論文 / 著書情報
Article / Book Information

題目(和文)	直列型待ち行列システムにおける定常分布の漸近的性質
Title(English)	Asymptotic Behavior of Stationary Distributions in Tandem Queueing Systems
著者(和文)	藤本衡
Author(English)	KOU FUJIMOT
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3407号, 授与年月日:1997年3月26日, 学位の種別:課程博士, 審査員:
Citation(English)	Degree:Doctor of Science, Conferring organization: Tokyo Institute of Technology, Report number:甲第3407号, Conferred date:1997/3/26, Degree Type:Course doctor, Examiner:
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

Asymptotic Behavior of Stationary Distributions in Tandem Queueing Systems

Kou Fujimoto

Submitted in partial fulfillment of

the requirement for the degree of

DOCTOR OF SCIENCE

Supervised by

Professor Naoki Makimoto and Professor Yukio Takahashi

Department of Information Science

Tokyo Institute of Technology

Ookayama, Meguro-ku

Tokyo 152, Japan

October 1996

Contents

Acknowledgements

I owe a special debt of gratitude to Professor Yukio Takahashi for continuing guidance over this dissertation. I wish to thank Assistant Professor Naoki Makimoto for his encouragement and helpful suggestion especially for theoretical side. I also would like to express my thanks to all other members of laboratories of Professor Takahashi's, Professor Kojima's, Associate Professor Yanagida's.

I am deeply indebted to Mr. Akira Sato of Fujitsu Ltd. and Professor Song Yu of Fukuoka Institute of Technology, who introduced me to this object of study and taught me a variety of techniques for the numerical experiments.

I am grateful to my parents, my sister and her husband, and my grandparents for their support.

Chapter 1

Tandem Queueing Systems: Background and Model Description

1.1 Introduction

There are many studies on various queueing models. Above all, tandem queueing systems are very important both in theory and applications. In the theoretical aspect, tandem queueing systems are basic models of queueing networks as well as the extension of single queue models. To understand the complex behaviors of queueing networks, we should know about tandem queueing systems more as a first step. For applications, tandem queueing systems themselves have been applied to many practical systems such as production lines and point-to-point communications.

Recently applications of queueing networks explosively increase as computer and com-

munication networks are developed and spread in various fields. Hence the understanding of the queueing networks, and the understanding of the tandem queueing systems as building blocks, become very important. However, the stationary state probabilities of tandem queueing systems, or even basic properties of them, are scarcely known except for simple models which have product-form solutions. This is the reason why the author took interest in the theory of tandem queueing systems.

On the other hand, analysis of general single queues has been developed greatly by virtue of the theory of matrix-geometric form solutions [30]. One of the most important result is the geometric decay property of the tail of the stationary distribution. That is, if a single queue has a matrix-geometric form solution, its stationary queue-length probability $\pi(n)$ is asymptotically of the form as $C\eta^n$. This property is very useful, for example, on the numerical calculation of the stationary state probabilities and on the discussion of tail probabilities for estimating very small loss probabilities (e.g. less than 10^{-9}) of the corresponding finite queues.

The aim of this thesis is to extend this result to the tandem queueing systems. In this chapter we look the background of the research, especially on the tandem queueing systems and the related topics, and summarize the subsequent chapters.

1.2 Tandem Queueing Systems

Before introducing our model, we describe the class of tandem queueing systems and categorize it. The basic model is illustrated in Figure 1.1.

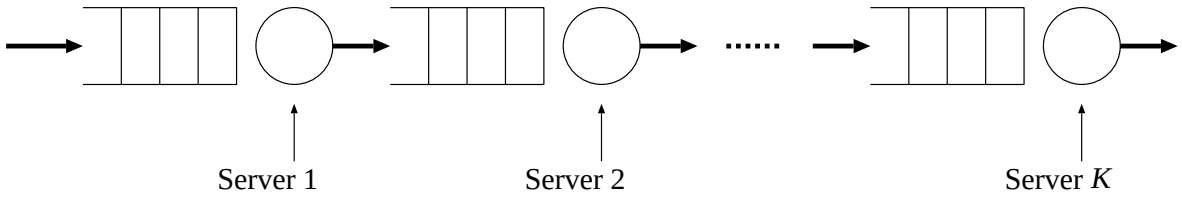


Figure 1.1: A tandem queueing system

A tandem queueing system consists of $K(\geq 2)$ queues arranged in series. Each queue has one or more servers and a buffer. Customers arrive at the first queue to be served there, proceed through all queues in order where services are done, and then leave the system after the Kth service. There are some assumptions which are common in most of the tandem queueing systems.

- The system is operated at steady-state condition. (This includes the case where there exist an infinite number of customers in the buffer of the first queue. In such a case, if we observe the behavior of the remaining queues, they are operated under the steady-state condition.)
- No customers are abandoned.
- Interarrival times and service times at each queue are independent.

- The transit times between queues are ignored.

Besides them, some general assumptions are seen in many articles:

Arrival process: The Poisson arrival is assumed in many studies. For more complex arrivals, renewal processes are employed.

Service distributions: Though exponential distribution predominates, phase-type distributions are also considered in recent studies.

Queueing discipline: All customers are served according to the first-come first-served (FCFS) discipline in most cases. The last-come first-served (LCFS) and processor-sharing (PS) disciplines are also considered in some studies.

Buffer capacity: The capacity of the buffer of the first queue is commonly assumed to be infinite. For the intermediate buffers, finite capacity is often assumed since most practical systems have only finite room for customers. However, infinite buffer is sometimes assumed since it offers good approximations to sufficiently large capacity of buffer, and since it is more suitable for analysis.

Blocking type: With finite buffers, there must be blockings. Usually, either of production or communication blockings are assumed.

Reliability of Servers: There may be breakdowns of servers, which cause extra blockings.

Since tandem queueing systems applied to practical problems are often very large and complicated to solve, many approximation methods have been proposed (see Altiok [1] or Song and Takahashi [41] for example).

Many exact and approximate analyses have been done for the cases with finite buffers. For details, see a recent nice survey by Papadopoulos and Heavey [34]. Among them, we have to note that the theory of matrix-geometric form solution is applied to tandem queueing systems with buffers of finite capacity (see Latouche and Neuts [21] for example). This enables us to evaluate the system performances numerically, or to analyze the asymptotic properties of its stationary distribution. However, this theory cannot be directly applied to the cases with infinite-capacity buffers.

1.2.1 Tandem Queueing Systems with Infinite buffer capacity

Though the case with infinite capacity is less popular than that with finite capacity, there are a number of studies on it. Among others, Jackson networks (see [7] for example) can be used to evaluate tandem queueing systems with infinite capacity of buffers, a Poisson arrival and exponential services. This is because tandem queueing systems with these assumptions have *product-form* stationary probabilities, and each queue can be analyzed, in some sense, as if they are statistically independent.

BCMP networks [2] also provide a strong tool to evaluate tandem queueing systems. BCMP networks can be applied to the tandem queueing systems with multiple types of customers while Jackson networks only admit a single type. Moreover, if the servers are

infinite or if the queueing discipline is PS or LCFS with preemptive-resume, BCMP networks admit service times subjecting to phase-type distributions. In both Jackson and BCMP networks, the stationary distributions are of the product form.

There are also many other studies. For example, Boxma [4, 5] considered two queues in series with the Poisson arrival and customerwise identical service times. He showed the necessary and sufficient condition for the stationarity of a tandem queueing system, and obtained explicit expressions for the stationary distributions of the sojourn times and the waiting times at the second queue, etc. Moreover, asymptotic and numerical results were obtained from these results.

Le Gall [22] extended it with three or more queues and renewal arrival. An explicit form of the stationary distribution of overall sojourn time was derived, and many examples were shown.

Karpelevitch and Kreinin [16] derived the Laplace transform of the stationary joint waiting time distribution and the generating function of joint queue length distribution at the arrival epoch of a Jackson-type tandem queueing system with two stages.

Katayama [17] studied on the mean sojourn times in a multi-stage tandem queueing system with the Poisson arrival and general service times, served by a single server with a cyclic switching rule. He derived the mean sojourn times, as well as the upper and lower bounds of the mean sojourn times and mean waiting times at the first queue for workload conserving switching rules.

Miyazawa [29] considered a two-stage tandem queueing system with stationary inputs

and general service time distributions. Some formulae on the joint queue-length probabilities and their expectations were presented.

For the two-stage tandem queueing system with exponential servers and renewal arrivals, Ganesh and Anantharam [13] proved that the marginal queue-length distribution of the second stage has three types of geometric tails.

Despite of these studies, scarcely known are the properties of basic tandem queueing systems with phase-type service distributions under the FCFS discipline. The analysis of such a tandem queueing system will give a good insight to more complex queueing networks and also will provide a good building block for approximate analyses of general queueing networks.

1.2.2 Model Assumptions

In this thesis, we consider a tandem queueing system with the following assumptions.

- Interarrival times between successive customers are random variables subjecting to a common phase-type distribution.
- Service times for customers at each server are random variables subjecting to a server-specific phase-type distribution.
- All the interarrival and service times are mutually independent.
- Queueing discipline is the FCFS.

- All buffers are of infinite capacity, and then no blockings occur.
- Servers do not break down.

If we denote the number of customers at the kth stage by n_k, the phase of the interarrival time distribution by i_0 and the phase of the service time distribution at the kth stage by i_k, then the tandem queueing system can be regarded as a continuous-time Markov chain whose state is represented as $(n_1, n_2, \ldots, n_K; i_0, i_1, i_2, \ldots, i_K)$.

