widehat{\pi}_i^{(1)}(z) + z\mu_i\frac{\mathrm{d}}{\mathrm{d}z}\widehat{\pi}_i^{(1)}(z) = \lambda_i\widehat{X}_i(z)\widehat{\pi}_i^{(1)}(z) + \mu_i\frac{\mathrm{d}}{\mathrm{d}z}\widehat{\pi}_i^{(1)}(z) \\ + \gamma_i\widehat{\pi}_{p,i}^{(1)}(z) + \sum_{j\in\mathbb{D}} q_{j,i}\widehat{\pi}_j^{(1)}(z).\end{aligned}$$

Rearranging the above, we obtain

$$(1-z)\frac{\mathrm{d}}{\mathrm{d}z}\widehat{\boldsymbol{\pi}}^{(1)}(z)\boldsymbol{M} = \widehat{\boldsymbol{\pi}}^{(1)}(z)\boldsymbol{\Lambda}(\boldsymbol{I}-\widehat{\boldsymbol{X}}(z)) - (\widehat{\boldsymbol{\pi}}_p^{(1)}(z)-\widehat{\boldsymbol{\pi}}^{(1)}(z))\boldsymbol{\Gamma} - \widehat{\boldsymbol{\pi}}^{(1)}(z)\boldsymbol{Q}.$$

In the above differential equation, Replacing $\boldsymbol{\Lambda}$ with $N\boldsymbol{\Lambda}$, $\boldsymbol{Q}$ with $N^{\alpha}\boldsymbol{Q}$, $\boldsymbol{\Gamma}$ with $N\boldsymbol{\Gamma}$, and $(\boldsymbol{I}-\boldsymbol{P})$ with $N^{-1}(\boldsymbol{I}-\boldsymbol{P})$, we obtain (3.8) for any $|z| \le 1$. □

Using Lemma 3.1 and $\mathsf{E}[X_i^2] < \infty$, we obtain Lemma 3.2.

Lemma 3.2 *For any $i \in \mathbb{D}$, there exists some $C > 0$ such that*

$$\mathsf{E}[(L^{(N)}N^{-1})^2 I(J^{(N)}=i)] \le C, \qquad \textit{for all } N \ge 1.$$

The proof of Lemma 3.2 is given in Appendix B.

3.3.2 Law of large numbers

Next, we show the law of large numbers.

Lemma 3.3 *$L^{(N)}N^{-1}$ converges in probability to ρ as $N \to \infty$, where ρ is given by (3.4).*

Proof. Using Lemma 3.1, we first construct the differential equation for the moment generating function (MGF) of $L^{(N)}N^{-1}$. We define, for $\theta \in (-\infty, 0]$,

$$z(\theta) := \mathrm{e}^{\theta N^{-1}}.$$

It should be noted that $z(\theta)$ implicitly depends on N. We also define

$$\widetilde{\boldsymbol{\pi}}^{(N)}(\theta) := \widehat{\boldsymbol{\pi}}^{(N)}(z(\theta)),$$
$$\widetilde{\boldsymbol{\pi}}_p^{(N)}(\theta) := \widehat{\boldsymbol{\pi}}_p^{(N)}(z(\theta)).$$

Note that the sum of all elements of $\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)$ is equivalent to the MGF of $L^{(N)}N^{-1}$. Substituting $z = z(\theta)$ into (3.8), we have

$$\begin{aligned}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{Q} = {} & N^{-\alpha}(z(\theta)-1)\frac{\mathrm{d}}{\mathrm{d}z}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{M} \\ & - N^{1-\alpha}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{\Lambda}\big[\widehat{\boldsymbol{X}}(z(\theta)) - \boldsymbol{I}\big] \\ & - N^{1-\alpha}\{\widetilde{\boldsymbol{\pi}}_p^{(N)}(\theta) - \widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\}\boldsymbol{\Gamma}. \end{aligned} \tag{3.11}$$

Note that

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}z(\theta)\frac{\mathrm{d}}{\mathrm{d}z}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta) = N^{-1}z(\theta)\frac{\mathrm{d}}{\mathrm{d}z}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta).$$

Applying this equation to (3.11), we obtain the following differential equation.

$$\begin{aligned}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{Q} = N^{1-\alpha}(1-z(\theta)^{-1})\frac{\mathrm{d}}{\mathrm{d}\theta}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{M} \\ - N^{1-\alpha}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{\Lambda}\big[\widehat{\boldsymbol{X}}(z(\theta)) - \boldsymbol{I}\big] \\ - N^{1-\alpha}\{\widetilde{\boldsymbol{\pi}}_p^{(N)}(\theta) - \widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\}\boldsymbol{\Gamma}. \end{aligned} \tag{3.12}$$

To prove the convergence in probability, we show that the MGF of $L^{(N)}N^{-1}$ converges pointwise to that of ρ for any $\theta \in (-\infty, 0]$. To this end, we evaluate each term in the right hand side of (3.12). Using the first order Maclaurin expansion of $z(\theta)^{-1}$, we have

$$1 - z(\theta)^{-1} = N^{-1}\theta + O(N^{-2}). \tag{3.13}$$

By the dominated convergence theorem, it follows from $\mathsf{E}[X_i] < \infty$ and (3.13) that

$$\widehat{\boldsymbol{X}}(z(\theta)) - \boldsymbol{I} = N^{-1}\theta\overline{\boldsymbol{X}} + o(N^{-1}). \tag{3.14}$$

In addition, the third term in the right hand side of (3.12) can be expressed as follows.

$$\{\widetilde{\boldsymbol{\pi}}_p^{(N)}(\theta) - \widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\}\boldsymbol{\Gamma} = -N^{-1}\theta\frac{\mathrm{d}}{\mathrm{d}\theta}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{B} + o(N^{-1}), \tag{3.15}$$

where $\boldsymbol{B} := \boldsymbol{\Gamma}(\boldsymbol{I} - \boldsymbol{P})$. The derivation of (3.15) is presented in Appendix C. Applying (3.13)–(3.15) to the right hand side of (3.12), we obtain

$$\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{Q} = N^{-\alpha}\theta\frac{\mathrm{d}}{\mathrm{d}\theta}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)(\boldsymbol{M} + \boldsymbol{B}) - N^{-\alpha}\theta\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{\Lambda}\overline{\boldsymbol{X}} + o(N^{-\alpha}). \tag{3.16}$$

Right-multiplying (3.16) by $\theta^{-1}N^{\alpha}\boldsymbol{e}$ yields

$$0 = \left[\frac{\mathrm{d}}{\mathrm{d}\theta}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)(\boldsymbol{M} + \boldsymbol{B}) - \widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{\Lambda}\overline{\boldsymbol{X}}\right]\boldsymbol{e} + o(1). \tag{3.17}$$

Furthermore, right-multiplying (3.16) by $\boldsymbol{F}$ and using (3.2) yields

$$\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)(\boldsymbol{T} - \boldsymbol{I}) = N^{-\alpha}\theta\left[\frac{\mathrm{d}}{\mathrm{d}\theta}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)(\boldsymbol{M} + \boldsymbol{B}) - \widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{\Lambda}\overline{\boldsymbol{X}}\right]\boldsymbol{F} + o(N^{-\alpha}). \tag{3.18}$$

Taking the limits as $N \to \infty$ in (3.18) and the derivative of (3.18) yields

$$\begin{aligned}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{T} - \widetilde{\boldsymbol{\pi}}^{(N)}(\theta) = o(1), \\ \frac{\mathrm{d}}{\mathrm{d}\theta}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{T} - \frac{\mathrm{d}}{\mathrm{d}\theta}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta) = o(1).\end{aligned}$$

Applying these results to (3.17), we obtain

$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{T}(\boldsymbol{M} + \boldsymbol{B})\boldsymbol{e} - \widetilde{\boldsymbol{\pi}}^{(N)}(\theta)\boldsymbol{T}\boldsymbol{\Lambda}\overline{\boldsymbol{X}}\boldsymbol{e} + o(1),$$

that is,

$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}\left[\mathrm{e}^{L^{(N)}N^{-1}\theta}\right] \cdot \boldsymbol{\tau}(\boldsymbol{M}+\boldsymbol{B})\boldsymbol{e} - \mathsf{E}\left[\mathrm{e}^{L^{(N)}N^{-1}\theta}\right] \cdot \boldsymbol{\tau}\boldsymbol{\Lambda}\overline{\boldsymbol{X}}\boldsymbol{e} + o(1), \tag{3.19}$$

By definition of ρ, (3.19) is rewritten as follows.

$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}\left[\mathrm{e}^{L^{(N)}N^{-1}\theta}\right] - \rho \cdot \mathsf{E}\left[\mathrm{e}^{L^{(N)}N^{-1}\theta}\right] + o(1).$$

From the above equation, the MGF of $L^{(N)}N^{-1}$ converges pointwise to $\exp(\rho\theta)$ for any $\theta \in (-\infty, 0]$. Note that $\exp(\rho\theta)$ is the MGF of ρ. Therefore, it follows from Lévy's continuity theorem (see e.g. [67, Chapter 18.1]) that $L^{(N)}N^{-1}$ converges in distribution to ρ as $N \to \infty$, and thus in probability. □

Using Lemmas 3.2 and 3.3, we obtain Lemma 3.4.

Lemma 3.4 *For any $i \in \mathbb{D}$ and $\theta \in (-\infty, \infty)$, it follows that*

$$0 = \lim_{N\to\infty} \mathsf{E}\left[\left(L^{(N)}N^{-1} - \rho\right) \cdot \mathrm{e}^{\mathrm{i}L^{(N)}N^{-1}\theta} I(J^{(N)} = i)\right],$$
$$0 = \lim_{N\to\infty} \mathsf{E}\left[\left((L^{(N)}N^{-1})^2 - \rho^2\right) \cdot \mathrm{e}^{\mathrm{i}L^{(N)}N^{-1}\theta} I(J^{(N)} = i)\right].$$

The proof for Lemma 3.4 is given in Appendix D. We use Lemma 3.4 to show Theorem 3.1.

3.3.3 Proof for the central limit theorem

Finally, we show the CLT for the stationary queue length (i.e.,Theorem 3.1) using Lemmas 3.1–3.4.

Proof of Theorem 3.1. Using Lemma 3.1, we first construct the differential equation for the characteristic function (CF) of (3.3). We define

$$z(\theta) := \exp(\mathrm{i}N^{-1+\beta/2}\theta), \qquad \theta \in \mathbb{R}.$$

It should be noted that $z(\theta)$ implicitly depends on N. We also define

$$\boldsymbol{\pi}_*^{(N)}(\theta) := \exp(-\mathrm{i}N^{\beta/2}\rho\theta) \cdot \widehat{\boldsymbol{\pi}}^{(N)}(z(\theta)),$$
$$\boldsymbol{\pi}_{p,*}^{(N)}(\theta) := \exp(-\mathrm{i}N^{\beta/2}\rho\theta) \cdot \widehat{\boldsymbol{\pi}}_p^{(N)}(z(\theta)).$$

Note that the sum of all elements of $\boldsymbol{\pi}_*^{(N)}(\theta)$ coincides with the CF of (3.3). Substituting $z = z(\theta)$ into (3.8) and multiplying the result by $\exp(-\mathrm{i}N^{\beta/2}\rho\theta)$ yields

$$N^{\alpha}\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{Q} = \mathrm{e}^{-\mathrm{i}N^{\beta/2}\rho\theta}(z(\theta)-1)\frac{\mathrm{d}}{\mathrm{d}z}\widehat{\boldsymbol{\pi}}^{(N)}(z(\theta))\boldsymbol{M}$$
$$- N\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{\Lambda}\left[\widehat{\boldsymbol{X}}(z(\theta)) - \boldsymbol{I}\right] - N\{\boldsymbol{\pi}_{p,*}^{(N)}(\theta) - \boldsymbol{\pi}_*^{(N)}(\theta)\}\boldsymbol{\Gamma}. \tag{3.20}$$

Note here that

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\boldsymbol{\pi}_*^{(N)}(\theta) = -\mathrm{i}N^{\beta/2}\rho\boldsymbol{\pi}_*^{(N)}(\theta) + \mathrm{i}N^{-1+\beta/2}\mathrm{e}^{-\mathrm{i}\rho N^{\beta/2}\theta}z(\theta)\frac{\mathrm{d}}{\mathrm{d}z}\widehat{\boldsymbol{\pi}}^{(N)}(z(\theta)).$$

Applying this equation to (3.20), we obtain the following differential equation.

$$\begin{aligned} N^{\alpha}\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{Q} = &-\mathrm{i}N^{1-\beta/2}(1-z(\theta)^{-1})\frac{\mathrm{d}}{\mathrm{d}\theta}\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{M} + N\rho(1-z(\theta)^{-1})\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{M} \\ &- N\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{\Lambda}\big[\widehat{\boldsymbol{X}}(z(\theta)) - \boldsymbol{I}\big] - N\{\boldsymbol{\pi}_{p,*}^{(N)}(\theta) - \boldsymbol{\pi}_*^{(N)}(\theta)\}\boldsymbol{\Gamma}. \end{aligned} \tag{3.21}$$

To show the convergence in distribution, we evaluate each term in the right hand side of (3.21). Using the second order Maclaurin expansion of $z(\theta)^{-1}$ yields

$$1 - z(\theta)^{-1} = \mathrm{i}N^{-1+\beta/2}\theta + N^{-2+\beta}\frac{1}{2}\theta^2 + O(N^{-3+3\beta/2}).$$

which leads to

$$-\mathrm{i}N^{1-\beta/2}(1-z(\theta)^{-1}) = \theta + o(1), \tag{3.22}$$

$$N(1-z(\theta)^{-1}) = \mathrm{i}N^{\beta/2}\theta + N^{-1+\beta}\frac{1}{2}\theta^2 + o(1). \tag{3.23}$$

By the dominated convergence theorem, it follows from $\mathsf{E}[X_i^2] < \infty$ and (3.23) that

$$N\big[\widehat{\boldsymbol{X}}(z(\theta)) - \boldsymbol{I}\big] = \mathrm{i}N^{\beta/2}\theta\overline{\boldsymbol{X}} - N^{-1+\beta}\frac{1}{2}\theta^2\underline{\boldsymbol{X}} + o(1). \tag{3.24}$$

The forth term in the right hand side of (3.21) can be expressed as follows.

$$\begin{aligned} N\{\boldsymbol{\pi}_*^{(N)}(\theta) - \boldsymbol{\pi}_{p,*}^{(N)}(\theta)\}\boldsymbol{\Gamma} = &\theta\frac{\mathrm{d}}{\mathrm{d}\theta}\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{B} + \mathrm{i}N^{\beta/2}\rho\theta\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{B} \\ &+ N^{-1+\beta}\frac{1}{2}\theta^2\boldsymbol{\pi}_*^{(N)}(\theta)\big[\rho\boldsymbol{B} + \rho^2\big[\boldsymbol{I}-\boldsymbol{P}\big]\boldsymbol{B}\big] + o(1), \end{aligned} \tag{3.25}$$

where $\boldsymbol{B} := \boldsymbol{\Gamma}(\boldsymbol{I}-\boldsymbol{P})$. The derivation of (3.25) is shown in Appendix E. Substituting (3.22)– (3.25) into the right hand side of (3.21), we obtain

$$\begin{aligned} N^{\alpha}\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{Q} = &\theta\frac{\mathrm{d}}{\mathrm{d}\theta}\boldsymbol{\pi}_*^{(N)}(\theta)\big[\boldsymbol{M}+\boldsymbol{B}\big] - \mathrm{i}N^{\beta/2}\theta\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{Y}_1 \\ &+ N^{-1+\beta}\frac{1}{2}\theta^2\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{Y}_2 + o(1), \end{aligned} \tag{3.26}$$

where $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are given by

$$\begin{aligned} \boldsymbol{Y}_1 &= \boldsymbol{\Lambda}\overline{\boldsymbol{X}} - \rho\boldsymbol{M} - \rho\boldsymbol{B}, \\ \boldsymbol{Y}_2 &= \boldsymbol{\Lambda}\underline{\boldsymbol{X}} + \rho\boldsymbol{M} + \rho\boldsymbol{B} + \rho^2\big[\boldsymbol{I}-\boldsymbol{P}\big]\boldsymbol{B}. \end{aligned}$$

Right-multiplying (3.26) by $\boldsymbol{e}$ yields

$$\begin{aligned} 0 = &\frac{\mathrm{d}}{\mathrm{d}\theta}\boldsymbol{\pi}_*^{(N)}(\theta)\big[\boldsymbol{M}+\boldsymbol{B}\big]\boldsymbol{e} - \mathrm{i}N^{\beta/2}\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{Y}_1\boldsymbol{e} \\ &+ N^{-1+\beta}\frac{1}{2}\theta\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{Y}_2\boldsymbol{e} + o(1). \end{aligned} \tag{3.27}$$

In addition, right-multiplying (3.26) by $N^{-\alpha}\boldsymbol{F}$ and using (3.2) yields

$$\begin{aligned}\boldsymbol{\pi}_*^{(N)}(\theta)(\boldsymbol{T}-\boldsymbol{I}) = N^{-\alpha}\theta\frac{\mathrm{d}}{\mathrm{d}\theta}\boldsymbol{\pi}_*^{(N)}(\theta)\big[\boldsymbol{M}+\boldsymbol{B}\big]\boldsymbol{F}\\ - \mathrm{i}N^{-\alpha+\beta/2}\theta\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{Y}_1\boldsymbol{F}\\ + N^{-1-\alpha+\beta}\frac{1}{2}\theta^2\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{Y}_2\boldsymbol{F}+o(1).\end{aligned} \tag{3.28}$$

Note that $\boldsymbol{F}=\boldsymbol{D}+\boldsymbol{T}$. Taking the limits as $N\to\infty$ in (3.28) and the derivative of (3.28) yields

$$\begin{aligned}\boldsymbol{\pi}_*^{(N)}(\theta) &= \boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{T}+o(1),\\ N^{\beta/2}\boldsymbol{\pi}_*^{(N)}(\theta) &= N^{\beta/2}\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{T}+\mathrm{i}N^{-\alpha+\beta}\theta\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{Y}_1\big[\boldsymbol{D}+\boldsymbol{T}\big]+o(1),\\ \frac{\mathrm{d}}{\mathrm{d}\theta}\boldsymbol{\pi}_*^{(N)}(\theta) &= \frac{\mathrm{d}}{\mathrm{d}\theta}\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{T}+o(1).\end{aligned}$$