If the number of queues K is small ($K = 2$ or 3) and the traffic intensities at queues are low, we can get the stationary probabilities and related performance measures by numerical computations. In fact, we calculated a number of cases as will be reported in Chapters 2 and 6 of this thesis. However, the numerical analysis is not an easy task, and is limited in size of models. It is desirable to know some basic properties which hold in more general tandem queueing systems. In this thesis, we focus on the asymptotic properties of the joint queue-length distribution, and study them both numerically and theoretically.

1.3 Summaries of Subsequent Chapters

In this section, we briefly summarize discussions and results in the subsequent chapters.

Chapter 2: Numerical Experiments on Tail Behavior of Stationary Distributions in Two-Stage Tandem Queueing Systems

We rarely have handholds to develop the theorems on the asymptotic form of the stationary distributions in our tandem queueing system. So we start with making an extensive numerical experiments on our model with two queues. We investigate the numerical results in detail to obtain conjectures on the asymptotic behavior of the stationary distribution. Then we find that decay rates of the tail of the stationary joint queue-length distribution are intimately related to the Laplace-Stieltjes transforms of interarrival and service time distributions.

We conjecture that the stationary probability $\pi(n_1, n_2; i_0, i_1, i_2)$ of two-stage tandem queueing system has geometric tails as follows:

1. For fixed traffic intensity ρ_1 of the first server, if the traffic intensity ρ_2 of the second server is less than a certain threshold $\tilde{\rho}_2$, there exist constants $\eta_1, \eta_2, c_0(i_0), c_1(i_1), c_2(i_2)$ and G such that

$$\pi(n_1, n_2; i_0, i_1, i_2) \sim G\ c_0(i_0)c_1(i_1)c_2(i_2)\eta_1^{n_1}\eta_2^{n_2},$$

as $n_1, n_2 \to \infty$ on a line $n_2 = an_1 + b$ with rational $a > 0$ and b. This asymptotic representation is also valid when $n_1 \to \infty$ with fixed n_2 and when $n_2 \to \infty$ with fixed n_1.

2. In the case $\rho_2 > \tilde{\rho}_2$, there exists a positive constant $\tilde{a}$ such that the decay rates are different between the cases $0 < a < \tilde{a}$ and $a > \tilde{a}$ for the slope a of the line on which

n_1 and n_2 increase. We denote the two sets of constants corresponding to these two cases as $\{\eta_1, \eta_2, c_0(i_0), c_1(i_1), c_2(i_2), G\}$ and $\{\overline{\eta}_1, \overline{\eta}_2, \overline{c}_0(i_0), \overline{c}_1(i_1), \overline{c}_2(i_2), \overline{G}\}$.

(a) When $n_1, n_2 \to \infty$ on a line $n_2 = an_1 + b$ with rational a and b such that $0 < a < \tilde{a}$,

$$\pi(n_1, n_2; i_0, i_1, i_2) \sim G\ c_0(i_0)c_1(i_1)c_2(i_2)\eta_1^{n_1}\eta_2^{n_2}.$$

This asymptotic representation is also valid when $n_1 \to \infty$ with fixed n_2.

(b) When $n_1, n_2 \to \infty$ on a line $n_2 = an_1 + b$ with rational a and b such that $a > \tilde{a}$,

$$\pi(n_1, n_2; i_0, i_1, i_2) \sim \overline{G}\ \overline{c}_0(i_0)\overline{c}_1(i_1)\overline{c}_2(i_2)\overline{\eta}_1^{n_1}\overline{\eta}_2^{n_2}.$$

This asymptotic representation is also valid when $n_2 \to \infty$ with fixed n_1.

3. Most of the constants above are determined by equations given in Section 2.5. Unfortunately we do not yet have any equations to determine the values of the multiplicative coefficients G and $\overline{G}$.

Chapter 3: Asymptotic Properties in Quasi-Birth-and-Death Processes with a Countable Number of Phases

The conjecture in Chapter 2 can be regarded as the extension of the geometric decay property of the queue length distribution in $GI/PH/c$ queue [33, 45], which was proved by using Neuts' theory of matrix-geometric form solution [30]. However, the quasi-birth-and-death process derived from our two-stage tandem queueing system is not covered by

Neuts' theory, since the number of states in each level is countably infinite. Therefore, we need a new theory.

We give a sufficient condition that the tail of the stationary distribution decays geometrically in a quasi-birth-and-death process with a countable number of states in each level. This is a sharpened result of the matrix-geometric extension given by Miller [28] and Ramaswami and Taylor [36]. This result will be applicable to many stochastic models which have never been analyzed so far because of their complexities.

Chapter 4: Asymptotic Properties of Stationary Distributions in Two-Stage Tandem Queueing Systems

To apply the theorems proved in Chapter 3, we show that the quasi-birth-and-death process derived from our tandem queueing system actually satisfies the assumptions of the theorems.

We prove the following

Theorem. For fixed n_1, i_0, i_1 and i_2, if $\eta_1 < \overline{\eta}_2$ then

$$\pi(n_1, n_2; i_0, i_1, i_2) \sim \overline{G}_1(n_1; i_0, i_1, i_2)\overline{\eta}_2^{n_2} \quad (n_2 \to \infty). \tag{1.1}$$

In this case, if $\overline{\eta}_1 < 1$ then

$$\overline{G}_1(n_1; i_0, i_1, i_2) \sim \overline{G}(i_0, i_1, i_2)\overline{\eta}_1^{n_1} \quad (n_1 \to \infty). \tag{1.2}$$

Here η_1, $\overline{\eta}_1$ and $\overline{\eta}_2$ are the constants discussed in Chapter 2.

We also prove a similar theorem for the case in which $n_1 \to \infty$ for fixed n_2. Though these theorems cover only some special cases of the conjecture, they will provide us a powerful tool for further studies and also a powerful tool for fast and accurate computations.

Chapter 5: Numerical Computation for Tandem Queueing Systems on a Parallel Computer

Numerical calculations of the stationary probabilities for two-stage models can be executed on an engineering workstation if the model is in a reasonable size and if a suitable numerical method is used. However, for three-stage models, it is very difficult to make such calculations on a workstation since the number of states (variables) to be handled becomes huge, say hundreds times of that in two-stage models. So the numerical experiments for three-stage models, which will be discussed in Chapter 6, and are planned to be done on a massive parallel computer.

On a massive parallel computer, data can be divided and allocated to multiple processor elements and computations can be done in parallel. Data transfers between processor elements are needed, however, and they cause considerable overhead. For the use of a massive parallel computer, we need a new theory and a new algorithm to exploit the strong points of the computer and to avoid the weak points of it. Namely, we have to find a way to allocate data suitably onto multiple processor elements, so that the computations can be done within each processor element as much as possible, and that the data transfers between processor elements become as small as possible.

In Chapter 5, we discuss such problems occurring in the use of a massive parallel computer, and propose a variation of the aggregation/disaggregation (A/D) method for effective calculations of large-scale Markov chains arising from three-stage tandem queueing models.

Chapter 6: Numerical Experiments on Tail Behavior of Stationary Distributions in Three-stage Tandem Queueing Systems

In Chapters 2 and 4, we find some tail properties of the stationary distribution in two-stage tandem queueing systems. They are a great progress, but not sufficient to understand basic properties of more complex queueing network models. To proceed one more step, we make experiments for three-stage tandem queueing models by using a massive parallel computer with a new variation of the A/D method proposed in Chapter 5.

By the limitation of available computation time and the limitation of the performance of the computer itself, we can do the experiments only for a limited number of models. However, the numerical results obtained show some asymptotic properties similar to those in two-stage models. More definitely, the tails of the stationary distribution of the queue lengths decay geometrically. In some cases with low traffic intensities, there is only one set of decay parameters, but in other cases there are two sets of decay parameters. We cannot find cases in which there are three or more sets of parameters. These decay parameters satisfy similar systems of equations to those appeared in two-stage models.

1.4 Remaining Problems

As mentioned above, we show some asymptotic properties of two- and three-stage tandem queueing systems in this thesis. There still remain, however, much more problems to be studied.

For two-stage models, our theorem does not cover the whole of the conjecture given in Chapter 2. We should prove the rest of the conjecture theoretically. We are preparing a theorem which gives upper bounds for decay rates when $n_2 \to \infty$ with fixed n_1. We should further study on the cases where n_1 and n_2 become large along a line $n_2 = an_1 + b$ with positive a.

The way to give the values of constants G and $\overline{G}$ is not given neither theoretically nor numerically. It is essential for making a quantitative estimation of the tail probability.

For three-stage models, our numerical experiments are not enough to make a sufficiently reliable conjecture. Especially, we are not much confidential that there exist only two sets of decay parameters. Then we must make more extensive experiments as soon as possible. The theoretical consideration must be done, too.

The aim of this study is to have a deep insight on more and comprehensive queueing networks than BCMP networks and to propose suitable approximate and exact numerical methods for them. Though the results two- and three-stage tandem queueing systems are toward the aim, we must make more efforts to extend the results into more general queueing networks.

Chapter 2

Numerical Experiments on Tail Behavior of Stationary Distributions in Two-Stage Tandem Queueing Systems

2.1 Introduction

Tandem queueing systems are basic models in the queueing theory and have been studied for a long time. However, the stationary state probabilities, or even basic properties of them, are scarcely known except for some simple cases with product form solutions. In this chapter, we observe the tail behavior of the stationary joint queue-length distribution in a

two-stage tandem queueing system $PH/PH/1 \to /PH/1$ with buffers of infinite capacity.