Substituting these equations into (3.27), we obtain

$$\begin{aligned}0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\pi_*^{(N)}(\theta)\boldsymbol{\tau}\big[\boldsymbol{M}+\boldsymbol{B}\big]\boldsymbol{e} - \mathrm{i}N^{\beta/2}\pi_*^{(N)}(\theta)\boldsymbol{\tau}\boldsymbol{Y}_1\boldsymbol{e}\\ + N^{-\alpha+\beta}\theta\pi_*^{(N)}(\theta)\boldsymbol{\tau}\boldsymbol{Y}_1\boldsymbol{D}\boldsymbol{Y}_1\boldsymbol{e}\\ + N^{-\alpha+\beta}\theta\pi_*^{(N)}(\theta)\boldsymbol{\tau}\boldsymbol{Y}_1\boldsymbol{e}\cdot\boldsymbol{\tau}\boldsymbol{Y}_1\boldsymbol{e} + N^{-1+\beta}\frac{1}{2}\theta\pi_*^{(N)}(\theta)\boldsymbol{\tau}\boldsymbol{Y}_2\boldsymbol{e}+o(1),\end{aligned} \tag{3.29}$$

where $\pi_*^{(N)}(\theta)$ denotes the CF of (3.3); that is, $\pi_*^{(N)}(\theta)=\boldsymbol{\pi}_*^{(N)}(\theta)\boldsymbol{e}$. It is easy to see that

$$\begin{aligned}\boldsymbol{\tau}\boldsymbol{Y}_1\boldsymbol{e} &= \boldsymbol{\tau}\boldsymbol{\Lambda}\overline{\boldsymbol{X}}\boldsymbol{e}-\rho\cdot\boldsymbol{\tau}(\boldsymbol{M}+\boldsymbol{\Gamma}(\boldsymbol{i}-\boldsymbol{P}))\boldsymbol{e}\\ &= \boldsymbol{\tau}\boldsymbol{\Lambda}\overline{\boldsymbol{X}}\boldsymbol{e}-\frac{\boldsymbol{\tau}\boldsymbol{\Lambda}\overline{\boldsymbol{X}}\boldsymbol{e}}{\boldsymbol{\tau}(\boldsymbol{M}+\boldsymbol{\Gamma}(\boldsymbol{i}-\boldsymbol{P}))\boldsymbol{e}}\cdot\boldsymbol{\tau}(\boldsymbol{M}+\boldsymbol{\Gamma}(\boldsymbol{I}-\boldsymbol{P}))\boldsymbol{e}\\ &= 0.\end{aligned}$$

Applying this equation to (3.29), we obtain

$$\begin{aligned}0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\pi_*^{(N)}(\theta)\boldsymbol{\tau}\big[\boldsymbol{M}+\boldsymbol{B}\big]\boldsymbol{e} + N^{-\alpha+\beta}\theta\pi_*^{(N)}(\theta)\boldsymbol{\tau}\boldsymbol{Y}_1\boldsymbol{D}\boldsymbol{Y}_1\boldsymbol{e}\\ + N^{-1+\beta}\frac{1}{2}\theta\pi_*^{(N)}(\theta)\boldsymbol{\tau}\boldsymbol{Y}_2\boldsymbol{e}+o(1).\end{aligned} \tag{3.30}$$

By definition of $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, (3.30) can be rewritten as follows.

$$0 = \frac{\mathrm{d}}{\mathrm{d}\theta}\pi_*^{(N)}(\theta)+N^{-\alpha+\beta}\sigma_1^2\theta\pi_*^{(N)}(\theta)+N^{-1+\beta}\sigma_2^2\theta\pi_*^{(N)}(\theta)+o(1),$$

which leads to

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\pi_*^{(N)}(\theta) = -\sigma^2\theta\pi_*^{(N)}(\theta)+o(1).$$

From the above equation, the CF of (3.3) converges pointwise to $\exp(-\sigma^2\theta^2/2)$ for any $\theta\in\mathbb{R}$. Note that $\exp(-\sigma^2\theta^2/2)$ is the CF of the normal distribution with mean zero and variance σ^2. Therefore, it follows from Lévy's continuity theorem (see e.g. [67, Chapter 18.1]) that (3.3) converges in distribution to the normal distribution with mean zero and variance σ^2 as $N\to\infty$. □

3.4 Stability condition

In this section, we present the stability condition of the queueing model described in Section 3.2. In Theorem 3.1, we established the CLT under the assumption that the second moment of the batch-size distribution is finite. The following theorem implies that the Markov chain $\{(J^{(N)}(t), L^{(N)}(t))\}$ has a unique stationary distribution even if the first moment of the batch-size distribution is not finite.

Theorem 3.2 (Stability condition) *The irreducible Markov chain* $\{(J^{(N)}(t), L^{(N)}(t))\}$ *has a unique stationary distribution if and only if*

$$\mathsf{E}[\log X_i] < \infty, \qquad \textit{for any } i \in \mathbb{D}. \tag{3.31}$$

Proof. Supposing that (3.31) holds. We show the sufficiency of Theorem 3.2. In order to use Lemma 2.1, we define $\varphi(i,k)$ and $\psi(i,k)$ as follows.

$$\begin{aligned} \varphi(i,k) &= \log(k+1), && (i,k) \in \mathbb{D} \times \mathbb{Z}_+, \\ \psi(i,k) &= \sum_{(j,\ell)\in\mathbb{D}\times\mathbb{Z}_+} \varphi(i,k) \cdot u(i,k;j,\ell), && (i,k) \in \mathbb{D} \times \mathbb{Z}_+. \end{aligned}$$

From (3.1), we then have, for $i \in \mathbb{D}$ and $k = 0$,

$$\psi(i,0) = \sum_{\ell=1}^{\infty} \lambda_i x_\ell \log(\ell+1) < \infty, \tag{3.32}$$

where the inequality holds due to (3.31). We also have, for $i \in \mathbb{D}$ and $k \geq 1$,

$$\begin{aligned} \psi(i,k) &= \sum_{\ell=1}^{\infty} \lambda_i x_\ell \log\left(\frac{k+\ell+1}{k+1}\right) + k\mu_i \log\left(\frac{k}{k+1}\right) \\ &\quad + \sum_{\ell=0}^{k} \gamma_i \binom{k}{\ell} (1-p_i)^\ell p_i^{k-\ell} \log\left(\frac{\ell+1}{k+1}\right) \\ &\leq \sum_{\ell=1}^{\infty} \lambda_i x_{i,\ell} \log\left(1+\frac{\ell}{k+1}\right) + k\mu_i \log\left(1-\frac{1}{k+1}\right). \end{aligned} \tag{3.33}$$

Note here that, for $k, \ell \in \mathbb{N}$,

$$\log\left(1+\frac{\ell}{k+1}\right) = \log\ell + \log\left(\frac{1}{\ell}+\frac{1}{k+1}\right) \leq \log\ell + \log 2. \tag{3.34}$$

In addition, it follows from (2.9) that there exists some $\delta > 0$ such that

$$k\mu \log\left(1-\frac{1}{k+1}\right) \leq -2\delta, \qquad \text{for all } k \geq 1. \tag{3.35}$$

Applying (3.34) and (3.35) to (3.33), we obtain

$$\psi(i,k) \le \lambda_i \mathsf{E}[\log X_i] + \lambda_i \log 2 - 2\delta < \infty. \tag{3.36}$$

In addition, by dominated convergence theorem, we have

$$\lim_{k\to\infty} \sum_{\ell=1}^{\infty} \log\Big(1 + \frac{\ell}{k+1}\Big) x_{i,\ell} = 0,$$

and thus there exists some $K := K_\delta \in \mathbb{N}$ such that, for all $k = K+1, K+2, \ldots,$

$$\lambda_i \sum_{\ell=1}^{\infty} \log\Big(1 + \frac{\ell}{k+1}\Big) x_{i,\ell} \le \delta.$$

Applying (3.35) and the above inequality to (3.33), we obtain

$$\psi(i,k) \le -\delta, \quad k = K+1, K+2, \ldots. \tag{3.37}$$

Using Lemma 2.1, it follows from (3.32), (3.36) and (3.37) that the irreducible Markov chain $\{(J^{(N)}(t), L^{(N)}(t))\}$ has a unique stationary distribution.

On the other hand, supposing that $\{(J^{(N)}(t), L^{(N)}(t))\}$ has a unique stationary distribution $\boldsymbol{\pi} = (\pi^{(N)}_{i,n})_{i\in\mathbb{D}, n\in\mathbb{Z}_+}$. Let $\widehat{\boldsymbol{\pi}}^{(N)}(z)$ and $\widehat{\boldsymbol{\pi}}^{(N)}_p(z)$ denote the row vectors defined by (3.6) and (3.7). From Lemma 3.1, $\widehat{\boldsymbol{\pi}}^{(N)}(z)$ and $\widehat{\boldsymbol{\pi}}^{(N)}_p(z)$ satisfy (3.8). Right-multiplying (3.8) by $(1-z)^{-1}\boldsymbol{e}$ yields

$$N\sum_{i\in\mathbb{D}} \lambda_i \frac{1-\widehat{X}_i(z)}{1-z} \widehat{\pi}^{(N)}_i(z) = \sum_{i\in\mathbb{D}} \mu_i \frac{\mathrm{d}}{\mathrm{d}z} \widehat{\pi}^{(N)}_i(z) + N\sum_{i\in\mathbb{D}} \gamma_i \frac{\widehat{\pi}^{(N)}_{p,i}(z) - \widehat{\pi}^{(N)}_i(z)}{1-z}.$$

Applying $\widehat{\pi}^{(N)}_i(1) \ge \widehat{\pi}^{(N)}_{p,i}(z)$, $0 \le z \le 1$, into the above equation, we have

$$N\sum_{i\in\mathbb{D}} \lambda_i \frac{1-\widehat{X}_i(z)}{1-z} \widehat{\pi}^{(N)}_i(z) \le \sum_{i\in\mathbb{D}} \mu_i \frac{\mathrm{d}}{\mathrm{d}z} \widehat{\pi}^{(N)}_i(z) + N\sum_{i\in\mathbb{D}} \gamma_i \frac{\widehat{\pi}^{(N)}_i(1) - \widehat{\pi}^{(N)}_i(z)}{1-z}. \tag{3.38}$$

Using Lagrange's mean value theorem, there exist $\theta \in (0,1)$ such that

$$\widehat{\pi}^{(N)}_i(1) = \widehat{\pi}^{(N)}(z) + (1-z)\frac{\mathrm{d}}{\mathrm{d}z}\widehat{\pi}^{(N)}(z + \theta(1-z)), \quad \text{for any } z \in [0,1].$$

Substituting this equation into (3.38) yields

$$N\sum_{i\in\mathbb{D}} \lambda_i \frac{1-\widehat{X}_i(z)}{1-z} \widehat{\pi}^{(N)}_i(z) \le \sum_{i\in\mathbb{D}} \mu_i \frac{\mathrm{d}}{\mathrm{d}z} \widehat{\pi}^{(N)}_i(z) N\sum_{i\in\mathbb{D}} \gamma_i \frac{\mathrm{d}}{\mathrm{d}z}\widehat{\pi}^{(N)}(z + \theta(1-z)).$$

Integrating this inequality over $z \in [0,1]$, we have

$$\begin{aligned} N\sum_{i\in\mathbb{D}} \lambda_i \int_0^1 \frac{1-\widehat{X}_i(z)}{1-z} \widehat{\pi}^{(N)}_i(z)\mathrm{d}z \le \sum_{i\in\mathbb{D}} \mu_i \left\{\widehat{\pi}^{(N)}_i(1) - \widehat{\pi}^{(N)}_i(0)\right\} \\ + N\sum_{i\in\mathbb{D}} \gamma_i \left\{\widehat{\pi}^{(N)}_i(1) - \widehat{\pi}^{(N)}_i(\theta)\right\}. \end{aligned} \tag{3.39}$$

Note here that

$$\int_0^1 \frac{1-\widehat{X}_i(z)}{1-z}\widehat{\pi}_i^{(N)}(z)\mathrm{d}z \geq \widehat{\pi}_i^{(N)}(0)\cdot\int_0^1 \frac{1-\widehat{X}_i(z)}{1-z}\mathrm{d}z = \widehat{\pi}_i^{(N)}(0)\cdot \mathsf{E}\left[\int_0^1 \sum_{k=1}^{X_i} z^{k+1}\mathrm{d}z\right]$$
$$= \widehat{\pi}_i^{(N)}(0)\cdot \mathsf{E}\left[\sum_{k=1}^{X_i}\frac{1}{k}\right] \geq \widehat{\pi}_i^{(N)}(0)\cdot \mathsf{E}\left[\log X_i\right].$$

Combining (3.39) and the above inequality, we obtain

$$N\sum_{i\in\mathbb{D}}\lambda_i \mathsf{E}\left[\log X_i\right]\widehat{\pi}_i^{(N)}(0) \leq \sum_{i\in\mathbb{D}}\mu_i\left\{\widehat{\pi}_i^{(N)}(1)-\widehat{\pi}_i^{(N)}(0)\right\}$$
$$+ N\sum_{i\in\mathbb{D}}\gamma_i\left\{\widehat{\pi}_i^{(N)}(1)-\widehat{\pi}_i^{(N)}(\theta)\right\}. \tag{3.40}$$

Because of the irreducibility of $\{(J^{(N)}(t), L^{(N)}(t))\}$, $\boldsymbol{\pi}$ is strictly positive, and thus it follows that $\mathbf{0} < \boldsymbol{\pi}(0) < \widehat{\boldsymbol{\pi}}(z)$ for $z \in (0,1]$. Therefore, (3.40) implies that (3.31) holds. □

3.5 Numerical results

As mentioned in Remark 3.2, using the CLT, we obtained the approximation for the stationary queue length distribution of our model with large arrival rates as follows.

$$\mathsf{P}(L^{(N)} \leq x) \simeq \Phi_{(0,\sigma^2)}\left(N^{\beta/2}\left(\frac{x}{N}-\rho\right)\right), \qquad x\in\mathbb{R}. \tag{3.41}$$

In this section, we compare the approximated and simulated queue length distributions so as to confirm the accuracy of (3.41). Figures 3.3 and 3.4 show the comparison of these distributions, where the parameters except N and α are fixed as follows.

$$d=2, \quad \boldsymbol{Q}=\begin{pmatrix} -0.2, & 0.2 \\ 0.1, & 0.1 \end{pmatrix}, \quad (\lambda_1,\lambda_2)=(1,2), \quad X_1, X_2 \sim U\{1,10\},$$
$$(\mu_1,\mu_2)=(2,2), \quad (\gamma_1,\gamma_2)=(0.0001, 0.0002), \quad (p_1,p_2)=(0.2,0.1).$$

Figures 3.3 demonstrates the comparison with $\alpha = 2\ (>1)$ and $N = 2, 5, 50$. We observe that the distribution by simulation is very close to that presented in (3.41) even for a relatively small N. This suggests that the weak convergence shown in Theorem 3.1 is very fast when α is strictly larger than one.

Figures 3.4 shows the comparison with $\alpha = 0.5\ (<1)$ and $N = 2, 50, 10^4$. We observe that the simulated distribution and the normal distribution presented by (3.41) are not close even when N is large. This implies that the weak convergence in Theorem 3.1 is very slow when α is strictly smaller than one.