In the ordinary one-stage queue $PH/PH/c$ with traffic intensity $\rho < 1$, it is shown that the stationary distribution has a geometric tail [45]. Namely, if we let $x(n; i_0, i_1)$ be the stationary probability that there are n customers in the system while the states (phases) of arrival and service processes are i_0 and i_1 respectively, then

$$x(n; i_0, i_1) \sim G c_0(i_0) c_1(i_1) \eta^n, \quad n \to \infty, \tag{2.1}$$

and hence

$$\frac{x(n+1; i_0, i_1)}{x(n; i_0, i_1)} \sim \eta, \quad n \to \infty, \tag{2.2}$$

where $\eta, G, c_0(i_0)$ and $c_1(i_1)$ are constants and $\sim$ indicates that the ratio of both sides tends to 1.

The decay rate η is given in the following manner. Let $T^*(s)$ and $S^*(s)$ be the Laplace-Stieltjes transforms (LSTs) of the interarrival and service time distributions, respectively, and let ω be the unique positive solution of the equation

$$T^*(s) S^*(-cs) = 1. \tag{2.3}$$

Then $\eta = T^*(\omega)$.

This geometric decay property is very useful, for example, on the computation of the stationary state probabilities and on the discussion of tail probabilities for estimating very small loss probabilities (e.g. less than 10^{-9}) of the corresponding finite queue.

The problem here is to see whether a similar geometric tail property holds or not in two-stage tandem queueing systems.

The marginal queue-length distribution of the first stage clearly has a geometric tail, since the behavior of the first stage is not affected by that of the second stage. Our concern is the tail property of the joint queue-length distribution of the first and the second stages or the state probabilities in the steady state.

To see it, we make extensive numerical experiments though the types of models are limited to simple ones because of the limitation of the sizes of computable models. We scrutinize the results and find two types of geometric decay depending on the traffic intensities of the first and second stages. To the author's knowledge, this fact has never reported so far in the literature. Based on the observations of the numerical results, we give a conjecture on the geometric decay together with systems of equations which determine the parameters in the conjecture. The author thinks that the property stated in the conjecture will not only be useful for practical computation or simulation of two stage tandem queueing systems, but also will provide a key to further theoretical researches for tandem queueing systems.

This chapter is organized as follows. In Section 2.2, we describe our tandem queueing model and present our conjecture on geometric tail of the stationary distribution. In Section 2.3, we explain our numerical experiments briefly. Section 2.4 presents various numerical results which show the tail properties we conjecture in Section 2.2. We discuss, in Section 2.5, some equations which determine parameters used in the conjecture.

2.2 The Model and the Conjecture

Here we introduce our two-stage tandem queueing model and give a conjecture on the geometric tail. We also show some numerical results which support the conjecture in a variety of cases.

We denote by $PH(\boldsymbol{a}, \boldsymbol{\Phi})$ a phase-type distribution with initial probability vector $\boldsymbol{a}$ and transition rate matrix $\boldsymbol{\Phi}$.

2.2.1 Two-Stage Tandem Queueing System $\boldsymbol{PH/PH/1\rightarrow\ /PH/1}$

We consider an open, two-stage tandem queueing system (Figure 2.1). Customers arrive at the first stage to be served there, move to the second to be served there again, and then go out of the system. Customers are served according to first-come first-served (FCFS) discipline at each stage. The kth stage ($k = 1, 2$) has a single server and a buffer of infinite capacity, so that no loss or blocking occurs. Interarrival times of customers are independent and identically distributed (i.i.d.) random variables subjecting to a phase-type distribution $PH(\boldsymbol{\alpha}, \boldsymbol{T})$. Service times at the kth stage are also i.i.d. variables subjecting to a phase-type distribution $PH(\boldsymbol{\beta}_k, \boldsymbol{S}_k)$. The interarrival and service times are assumed to be mutually independent.

The state of the system is represented by a quintuple $(n_1, n_2; i_0, i_1, i_2)$, where i_0 is the phase of the arrival process, i_k is the phase of the service process at the kth stage, and n_k is the number of customers in the kth stage ($k = 1, 2$). Then the system behaves as a

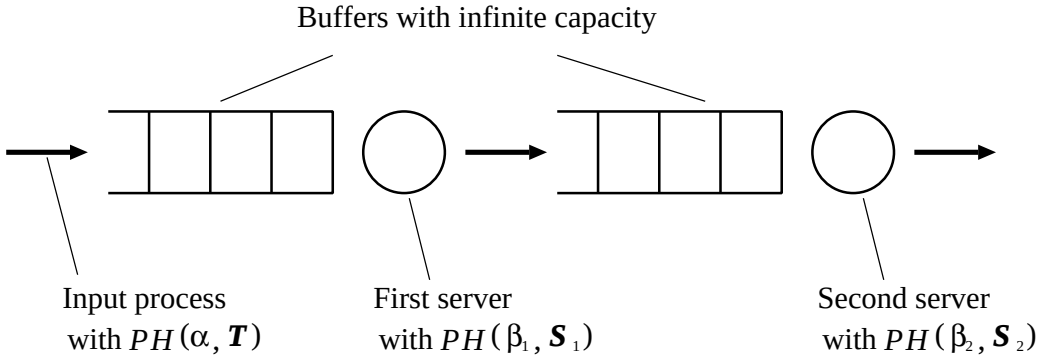


Figure 2.1: Two-stage tandem queueing system

continuous-time Markov chain.

We denote the traffic intensity at the kth stage by $\rho_k = \lambda/\mu_k$ where $1/\lambda$ is the mean interarrival time and $1/\mu_k$ is the mean service time at the kth stage ($k = 1, 2$). We assume $\rho_1, \rho_2 < 1$ so that the chain is stable and has stationary probabilities $x(n_1, n_2; i_0, i_1, i_2)$.

2.2.2 Geometric Decay Property from the Numerical Experiments and the Conjecture

The tail properties extracted from the numerical results are roughly summarized as follows.

For a given traffic intensity ρ_1 of the first stage, there exists a threshold $\tilde{\rho}_2$ for the traffic intensity ρ_2 of the second stage, and if $\rho_2 < \tilde{\rho}_2$, then the joint queue-length probability $p(n_1, n_2)$ is asymptotically of the geometric form

$$p(n_1, n_2) \sim G\ \eta_1^{n_1} \eta_2^{n_2} \quad (n_1, n_2 \to \infty). \tag{2.4}$$

If $\rho_2 > \tilde{\rho}_2$, then $p(n_1, n_2)$ decays in a similar manner, but the coefficient G and the decay rates η_1, η_2 are different between the cases with $n_2 < \tilde{a}n_1$ and with $n_2 > \tilde{a}n_1$ for a certain positive value $\tilde{a}$. Moreover, the conditional probability of phases $y(i_0, i_1, i_2 \mid n_1, n_2) = x(n_1, n_2; i_0, i_1, i_2)/p(n_1, n_2)$ is asymptotically independent of n_1 and n_2 in each case, and hence the stationary distribution has geometric tail.

To describe the geometric decay property more formally, however, we should clarify the way of making n_1 and n_2 large in (2.4). Here we consider the case in which n_1 and n_2 increase on a line $n_2 = an_1 + b$. To ensure that there exist infinitely many points (n_1, n_2) on the line, the coefficient a should be positive and rational and the constant b rational. As extreme cases, we also consider the case in which $n_1 \to \infty$ with fixed n_2 and the case in which $n_2 \to \infty$ with fixed n_1.

The conjecture we make is formally stated as follows.

Conjecture For fixed ρ_1, there exists a threshold $\tilde{\rho}_2$ and the behavior of $x(n_1, n_2; i_0, i_1, i_2)$ is different between the cases $\rho_2 < \tilde{\rho}_2$ and $\rho_2 > \tilde{\rho}_2$.

1. In the case $\rho_2 < \tilde{\rho}_2$, there exist constants $\eta_1, \eta_2, c_0(i_0), c_1(i_1), c_2(i_2)$ and G such that

$$x(n_1, n_2; i_0, i_1, i_2) \sim G\, c_0(i_0)c_1(i_1)c_2(i_2)\eta_1^{n_1}\eta_2^{n_2}$$

   as $n_1, n_2 \to \infty$ on a line $n_2 = an_1 + b$ with rational $a > 0$ and b. This asymptotic representation is also valid when $n_1 \to \infty$ with fixed n_2 and when $n_2 \to \infty$ with fixed n_1.

2. In the case $\rho_2 > \tilde{\rho}_2$, there exists a positive constant $\tilde{a}$ such that the decay rates are different between the cases $0 < a < \tilde{a}$ and $a > \tilde{a}$ for the slope a of the line on which n_1 and n_2 increase. We denote the two sets of constants corresponding to these two cases as $\{\eta_1, \eta_2, c_0(i_0), c_1(i_1), c_2(i_2), G\}$ and $\{\overline{\eta}_1, \overline{\eta}_2, \overline{c}_0(i_0), \overline{c}_1(i_1), \overline{c}_2(i_2), \overline{G}\}$.

   (a) When $n_1, n_2 \to \infty$ on a line $n_2 = an_1 + b$ with rational a and b such that $0 < a < \tilde{a}$,

$$x(n_1, n_2; i_0, i_1, i_2) \sim G\ c_0(i_0)c_1(i_1)c_2(i_2)\eta_1^{n_1}\eta_2^{n_2}.$$

   This asymptotic representation is also valid when $n_1 \to \infty$ with fixed n_2.

   (b) When $n_1, n_2 \to \infty$ on a line $n_2 = an_1 + b$ with rational a and b such that $a > \tilde{a}$,

$$x(n_1, n_2; i_0, i_1, i_2) \sim \overline{G}\ \overline{c}_0(i_0)\overline{c}_1(i_1)\overline{c}_2(i_2)\overline{\eta}_1^{n_1}\overline{\eta}_2^{n_2}.$$

   This asymptotic representation is also valid when $n_2 \to \infty$ with fixed n_1.

3. The constants above are determined by equations given in the latter sections as indicated by the equation numbers:

$$\eta_1, \eta_2 \ \ldots\ldots\ldots\ldots\ldots\ldots (2.11)$$

$$\overline{\eta}_1, \overline{\eta}_2 \ \ldots\ldots\ldots\ldots\ldots\ldots (2.13)$$

$$\tilde{\rho}_2 \ \ldots\ldots\ldots\ldots\ldots\ldots (2.14)$$

$$\tilde{a} \ \ldots\ldots\ldots\ldots\ldots\ldots (2.8), (2.15)$$

$$c_k(i_k), \overline{c}_k(i_k)\ (k = 0, 1, 2)\ \ldots (2.16)$$

Unfortunately we do not yet have any equations to determine the values of the multiplicative coefficients G and $\overline{G}$. On the point, see comments at the end of Section 2.5.

2.2.3 Numerical Test for the Conjecture

To see if the asymptotic properties stated in the conjecture above hold or not, we tabulate the values of the ratios

$$g(n_1, n_2) = \frac{p(n_1, n_2)}{\eta_1^{n_1} \eta_2^{n_2}} \quad \text{and} \quad \overline{g}(n_1, n_2) = \frac{p(n_1, n_2)}{\overline{\eta}_1^{n_1} \overline{\eta}_2^{n_2}}$$

in Tables 2.1 and 2.2 for eight types of models with selected pair of traffic intensities (ρ_1, ρ_2) and selected points (n_1, n_2) lying on lines $n_2 = 4n_1 - 5$ and $n_2 = (n_1 + 5)/4$. All these values are extracted from the results of numerical experiments described in Section 2.3.

These tables show that each row certainly converges to a positive limit, even though the speed of convergence is much slower in a few cases (see Table 2.2(b)). In Table 2.1, the ratios along two different lines seem to converge to a common limit in each model. This corresponds to the first statement of the conjecture. In Table 2.2, these ratios seem to converge to different limits in each model. This agrees with the second statement of the conjecture. These numerical results support the conjecture on the geometric decay properties of the joint queue-length distribution. The asymptotic independence of phases is shown in Tables 2.4 and 2.5 in Section 2.4.3 for a particular model $E_2/H_2/1 \to /E_2/1$ with $\rho_1 = 0.6$ and $\rho_2 = 0.8$.

The author tested the conjecture for more than 1,000 cases. There exist a small number

Table 2.1: Geometric decay of $p(n_1, n_2)$: the case $\rho_2 < \tilde{\rho}_2$

In each model, the upper row represents $g(n_1, n_2)$ for n_1 and n_2 such that $n_2 = (n_1 + 5)/4, n_1 = 15, 35, \ldots, 95$, and the lower row represents it for n_1 and n_2 such that $n_2 = 4n_1 - 5, n_1 = 5, 10, \ldots, 25$. The traffic intensities ρ_1 and ρ_2 are selected so that $\rho_2 < \tilde{\rho}_2$.

(n_1, n_2)	(5,15)	(10,35)	(15,55)	(20,75)	(25,95)
	(15,5)	(35,10)	(55,15)	(75,20)	(95,25)
$M/E_2/1 \to /E_2/1$	1.4659	1.4725	1.4725	1.4725	1.4725
$\rho_1 = 0.60, \rho_2 = 0.35$	1.4230	1.4682	1.4722	1.4725	1.4725
$M/H_2/1 \to /E_2/1$	0.2773	0.2788	0.2788	0.2788	0.2788
$\rho_1 = 0.60, \rho_2 = 0.40$	0.2798	0.2788	0.2788	0.2788	0.2788
$E_2/E_2/1 \to /E_2/1$	3.0959	3.1120	3.1121	3.1121	3.1121
$\rho_1 = 0.60, \rho_2 = 0.40$	3.1113	3.1122	3.1122	3.1122	3.1122
$H_2/E_2/1 \to /E_2/1$	0.8375	0.8397	0.8397	0.8397	0.8397
$\rho_1 = 0.60, \rho_2 = 0.40$	0.7311	0.8042	0.8289	0.8365	0.8387
$M/E_2/1 \to /H_2/1$	0.7026	0.7026	0.7027	0.7027	0.7027
$\rho_1 = 0.60, \rho_2 = 0.20$	0.6883	0.7018	0.7026	0.7027	0.7027
$M/H_2/1 \to /H_2/1$	0.1538	0.1538	0.1538	0.1538	0.1538
$\rho_1 = 0.60, \rho_2 = 0.40$	0.1570	0.1542	0.1538	0.1538	0.1538
$E_2/H_2/1 \to /E_2/1$	0.4521	0.4557	0.4558	0.4558	0.4558
$\rho_1 = 0.60, \rho_2 = 0.40$	0.4564	0.4558	0.4558	0.4558	0.4558
$E_4/M/1 \to /H_2/1$	0.4924	0.4923	0.4923	0.4923	0.4923
$\rho_1 = 0.60, \rho_2 = 0.40$	0.4991	0.4923	0.4923	0.4923	0.4923

of cases in which the convergence of $g(n_1, n_2)$ and/or $\overline{g}(n_1, n_2)$ cannot be judged from the numerical results of $p(n_1, n_2)$ with $n_1, n_2 \leq 100$. The author thinks that, if we can calculate $p(n_1, n_2)$ for larger n_1 and n_2, we will be able to see the convergence numerically. Except these slow converging cases, $g(n_1, n_2)$ and $\overline{g}(n_1, n_2)$ do converge numerically to certain limits in all the cases we tested.

Table 2.2: Geometric decay of $p(n_1, n_2)$: the case $\rho_2 > \tilde{\rho}_2$

a: faster convergence models

In each model, the upper row represents $g(n_1, n_2)$ for n_1 and n_2 such that $n_2 = (n_1 + 5)/4, n_1 = 15, 35, \ldots, 95$, and the lower row represents $\overline{g}(n_1, n_2)$ for n_1 and n_2 such that $n_2 = 4n_1 - 5, n_1 = 5, 10, \ldots, 25$. The traffic intensities ρ_1 and ρ_2 are selected so that $\rho_2 > \tilde{\rho}_2$ and $\tilde{a}$ is near to 1.

(n_1, n_2)	(5,15) (15,5)	(10,35) (35,10)	(15,55) (55,15)	(20,75) (75,20)	(25,95) (95,25)
$M/H_2/1 \rightarrow /E_2/1$	0.1026	0.1029	0.1022	0.1017	0.1014
$\rho_1 = 0.60, \rho_2 = 0.80$	0.1140	0.1184	0.1193	0.1193	0.1192
$E_2/E_2/1 \rightarrow /E_2/1$	0.9333	0.9351	0.9353	0.9353	0.9353
$\rho_1 = 0.60, \rho_2 = 0.70$	0.9141	0.9157	0.9157	0.9156	0.9154
$M/H_2/1 \rightarrow /H_2/1$	0.0790	0.0792	0.0789	0.0786	0.0784
$\rho_1 = 0.60, \rho_2 = 0.75$	0.0846	0.0866	0.0869	0.0868	0.0867
$E_2/H_2/1 \rightarrow /E_2/1$	0.2162	0.2053	0.2011	0.2002	0.2001
$\rho_1 = 0.60, \rho_2 = 0.85$	0.2776	0.2817	0.2804	0.2801	0.2800
$E_4/M/1 \rightarrow /H_2/1$	0.3389	0.3246	0.3235	0.3234	0.3234
$\rho_1 = 0.60, \rho_2 = 0.77$	0.4276	0.4203	0.4192	0.4191	0.4191

b: slower convergence models

This table shows cases in which the convergence is much slower. In each model, the upper row represents $g(n_1, n_2)$ for n_1 and n_2 such that $n_2 = (n_1+5)/4, n_1 = 115, 135, \ldots, 195$, and the lower row represents $\overline{g}(n_1, n_2)$ for n_1 and n_2 such that $n_2 = 4n_1 - 5, n_1 = 30, 35, \ldots, 50$. The traffic intensities ρ_1 and ρ_2 are selected so that $\rho_2 > \tilde{\rho}_2$ and $\tilde{a}$ is near to 1.

(n_1, n_2)	(115,30) (30,115)	(135,35) (35,155)	(155,40) (40,155)	(175,45) (45,175)	(195,50) (50,195)
$M/E_2/1 \rightarrow /E_2/1$	0.1800	0.1815	0.1829	0.1842	0.1855
$\rho_1 = 0.60, \rho_2 = 0.71$	0.0700	0.0671	0.0647	0.0627	0.0609
$H_2/E_2/1 \rightarrow /E_2/1$	0.0957	0.0952	0.0948	0.0947	0.0946
$\rho_1 = 0.60, \rho_2 = 0.70$	0.0206	0.0180	0.0160	0.0142	0.0128
$M/E_2/1 \rightarrow /H_2/1$	0.1360	0.1301	0.1294	0.1298	0.1305
$\rho_1 = 0.60, \rho_2 = 0.70$	0.1043	0.1047	0.1050	0.1052	0.1055

2.3 Numerical Experiments

To see the tail behavior of the joint queue-length distribution, we made extensive numerical experiments for a variety of models. Specifically, we calculated the stationary state probabilities and drew graphs to see the characteristics of the tail behavior. We tested various types of models with various traffic intensities ρ_1 and ρ_2. Among them, for the 8 types of models listed below, we tested systematically with $\rho_1 = .2, .3, \ldots, .9$ and $\rho_2 = .2, .3, \ldots, .9$, and saw the changes of the tail behaviors by the traffic intensities in detail:

- $M/E_2/1 \rightarrow /E_2/1$

ρ_1	ρ_2			
0.20	0.15	0.18	0.20	0.22
	0.30	0.40	0.60	0.80
0.40	0.20	0.40	0.60	0.80
0.60	0.10	0.15	*0.20	0.25
	0.30	0.35	*0.40	0.45
	0.50	0.55	*0.60	0.65
	0.70	0.75	*0.80	0.85
	0.90			
0.80	0.20	0.40	0.60	0.80

- $M/H_2/1 \rightarrow /E_2/1$

ρ_1	ρ_2			
0.60	0.20	0.40	0.60	0.80

- $E_2/E_2/1 \rightarrow /E_2/1$

ρ_1	ρ_2			
0.60	0.20	0.40	0.60	0.80

- $M/H_2/1 \rightarrow /H_2/1$

ρ_1	ρ_2			
0.20	0.15	0.18	0.20	0.22
	0.25	0.30	0.40	0.60
0.60	0.10	0.15	*0.20	0.25
	0.30	0.35	*0.40	0.45
	0.50	0.55	*0.60	0.65
	0.70	0.75	*0.80	0.85
	0.90			

- $M/E_2/1 \rightarrow /H_2/1$

ρ_1	ρ_2			
0.60	0.20	0.40	0.44	0.46
	0.48	0.50	0.52	0.56
	0.60	0.70	0.80	

- $E_4/M/1 \rightarrow /H_2/1$

ρ_1	ρ_2			
0.60	0.20	0.40	0.60	0.80

The cases with * are calculated with truncations at $\nu = 150$. Other case are calculated with truncations at $\nu = 100$.