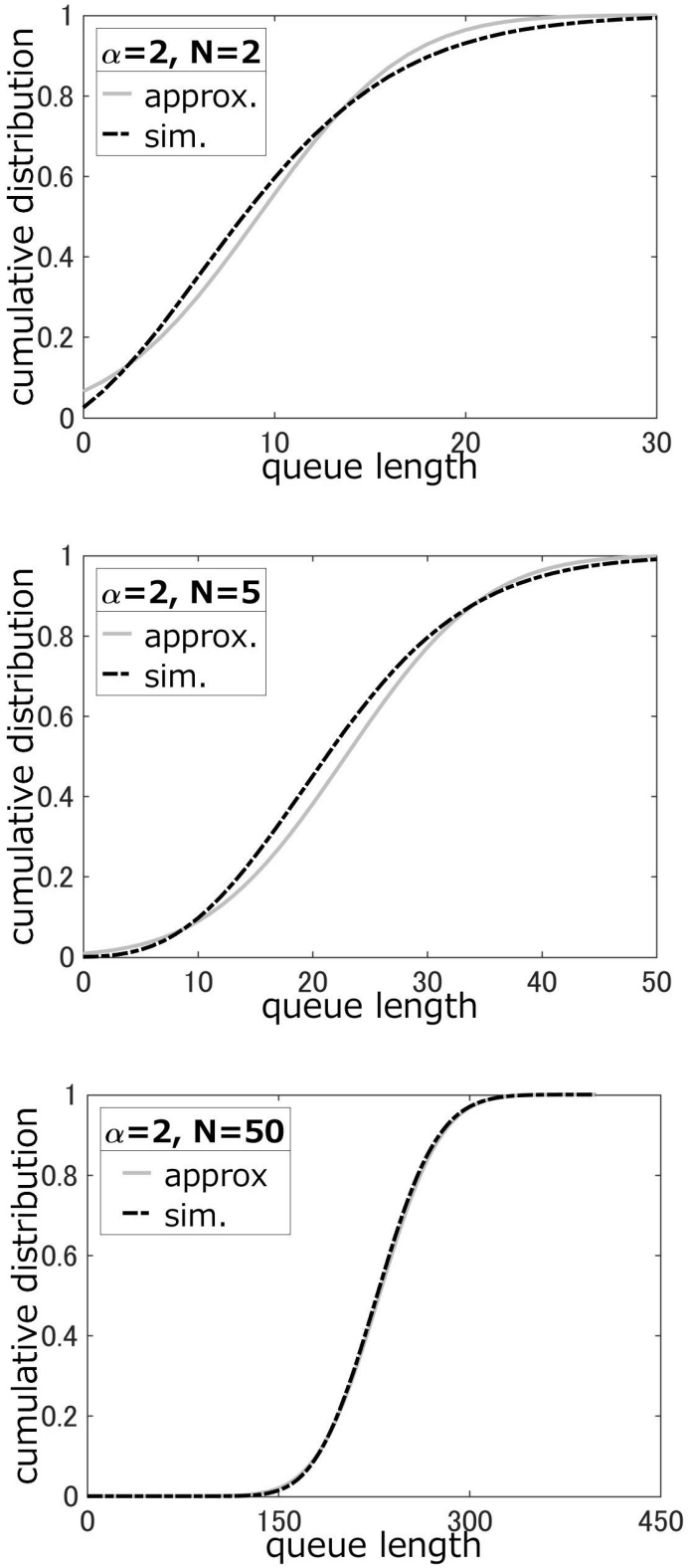

Figure 3.3: Simulated and approximated queue length distributions: $\alpha = 2$

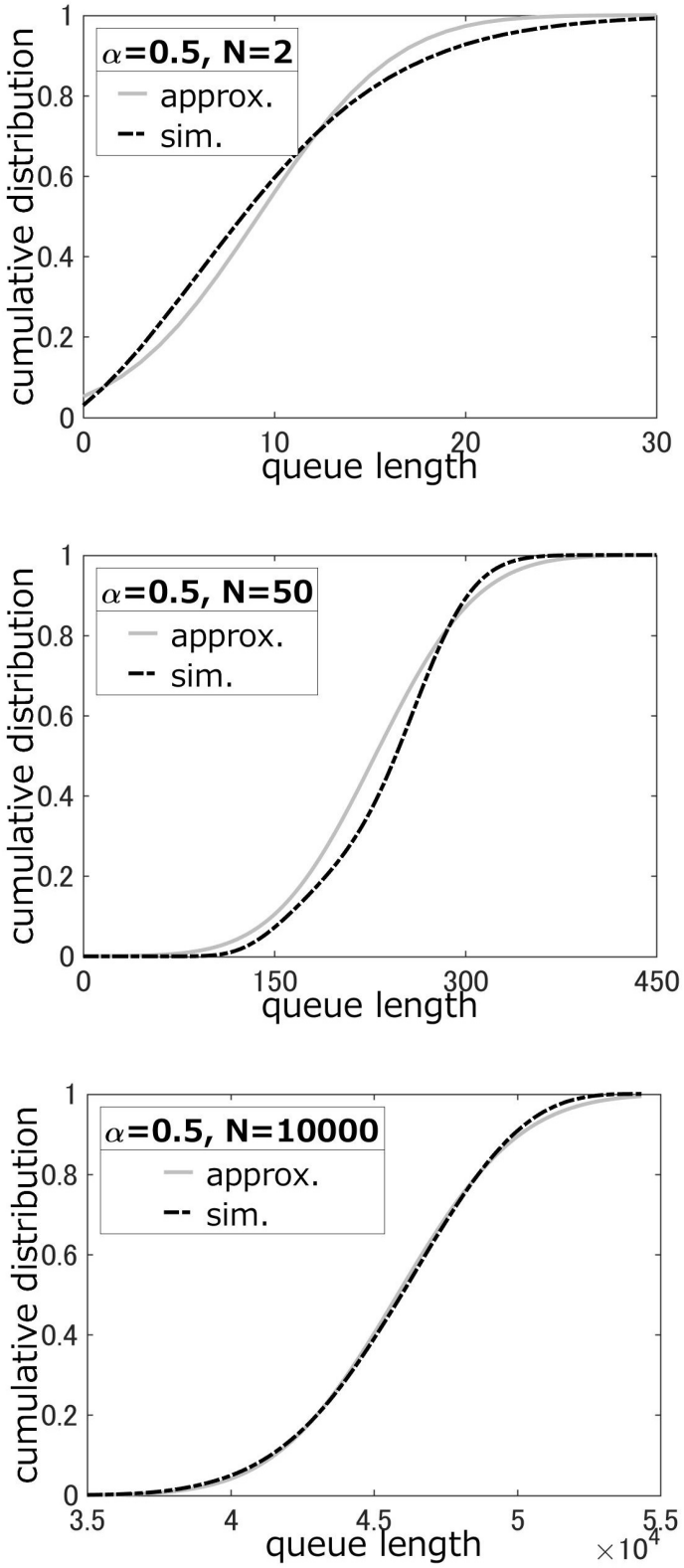

Figure 3.4: Simulated and approximated queue length distributions: $\alpha = 0.5$

3.6 Conclusion

In this chapter, we studied the Markov-modulated $\mathrm{M}^X/\mathrm{M}/\infty$ queue with binomial catastrophes. We focused on the scaling model such that the arrival rates are scaled by a factor N, the transition rates of the background process being scaled by N^α for a scaling coefficient α, the occurrence rates of binomial catastrophes being scaled by N and the removal probabilities being scaled by N^{-1}. Under this scaling regime, we established the central limit theorem for the stationary queue length distribution. Using the derived central limit theorem, we obtained the approximation of the stationary queue length distribution with large arrival rates.

In future work, we would like to show the central limit theorem under other heavy traffic regimes. In particular, we are interested in a regime such that the transition rate of the background process is scaled by $N^{\alpha'}$, where α' is introduced anew in addition to α. Furthermore, we would like to study the behavior of the our model without the assumption that the second moment of the batch size is finite. We also are interested in properties of the queue length process of our model at the transient state. We predict that the CLT of the queue length process holds; that is, the centered and normalized queue length process with some transformation converges to an Ornstein-Uhlenbeck process.

Chapter 4

Batch Arrival Single-erver Queue with Variable Service Speed

4.1 Introduction

In recent years, variable-speed CPUs have become popular because they can reduce energy consumption while maintaining on acceptable transmission delay (response time) for jobs. Thus, many researchers have studied queueing models with variable service speed [4, 40, 54, 64]. Variable-speed CPUs can be automatically adjusted in terms of its speed according to the workload or the number of jobs in the system. Working at high speed reduces transmission delay, but increase energy consumption. Thus, in general, a variable-speed CPU processes at high speed when the workload is large and reduces its speed accordingly when the workload is low. By such a way, a variable-speed CPU balance energy consumption and transmission delay.

CPUs still consume approximately 60% of their peak consumption processing a job even while not processing jobs [5]. Thus, a simple idea for reducing energy is to adopt *on-off policy*: that is, the server is turned off immediately after the system becomes empty, and the OFF server is reactivated immediately after a new job arrives at the empty system. However, a setup time is needed in order to reactivate the OFF server. Servers cannot process jobs during the setup, but consume energy. Thus, turning off the server does not always reduce energy consumption though increases the transmission delay.

We now consider a batch arrival single-server queue with variable service speed and the on-off policy, which is motivated by data centers with a variable-speed and power-aware CPU. We assume that customers arrive at the system in batches according to a Poisson process. Service requirements of customers in a batch is assumed to be independent and identically distributed (i.i.d.) with an exponential distribution. The service speed of the server is instantaneously adapted according to the queue length. In particular, we consider that the service speed changes in proportion to the queue length. Furthermore, setup times

are assumed to be i.i.d. with an exponential distribution. In this thesis, the above queueing model is referred to as the M^X/M/1/SET-VARI queue, where SET and VARI stand for setup (on-off policy) and variable service speed, respectively. It should be noted that the queue length of the M^X/M/1/SET-VARI queue is identical to that of the $M^X/M/\infty$ queue with the on-off policy.

In this chapter, we first obtain the stability condition of the M^X/M/1/SET-VARI queue. We show that the stability condition of our model is that the logarithmic moment of the batch size is finite. Interestingly, the system can be stable even if the mean batch size is infinite. Second, we derive the probability generating function (PGF) of the stationary queue length of our queueing model. Third, we derive the Laplace-Stieltjes transform (LST) of the stationary sojourn time distribution in term of infinite series form involving infinite dimensional matrices. The derivation of the sojourn time distribution is challenging because the sojourn time of a tagged customer depends on not only the state of the system upon arrival but also on the batches arriving after it. Therefore, the sojourn time distribution cannot be obtained directly from the queue length distribution via the distributional Little's law [36].

Our model extends the one proposed by Lu et al. [46]. They considered an M/M/1/SET-VARI queue. In [46], the stationary queue length distribution was derived in terms of infinite series. From the queue length distribution, the mean response time is obtained via Little's law and the mean power consumption is obtained. These metrics are used in [46] to find the energy-response trade-off. However, the sojourn time distribution was not considered in [46]. Adan and D'Auria [1] considered a single-server queue in which customers arrive according to a Poisson process, the service requirements of customers follow the exponential distribution and the service rate of the server is controlled by a threshold. They derived the stationary queue length distribution and the LST of the sojourn time distribution in explicit form. The sojourn time distribution of our model is derived using the first step analysis which is also adopted by Adan and D'Auria [1]. The difference is that the underlying Markov chain in Adan and D'Auria [1] is homogeneous after a threshold while our underlying Markov chain is spatially nonhomogeneous. As a result, the former allows explicit expression while our formulae involve inverse mappings of infinite matrices.

The remainder of this chapter is organized as follows. In Section 4.2, we describe the M^X/M/1/SET-VARI queue in detail. In Section 4.3, we present the stability condition. In Section 4.4, we derive the PGF of the queue length. In Section 4.5, we obtain the LST of the sojourn time distribution. In Section 4.6, we show numerical experiments showing the energy-performance trade-off. Finally, Section 4.7 is devoted to concluding remarks and future work.

4.2 Model description

In this section, we describe our queueing model; that is, the M^X/M/1/SET-VARI queue. We consider the single-server queue operating under the first come first served (FCFS) service discipline and an infinite buffer space. Batches of customers arrive at the system according to the Poisson process with rate $\lambda \in (0, \infty)$. Batch sizes (i.e., the numbers of customers in batches) are i.i.d. with random variable X on $\mathbb{N}$. We define $x_k = \mathsf{P}(X = k)$, $k \in \mathbb{N}$, and $\widehat{X}(z) := \sum_{k=1}^{\infty} x_k z^k$, $|z| \leq 1$.

The special feature of our model is that the service speed of the server changes in proportion to the queue length. Service requirements of customers are i.i.d. with the exponential distribution having mean 1. When there is one customer in the system, the amount of service provided by the server per unit time is $\mu \in (0, \infty)$ When there are n customers in the system, the speed of the server is scaled up to $n\mu$. Thus, when there are n customers in the system, the residual sojourn time of the ongoing customer follows the exponential distribution having mean $1/(n\mu)$.

Moreover, our model adopts the on-off policy. The server is turned off immediately after the system becomes empty, and the OFF server is reactivated immediately after a new batch arrives at the empty system. However, the setup time is needed to reactivate the OFF server, which implies that a batch arriving at the empty system has to wait until the setup time finishes. Setup times are i.i.d. with the exponential distribution having mean $1/\alpha \in (0, \infty)$. The server cannot process customers during the setup time, but consumes energy.

Let $L(t)$ denote the queue length (the number of server in the system) at time t. Let also $J(t)$ denote the state of server at time t: $J(t) = 0$ when the server is off or in the setup, and $J(t) = 1$ when the server is processing a customer. Under the current setting, the joint stochastic process $\{Z(t) := (J(t), L(t)); t \in \mathbb{R}_+\}$ is a continuous-time Markov chain with state space $\mathbb{S} = \{(0, k);\ k \in \mathbb{Z}_+\} \cup \{(1, k);\ k \in \mathbb{N}\}$. We assume that x_1 is strictly positive. It then follows that the Markov chain $\{Z(t)\}$ is irreducible. Let denote $\boldsymbol{Q} := (q(i, k; j, \ell))_{(i,k;j,\ell)\in\mathbb{S}\times\mathbb{S}}$ as the infinitesimal generator of $\{Z(t)\}$, which is given by

$$q(i,k;j,\ell) = \begin{cases} -\lambda, & i = j = 0,\ k = \ell = 0, \\ -\lambda - \alpha, & i = j = 0,\ k = \ell \geq 1, \\ -\lambda - k\mu, & i = j = 1,\ k = \ell, \\ \lambda x_{\ell-k}, & i = j,\ k < \ell, \\ k\mu, & i = j = 1,\ k = \ell + 1 \geq 2, \\ k\mu, & i = 1,\ j = 0\ , k = 1,\ \ell = 0, \\ \alpha, & i = 0,\ j = 1,\ k = \ell, \\ 0, & otherwise. \end{cases}$$

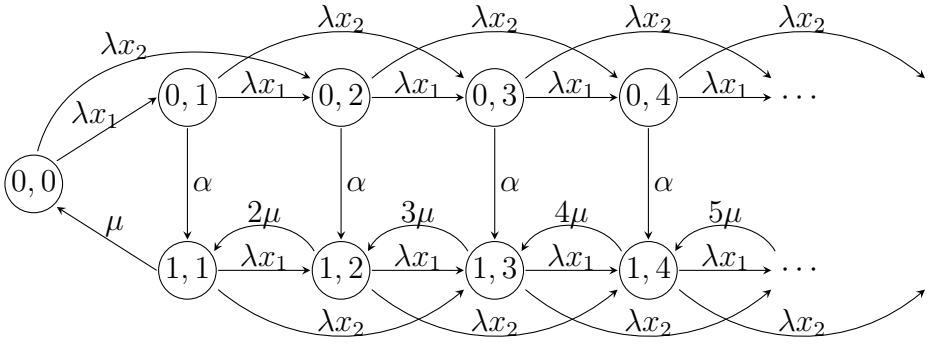

Figure 4.1: Transition diagram of the M^X/M/1/SET-VARI queue: $x_1 + x_2 = 1$

Figure 4.1 shows the transition diagram of the Markov chain $\{Z(t)\}$ for a special case in which the maximum batch size is two.

Remark 4.1 As mentioned in Section 4.1, the queue length of the M^X/M/1/SET-VARI queue is identical to that of the M^X/M/∞ queue with the on-off policy. However, the sojourn time distributions of these two models may be different because the sojourn time distribution of a tagged customer of the latter is determined upon its arrival while that of the former is affected by future arrivals. Some researchers have studied the M^X/M/∞ queue without the on-off policy. For example, Shanbhag [61] derived moment generating functions of some performance measures, e.g., the queue length and the sojourn time.

4.3 Stability condition

In this section, we derive the stability condition of the M^X/M/1/SET-VARI queue. Note that an irreducible and regular continuous-time Markov chain is positive recurrent if and only if it has a stationary distribution (i.e., the limiting distribution) [11]. In addition, if an irreducible and regular Markov chain is positive recurrent, its stationary distribution is unique and positive. Theorem 4.1 shows the stability condition of our model because the Markov chain $\{Z(t)\}$ is irreducible and regular.

Theorem 4.1 (Stability condition) *The Markov chain $\{Z(t); t \in \mathbb{R}_+\}$ has a unique stationary distribution if and only if the following inequality holds.*

$$\mathsf{E}[\log X] < \infty. \tag{4.1}$$

We emphasize that the addition of the on/off policy does not change the stability of the system because Cong [17] showed that the stability condition of the M^X/M/∞ queue (without the on-off policy), which is a special case of our model, is also that the logarithmic

moment of the batch-size distribution is finite. This is intuitively clear because the effect of setup times is likely to disappear when there exist many customers in the system.

Proof of Theorem 4.1. Supposing that (4.1) holds. Using Lemma 2.1, we show that $\{Z(t)\}$ is ergodic. To this end, we define $\varphi(i,k)$ and $\psi(i,k)$ as

$$\begin{aligned} \varphi(i,k) &= \log(k+1), && (i,k) \in \mathbb{S}, \\ \psi(i,k) &= \sum_{(j,\ell)\in\mathbb{S}} \varphi(j,\ell) \cdot q(i,k;j,\ell), && (i,k) \in \mathbb{S}. \end{aligned}$$

We then have

$$\psi(0,0) = \sum_{\ell=1}^{\infty} \log(\ell+1) \cdot \lambda x_\ell = \lambda \mathsf{E}[\log(X+1)] < \infty, \tag{4.2}$$

where the inequality follows from (4.1). For $k \in \mathbb{N}$, we also have

$$\begin{aligned} \psi(0,k) &= -\log(k+1) \cdot (\lambda+\alpha) + \sum_{\ell=1}^{\infty} \log(k+\ell+1) \cdot \lambda x_\ell + \log(k+1) \cdot \alpha \\ &= \lambda \sum_{\ell=1}^{\infty} \log\left(1 + \frac{\ell}{k+1}\right) \cdot x_\ell. \end{aligned} \tag{4.3}$$

Note here that, for $k, \ell \in \mathbb{N}$,

$$\log\left(1 + \frac{\ell}{k+1}\right) = \log \ell + \log\left(\frac{1}{\ell} + \frac{1}{k+1}\right) \le \log \ell + \log 2. \tag{4.4}$$

Applying (4.4) to (4.3) and using (4.1), we obtain

$$\psi(0,k) \le \lambda \mathsf{E}[\log X] + \lambda \log 2 < \infty, \qquad k \in \mathbb{N}. \tag{4.5}$$

Furthermore, we have, for $k \in \mathbb{N}$,

$$\begin{aligned} \psi(1,k) &= -\log(k+1) \cdot (\lambda + k\mu) + \sum_{\ell=1}^{\infty} \log(k+\ell+1) \cdot \lambda x_\ell + \log k \cdot k\mu \\ &= \lambda \sum_{\ell=1}^{\infty} \log\left(1 + \frac{\ell}{k+1}\right) \cdot x_\ell + k\mu \cdot \log\left(1 - \frac{1}{k+1}\right). \end{aligned} \tag{4.6}$$

Note that, from (2.9), there exists some $\delta > 0$ such that

$$k\mu \log\left(1 - \frac{1}{k+1}\right) \le -2\delta, \qquad \text{for all } k \in \mathbb{N}.$$

Applying the above inequality and (4.4) into (4.6) yields

$$\psi(1,k) \le \lambda \mathsf{E}[\log X] + \lambda \log 2 - 2\delta < \infty, \qquad k \in \mathbb{N}, \tag{4.7}$$

where the last inequality follows from (4.1). In addition, by dominated convergence theorem, we have

$$\lim_{k\to\infty}\sum_{\ell=1}^{\infty}\log\Big(1+\frac{\ell}{k+1}\Big)x_\ell=0,$$

and thus there exists some $K:=K_\delta\in\mathbb{N}$ such that, for all $k=K+1,K+2,\ldots,$

$$\lambda\sum_{\ell=1}^{\infty}\log\Big(1+\frac{\ell}{k+1}\Big)x_\ell\le\delta.$$

Applying this inequality and (4.4) into (4.6), we obtain

$$\psi(1,k)\le-\delta,\qquad k=K+1,K+2,\ldots. \tag{4.8}$$

Using Lemma 2.1, it follows from (4.2), (4.5),(4.7) and (4.8) that $\{Z(t)\}$ is ergodic.