Here, for the two-phase hyperexponential distribution (H_2), we used the one with the density function of the form

$$s(x) = 0.2e^{-4\mu x} + 0.8e^{-\mu x}, \ x > 0.$$

The total number of cases tested exceeds one thousand. For the calculations of the stationary probabilities, we employed the aggregation/disaggregation method [37, 44]. Since our model has infinite number of states, we have to truncate the state space for both n_1 and n_2 in the calculations. However, in an iteration of the aggregation/disaggregation method, a new value of $x(n_1, n_2; i_0, i_1, i_2)$ is calculated from the current values of neighboring states $x(n_1 - 1, n_2; i_0, i_1, i_2)$, $x(n_1 + 1, n_2 - 1; i_0, i_1, i_2)$ and $x(n_1, n_2 + 1; i_0, i_1, i_2)$. Therefore, if we truncate the state space at $n_1 = \nu_1$ and $n_2 = \nu_2$, we have to estimate the values of $x(\nu_1 + 1, n_2 - 1; i_0, i_1, i_2)$, $1 \leq n_2 \leq \nu_2$ and $x(n_1, \nu_2 + 1; i_0, i_1, i_2)$, $0 \leq n_1 \leq \nu_1$. In our experiments, we estimated those values by assuming the geometric decay for these variables, namely, e.g., $x(\nu_1 + 1, n_2 - 1; i_0, i_1, i_2)$ was estimated as $\{x(\nu_1, n_2 - 1; i_0, i_1, i_2)\}^2 / x(\nu_1 - 1, n_2 - 1; i_0, i_1, i_2)$. Both of the truncation points ν_1 and ν_2 were set to 100 in most of the cases. So the number of states to be calculated was 40,000 for the models listed above with Poisson arrivals, and was 80,000 for ones above with other renewal arrivals. This number 80,000 is quite large and, by the author's experiences, it is very near to the limit of the size for the calculation of stationary probabilities of a Markov chain using today's engineering workstations.

The program was written in C and ran on a SONY NWS-3860 workstation. The

computational burden is practically $\mathcal{O}(N)$ with $N = \nu_1 \times \nu_2$, and it increases rapidly as $\rho_k \to 1$. Table 2.3 tabulates the CPU time for the computation of $E_2/H_2/1 \to /E_2/1$ with $\rho_1 = .6$ and $\rho_2 = .2, .4, .6$ and $.8$.

Table 2.3: The CPU time for $E_2/H_2/1 \to /E_2/1$ with $\rho_1 = 0.6$ and $\nu_1 = \nu_2 = 100$

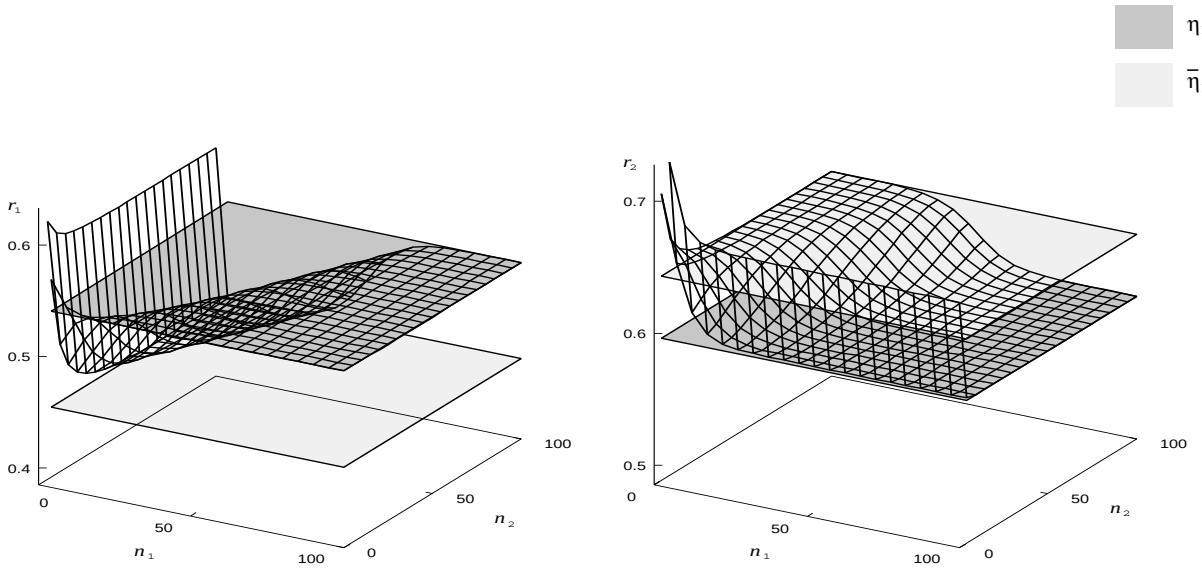
ρ_2	0.2	0.4	0.6	0.8
CPU time [sec.]	16	23	38	158

2.4 Tail Properties from the Numerical Experiments

The conjecture made in Section 2.2 is based on careful observations of the results of the numerical experiments explained in Section 2.3. Here we show a few results to indicate how the author reaches the conjecture. By the limitation of pages, we take the case of $E_2/H_2/1 \to /E_2/1$ with $\rho_1 = 0.6$ and $\rho_2 = 0.8$ as a typical example, and show its tail properties in detail. We start with observing the ratios of two neighboring joint probabilities of numbers of customers in the steady state.

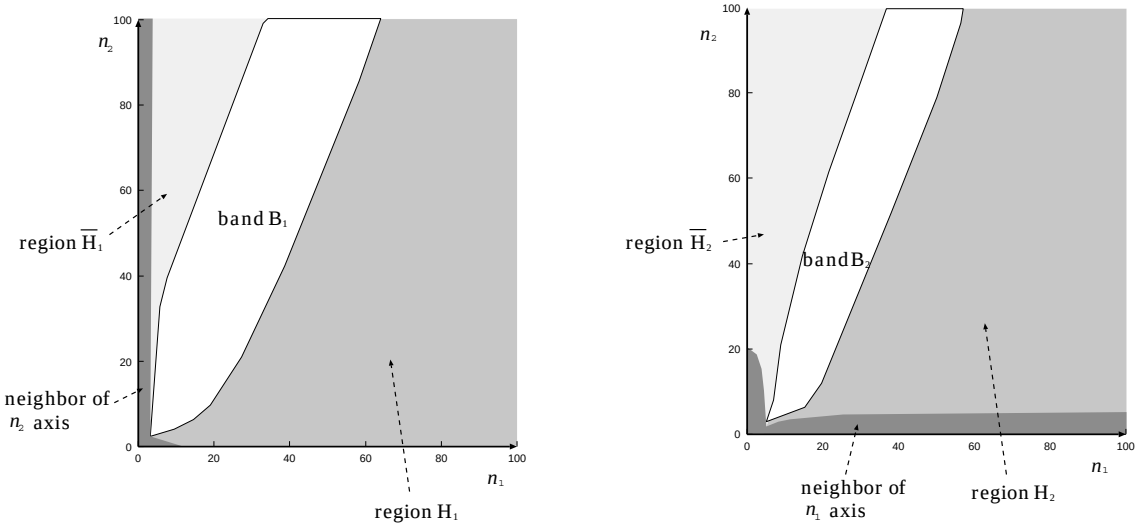
2.4.1 Decay Rates of the Joint Queue-length Probability

Let $p(n_1, n_2)$ be the joint probability that there exist n_k customers in the kth stage $(k = 1, 2)$ in the steady state. Namely, $p(n_1, n_2) = \sum_{i_0} \sum_{i_1} \sum_{i_2} x(n_1, n_2; i_0, i_1, i_2)$. We



a: $r_1(n_1, n_2) = \frac{p(n_1+1,n_2)}{p(n_1,n_2)}$ b: $r_2(n_1, n_2) = \frac{p(n_1,n_2+1)}{p(n_1,n_2)}$

Figure 2.2: r_k behavior in $E_2/H_2/1 \rightarrow /E_2/1$ $(\rho_1 = 0.6, \rho_2 = 0.8)$



a: r_1 b: r_2

Figure 2.3: Characterization of r_k surface

are interested in the ratios of neighboring $p(n_1, n_2)$'s:

$$r_1(n_1, n_2) = \frac{p(n_1+1, n_2)}{p(n_1, n_2)} \quad \text{and} \quad r_2(n_1, n_2) = \frac{p(n_1, n_2+1)}{p(n_1, n_2)}.$$

Figures 2.2a and 2.2b show graphs of $r_1(n_1, n_2)$ and $r_2(n_1, n_2)$. In each figure, the graph of $r_k(n_1, n_2)$ is a curved surface represented by a lattice. A dark gray plane indicates a constant η_k ($\eta_1 = 0.543, \eta_2 = 0.593$) and a light gray plane indicates another constant $\overline{\eta}_k$ ($\overline{\eta}_1 = 0.457, \overline{\eta}_2 = 0.640$). Both of these constants are given as solutions of systems of equations which will be presented in Section 2.5.

In Figure 2.2a, we see that $r_1(n_1, n_2)$ is relatively large very near the n_2 axis but it is in between η_1 and $\overline{\eta}_1$ in most of the region of (n_1, n_2). Especially $r_1(n_1, n_2)$ is close to η_1 in a region in which n_1 is relatively larger than n_2 and it is close to $\overline{\eta}_1$ in a region in which n_2 is relatively larger than n_1 though the latter might not be seen clearly from the figure.

Figure 2.3a shows regions of (n_1, n_2) in which $r_1(n_1, n_2)$ is close to η_1 or $\overline{\eta}_1$. In the dark gray region, labeled H_1, $r_1(n_1, n_2)$ is close to η_1, namely $|r_1(n_1, n_2) - \eta_1| < \varepsilon_1$ with $\varepsilon_1 = 0.1 \times |\eta_1 - \overline{\eta}_1|$, and in the light gray region, labeled $\overline{\mathrm{H}}_1$, $r_1(n_1, n_2)$ is close to $\overline{\eta}_1$, namely $|r_1(n_1, n_2) - \overline{\eta}_1| < \varepsilon_1$. (Here we take the particular value of ε_1 for the convenience of the explanation. We may take a smaller value if we need more accuracy in the subsequent approximations.) The band B_1 between H_1 and $\overline{\mathrm{H}}_1$ represents the region where $r_1(n_1, n_2)$ smoothly changes from $\eta_1 - \varepsilon_1$ to $\overline{\eta}_1 + \varepsilon_1$. We note that the region H_1 covers the n_1 axis while $\overline{\mathrm{H}}_2$ does not n_2 axis.

Similarly, in Figure 2.2b, we see that $r_2(n_1, n_2)$ is relatively large very near the n_1 axis

but it is in between η_2 and $\overline{\eta}_2$ in most of the region of (n_1, n_2). Especially $r_2(n_1, n_2)$ is close to η_2 in a region in which n_1 is relatively larger than n_2 and it is close to $\overline{\eta}_2$ in a region in which n_2 is relatively larger than n_1.

Figure 2.3b shows a decomposition of the n_1-n_2 plane by $r_2(n_1, n_2)$. The ratio $r_2(n_1, n_2)$ is close to η_2 in the dark gray region labeled H_2, and close to $\overline{\eta}_2$ in the light gray region labeled $\overline{\mathrm{H}}_2$. The band B_2 represents the region where $r_2(n_1, n_2)$ smoothly changes from $\eta_2 - \varepsilon_2$ to $\overline{\eta}_2 + \varepsilon_2$ where $\varepsilon_2 = 0.1 \times |\eta_2 - \overline{\eta}_2|$. In this case n_2-axis is included in $\overline{\mathrm{H}}_2$, but n_1-axis is not in H_2.

Figure 2.3a resembles 2.3b in some sense. H_1 mostly coincides with H_2, and $\overline{\mathrm{H}}_1$ does with $\overline{\mathrm{H}}_2$. Hence, $r_1(n_1, n_2) \approx \eta_1$ and $r_2(n_1, n_2) \approx \eta_2$ in the region $\mathrm{H}_1 \bigcap \mathrm{H}_2$, and $r_1(n_1, n_2) \approx \overline{\eta}_1$ and $r_2(n_1, n_2) \approx \overline{\eta}_2$ in the region $\overline{\mathrm{H}}_1 \bigcap \overline{\mathrm{H}}_2$, where "$\approx$" indicates that both sides are approximately equal. Hence

$$\frac{p(n_1 + l_1, n_2 + l_2)}{p(n_1, n_2)} \approx \begin{cases} \eta_1^{l_1} \eta_2^{l_2}, & \text{if } (n_1, n_2),\ (n_1 + l_1, n_2 + l_2) \in \mathrm{H}_1 \bigcap \mathrm{H}_2, \\ \overline{\eta}_1^{l_1} \overline{\eta}_2^{l_2}, & \text{if } (n_1, n_2),\ (n_1 + l_1, n_2 + l_2) \in \overline{\mathrm{H}}_1 \bigcap \overline{\mathrm{H}}_2. \end{cases} \tag{2.5}$$

The band B_2 lies almost on the same position as B_1, though the band width is a bit narrower. In these graphs, the bands B_1 and B_2 seem to keep their widths constant. More definitely, they are included in a region $\{(n_1, n_2) : \tilde{a} n_1 + \underline{b} < n_2 < \tilde{a} n_1 + \overline{b}\}$ bounded by two parallel lines with a common slope $\tilde{a} = 2.28$ and segments $\underline{b} = -48$ and $\overline{b} = 26$ as shown in Figure 2.4. The value $\tilde{a}$ of the slope will be discussed in Section 2.4.4 and in Section 2.5.

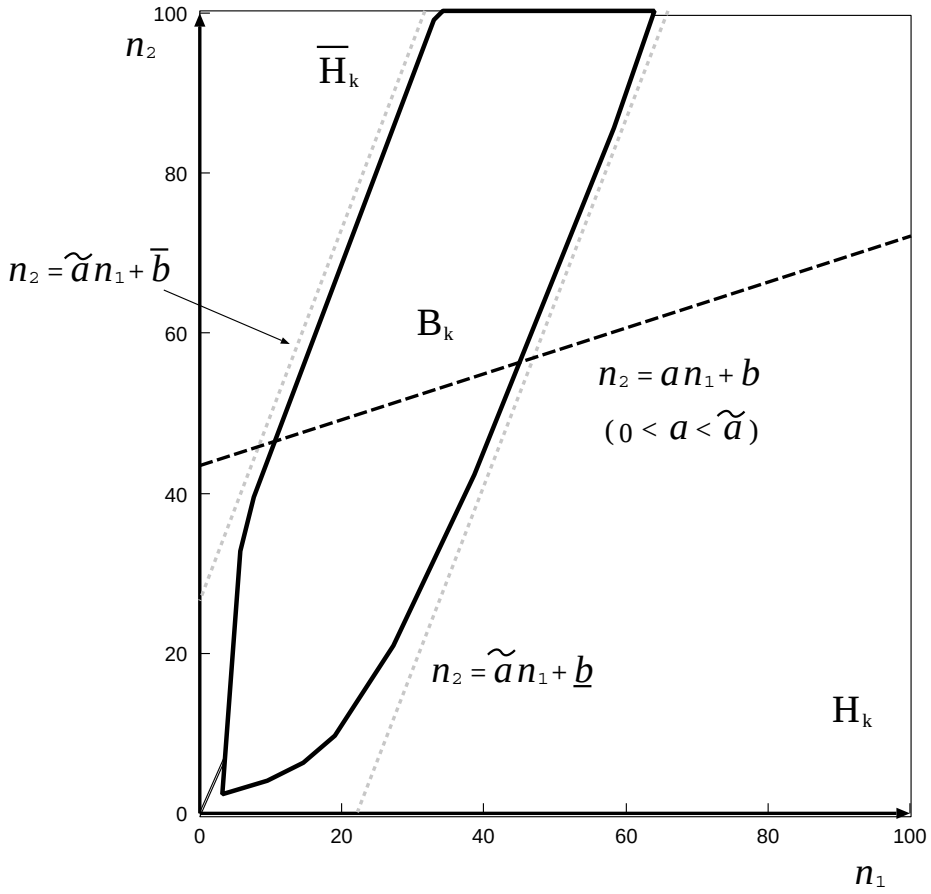


Figure 2.4: constraint lines for band B_k

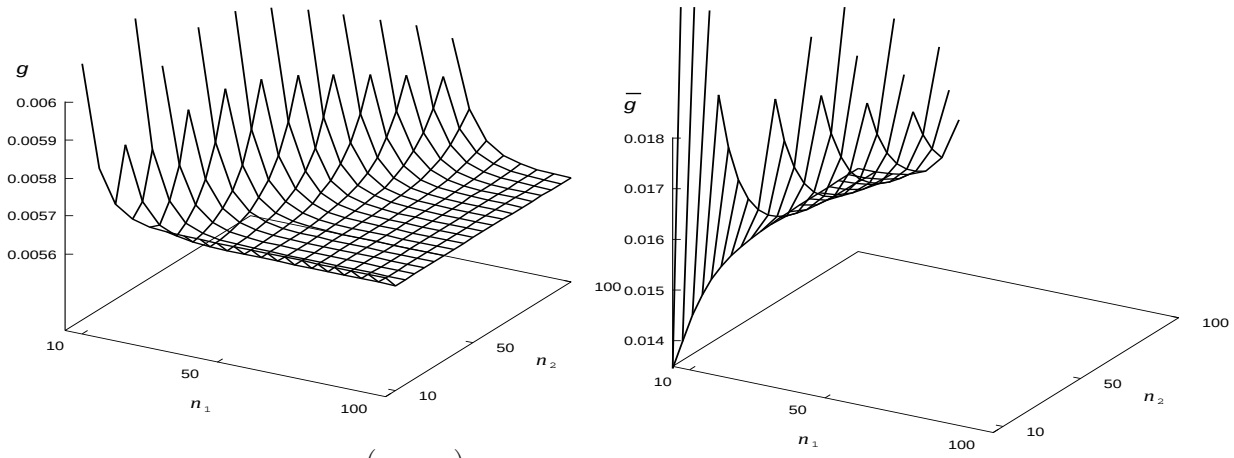
2.4.2 Geometric Form of the Joint Queue-length Probability

It seems that $p(n_1, n_2)$ is written approximately in a geometric form in the n_1-n_2 plane:

$$p(n_1, n_2) \approx \begin{cases} G\ \eta_1^{n_1} \eta_2^{n_2}, & \text{if } (n_1, n_2) \in \mathrm{H}_1 \bigcap \mathrm{H}_2, \\ \overline{G}\ \overline{\eta}_1^{n_1} \overline{\eta}_2^{n_2}, & \text{if } (n_1, n_2) \in \overline{\mathrm{H}}_1 \bigcap \overline{\mathrm{H}}_2, \end{cases} \tag{2.6}$$

where G and $\overline{G}$ are certain constants independent of n_1 and n_2.