On the other hand, supposing that $\{Z(t)\}$ is ergodic. We define $\boldsymbol{\pi}=(\pi_{i,k})_{(i,k)\in\mathbb{S}}$ as the stationary distribution of $\boldsymbol{Q}$. We define the generating functions $\widehat{\pi}_0(z)$ and $\widehat{\pi}_1(z)$ as follows, respectively.

$$\widehat{\pi}_0(z)=\sum_{k=0}^{\infty}\pi_{0,k}z^k,\qquad\widehat{\pi}_1(z)=\sum_{k=1}^{\infty}\pi_{1,k}z^k,\qquad|z|\le1.$$

We then have the following balance equations.

$$\lambda\pi_{0,0}=\mu\pi_{1,1}, \tag{4.9}$$

$$(\lambda+\alpha)\pi_{0,k}=\lambda\sum_{\ell=1}^{k}x_\ell\pi_{0,k-\ell},\qquad k\in\mathbb{N}, \tag{4.10}$$

$$(\lambda+k\mu)\pi_{1,k}=\alpha\pi_{0,k}+(1+k)\mu\pi_{1,k+1}+\lambda\sum_{\ell=1}^{k-1}x_\ell\pi_{1,k-\ell},\qquad k\in\mathbb{N}, \tag{4.11}$$

where the empty sum (i.e., summation from 1 to0) is defined as 0. Multiplying (4.11) by z and (4.11) by z^k and taking the sum over $k\in\mathbb{N}$ yields

$$\begin{aligned}&\lambda\sum_{k=1}^{\infty}\pi_{1,k}z^k+\mu z\sum_{k=1}^{\infty}\pi_{1,k}\big(z^k\big)'\\&\quad=\alpha\sum_{k=0}^{\infty}\pi_{0,k}z^k-\alpha\pi_{0,0}+\mu\sum_{k=1}^{\infty}\pi_{1,k}\big(z^k\big)'-\mu\pi_{1,1}+\lambda\sum_{\ell=1}^{\infty}x_\ell z^\ell\sum_{k=1}^{\infty}\pi_{1,k}z^k.\end{aligned}$$

Rearranging the above equation, we find that

$$\frac{\mathrm{d}}{\mathrm{d}z}\widehat{\pi}_1(z)=\frac{\lambda}{\mu}q(z)\widehat{\pi}_1(z)+\frac{\lambda}{\mu}q(z)\widehat{\pi}_0(z), \tag{4.12}$$

where $q(z)$ is defined by

$$q(z) = \frac{1 - \widehat{X}(z)}{1 - z} = \mathsf{E}\left[\sum_{\ell=0}^{X-1} z^{\ell}\right] = \sum_{\ell=0}^{\infty} \mathsf{P}(X > \ell) z^{\ell}. \tag{4.13}$$

Let $Q(z)$ denote the primitive function satisfying that $Q(0) = 0$; that is,

$$Q(z) = \int_0^z q(u)\mathrm{d}u = \mathsf{E}\left[\sum_{\ell=1}^{X} \frac{z^{\ell}}{\ell}\right] = \sum_{\ell=1}^{\infty} \mathsf{P}(X \geq \ell) \frac{z^{\ell}}{\ell}. \tag{4.14}$$

Using $Q(z)$, the solution of (4.12) is given by

$$\widehat{\pi}_1(z) = H(z) \mathrm{e}^{\frac{\lambda}{\mu} Q(z)}, \tag{4.15}$$

where $H(z)$ is some function which will be determined later. Differentiating (4.15) and substituting the result into (4.12), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}z} H(z) = \mathrm{e}^{-\frac{\lambda}{\mu} Q(z)} \frac{\lambda}{\mu} q(z) \widehat{\pi}_0(z).$$

It follows from $\widehat{\pi}_1(0) = 0$ and $Q(0) = 0$ that $H(0) = 0$. Therefore, we have

$$H(z) = \int_0^z \mathrm{e}^{-\frac{\lambda}{\mu} Q(u)} \frac{\lambda}{\mu} q(u) \widehat{\pi}_0(u) \mathrm{d}u.$$

Substituting this equation into (4.15), we obtain

$$\widehat{\pi}_1(z) = \mathrm{e}^{\frac{\lambda}{\mu} Q(z)} \left\{ \int_0^z \mathrm{e}^{-\frac{\lambda}{\mu} Q(u)} \frac{\lambda}{\mu} q(u) \widehat{\pi}_0(u) \mathrm{d}u \right\}. \tag{4.16}$$

It follows from (4.16) that

$$\begin{aligned}
1 &= \widehat{\pi}_0(1) + \widehat{\pi}_1(1) \\
&= \widehat{\pi}_0(1) + \mathrm{e}^{\frac{\lambda}{\mu} Q(1)} \left\{ \int_0^1 \mathrm{e}^{-\frac{\lambda}{\mu} Q(u)} \frac{\lambda}{\mu} q(u) \widehat{\pi}_0(u) \mathrm{d}u \right\} \\
&\geq \pi_{0,0} + \mathrm{e}^{\frac{\lambda}{\mu} Q(1)} \left\{ \int_0^1 \mathrm{e}^{-\frac{\lambda}{\mu} Q(u)} \frac{\lambda}{\mu} q(u) \pi_{0,0} \mathrm{d}u \right\} \\
&= \pi_{0,0} + \pi_{0,0} \cdot \mathrm{e}^{\frac{\lambda}{\mu} Q(1)} \cdot \left\{ 1 - \mathrm{e}^{-\frac{\lambda}{\mu} Q(1)} \right\} \\
&\geq \pi_{0,0} \cdot \frac{\lambda}{\mu} Q(1),
\end{aligned} \tag{4.17}$$

where the first inequality follows from $\widehat{\pi}_0(z) \geq \pi_{0,0}$, $0 \leq z \leq 1$, and the second inequality follows from $\mathrm{e}^x \leq x + 1$, $x \in \mathbb{R}$. Note here that, for $\ell \in \mathbb{N}$,

$$\sum_{k=1}^{\ell} \frac{1}{k} \geq \log \ell.$$

Combining (4.14) and the above inequality yields

$$Q(1) = \mathsf{E}\left[\sum_{k=1}^{X} \frac{1}{k}\right] \geq \mathsf{E}[\log X]. \tag{4.18}$$

Consequently, (4.17) and (4.18) implies that (4.1) holds. □

4.4 Queue length distribution

In this section, we consider the stationary queue length distribution, assuming the stability condition $\mathsf{E}[\log X] < \infty$. Let $\boldsymbol{\pi} = (\pi_{i,k})_{i,k\in\mathbb{S}}$ denote the stationary distribution of $\boldsymbol{Q}$. We define L as the random variable following the stationary queue length distribution. We also define $\widehat{\pi}(z)$ as the PGF of L; that is,

$$\widehat{\pi}(z) := \mathsf{E}[z^L] = \sum_{k=0}^{\infty} \pi_{0,k} z^k + \sum_{k=1}^{\infty} \pi_{1,k} z^k, \qquad |z| \le 1.$$

We derive the PGF $\widehat{\pi}(z)$ as Theorem 4.2.

Theorem 4.2 (PGF of the stationary queue length) *The probability generating function of the stationary queue length of the M^X/M/1/SET-VARI queue, denoted by $\widehat{\pi}(z)$, is given by, for $|z| \le 1$,*

$$\widehat{\pi}(z) = \frac{\dfrac{1}{\lambda+\alpha-\lambda\widehat{X}(z)} + \displaystyle\int_0^z \mathrm{e}^{\frac{\lambda}{\mu}\{Q(z)-Q(u)\}} \frac{\lambda}{\mu} q(u) \frac{1}{\lambda+\alpha-\lambda\widehat{X}(u)} \mathrm{d}u}{\dfrac{1}{\alpha} + \displaystyle\int_0^1 \mathrm{e}^{\frac{\lambda}{\mu}\{Q(1)-Q(u)\}} \frac{\lambda}{\mu} q(u) \frac{1}{\lambda+\alpha-\lambda\widehat{X}(u)} \mathrm{d}u}, \tag{4.19}$$

where $q(z)$ and $Q(z)$ are given by (4.13) *and* (4.14)*, respectively.*

Proof. Multiplying (4.10) by z^k taking the sum over $k \in \mathbb{N}$, and rearranging the result, we obtain

$$\widehat{\pi}_0(z) = \frac{\lambda+\alpha}{\lambda+\alpha-\lambda\widehat{X}(z)} \pi_{0,0}. \tag{4.20}$$

On the other hand, $\widehat{\pi}_1(z)$ is given by (4.16). Thus, from (4.20) and (4.16), we have

$$\widehat{\pi}(z) = \frac{\lambda+\alpha}{\lambda+\alpha-\lambda\widehat{X}(z)} \pi_{0,0} + \int_0^z \mathrm{e}^{\frac{\lambda}{\mu}\{Q(z)-Q(u)\}} \frac{\lambda}{\mu} q(u) \frac{\lambda+\alpha}{\lambda+\alpha-\lambda\widehat{X}(u)} \pi_{0,0} \mathrm{d}u. \tag{4.21}$$

Applying the normalizing condition (i.e., $\widehat{\pi}(1) = 1$) to (4.21) yields

$$\pi_{0,0} = \left\{ \frac{\lambda+\alpha}{\alpha} + \int_0^1 \mathrm{e}^{\frac{\lambda}{\mu}\{Q(1)-Q(u)\}} \frac{\lambda}{\mu} q(u) \frac{\lambda+\alpha}{\lambda+\alpha-\lambda\widehat{X}(u)} \mathrm{d}u \right\}^{-1}. \tag{4.22}$$

Substituting this equation into (4.21), we consequently obtain $\widehat{\pi}(z)$. □

Remark 4.2 As mentioned in Section 4.1, the queue length of our model is identical to that of the M^X/M/∞ queue with the on-off policy. In [61], the PGF of the stationary queue length distribution for the M^X/M/∞ queue (without the on-off policy), denoted by $\widehat{\pi}^*(z)$, was derived as

$$\widehat{\pi}^*(z) = \mathrm{e}^{\frac{\lambda}{\mu}\{Q(z)-Q(1)\}}, \qquad |z| \le 1. \tag{4.23}$$

It is easy to see that (4.19) tends to (4.23) as $\alpha \to \infty$.

Moreover, we can easily derive the average of the stationary queue length.

Corollary 4.1 (Average of the stationary queue length) *Assuming that* $\mathsf{E}[X] < \infty$*,* $\mathsf{E}[L]$ *is given as follows.*

$$\mathsf{E}[L] = \frac{\lambda \mathsf{E}[X]}{\mu} \left\{ \frac{\lambda + \alpha}{\alpha} \frac{\mu}{\alpha} \pi_{0,0} + 1 \right\},$$

where $\pi_{0,0}$ *is given by (4.22).*

Proof. Differentiating (4.21), we have

$$\frac{\mathrm{d}}{\mathrm{d}z} \widehat{\pi}(z) = \frac{\lambda(\lambda + \alpha) \widehat{X}'(z)}{(\lambda + \alpha - \lambda \widehat{X}(z))^2} \pi_{0,0} + \frac{\lambda}{\mu} q(z) \widehat{\pi}(z).$$

Taking the limit as $z \uparrow 1$ in the above equation yields

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}z} \widehat{\pi}(1) &= \frac{\lambda(\lambda + \alpha) \mathsf{E}[X]}{\alpha^2} \pi_{0,0} + \frac{\lambda}{\mu} \left\{ \lim_{z \uparrow 1} \frac{1 - \widehat{X}(z)}{1 - z} \right\} \\ &= \frac{\lambda(\lambda + \alpha) \mathsf{E}[X]}{\alpha^2} \pi_{0,0} + \frac{\lambda}{\mu} \mathsf{E}[X], \end{aligned}$$

where the second equality holds because of L'Hospital's rule. From the relation $\mathsf{E}[L] = \widehat{\pi}'(1)$, we complete to show Corollary 4.1. □

4.5 Sojourn time distribution

In the M^X/M/1/SET-VARI queue, the sojourn time of a tagged customer is affected by the batches arriving after it, because the server changes its service speed upon arrivals and departures of customers. This makes the analysis of the sojourn time distribution complex and challenging. In this chapter, we derive the LST for the sojourn time distribution. Note that the LST of a distribution function F on $\mathbb{R}_+$ is defined as $F^*(s) := \int_0^\infty \mathrm{e}^{-st} F(\mathrm{d}t)$. We assume that $\mathsf{E}[X] < \infty$ for the existence of the equilibrium batch-size distribution.

Theorem 4.3 (LST of the stationary sojourn time distribution) *The Laplace-Stieltjes transform of the sojourn time distribution, denoted by* $W^*(s)$*, is given as follow:*

$$\begin{aligned} W^*(s) = &\sum_{m=1}^{\infty} \frac{1}{\mathsf{E}[X]} \Big[\boldsymbol{\pi}_0 \boldsymbol{I}_m \boldsymbol{B} (\boldsymbol{I} - \boldsymbol{I}_m)(\boldsymbol{I} - \boldsymbol{\Lambda}^{(0)})^{-1} \boldsymbol{A} \Big] \Big[(\boldsymbol{I} - \boldsymbol{\Lambda}^{(1)})^{-1} \boldsymbol{M} \Big]^m \boldsymbol{e} \\ &+ \sum_{m=1}^{\infty} \frac{1}{\mathsf{E}[X]} \Big[\boldsymbol{\pi}_1 \boldsymbol{I}_m \boldsymbol{B} (\boldsymbol{I} - \boldsymbol{I}_m) \Big] \Big[(\boldsymbol{I} - \boldsymbol{\Lambda}^{(1)})^{-1} \boldsymbol{M} \Big]^m \boldsymbol{e}, \end{aligned}$$

where $\boldsymbol{\Lambda}^{(1)}$*,* $\boldsymbol{M}$*,* $\boldsymbol{\Lambda}^{(0)}$*,* $\boldsymbol{A}$*,* $\boldsymbol{B}$ *and* $\boldsymbol{I}_m$ *are given by (4.26), (4.27), (4.33), (4.34), (4.36) and (4.37), respectively.*

Proof. First, we consider the residual sojourn times of customers in the system when the server is processing a customer. Let $W_1(n,m)$ denote the residual sojourn time of the tagged customer given that the server is processing a customer, that there exist n customers in the system and that the tagged customer is in the mth position of the waiting line. Conditioning on the first step transitions, we have, for $n \geq m$,

$$W_1(n,m) = \frac{Y}{\lambda + n\mu} + \begin{cases} W_1(n-1,m-1), & \text{w.p. } \dfrac{n\mu}{\lambda+n\mu}, \\ W_1(n+k,m), & \text{w.p. } \dfrac{\lambda x_k}{\lambda+n\mu}, \qquad k \in \mathbb{N}, \end{cases} \tag{4.24}$$

where Y denotes the random variable following the exponential distribution with mean 1, and $W_1(n,0) = 0$ for $n \in \mathbb{N}$. Furthermore, let $W_1^*(n,m,s)$ denote the LST of $W_1(n,m)$; that is, $W_1^*(n,m,s) = \mathsf{E}[\exp(-sW_1(n,m))]$. Using (4.24), we obtain the following recursive formula of $W_1^*(n,m,s)$, for $n \geq m$.

$$W_1^*(n,m,s) = \frac{n\mu}{s+\lambda+n\mu} W_1^*(n-1,m-1,s) + \frac{\lambda}{s+\lambda+n\mu} \sum_{k=1}^{\infty} x_k W_1^*(n+k,m,s), \tag{4.25}$$

where $W_1^*(n,0,s) = 1$. We use the convention that $W_1^*(n,m,s) = 0$ for $n < m$.

To simplify the recursive formula (4.25) with two variables, n and m, we define the infinite matrices $\boldsymbol{\Lambda}^{(1)} := (\Lambda^{(1)}_{k,\ell})_{k,\ell\in\mathbb{N}}$ and $\boldsymbol{M} := (M_{k,\ell})_{k,\ell\in\mathbb{N}}$ as

$$\Lambda^{(1)}_{k,\ell} = \begin{cases} \dfrac{\lambda x_{\ell-k}}{s+\lambda+(k-1)\mu}, & 1 < k < \ell, \\ 0, & otherwise, \end{cases} \tag{4.26}$$

$$M_{k,\ell} = \begin{cases} \dfrac{(k-1)\mu}{s+\lambda+(k-1)\mu}, & 1 < k = \ell+1, \\ 0, & otherwise. \end{cases} \tag{4.27}$$

We also define the infinite column vector $\boldsymbol{W}_1^*(m,s)$ as

$$\boldsymbol{W}_1^*(m,s) = (W_1^*(0,m,s), W_1^*(1,m,s), W_1^*(2,m,s), \dots)^\top. \tag{4.28}$$

Rearranging (4.25) by using (4.26)–(4.28), we obtain the following recursive formula with only one variable.

$$\boldsymbol{W}_1^*(m,s) = \boldsymbol{M}\boldsymbol{W}_1^*(m-1,s) + \boldsymbol{\Lambda}^{(1)}\boldsymbol{W}_1^*(m,s), \qquad m \in \mathbb{N}. \tag{4.29}$$

We prove that the operator norm of infinite matrix $\boldsymbol{\Lambda}^{(1)}$, $\| \boldsymbol{\Lambda}^{(1)} \| = \sup_{\|\boldsymbol{z}\|_2=1} \| \boldsymbol{\Lambda}^{(1)}\boldsymbol{z} \|_2$, is strictly smaller than one. Indeed, we have, for all $\boldsymbol{z} = (z_1, z_2, \dots)^\top$ such that $\| \boldsymbol{z} \|_2 = 1$,

$$\begin{aligned} \| \boldsymbol{\Lambda}^{(1)}\boldsymbol{z} \|_2 &< \frac{\lambda}{\lambda+s}\Big(\sum_{i>1}\Big(\sum_{j>i} x_{j-i}z_j\Big)^2\Big)^{1/2} = \frac{\lambda}{\lambda+s}\Big(\sum_{i>1}\Big(\sum_{j\geq 1} x_j z_{j+i}\Big)^2\Big)^{1/2} \\ &\leq \frac{\lambda}{\lambda+s}\Big(\sum_{i>1}\sum_{j\geq 1} x_j\Big(z_{j+i}\Big)^2\Big)^{1/2} \leq \frac{\lambda}{\lambda+s}\Big(\sum_{j\geq 1} x_j ||\boldsymbol{z}||_2^2\Big)^{1/2} = \frac{\lambda}{\lambda+s}, \end{aligned}$$

where the second inequality holds because of Jensen's inequality. It then follows that

$$\| \boldsymbol{\Lambda}^{(1)} \| \leq \frac{\lambda}{\lambda + s} < 1.$$

Because $\| \boldsymbol{\Lambda}^{(1)} \|$ is strictly smaller than one, $(\boldsymbol{I} - \boldsymbol{\Lambda}^{(1)})$ has inverse mapping [37, Section 29, Theorem 8]. Thus, from (4.29), we have the following recurrence formula, for $m \in \mathbb{N}$.

$$\boldsymbol{W}_1^*(m, s) = (\boldsymbol{I} - \boldsymbol{\Lambda}^{(1)})^{-1} \boldsymbol{M} \boldsymbol{W}_1^*(m-1, s), \tag{4.30}$$

Note that it follows from $W_1^*(n, 0, s) = 1$ that $\boldsymbol{W}_1^*(0, s) = \boldsymbol{e}$. Thus, solving the recursive formula (4.30), we obtain

$$\boldsymbol{W}_1^*(m, s) = \{(\boldsymbol{I} - \boldsymbol{\Lambda}^{(1)})^{-1} \boldsymbol{M}\}^m \boldsymbol{e}. \tag{4.31}$$

Next, we consider the residual sojourn times of customers in the system when the server is not processing a customer. We define $W_0(n, m)$ as the residual sojourn time of the tagged customer given that the server is not processing a customer, that there exist n customers in the system, and that the tagged customer is in the mth position of the waiting line. Conditioning on the first step transitions, we have, for $n \geq m$,

$$W_0(n, m) = \frac{Y}{\lambda + \alpha} + \begin{cases} W_1(n, m), & \text{w.p. } \dfrac{\alpha}{\lambda + \alpha}, \\ W_0(n + k, m), & \text{w.p. } \dfrac{\lambda x_k}{\lambda + \alpha}, \end{cases} \qquad k \in \mathbb{N},$$

Furthermore, let $W_0^*(n, m, s)$ denote the LST of $W_0(n, m)$ for $n \in \mathbb{N}$ and $m \leq n$. In addition, we define the infinite column vector $\boldsymbol{W}_0^*(m, s)$ as

$$\boldsymbol{W}_0^*(m, s) = (W_0^*(0, m, s), W_0^*(1, m, s), W_0^*(2, m, s), \dots)^\top,$$

where $W_0^*(n, 0, s) = 1$, $n \in \mathbb{Z}_+$, and $W_0^*(n, m, s) = 0$, $n \in \mathbb{Z}_+$ and $m > n$. We use the convention that $W_0^*(n, m, s) = 0$ for $n < m$. As with the analysis for $\boldsymbol{W}_1^*(m, s)$, i.e., (4.24)–(4.31), we obtain

$$\boldsymbol{W}_0^*(m, s) = (\boldsymbol{I} - \boldsymbol{\Lambda}^{(0)})^{-1} \boldsymbol{A} \boldsymbol{W}_1^*(m, s), \tag{4.32}$$

where the infinite matrices $\boldsymbol{\Lambda}^{(0)} := (\Lambda_{k,\ell}^{(0)})_{k,\ell \in \mathbb{N}}$ and $\boldsymbol{A} := (A_{k,\ell})_{k,\ell \in \mathbb{N}}$ are given by

$$\Lambda_{k,\ell}^{(0)} = \begin{cases} \dfrac{\lambda x_{\ell - k}}{s + \lambda + \alpha}, & 1 < k < \ell, \\ 0, & otherwise, \end{cases} \tag{4.33}$$

$$A_{k,\ell} = \begin{cases} \dfrac{\alpha}{s + \lambda + \alpha}, & 1 < k = \ell, \\ 0, & otherwise. \end{cases} \tag{4.34}$$

Note that $(\boldsymbol{I}-\boldsymbol{\Lambda}^{(0)})$ has the inverse mapping, which can be proved as similar to the analysis for $(\boldsymbol{I}-\boldsymbol{\Lambda}^{(1)})$.

Finally, we derive the (unconditional) LST of the stationary sojourn time distribution. To this end, we express $W^*(s)$ using $W_1^*(n,m,s)$ and $W_0^*(n,m,s)$. We define $\tau(i,n,m)$ as the probability that the tagged customer is located in the mth position and the state of the system becomes (i,n) immediately after its arrival. Let $\mathcal{I} \in \{0,1\}$ denote the state of the system immediately before the tagged customer arrives at the system and let $\mathcal{L}_p$ denote the number of customers in the system just before the tagged customer arrives at the system. From PASTA [69], we have

$$\mathsf{P}(\mathcal{I}=i,\, \mathcal{L}_p=n)=\pi_{i,n}, \qquad (i,n)\in\mathbb{S}.$$

We define $\mathcal{P}$ as the position at which the tagged customer is located immediately after it enters the system. We also define $\widetilde{X}$ as the batch size in which the tagged customer belongs. Note that $\widetilde{X}$ follows the equilibrium distribution of X; that is,

$$\mathsf{P}(\widetilde{X}=k)=\frac{kx_k}{\mathsf{E}[X]}, \qquad k\in\mathbb{N}.$$

Using there random variables, $\tau(1,n,m)$ can be written as, for $n\geq m$,

$$\tau(1,n,m)=\sum_{k=n-m+1}^{n-1}\mathsf{P}(\mathcal{I}=1,\,\mathcal{L}_p=n-k,\,\widetilde{X}=k,\,\mathcal{P}=m).$$

Thus, we obtain, for $n\geq m$,

$$\begin{aligned}\tau(1,n,m)&=\sum_{k=n-m+1}^{n-1}\mathsf{P}(\mathcal{I}=i,\mathcal{L}_p=n-k)\cdot\mathsf{P}(\widetilde{X}=k)\mathsf{P}(\mathcal{P}=m|\mathcal{L}_p=n-k,\widetilde{X}=k)\\&=\sum_{k=n-m+1}^{n-1}\pi_{1,n-k}\frac{x_k}{\mathsf{E}[X]},\end{aligned}$$

where the empty sum (i.e., summation from one to zero) is defined as zero. Similar to the above, we obtain, for $n\in\mathbb{Z}_+$ and $n\geq m$,

$$\tau(0,n,m)=\sum_{k=n-m+1}^{n}\pi_{0,n-k}\frac{x_k}{\mathsf{E}[X]}.$$

Note here that $W_i^*(n,m,s)=0$ for $n<m$. Thus, using $W_1^*(n,m,s)$ and $W_0^*(n,m,s)$, the LST of the sojourn time distribution can be expressed as follows.

$$\begin{aligned}W^*(s)&=\sum_{n=1}^{\infty}\sum_{m=1}^{n}\tau(0,n,m)W_0^*(n,m,s)+\sum_{n=2}^{\infty}\sum_{m=2}^{n}\tau(1,n,m)W_1^*(n,m,s)\\&=\sum_{m=1}^{\infty}\sum_{n=m}^{\infty}\sum_{k=n-m+1}^{n}\pi_{0,n-k}\frac{x_k}{\mathsf{E}[X]}W_0^*(n,m,s)\\&\quad+\sum_{m=2}^{\infty}\sum_{n=m}^{\infty}\sum_{k=n-m+1}^{n-1}\pi_{1,n-k}\frac{x_k}{\mathsf{E}[X]}W_1^*(n,m,s).\end{aligned}\tag{4.35}$$

state		energy consumption
service	$(1,k)\ k \geq 1$	$K_{\text{service}} \times (k\mu)^2$
setup	$(0,k)\ k \geq 1$	$K_{\text{set}} \times \mu^2$
idle	$(0,0)$	0

Table 4.1: Energy consumption per unit time of the M^X/M/1/SET-VARI queue

It is obvious that the infinite series included in $W^*(s)$ converges. The reason is that $\sum_{n=1}^{\infty}\sum_{m=1}^{n}\tau(0,n,m) + \sum_{n=2}^{\infty}\sum_{m=2}^{n}\tau(1,n,m) = 1$ and $0 \leq W_i^*(n,m,s) \leq 1$ for $i = 1, 2$, $n \in \mathbb{N}$ and $1 \leq m \leq n$.

For a compact expression of (4.35), we define the infinite matrices $\boldsymbol{B} := (B_{k,\ell})_{k,\ell\in\mathbb{N}}$ and $\boldsymbol{I}_m := (I_{m,k,\ell})_{k,\ell\in\mathbb{N}}$, for $m \in \mathbb{N}$, as

$$B_{k,\ell} = \begin{cases} x_{\ell-k}, & 1 \leq k < \ell, \\ 0, & otherwise. \end{cases} \tag{4.36}$$

$$I_{m,k,\ell} = \begin{cases} 1, & 1 \leq k = \ell \leq m, \\ 0, & otherwise. \end{cases} \tag{4.37}$$

In addition, we define the infinite row vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ as

$$\boldsymbol{\pi}_0 = (\pi_{0,0}, \pi_{0,1}, \pi_{1,2}, \dots), \qquad \boldsymbol{\pi}_1 = (0, \pi_{1,1}, \pi_{1,2}, \dots).$$

Rearranging (4.35) by using these matrices and vectors, we obtain

$$W^*(s) = \sum_{m=1}^{\infty} \frac{\boldsymbol{\pi}_0}{\mathsf{E}[X]} \boldsymbol{I}_m \boldsymbol{B}(\boldsymbol{I} - \boldsymbol{I}_m)\boldsymbol{W}_0^*(m,s) + \sum_{m=1}^{\infty} \frac{\boldsymbol{\pi}_1}{\mathsf{E}[X]} \boldsymbol{I}_m \boldsymbol{B}(\boldsymbol{I} - \boldsymbol{I}_m)\boldsymbol{W}_1^*(m,s). \tag{4.38}$$

From (4.31), (4.32) and (4.38), we complete the proof of Theorem 4.3. □

Remark 4.3 The LST of the sojourn time distribution given in Theorem 4.3 is in series form involving infinite dimensional matrices. Therefore, an approximation is necessary for numerical calculation. In Section 4.6.4, for numerical experiments, we present a method to approximate $W^*(s)$. However, we have not yet been able to find a bound for the error. It is important future work to find an approximation method with guaranteed accuracy.

4.6 Numerical experiments

In this section, we present some numerical experiments for showing the transmission delay and the energy consumption of the M^X/M/1/SET-VARI queue. The energy consumption

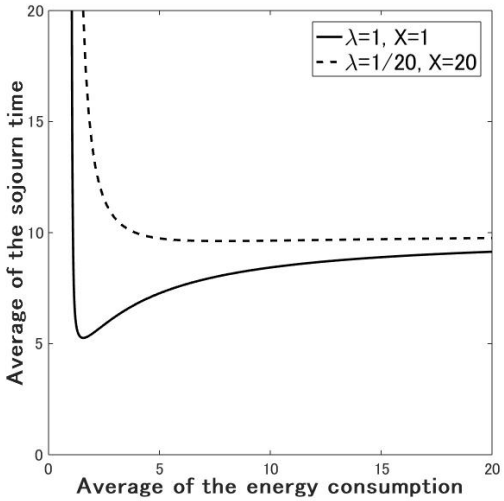

Figure 4.2: Energy performance of the M^X/M/1/SET-VARI queue

per unit time for each state of M^X/M/1/SET-VARI queue is assumed to be in Tables 4.1 [27, 46]. Note that the constants K_{service}, K_{set} and K_{idle} from Table 4.1 depend on the particular system. In this section, we assume that $K_{\text{service}} = K_{\text{set}} = 1$.

4.6.1 Energy performance of the M^X/M/1/SET-VARI queue

In this section, we observe the trade-off between the transmission delay (sojourn time) and the energy consumption in the M^X/M/1/SET-VARI queue. Using Theorem 4.2, the average sojourn time, denoted by $\mathsf{E}[W_{\text{v}}]$, and the average energy consumption, denoted by $\mathsf{E}[P_{\text{v}}]$, can be expressed as follows.

$$\mathsf{E}[W_{\text{v}}] = \frac{\lambda+\alpha}{\alpha^2}\pi_{0,0} + \frac{1}{\mu},$$
$$\mathsf{E}[P_{\text{v}}] = K_{\text{set}}\mu^2 \cdot \frac{\lambda}{\alpha}\pi_{0,0} + K_{\text{service}}\mu^2 \cdot \left\{ \frac{1}{2}\frac{\lambda}{\mu}\mathsf{E}[X(X+1)] + \frac{\lambda^2}{\mu}\mathsf{E}[X]^2\mathsf{E}[W_{\text{v}}] \right\},$$

where $\pi_{0,0}$ is given by (4.22).

We assume that $\alpha = 1$, $\mu \in (0, 3]$ and that λ and X change while keeping that the mean arrival rate of customers $\lambda\mathsf{E}[X]$ is equal to one. Figure 4.2 shows the relation between the average energy consumption (x-axis) and the average sojourn time (y-axis) of the M^X/M/1/SET-VARI queue. Figure 4.2 presents the results whose batch-size distributions are CONST(1) and CONST(20). Note that the case with CONST(1) implies customers arrive one by one. In the case with CONST(1), it can be seen that the average sojourn time can be reduced as the energy consumption increases. However, in the case with CONST(20), the average sojourn time is not necessarily reduced as the energy con-

state		energy consumption
service	$(1, k)\ k \geq 1$	$K_{\text{service}} \times \mu^2$
setup	$(0, k)\ k \geq 1$	$K_{\text{set}} \times \mu^2$
idle	$(0, 0)$	0

Table 4.2: Energy consumption per unit time of the M^X/M/1/SET-FIX queue

sumption increases. From these observation, the trade-off between the sojourn time and the energy consumption is not necessarily established.

4.6.2 Efficiency of the variable service speed

To observe the efficiency of the variable service speed of the M^X/M/1/SET-VARI queue, we compare with a queueing model such that the service speed is fixed. For the comparison, we consider the $M^X/M/1$/SET-FIX queue, which is the single-server queue such that the service speed is fixed and that the other settings are kept the same as the M^X/M/1/SET-VARI queue. In this subsection, the M^X/M/1/SET-VARI queue is referred to as *the variable speed queue*, and the $M^X/M/1$/SET-FIX queue is referred to as *the fixed speed queue*. The energy consumption for each state of the fixed speed queue is assumed to be in Table 4.2. Note that the constants K_{service} and K_{set} from Table 4.2 depend on the particular system. In this subsection, we assume that $K_{\text{service}} = K_{\text{set}} = 1$. Thus, the mean sojourn time of the fixed speed queue, denoted by $\mathsf{E}[W_{\text{f}}]$, and the mean energy consumption of the fixed speed queue, denoted by $\mathsf{E}[P_{\text{f}}]$, can be expressed as follows [4].

$$\mathsf{E}[W_{\text{f}}] = \frac{1}{\alpha} + \frac{1 + \mathsf{E}[X^2]/\mathsf{E}[X]}{2(\mu - \lambda\mathsf{E}[X])}$$
$$\mathsf{E}[P_{\text{f}}] = K_{\text{set}}\mu^2 \cdot \frac{\lambda}{\lambda + \alpha}\left(1 - \frac{\lambda\mathsf{E}[X]}{\mu}\right) + K_{\text{service}}\mu^2 \cdot \frac{\lambda\mathsf{E}[X]}{\mu}.$$

Under the same energy consumption, we compare the average sojourn times of the variable and fixed speed queues; that is, we compare $\mathsf{E}[W_{\text{v}}]$ with $\mathsf{E}[W_{\text{f}}]$ under the condition that $\mathsf{E}[P_{\text{v}}] = \mathsf{E}[P_{\text{f}}]$. The procedure of numerical experiments is as follows. First, fixing the value of $\mu \in (0, 3]$, we compute the average energy consumption per unit time of the variable speed queue, denoted by $\mathsf{E}[P_{\text{v}}]$, and the average sojourn time of the variable speed queue $\mathsf{E}[W_{\text{v}}]$. Let $\mu_{\text{f}}(A)$ denote the unique service rate which realizes the average energy consumption A in the fixed queue. Next, we compute $\mu_{\text{f}}(\mathsf{E}[P_{\text{v}}])$ by

$$\mu_{\text{f}}(\mathsf{E}[P_{\text{v}}]) = \frac{1}{\lambda\mathsf{E}[X]}\left\{\frac{\lambda + \alpha}{\alpha}\mathsf{E}[P_{\text{v}}] - \frac{\lambda}{\alpha}\right\}.$$

As a result, the energy consumption for both models is kept the same. Finally, we compute the average sojourn time of the fixed speed queue under the assumption that $\mu = \mu_{\text{f}}(\mathsf{E}[P_{\text{v}}])$.