To justify this from numerical results, we draw graphs of the ratios $g(n_1, n_2)$ and $\overline{g}(n_1, n_2)$ defined in Section 2.2.2. Figure 2.5a shows that $g(n_1, n_2)$ almost coincides with



a: $g(n_1, n_2) = \dfrac{p(n_1, n_2)}{\eta_1^{n_1}\eta_2^{n_2}}$

b: $\overline{g}(n_1, n_2) = \dfrac{p(n_1, n_2)}{\overline{\eta}_1^{n_1}\overline{\eta}_2^{n_2}}$

Figure 2.5: Behaviors of $g(n_1, n_2)$ and $\overline{g}(n_1, n_2)$ in $E_2/H_2/1 \rightarrow /E_2/1$ with $\rho_1 = 0.6, \rho_2 = 0.8$

a constant $G = 5.67 \times 10^{-3}$ when $(n_1, n_2) \in \mathrm{H}_1 \bigcap \mathrm{H}_2$, and Figure 2.5b shows that $\overline{g}(n_1, n_2)$ mostly coincides with a constant $\overline{G} = 1.51 \times 10^{-2}$ $(\neq G)$ when $(n_1, n_2) \in \overline{\mathrm{H}}_1 \bigcap \overline{\mathrm{H}}_2$. Note that, in these graphs, we cut the region where $n_1 < 5$ or $n_2 < 5$ to make the graphs easier to see the behavior when n_1 and n_2 are large. Hereafter we will use this convention in all graphs except Figure 2.9.

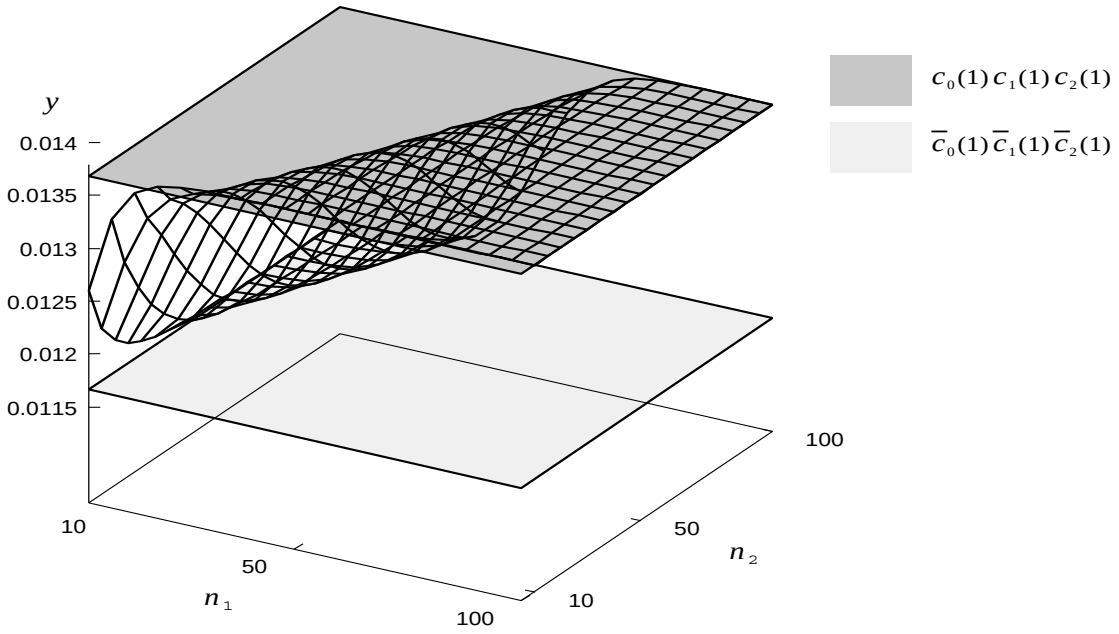


Figure 2.6: Behavior of $y(1,1,1|n_1,n_2)$ in $E_2/H_2/1 \rightarrow /E_2/1$ with $\rho_1 = 0.6, \rho_2 = 0.8$

2.4.3 Approximate Independence of Phases

The individual state probabilities $x(n_1,n_2;i_0,i_1,i_2)$ satisfy similar properties to those of $p(n_1,n_2)$ above. Hence it is expected that the conditional probability of phases

$$y(i_0,i_1,i_2|n_1,n_2) = \frac{x(n_1,n_2;i_0,i_1,i_2)}{p(n_1,n_2)}$$

almost coincides with a constant in $\mathrm{H}_1 \bigcap \mathrm{H}_2$, and with another constant in $\overline{\mathrm{H}}_1 \bigcap \overline{\mathrm{H}}_2$. In fact, Figure 2.6 shows that $y(1,1,1|n_1,n_2)$ behaves like $r_k(n_1,n_2)$. That is, $y(i_0,i_1,i_2|n_1,n_2)$ coincides with a constant $c(i_0,i_1,i_2)$ in $\mathrm{H}_1 \bigcap \mathrm{H}_2$, and with the other constant $\overline{c}(i_0,i_1,i_2)$ in $\overline{\mathrm{H}}_1 \bigcap \overline{\mathrm{H}}_2$. Furthermore, we can see that both $c(i_0,i_1,i_2)$ and $\overline{c}(i_0,i_1,i_2)$ are decomposed into

three components:

$$y(i_0, i_1, i_2|n_1, n_2) \approx \begin{cases} c_0(i_0)c_1(i_1)c_2(i_2) & \text{if } (n_1, n_2) \in \mathrm{H}_1 \bigcap \mathrm{H}_2, \\ \overline{c}_0(i_0)\overline{c}_1(i_1)\overline{c}_2(i_2) & \text{if } (n_1, n_2) \in \overline{\mathrm{H}}_1 \bigcap \overline{\mathrm{H}}_2. \end{cases} \tag{2.7}$$

Here $c_0(i_0)$ or $\overline{c}_0(i_0)$ can be regarded as the asymptotic conditional probability of the phase i_0 of the arrival process, and $c_k(i_k)$ or $\overline{c}_k(i_k)$ $(k = 1, 2)$ as the asymptotic conditional probability of the phase i_k of the service process at the kth stage. This property can be validated by the numerical results of $y(i_0, i_1, i_2 \mid n_1, n_2)$ together with the values of $c_k(i_k)$'s and $\overline{c}_k(i_k)$'s given by (2.16) in Section 2.5. The values of them for $E_2/H_2/1 \to /E_2/1$ are listed in Tables 2.4 and 2.5. It is shown that $y(i_0, i_1, i_2 \mid 90, 10)$ almost coincides with $c_0(i_0)c_1(i_1)c_2(i_2)$, and that $y(i_0, i_1, i_2 \mid 10, 90)$ almost coincides with $\overline{c}_0(i_0)\overline{c}_1(i_1)\overline{c}_2(i_2)$ for all combinations of (i_0, i_1, i_2). Note that $(90, 10) \in \mathrm{H}_1 \cap \mathrm{H}_2$ while $(10, 90) \in \overline{\mathrm{H}}_1 \cap \overline{\mathrm{H}}_2$.

Table 2.4: $c_k(i_k)$ and $\overline{c}_k(i_k)$ in $E_2/H_2/1 \to /E_2/1$ with $\rho_1 = 0.6$, $\rho_2 = 0.8$

i_k	$c_0(i_0)$	$c_1(i_1)$	$c_2(i_2)$	$\overline{c}_0(i_0)$	$\overline{c}_1(i_1)$	$\overline{c}_2(i_2)$
1	0.0576	0.0546	0.4351	0.5967	0.0440	0.4444
2	0.4243	0.9454	0.5649	0.4033	0.9560	0.5556

2.4.4 Variation of Regions and Bands

Now we shall see how the regions H_k, $\overline{\mathrm{H}}_k$ and the band B_k vary according to the traffic intensities ρ_1 and ρ_2.