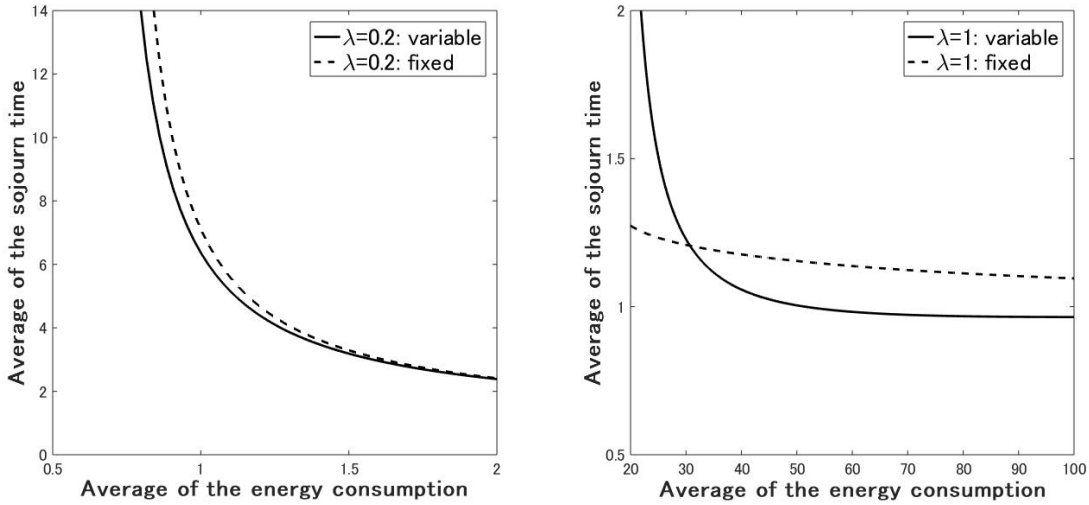

Figure 4.3: Efficiency of the variable service speed

state		energy consumption
service	$(1, k)\ k \geq 1$	$K_{\text{service}} \times \mu^2$
idle	$(0, 0)$	$K_{\text{idle}} \times \mu^2$

Table 4.3: Energy consumption per unit time of the M^X/M/1-VARI queue

In this numerical experiment, we compute the average sojourn time from the average number of customers in the system using Little's formula [36].

We assume that X following Binom$(9, 1/6)$ and $\alpha = 1$. The left-hand side of Figures 4.3 presents the comparison with $\lambda = 0.2$. We can observe that the variable speed queue can realize a shorter average sojourn time than the fixed speed queue when they consume the same energy. The right-hand side of Figures 4.3 presents the comparison with $\lambda = 1$, which is the situation that a relatively many batches arrive in comparison with the case of the left-hand side. We can observe that the variable speed queue does not necessary have better performance than the fixed speed queue. When energy consumption can be large to some extent, the variable speed queue can reduce the average sojourn time rather than the fixed speed queue. On the other hand, when we want to reduce energy consumption as much as possible, the fixed speed queue can reduce the average sojourn time rather than the variable speed queue.

4.6.3 Efficiency of the on-off policy

To observe the efficiency of the on-off policy of the M^X/M/1/SET-VARI queue, we compare with a queueing model without the on-off policy. For the comparison, we consider

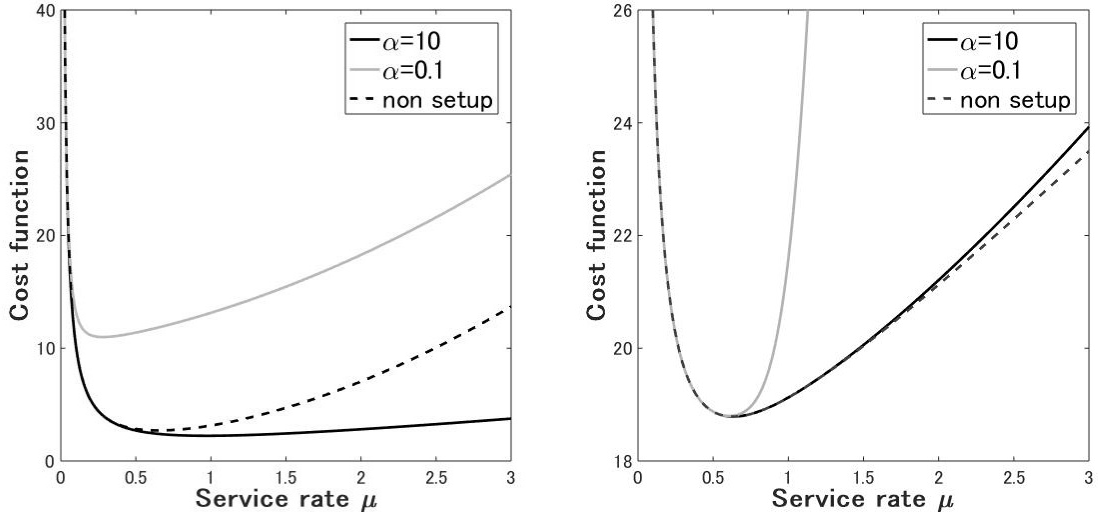

Figure 4.4: Efficiency of the on-off policy

the M^X/M/1-VARI queue, which is the single-server queue such that the server is not turned off when the system become empty and that the other settings are kept the same as the M^X/M/1/SET-VARI queue. In this subsection, the M^X/M/1/SET-VARI queue is referred to as *the on-off queue*, and the M^X/M/1-VARI queue is referred to as *the always ON queue*. The queue length process of the always ON queue is equivalent to that of the M^X/M/∞ queue. The energy consumption for each state of the always on queue is assumed to be in Table 4.3. Note that the constants $K_{\rm service}$ and $K_{\rm set}$ from Table 4.3 depend on the particular system. In this section, we assume that $K_{\rm service} = K_{\rm set} = 1$. Thus, the mean sojourn time of the always ON queue, denoted by $\mathsf{E}[W_{\rm on}]$, and the mean energy consumption of the always ON queue, denoted by $\mathsf{E}[P_{\rm on}]$, can be expressed as follows.

$$\mathsf{E}[W_{\rm on}] = \frac{1}{\mu},$$
$$\mathsf{E}[P_{\rm on}] = K_{\rm idle}\mu^2 \cdot \mathrm{e}^{-\frac{\lambda}{\mu}\mathsf{E}[X]} + K_{\rm service}\mu^2 \cdot \left\{ \frac{1}{2}\frac{\lambda}{\mu}\mathsf{E}[X(X+1)] + \left(\frac{\lambda}{\mu}\mathsf{E}[X]\right)^2 \right\}.$$

We now consider the performance metric z:

$$z = (\text{average sojourn time}) + \beta \cdot (\text{average energy consumption}), \tag{4.39}$$

where $\beta > 0$ is the constant which controls the ratio between the sojourn time and the energy consumption. We call the performance metric defined by (4.39) the cost function.

In Figures 4.4, assuming that X following Binom$(9, 1/6)$ and $\mu \in (0, 3]$, we show the values of the cost function (4.39) of the on-off queue with $\alpha = 10$ (shorter setup time), the on-off queue with $\alpha = 0.1$ (longer setup time) and the always ON queue. The left-hand side of Figures 4.4 presents the comparison with $\lambda = 0.1$ and $\beta = 2$. We can observe

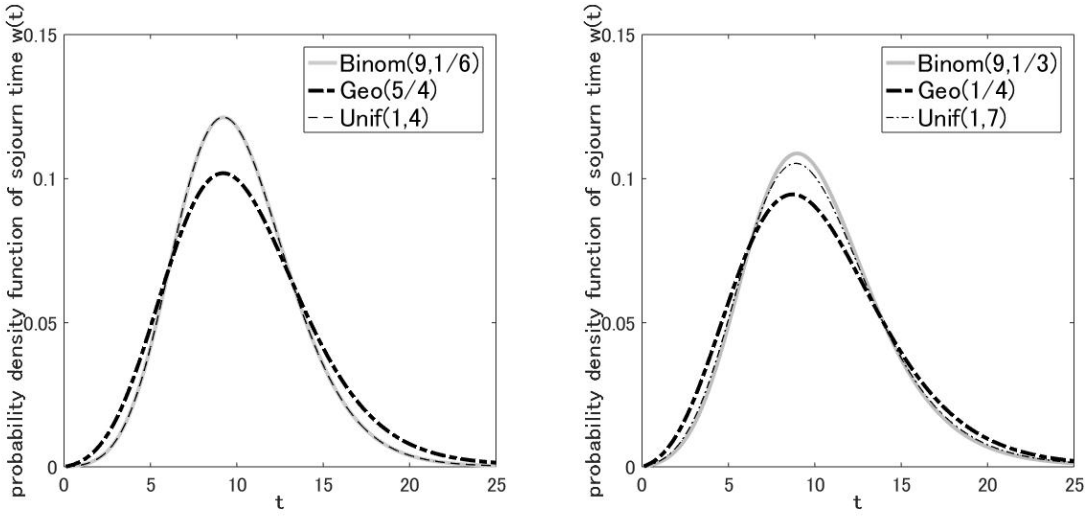

Figure 4.5: Probability density function of the stationary sojourn time

that the on-off queue with $\alpha = 10$ is better performance than the always ON queue. On the other hand, the on-off queue with $\alpha = 0.1$ does not have better performance than the always ON queue. This observation can be explained as follows. In the case of the shorter setup time ($\alpha = 10$), the on-off queue saves the energy due to the on-off policy when the system is empty and the average sojourn time is not affected much by the setup time. But, in the case of the longer setup time ($\alpha = 0.1$), the average sojourn time is heavily influenced by the setup time. The increase in the average sojourn time has bigger impact than the decrease in the energy consumption.

The right-hand side of Figures 4.4 presents the comparison with $\lambda = 5$ and $\beta = 0.1$, which is the situation that a relatively many batches arrive in comparison with the case of the left-hand side. We can observe that the always ON queue has better performance than the on-off queue in both cases: the short setup time ($\alpha = 10$) and the long setup time ($\alpha = 0.1$). This observation means that even if the setup times are short, the server should not be turned off under a heavy traffic situation. However, the minimum values of the cost functions of three queues are almost the same, which implies that the performance of the three queues are almost the same when running the system with the optimal μ.

4.6.4 Sojourn time distribution

In this section, we show the probability density function of the stationary sojourn time distribution of the M^X/M/1/SET-VARI queue by numerically inverting the Laplace-Stieltjes transform. In what follows, we assume that $\alpha = 0.1$, $\mu = 0.1$ and $\lambda \mathsf{E}[X] = 1$. The LST of the sojourn time distribution is given in Theorem 4.3 but it is in series form involving infinite dimensional matrices. Therefore, its approximation is necessary for numerical calculation. We present the procedure to compute the LST of the sojourn time distribution,

$W^*(s)$. First, we truncate the infinite vectors.

We present the procedure to compute the LST of the sojourn time distribution, $W^*(s)$. First, we truncate the infinite vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ to the vectors of their first (N^*+1) elements where the constant N^* is determined by

$$N^* = \inf\left\{n \in \mathbb{N};\, 1 - \sum_{j=0}^{n} \pi_{0,j} - \sum_{j=1}^{n} \pi_{1,j} < 10^{-4}\right\}.$$

This is equivalent to disregarding the states with more than N^* customers in the system whose probability is 10^{-4}. We compute $\pi_{0,0}$ by (4.22) and $\pi_{i,j}$, $i = 0, 1$, $j = 1, \ldots, N^*$ by (4.10) and (4.11). In addition, we truncate the infinite matrices appearing in $W^*(s)$ to their $N^* \times N^*$ north-west corner matrices. We compute each element of the infinite matrices by (4.26), (4.27), (4.33), (4.34), (4.36) and (4.37).

Next, we present the procedure to compute the value of the sojourn time distribution for $t \in (0, T/2]$ by numerically inverting the Laplace-Stieltjes transform [22] for fixed $T > 0$. The function $w(t)$ are defined as follows.

$$\begin{aligned} w(t) = {} & \frac{\mathrm{e}^{\frac{6}{T}t}}{T} \mathrm{Re}\left\{W^*\left(\frac{6}{T}\right)\right\} + \frac{2\mathrm{e}^{\frac{6}{T}t}}{T} \sum_{k=1}^{K} \mathrm{Re}\left\{W^*\left(\frac{6}{T} + \mathrm{i}k\frac{2\pi}{T}\right)\right\} \cos\left(k\frac{2\pi}{T}t\right) \\ & - \frac{2\mathrm{e}^{\frac{6}{T}t}}{T} \sum_{k=1}^{K} \mathrm{Im}\left\{W^*\left(\frac{6}{T} + \mathrm{i}k\frac{2\pi}{T}\right)\right\} \sin\left(k\frac{2\pi}{T}t\right). \end{aligned}$$

In our numerical experiments, we use the value of $w(t)$ as the sojourn time distribution. In this section, we set $K = 500$ and $T = 500$.

In Figures 4.5, we investigate the impact of the batch size distribution on the sojourn time distribution. The left-hand side of Figures 4.5 presents the sojourn time distribution for $\lambda = 0.4$ and $\mathsf{E}[X] = 2.5$, while the right-hand side of Figures 4.5 shows that for $\lambda = 0.25$ and $\mathsf{E}[X] = 4$ and $\mu = 0.1$. Note that in both figures $\lambda\mathsf{E}[X] = 1$. We observe that the curves of Binom$(9, 1/6)$ and Uni$\{1, 4\}$ almost coincide. The values of second, third and fourth moments are 7.5, 25.8 and 99.2 in Binom$(9, 1/6)$, and 7.5, 25.0 and 113.5 in Uni$\{1, 4\}$. On the other hand, the values of second, third and fourth moments are 10.0, 58.8 and 480.0 in Geo$(1/2.5)$. This suggests that high order moments (roughly fourth or higher) have less influence in the sojourn time distribution. Compared with the left-hand side of Figures 4.5, the curves of binomial distribution, uniform distribution and geometric distribution are different in the right -hand side of Figures 4.5. The second moments are 18.0 in Binom$(9, 1/3)$, 20.0 in Uni$\{1, 4\}$ and 38.0 in Geo$(1/2.5)$. This suggests that the second moment of the batch size has a significant impact on the sojourn time distribution.

4.6.5 Variance of the sojourn time

In this section, using Theorem 4.3, we calculate the variance of the stationary sojourn time distribution of the M^X/M/1/SET-VARI queue. We assume that $\alpha = 0.1$, $\mu \in (0, 0.5]$ and

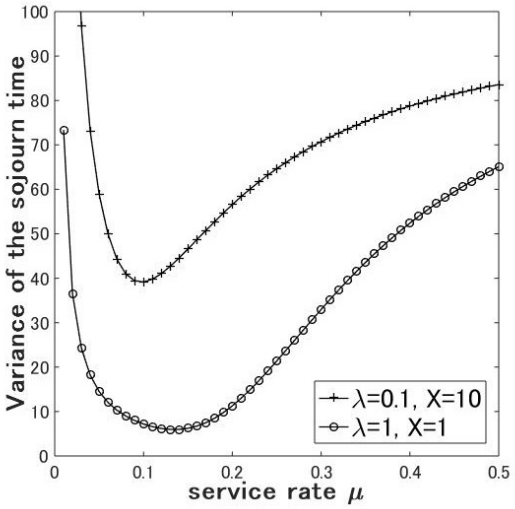

Figure 4.6: Variance of the sojourn time

that λ and X change while keeping that the mean arrival rate of customers $\lambda \mathsf{E}[X]$ is equal to one. Figure 4.6 presents the results whose batch-size distributions are CONST(10) and CONST(1). Note that the case with CONST(1) implies customers arrive one by one.

We can observe that the case with CONST(10) is larger than that with CONST(1). This observation can be explained as follows. In the case with CONST(10), the variance increases due to the fact that customers in a batch at the same time but receive service at different times. We can also observe that when μ is large, the variance increases as μ increases. This observation can be explained as follows. The difference in service speed due to the difference in the number of customers in the system increases as μ increases.

4.7 Conclusion

In this chapter, we studied the M^X/M/1/SET-VARI queue. We derived the PGF of the stationary queue length distribution in an integral form. Furthermore, we derived the LST of the stationary sojourn time distribution, which was obtained in series form involving infinite-dimensional matrices. Through numerical experiments, we observed some insights into the energy performance of the M^X/M/1/SET-VARI queue.

In future works, we would like to consider the queueing models such that the service speed changes according to various rules not only the case that in proportion to the queue length. As a more realistic model, we would like to consider the case that there is an upper limit on the change in the service speed. We are interested in the case that the service speed changes by thresholds and the case that the service speed is controlled by the workload in the system. We would like to know the stochastic properties and propose a better control strategy of service speeds by using analytically and numerically results.

Chapter 5

Conclusion

5.1 Summary

This thesis studied infinite-server queues and related models. In Chapter 2, we analyzed the stability for batch arrival infinite-server queues. First, we showed that the stability condition of BMAP/M/∞ queues is that the logarithmic moment of batch-size distribution is finite. We extended this result to the multiclass BMAP/M/∞ queues. Next, we showed the stability condition of GI^X/GI/∞ queues is that the maximum service time in the batch has a finite mean. Furthermore, we presented a tractable sufficient condition for the stability of GI^X/GI/∞ queues. We also proved that the stability condition of GI^X/GI/∞ queues is that the logarithmic moment of the batch-size distribution is finite, provided that the service time distribution has an exponential tail.

Chapter 3 considered the Markov-modulated M^X/M/∞ queue with binomial catastrophes, which is a batch arrival infinite-server queue such that customers may or may not leave the system without completing service due to accidents. We analyzed the scaling model of this model under a heavy traffic regime because it is difficult to exactly analyze Markov-modulated queues. We then established the central limit theorem (CLT) for the stationary queue length distribution; that is, the centered and normalized stationary queue length distribution converges in distribution to a normal distribution. Using the CLT, we obtained the approximation of the stationary queue length distribution with large arrival rates. We presented some numerical results to confirm the accuracy of this approximation.

Chapter 4 studied a batch arrival single-server queues with variable service speed and the on-off policy. In particular, we assumed that the service speed changes in proportion to the queue length. It should be noted that the queue length process of this single-server queue is identical to that of an infinite-server queue. We first presented the stability condition for this queueing model. We derived the probability generating function of the stationary queue length in an integral form. Furthermore, we derived the Laplace-Stieltjes transform of the stationary sojourn time distribution. Through numerical experiments, we observed some insights into the sojourn time and the energy performance.

5.2 Directions of future works

In Chapter 2, we considered the stability condition for batch arrival infinite-server queues. It is very important to derive the stability condition in a form that makes it easy to check whether the condition is satisfied. Thus, we purpose to derive physically and interpretable stability conditions for more general batch arrival infinite-server queues. We derived the stability condition of $GI^X/GI/\infty$ queues, but derived by the physically and interpretable form only when the service time distributions have exponential tails. We then would like to derive a physically and interpretable stability condition of $GI^X/GI/\infty$ queues without additional conditions. We predict that the condition of Corollary 4.1 is not only the sufficient condition but also the necessary condition.

$GI^X/GI/\infty$ queues are very general batch arrival infinite-server queues and used in many researches. However, there are restrictions that independence of inter-arrival times, batch sizes, and service times. Thus, we would like to derive the stability condition for batch arrival infinite-server queues such that there exist correlations between inter-arrival times, batch sizes, and service times.

In Chapter 3, we studied an Markov-modulated infinite-server queues with batch arrivals and binomial catastrophe, and shown the central limit theorem for the stationary queue length. We would like to consider the behavior of our model without the assumption that the second moment of the batch size is finite. We also would like to show the central limit theorem under other heavy traffic regime because the heavy traffic regime considered in this thesis is quite restrictive. Especially, we are interested in the regime such that the transition rate of the background process is scaled by $N^{\alpha'}$, where α' is a newly introduced coefficient in addition to α. Knowing the behavior of a scaling model under various regimes will lead to understanding the stochastic properties of the original (non-scaling) model. Furthermore, we are interested in properties of the queue length process of this model in the transient state. We predict that the central limit theorem of the stochastic process version for the queue length process hold; that is, a scaled queue length process with some transformation converges to an Ornstein-Uhlenbeck process.

In Chapter 4, we studied a batch arrival single server queues whose service speed of the server changes in proportional to the queue length. In future works, we want to consider the effect of each parameter on the energy performance either analytically or numerically. We would like to consider the queueing models such that the service speed changes according to various rules not only the case that in proportion to the queue length. As a more realistic model, we would like to consider the case that there is an upper limit on the change in the service speed. We are interested in the case that the service speed changes by thresholds. We would also like to consider the case that the service speed is controlled by the workload in the system. Furthermore, we would like to propose a better control strategy of service speeds by using analytically and numerically results.

Appendix

A Supplement proof for Theorem 2.2

This appendix devotes to show the pathwise ordered relation (2.20). Let T_n, $n \in \mathbb{N}$, denote the nth arrival time of batches from MBMAP $\{\boldsymbol{D}(0), \boldsymbol{D}_\nu(k); \nu \in \mathbb{K}, k \in \mathbb{N}\}$, where

$$0 < T_1 < T_2 < \cdots .$$

Let c_n and B_n, $n \in \mathbb{N}$, denote the class and batch size,respectively, of the batch arriving at time T_n. Furthermore, let $\{U_m; m \in \mathbb{N}\}$ denote a sequence of i.i.d. random variables with a uniform distribution on the interval $(0, 1)$. We then define S_m, $\overline{S}_m$ and $\underline{S}_m$, $m \in \mathbb{N}$, as random variables such that, for $A_{n-1} + 1 \le m \le A_n$ and $n \in \mathbb{N}$,

$$S_m = -\frac{1}{\mu_{c_n}} \log U_m, \tag{A.1}$$

$$\overline{S}_m = -\frac{1}{\mu_{\min}} \log U_m, \tag{A.2}$$

$$\underline{S}_m = -\frac{1}{\mu_{\max}} \log U_m, \tag{A.3}$$

where $A_0 = 0$ and $A_n = \sum_{k=1}^{n} B_k$ for $n \in \mathbb{N}$. It follows from (A.1)–(A.3) that

$$\mathsf{P}(S_m \le x) = 1 - \exp\{-\mu_{c_n} x\}, \qquad x \in \mathbb{R}_+, \tag{A.4}$$

$$\mathsf{P}(\overline{S}_m \le x) = 1 - \exp\{-\mu_{\min} x\}, \qquad x \in \mathbb{R}_+, \tag{A.5}$$

$$\mathsf{P}(\underline{S}_m \le x) = 1 - \exp\{-\mu_{\max} x\}, \qquad x \in \mathbb{R}_+. \tag{A.6}$$

In addition, since $\mu_{\min} \le \mu_{c_n} \le \mu_{\max}$, we have

$$\underline{S}_m \le S_m \le \overline{S}_m, \qquad m \in \mathbb{N}. \tag{A.7}$$

Based on (A.4)–(A.6), we assume that $\{S_m; A_{n-1} + 1 \le m \le A_n\}$, $\{\overline{S}_m; A_{n-1} + 1 \le m \le A_n\}$ and $\{\underline{S}_m; A_{n-1} + 1 \le m \le A_n\}$ are the service times of the customers in the nth batch arriving at the original MBMAP$_K$/M$_K$/∞ queue, Queues 1 and 2,

respectively. We then fix $|\boldsymbol{L}(t)|$, $L^{(1)}(t)$ and $L^{(2)}(t)$, $t \geq 0$ such that

$$|\boldsymbol{L}(t)| = \sum_{n=1}^{\infty} \sum_{m=A_{n-1}+1}^{A_n} I(T_n \leq t < T_n + S_m), \tag{A.8}$$

$$L^{(1)}(t) = \sum_{n=1}^{\infty} \sum_{m=A_{n-1}+1}^{A_n} I(T_n \leq t < T_n + \overline{S}_m), \tag{A.9}$$

$$L^{(2)}(t) = \sum_{n=1}^{\infty} \sum_{m=A_{n-1}+1}^{A_n} I(T_n \leq t < T_n + \underline{S}_m). \tag{A.10}$$

It is easy to see that $\{|\boldsymbol{L}(t)|\}$, $\{L^{(1)}(t)\}$ and $\{L^{(2)}(t)\}$ can be considered the total queue length processes of the original MBMAP$_K$/M$_K$/∞ queue, Queues 1 and 2, respectively, which are fed by the common MBMAP. Furthermore, combining (A.7) with (A.8)–(A.10), we obtain the pathwise ordered relation (2.20) between $\{|\boldsymbol{L}(t)|\}$, $\{L^{(1)}(t)\}$ and $\{L^{(2)}(t)\}$.

B Proof for Lemma 3.2

For the simplicity of expressions, we define, for $i \in \mathbb{D}$, $k \in \mathbb{Z}_+$, and $z \in [0,1]$,

$$\widehat{\pi}_i^{(k,N)}(z) := \frac{\mathrm{d}^k}{\mathrm{d}z^k}\widehat{\pi}_i^{(N)}(z), \qquad \widehat{\pi}_{p,i}^{(k,N)}(z) := \frac{\mathrm{d}^k}{\mathrm{d}z^k}\left\{\widehat{\pi}_{p,i}^{(N)}(z) - \widehat{\pi}_i^{(N)}(z)\right\},$$

and

$$q_i(z) := \frac{1-\widehat{X}_i(z)}{1-z}, \qquad q_i^{(k)}(z) := \frac{\mathrm{d}^k}{\mathrm{d}z^k} q_i(z).$$

To complete the proof of Lemma 3.2, we show that

$$N^{-k}\widehat{\pi}_i^{(k,N)}(1) = o(N), \qquad \text{for any } i \in \mathbb{D} \text{ and } k = 1, 2. \tag{B.11}$$

Right-multiplying (3.8) by $(1-z)^{-1}N^{-1}\boldsymbol{e}$, we have

$$N^{-1}\sum_{i\in\mathbb{D}} \widehat{\pi}_i^{(1,N)}(z)\mu_i = \sum_{i\in\mathbb{D}} \widehat{\pi}_i^{(N)}(z)\lambda_i q_i^{(0)}(z) - \sum_{i\in\mathbb{D}} \frac{\widehat{\pi}_{p,i}^{(0,N)}(z)}{1-z}\gamma_i. \tag{B.12}$$

Differentiating both sides of (B.12) yields

$$\begin{aligned} N^{-1}\sum_{i\in\mathbb{D}} \widehat{\pi}_i^{(2,N)}(z)\mu_i = &\sum_{i\in\mathbb{D}} \lambda_i \left\{\widehat{\pi}_i^{(1,N)}(z)q_i^{(0)}(z) + \widehat{\pi}_i^{(N)}(z)q_i^{(1)}(z)\right\} \\ &- \sum_{i\in\mathbb{D}} \left\{\frac{\widehat{\pi}_{p,i}^{(1,N)}(z)}{1-z} + \frac{\widehat{\pi}_{p,i}^{(0,N)}(z)}{(1-z)^2}\right\}\gamma_i. \end{aligned} \tag{B.13}$$

Note here that, for any $i \in \mathbb{D}$, $k = 0, 1$, and $z \in [0,1]$,

$$\widehat{\pi}_{p,i}^{(k,N)}(z) \geq 0,$$

and

$$q_i^{(0)}(z) = \mathsf{E}\left[\sum_{\ell=0}^{X_i-1} z^\ell\right] \le \mathsf{E}\left[\sum_{\ell=0}^{X_i-1} 1\right] = \mathsf{E}[X_i],$$
$$q^{(1)}(z) = \mathsf{E}\left[\sum_{\ell=1}^{X_i-1} \ell z^{\ell-1}\right] \le \mathsf{E}\left[\sum_{\ell=1}^{X_i-1} X_i\right] \le \mathsf{E}[X_i^2].$$

Applying these inequalities to (B.12) and (B.13), respectively, we have

$$N^{-1}\sum_{i\in\mathbb{D}} \widehat{\pi}_i^{(1,N)}(z)\mu_i \le \sum_{i\in\mathbb{D}} \lambda_i \widehat{\pi}_i^{(N)}(z)\mathsf{E}[X_i], \tag{B.14}$$
$$N^{-1}\sum_{i\in\mathbb{D}} \widehat{\pi}_i^{(2,N)}(z)\mu_i \le \sum_{i\in\mathbb{D}} \lambda_i \left\{\widehat{\pi}_i^{(1,N)}(z)\mathsf{E}[X_i] + \widehat{\pi}_i^{(N)}(z)\mathsf{E}[X_i^2]\right\}. \tag{B.15}$$

Taking the limit as $z \uparrow 1$ in (B.14) and (B.15) and using $\mathsf{E}[X_i^2] < \infty$, we obtain (B.11).

C Supplement proof for Lemma 3.3

This appendix devotes to show (3.15). In this appendix, we note z as abbreviation for $z(\theta)$. Let $f_i(N)$, $i \in \mathbb{D}$, denote the i-th element of the left hand side of (3.15); that is,

$$f_i(N) = N\mathsf{E}[\{N^{-1}\overline{p}_i + z(\theta)(1 - N^{-1}\overline{p}_i)\}^{L^{(N)}}\delta_i] - N\mathsf{E}[z(\theta)^{L^{(N)}}\delta_i],$$

where $\delta_i = I(J^{(N)} = i)$. Rearranging the above, we obtain, for any $i \in \mathbb{D}$,

$$f_i(N) = N\overline{p}_i\mathsf{E}\left[L^{(N)}N^{-1}\sum_{k=1}^{L^{(N)}} \frac{A_{k-1}}{k}\left\{z(\theta)^{L^{(N)}-k} - z(\theta)^{L^{(N)}}\right\}\delta_i\right], \tag{C.16}$$

where the random variable A_k is given by

$$A_k = \binom{L^{(N)}-1}{k}\{N^{-1}\overline{p}_i\}^k\{1 - N^{-1}\overline{p}_i\}^{L^{(N)}-1-k}.$$

We estimate $f_i(N)$. We have, for $1 \le k \le n$ and $x \in [0,1]$,

$$\begin{aligned} x^{n-k} - x^n &= (x^{-1}-1)x^n k\sum_{\ell=1}^{k-1} 1 + (x^{-1}-1)x^n\sum_{\ell=1}^{k-1}(x^{-\ell}-1) \\ &= (x^{-1}-1)x^{L^{(N)}}k + (x^{-1}-1)^2\sum_{\ell=1}^{k-1}\sum_{h=0}^{\ell-1} x^{n-h} \\ &\le (x^{-1}-1)x^n k + (x^{-1}-1)^2 k(k-1). \end{aligned}$$

Applying this inequality to (C.16), we have

$$\begin{aligned} f_i(N) \le{}& N(z(\theta)^{-1}-1)\overline{p}_i \cdot \mathsf{E}\left[L^{(N)}N^{-1}z(\theta)^{L^{(N)}}\sum_{k=1}^{L^{(N)}} A_{k-1}\delta_i\right] \\ &+ N(z(\theta)^{-1}-1)^2\overline{p}_i \cdot \mathsf{E}\left[L^{(N)}N^{-1}\sum_{k=1}^{L^{(N)}}(k-1)A_{k-1}\delta_i\right], \end{aligned} \tag{C.17}$$

Note here that

$$\sum_{k=1}^{L^{(N)}} A_{k-1} = 1,$$
$$\sum_{k=1}^{L^{(N)}} kA_{k-1} = N^{-1}\overline{p}_i \cdot (L^{(N)} - 1) \leq N^{-1}\overline{p}_i \cdot L^{(N)}.$$

Combining these relations and (C.17) yields

$$\begin{aligned} f_i(N) &\leq N\overline{p}_i(z(\theta)^{-1} - 1)\mathsf{E}[L^{(N)}N^{-1}z(\theta)^{L^{(N)}}\delta_i] + N\overline{p}_i^2(z(\theta)^{-1} - 1)^2\mathsf{E}[(L^{(N)}N^{-1})^2\delta_i] \\ &= N\overline{p}_i(z(\theta)^{-1} - 1)\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}[z(\theta)^{L^{(N)}}\delta_i] + N\overline{p}_i^2(z(\theta)^{-1} - 1)^2\mathsf{E}[(L^{(N)}N^{-1})^2\delta_i]. \end{aligned}$$

Therefore, from Lemma 3.2 and (3.13), we obtain, for any $i \in \mathbb{D}$,

$$f_i(N) \leq -\overline{p}_i\theta\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}[z(\theta)^{L^{(N)}}I(J^{(N)} = i)] + o(N). \tag{C.18}$$

On the other hand, applying $z^{-k} - 1 \geq k\theta N^{-1}$ to (C.16), we obtain

$$\begin{aligned} f_i(N) &\geq N\overline{p}_i\mathsf{E}\left[L^{(N)}N^{-1}\sum_{k=1}^{L^{(N)}} A_{k-1}\frac{1}{k} \cdot z(\theta)^{L^{(N)}} \cdot k\theta N^{-1}\right] \\ &= N\overline{p}_i\mathsf{E}[L^{(N)}N^{-1}z(\theta)^{L^{(N)}}] \\ &= -\overline{p}_i\theta\frac{\mathrm{d}}{\mathrm{d}\theta}\mathsf{E}[z(\theta)^{L^{(N)}}\delta_i]. \end{aligned} \tag{C.19}$$

Consequently, (3.15) follows from (C.18) and (C.19).

D Proof for Lemma 3.4

For $k = 1, 2$ and $i \in \mathbb{D}$, we define $B_k(i)$ as

$$B_k(i) = \mathsf{E}[(L^{(N)}N^{-1})^k\mathrm{e}^{\mathrm{i}L^{(N)}N^{-1}\theta}I(J^{(N)} = i)] - \rho^k\mathsf{E}[\mathrm{e}^{\mathrm{i}L^{(N)}N^{-1}\theta}I(J^{(N)} = i)].$$

We then have, for $k = 1, 2$ and $i \in \mathbb{D}$,

$$\begin{aligned} \left|B_k(i)\right| &\leq \mathsf{E}[|(L^{(N)}N^{-1})^k - \rho^k||\mathrm{e}^{\mathrm{i}L^{(N)}N^{-1}\theta}|I(J^{(N)} = i)] \\ &\leq \mathsf{E}[|(L^{(N)}N^{-1})^k I(J^{(N)} = i) - \rho^k I(J^{(N)} = i)|] \\ &\leq \mathsf{E}[|(L^{(N)}N^{-1})^k - \rho^k|]. \end{aligned} \tag{D.20}$$

From Lemma 3.3 and [49, Corollary 2], $(L^{(N)}N^{-1})^k$ converges in probability to ρ^k for $k = 1, 2$. In addition, it follows from Lemma 3.2 that $\{\mathsf{E}[(L^{(N)}N^{-1})^k]\}_{N\geq 1}$ is uniformly integrable for $k = 1, 2$. Thus, for $k = 1, 2$, $(L^{(N)}N^{-1})^k$ converges in mean to ρ^k; that is,

$$\lim_{N\to\infty}\mathsf{E}[|(L^{(N)}N^{-1})^k I(J^{(N)} = i) - \rho^k|] = 0. \tag{D.21}$$

Combining (D.20) and (D.21) yields

$$\lim_{N\to\infty}\left|B_k(i)\right| = 0, \quad \text{for } k = 1, 2 \text{ and } i \in \mathbb{D}.$$

E Supplement proof for Theorem 3.1

This appendix devotes to show (3.25). We define $g_i(N)$ and $h_i(N)$, $i \in \mathbb{D}$, as follows.

$$g_i(N) = \gamma_i \left\{ \widehat{\pi}_i^{(N)}(z(\theta)) - \widehat{\pi}_{p,i}^{(N)}(z(\theta)) \right\},$$
$$h_i(N) = \mathrm{i}N^{\beta/2}\gamma_i \overline{p}_i \theta \frac{\mathrm{d}}{\mathrm{d}z}\widehat{\pi}_i^{(N)}(z(\theta)) + N^{-1+\beta}\frac{1}{2}\{\rho\gamma_i\overline{p}_i + \rho^2\gamma_i\overline{p}_i^2\}\theta^2\widehat{\pi}_i^{(N)}(z(\theta)).$$

To complete the proof of (3.25), we show the following relation.

$$\left| \mathrm{e}^{-\mathrm{i}N^{\beta/2}\rho\theta} g_i(N) - \mathrm{e}^{-\mathrm{i}N^{\beta/2}\rho\theta} h_i(N) \right| = o(1), \tag{E.22}$$

We define $\delta_i = I(J^{(N)} = i)$. We have

$$\begin{aligned}
h_i(N) &= \mathrm{i}N^{\beta/2}\gamma_i\overline{p}_i\theta \mathsf{E}\left[L^{(N)} N^{-1} z(\theta)^{L^{(N)}} \delta_i \right] \\
&\quad + N^{-1+\beta}\frac{1}{2}\left\{\rho\gamma_i\overline{p}_i + \rho^2\gamma_i\overline{p}_i^2\right\}\theta^2 \mathsf{E}\left[z(\theta)^{L^{(N)}}\delta_i \right] \\
&= \mathrm{i}N^{\beta/2}\gamma_i\theta \mathsf{E}\left[\overline{p}_i L^{(N)} N^{-1} \cdot z(\theta)^{L^{(N)}}\delta_i \right] \\
&\quad + N^{-1+\beta}\frac{1}{2}\gamma_i\theta^2 \mathsf{E}\left[\left\{ \overline{p}_i L^{(N)} N^{-1} + \left(\overline{p}_i L^{(N)} N^{-1}\right)^2 \right\} \cdot z(\theta)^{L^{(N)}}\delta_i \right] + o(1),
\end{aligned} \tag{E.23}$$

where the last equation follows from Lemma 3.4. In addition, we also have

$$g_i(N) = N\gamma_i \mathsf{E}\Big[z(\theta)^{L^{(N)}} \sum_{k=1}^{L^{(N)}} (1 - z(\theta)^{-k}) C_k \delta_i \Big], \tag{E.24}$$

where $\delta_i := 1_{\{J^{(N)}=i\}}$ and C_k is given by

$$C_k = \binom{L^{(N)}}{k} \{N^{-1}\overline{p}_i\}^k \{1 - N^{-1}\overline{p}_i\}^{L^{(N)}-k}.$$

We define $D_k(\theta)$ as

$$D_k(\theta) = 1 - z(\theta)^{-k} - \mathrm{i}N^{-1+\beta/2}k\theta - N^{-2+\beta}\frac{k^2}{2}\theta^2.$$

Using $D_k(\theta)$, (E.24) can be rewritten as follows.

$$\begin{aligned}
g_i(N) &= N\gamma_i \mathsf{E}\left[z(\theta)^{L^{(N)}} \sum_{k=1}^{L^{(N)}} D_k(\theta) C_k \delta_i \right] + N\gamma_i \mathsf{E}\left[z(\theta)^{L^{(N)}} \sum_{k=1}^{L^{(N)}} k C_k \delta_i \right] \\
&\quad + N\gamma_i \mathsf{E}\left[z(\theta)^{L^{(N)}} \sum_{k=1}^{L^{(N)}} k^2 C_k \delta_i \right],
\end{aligned} \tag{E.25}$$

Note here that

$$\sum_{k=1}^{L^{(N)}} k C_k = \overline{p}_i L^{(N)} N^{-1},$$
$$\sum_{k=1}^{L^{(N)}} k^2 C_k = \overline{p}_i L^{(N)} N^{-1} + (\overline{p}_i L^{(N)} N^{-1})^2 + o(1).$$

Applying these relations to (E.25) yields

$$g_i(N) = N\gamma_i \mathsf{E}\left[z(\theta)^{L^{(N)}} \sum_{k=1}^{L^{(N)}} D_k(\theta) C_k \delta_i \right] + N\gamma_i \mathsf{E}\left[\overline{p}_i L^{(N)} N^{-1} \cdot z(\theta)^{L^{(N)}} \delta_i \right] + N\gamma_i \mathsf{E}\left[\left\{ \overline{p}_i L^{(N)} N^{-1} + \left(\overline{p}_i L^{(N)} N^{-1} \right)^2 \right\} \cdot z(\theta)^{L^{(N)}} \delta_i \right] + o(1), \quad \text{(E.26)}$$

Combining (E.23) and (E.26), we obtain

$$g_i(N) - h_i(N) = \mathsf{E}\left[z(\theta)^{L^{(N)}} \sum_{k=1}^{L^{(N)}} D_k(\theta) C_k \delta_i \right] + o(1). \quad \text{(E.27)}$$

Using the triangle inequality and $|z(\theta)| = |\mathrm{e}^{\mathrm{i}N^{\beta/2}\rho\theta}| = 1$, it follows from (E.27) that

$$\begin{aligned} |g_i(N) - h_i(N)| &\leq N\gamma_i \mathsf{E}\left[\left| z(\theta)^{L^{(N)}} \right| \cdot \sum_{k=1}^{L^{(N)}} |D_k(\theta)| \, C_k \delta_i \right] + o(1) \\ &= N\gamma_i \mathsf{E}\left[\sum_{k=1}^{L^{(N)}} |D_k(\theta)| \, C_k \delta_i \right] + o(1). \end{aligned} \quad \text{(E.28)}$$

Applying Lemma 3.2 and $|\mathrm{e}^{\mathrm{i}N^{\beta/2}\rho\theta}| = 1$ to the above inequality yields

$$|g_i(N) - h_i(N)| \leq N\gamma_i \mathsf{E}\left[\sum_{k=1}^{L^{(N)}} |D_k(\theta)| \, C_k \delta_i \right] + o(1). \quad \text{(E.29)}$$

Note here that, using the second order Maclaurin expansion of $z^{-k}(\theta)$, we have

$$z^{-k}(\theta) = 1 - \mathrm{i}N^{-1+\beta/2} k\theta - N^{-2+\beta} \frac{k^2}{2} \theta^2 + o(N^{-1}),$$

which leads to

$$D_k(\theta) = o(N^{-1}). \quad \text{(E.30)}$$

Note also that

$$\sum_{k=1}^{L^{(N)}} C_k \leq \sum_{k=0}^{L^{(N)}} C_k = 1. \quad \text{(E.31)}$$

Applying (E.30) and (E.31) to (E.29), we complete the proof of (E.22).

Bibliography

[1] I. Adan and B. D'Auria. Sojourn time in a single-server queue with threshold service rate control. *SIAM Journal on Applied Mathematics*, 76(1):197–216, 2016.

[2] I. F. Akyildiz and X. Wang. A survey on wireless mesh networks. *IEEE Communications magazine*, 43(9):23–30, 2005.

[3] S. Asmussen and G. Koole. Marked point processes as limits of Markovian arrival streams. *Journal of Applied Probability*, 30(2):365–372, 1993.

[4] Y. Baba. The M^X/M/1 queue with multiple working vacation. *American Journal of Operations Research*, 2(2):217–224, 2012.

[5] L. A. Barroso and U. Hölzle. The case for energy-proportional computing. *Computer*, 40(12):33–37, 2007.

[6] M. Baykal-Gursoy and W. Xiao. Stochastic decomposition in M/M/∞ queues with Markov modulated service rates. *Queueing Systems*, 48(1–2):75–88, 2004.

[7] O. Berman and E. Kim. Stochastic models for inventory management at service facilities. *Stochastic Models*, 15(4):695–718, 1999.

[8] J. Blom and M. Mandjes. A large-deviations analysis of Markov-modulated infinite-server queues. *Operations Research Letters*, 41(3):220–225, 2013.

[9] J. Blom, K. De Turck, and M. Mandjes. Analysis of Markov-modulated infinite-server queues in the central-limit regime. *Probability in the Engineering and Informational Sciences*, 29(3):433–459, 2015.

[10] A. Brandt and H. Sulanke. On the GI/M/∞ queue with batch arrivals of constant size. *Queueing Systems*, 2(2):187–200, 1987.

[11] P. Brémaud. *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, Volume 31. Springer Science & Business Media, 1999.

[12] L. Breuer. *From Markov jump processes to spatial queues*. Springer Science & Business Media, 2003.

[13] P. J. Brockwell, J. Gani, and S. I. Resnick. Birth, immigration and catastrophe processes. *Advances in Applied Probability*, 14(4):709–731, 1982.

[14] U. Chatterjee and S.P. Mukherjee. On the non-homogeneous service system $M^X/G/\infty$. *European Journal of Operational Research*, 38(2):202–207, 1989.

[15] B. D. Choi and K. K. Park. The $M^k/M/\infty$ queue with heterogeneous customers in a batch. *Journal of Applied Probability*, 29(2):477–481, 1992.

[16] Gautam Choudhury. An $M^X/G/1$ queueing system with a setup period and a vacation period. *Queueing Systems*, 36(1-3):23–38, 2000.

[17] T. D. Cong. On the $M^X/G/\infty$ queue by heterogeneous customers in a batch. *Journal of Applied probability*, 31(1):280–286, 1994.

[18] P. Coolen-Schrijner and E. A. van Doorn. The deviation matrix of a continuous-time Markov chain. *Probability in the Engineering and Informational Sciences*, 16(3):351–366, 2002.

[19] A. Di Crescenzo, V. Giorno, A. G. Nobile, and L. M. Ricciardi. On the M/M/1 queue with catastrophes and its continuous approximation. *Queueing Systems*, 43(4):329–347, 2003.

[20] A. Daw and J. Pender. On the distributions of infinite server queues with batch arrivals. *Queueing Systems*, 91(3-4):367–401, 2019.

[21] L. de Haan and U. Stadtmüller. Dominated variation and related concepts and tauberian theorems for laplace transforms. *Journal of mathematical analysis and applications*, 108(2):344–365, 1985.

[22] F. Durbin. Numerical inversion of laplace transforms: an efficient improvement to dubner and abate's method. *The Computer Journal*, 17(4):371–376, 1974.

[23] L. Elefteriadou, R. P. Roess, and W. R. McShane. Probabilistic nature of breakdown at freeway merge junctions. *Transportation Research Record*, 1484:80–89, 1995.

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

[25] A. K. Erlang. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. *Post Office Electrical Engineer's Journal*, 10:189–197, 1917.

[26] G. I. Falin and J. G. C. Templeton. *Retrial Queues*. Chapman & Hall, London, 1997.

[27] A. Gandhi, M. Harchol-Balter, and I. Adan. Server farms with setup costs. *Performance Evaluation*, 67(11):1123–1138, 2010.

[28] D. F. Holman, M. L. Chaudhry, and B. R. K. Kashyap. On the number in the system $GI^X/M/\infty$. *Sankhyā: The Indian Journal of Statistics, Series A*, 44(2):294–297, 1982.

[29] D. F. Holman, M. L. Chaudhry, and B. R. K. Kashyap. On the service system $M^X/G/\infty$. *European Journal of Operational Research*, 13(2):142–145, 1983.

[30] S. Hur and S. J. Paik. The effect of different arrival rates on the N-policy of M/G/1 with server setup. *Applied Mathematical Modelling*, 23(4):289–299, 1999.

[31] J. L. W. V. Jensen. On the convex functions and inequalities between mean values. *Acta Mathematica*, 30(1):175–193, 1906.

[32] N. Kaplan. Limit theorems for a GI/G/∞ queue. *The Annals of Probability*, 3(5):780–789, 1975.

[33] S. Kapodistria, T. Phung-Duc, and J. Resing. Linear birth/immigration-death process with binomial catastrophes. *Probability in the Engineering and Informational Sciences*, 30(1):79–111, 2016.

[34] J. Keilson and A. Seidmann. M/G/∞ with batch arrivals. *Operations Research Letters*, 7(5):219–222, 1988.

[35] J. Keilson and A. Seidmann. M/G/∞ with batch arrivals. *Operations Research Letters*, 7(5):219–222, 1988.

[36] J. Keilson and L. D. Servi. A distributional form of Little's law. *Operations Research Letters*, 7(5):223–227, 1988.

[37] A. N. Kolmogorov and S. V. Fomin. *Elements of the theory of functions and functional analysis*. Courier Corporation, 1957.

[38] B. Krishna Kumar and D. Arivudainambi. Transient solution of an M/M/1 queue with catastrophes. *Computers & Mathematics with Applications*, 40(10-11):1233–1240, 2000.

[39] B. Krishna Kumar, A. Vijayakumar, and S. Sophia. Transient analysis for state-dependent queues with catastrophes. *Stochastic Analysis and Applications*, 26(6):1201–1217, 2008.

[40] O. J. Boxmaand I. A. Kurkova. The M/G/1 queue with two service speeds. *Advances in Applied Probability*, 33(2):520–540, 2001.

[41] G. Latouche and V. Ramaswami. *Introduction to matrix analytic methods in stochastic modeling*, Volume 5. SIAM, 1999.

[42] L. Liu, B. R. K. Kashyap, and J. G. C. Templeton. On the $GI^X/G/\infty$ system. *Journal of Applied Probability*, 27(3):671–683, 1990.

[43] L. Liu, B. R. K. Kashyap, and J. G. C. Templeton. Queue lengths in the $GI^{X_n}/M^R/\infty$ service system. *Queueing Systems*, 22(1–2):129–144, 1996.

[44] L. Liu and J. G. C. Templeton. The $GR^{X_n}/G_n/\infty$ system: System size. *Queueing Systems*, 8(1):323–356, 1991.

[45] R. M. Loynes. The stability of a queue with non-independent inter-arrival and service times. In *Proceedings of Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 58, 497–520. Cambridge University Press, 1962.

[46] X. Lu, S. Aalto, and P. Lassila. Performance-energy trade-off in data centers: Impact of switching delay. In *Proceedings of Proceedings of 22nd IEEE ITC Specialist Seminar on Energy Efficient and Green Networking (SSEEGN)*, 50–55, 2013.

[47] D. M. Lucantoni. New results on the single server queue with a batch Markovian arrival process. *Communications in Statistics. Stochastic Models*, 7(1):1–46, 1991.

[48] D. M. Lucantoni, K. S. Meier-Hellstern, and M. F. Neuts. A single-server queue with server vacations and a class of non-renewal arrival processes. *Advances in Applied Probability*, 22(3):676–705, 1990.

[49] H. B. Mann and A. Wald. On stochastic limit and order relationships. *The Annals of Mathematical Statistics*, 14(3):217–226, 1943.

[50] W. A. Massey. The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems*, 21(2-4):173–204, 2002.

[51] H. Masuyama and T. Takine. Analysis of an infinite-server queue with batch markovian arrival streams. *Queueing Systems*, 42(3):269–296, 2002.

[52] S. Mittal. A survey of techniques for improving energy efficiency in embedded computing systems. *International Journal of Computer Aided Engineering and Technology*, 6(4):440–459, 2014.

[53] A. Nazarov and G. Baymeeva. The $M/G/\infty$ queue in random environment. In *Proceedings of International Conference on Information Technologies and Mathematical Modelling 2014*, Volume 487 of *Communications in Computer and Information Science*, 312–324. Springer, 2014.

[54] R. Núñez-Queija. A queueing model with varying service rate for ABR. In *Proceedings of International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, 93–104. Springer, 1998.

[55] C. A. O'cinneide and P. Purdue. The M/M/∞ queue in a random environment. *Journal of Applied Probability*, 23(1):175–184, 1986.

[56] A. G. Pakes and N. Kaplan. On the subcritical Bellman-Harris process with immigration. *Journal of Applied Probability*, 11(4):652–668, 1974.

[57] T. Phung-Duc. Exact solutions for M/M/c/setup queues. *Telecommunication Systems*, 64(2):309–324, 2017.

[58] V. Ramaswami and M. F. Neuts. Some explicit formulas and computational methods for infinite-server queues with phase-type arrivals. *Journal of Applied Probability*, 17(2):498–514, 1980.

[59] U. L. Rohde. *Digital PLL frequency synthesizers: Theory and design*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[60] D. N. Shanbhag. On infinite server queues with batch arrivals. *Journal of Applied Probability*, 3(1):274–279, 1966.

[61] D. N. Shanbhag. On infinite server queues with batch arrivals. *Journal of Applied Probability*, 3(1):274–279, 1966.

[62] E. Le Sueur and G. Heiser. Dynamic voltage and frequency scaling: The laws of diminishing returns. In *Proceedings of Proceedings of the 2010 international conference on Power aware computing and systems*, 1–8, 2010.

[63] L. Takács. Queues with infinitely many servers. *RAIRO-Operations Research-Recherche Opérationnelle*, 14(2):109–113, 1980.

[64] T. Takine. Single-server queues with Markov-modulated arrivals and service speed. *Queueing Systems*, 49(1):7–22, 2005.

[65] L. B. Toktay, L. M. Wein, and S. A. Zenios. Inventory management of remanufacturable products. *Management science*, 46(11):1412–1426, 2000.

[66] H. Toyoizumi. Infinite-server queues with large fluctuation in arrival processes. In *Proceedings of the 2018 Fall National Conference of the Operations Research Society of Japan*, 68–69, 2018.

[67] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[68] T. Van Woensel and N. Vandaele. Modeling traffic flows with queueing models: a review. *Asia-Pacific Journal of Operational Research*, 24(4):435–461, 2007.

[69] R. W. Wolff. Poisson arrivals see time averages. *Operations Research*, 30(2):223–231, 1982.

[70] D. Xu, X. Liu, and A. V. Vasilakos. Traffic-aware resource provisioning for distributed clouds. *IEEE Cloud Computing*, 2(1):30–39, 2015.

[71] M. Yajima and H. Masuyama. Stability analysis of $GI^X/GI/\infty$ queues. In *Proceedings of Proceedings of the 14th International Conference on Queueing Theory and Network Applications*, 2019.

[72] M. Yajima and T. Phung-Duc. Batch arrival single-server queue with variable service speed and setup time. *Queueing Systems*, 86(3-4):241–260, 2017.

[73] M. Yajima and T. Phung-Duc. A central limit theorem for a markov-modulated infinite-server queue with batch poisson arrivals and binomial catastrophes. *Performance Evaluation*, 129:2–14, 2019.

[74] M. Yajima, T. Phung-Duc, and H. Masuyama. The stability condition of BMAP/M/∞ queues. In *Proceedings of Proceedings of the 11th International Conference on Queueing Theory and Network Applications*, page 5. ACM, 2016.