Figure 2.7 shows the graphs of $r_1(n_1, n_2)$ for the model $E_2/H_2/1 \to /E_2/1$ with fixed $\rho_1 = 0.6$ and varying $\rho_2 = 0.4 \sim 0.9$. When ρ_2 is small, $\overline{\eta}_1$ is far below η_1 and $r_1(n_1, n_2)$

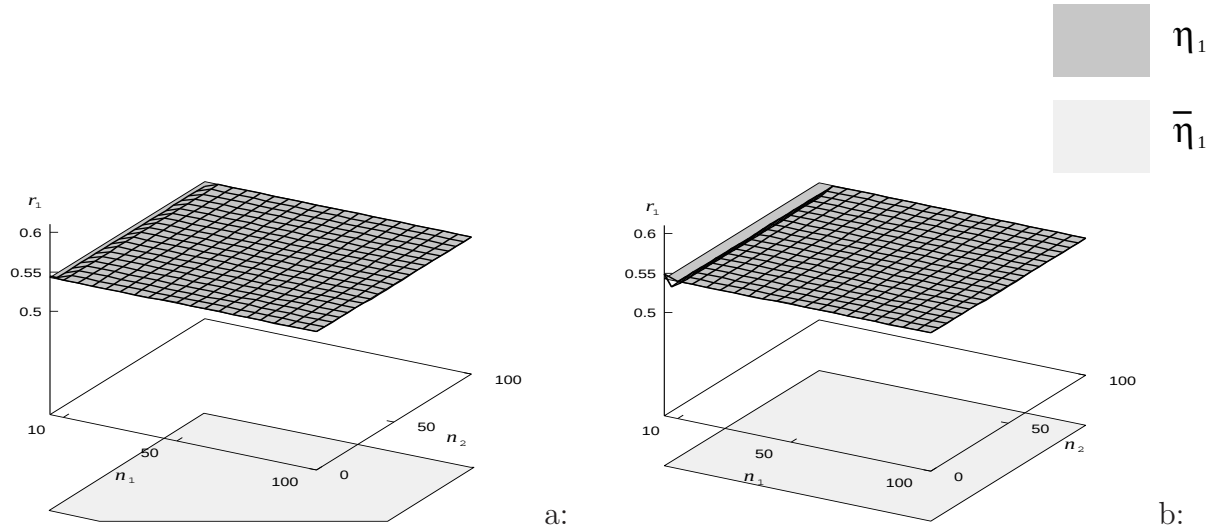
Table 2.5: $c_k(i_k)$ and $\overline{c}_k(i_k)$ in $E_2/H_2/1 \to /E_2/1$ with $\rho_1 = 0.6$, $\rho_2 = 0.8$

(i_0, i_1, i_2)	$c_0(i_0)c_1(i_1)c_2(i_2)$	$y(i_0, i_1, i_2 \mid 90, 10)$	$\overline{c}_0(i_0)\overline{c}_1(i_1)\overline{c}_2(i_2)$	$y(i_0, i_1, i_2 \mid 10, 90)$
$(1,1,1)$	1.368×10^{-2}	1.368×10^{-2}	1.166×10^{-2}	1.166×10^{-2}
$(1,1,2)$	1.776×10^{-2}	1.776×10^{-2}	1.457×10^{-2}	1.458×10^{-2}
$(1,2,1)$	2.368×10^{-1}	2.368×10^{-1}	2.535×10^{-1}	2.535×10^{-1}
$(1,2,2)$	3.074×10^{-1}	3.074×10^{-1}	3.169×10^{-1}	3.169×10^{-1}
$(2,1,1)$	1.008×10^{-2}	1.008×10^{-2}	7.882×10^{-3}	7.886×10^{-3}
$(2,1,2)$	1.309×10^{-2}	1.309×10^{-2}	9.853×10^{-3}	9.858×10^{-3}
$(2,2,1)$	1.745×10^{-1}	1.745×10^{-1}	1.714×10^{-1}	1.714×10^{-1}
$(2,2,2)$	2.266×10^{-1}	2.266×10^{-1}	2.142×10^{-1}	2.142×10^{-1}

coincides with η_1 almost on the whole n_1-n_2 plane. This means that H_1 covers the whole n_1-n_2 plane. The larger ρ_2 becomes, the closer $\overline{\eta}_1$ comes to η_1, and when $\rho_2 = 0.8$ the region $\overline{\mathrm{H}}_1$ appears. $\overline{\mathrm{H}}_1$ becomes larger than H_1 when $\rho_2 = 0.9$.

Figure 2.8 shows the corresponding graphs of $r_2(n_1, n_2)$. In these graphs we also see that $\overline{\mathrm{H}}_2$ appears only when $\rho_2 = 0.8$ and 0.9. Now we shall see the movement of the planes η_2 and $\overline{\eta}_2$. When ρ_2 is small, the plane $\overline{\eta}_2$ is far below the plane η_2. As ρ_2 becomes larger the plane $\overline{\eta}_2$ becomes closer to the plane η_2, but still below the plane η_2 for $\rho_2 \leq 0.7$. When ρ_2 becomes to 0.8, the plane $\overline{\eta}_2$ comes above the plane η_2, and at the same time the region $\overline{\mathrm{H}}_2$ appears. If we denote by $\tilde{\rho}_2$ the value of ρ_2 at which $\eta_2 = \overline{\eta}_2$, this seems to indicate that the region $\overline{\mathrm{H}}_2$ disappears when ρ_2 is less than $\tilde{\rho}_2$ and $\overline{\mathrm{H}}_2$ appears when ρ_2 exceeds $\tilde{\rho}_2$.

$\overline{\mathrm{H}}_1$ seems to appear at the same time as $\overline{\mathrm{H}}_2$. We will see this more in detail. Figure 2.9 shows the movement of the bands B_1 and B_2. The bands bounded by thin lines indicate



a: $\rho_2 = 0.4$ b: $\rho_2 = 0.5$

c: $\rho_2 = 0.6$ d: $\rho_2 = 0.7$

e: $\rho_2 = 0.8$ f: $\rho_2 = 0.9$

Figure 2.7: r_1 behavior in $E_2/H_2/1 \rightarrow /E_2/1$ $(\rho_1 = 0.6)$

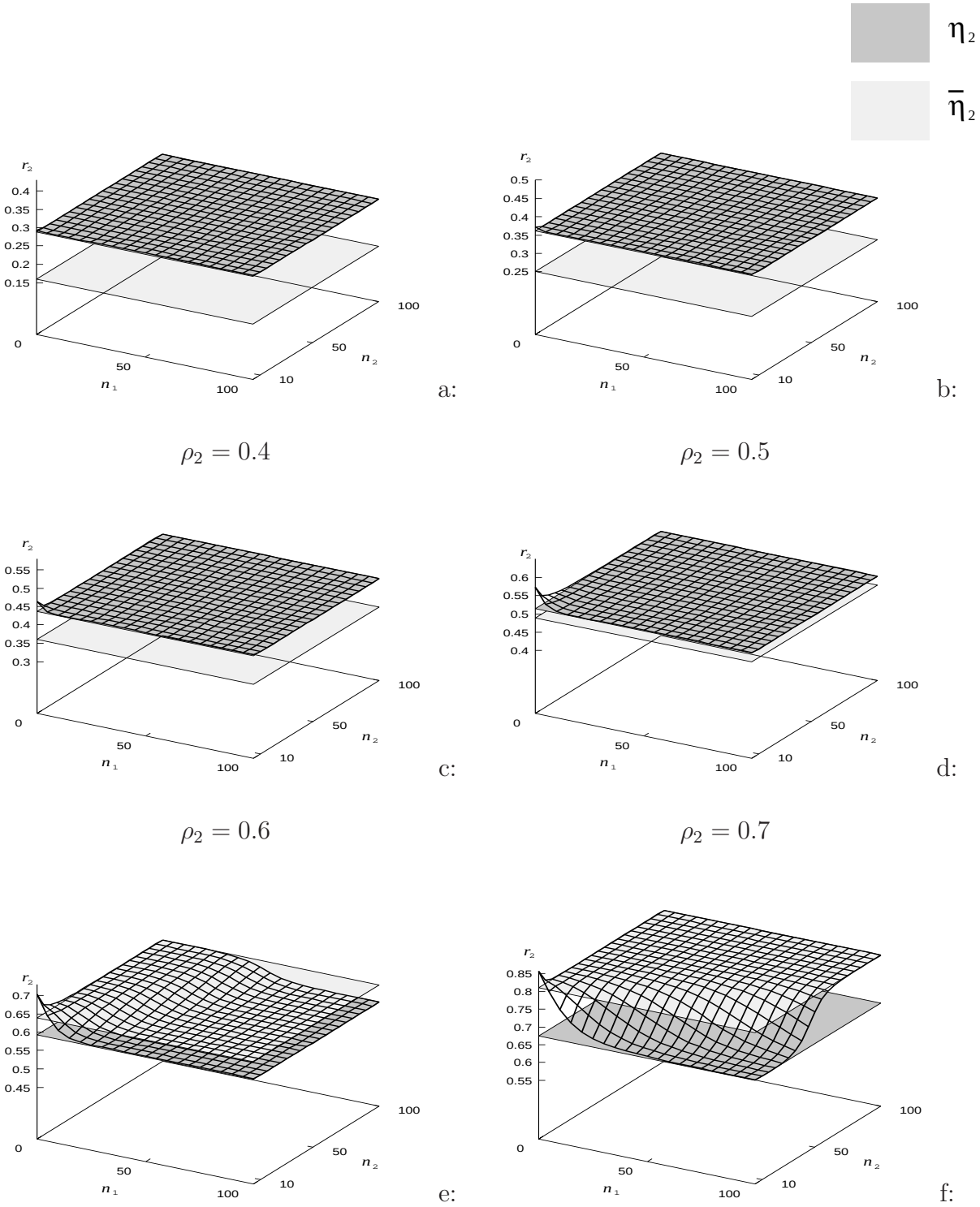


Figure 2.8: r_2 behavior in $E_2/H_2/1 \rightarrow /E_2/1$ $(\rho_1 = 0.6)$

B_1 for $\rho_2 = 0.75 \sim 0.90$, and those bounded by bold lines indicate B_2.

It is likely that these behaviors of the bands are related with those of η_k and $\overline{\eta}_k$. In Figure 2.5, both $g(n_1, n_2)$ and $\overline{g}(n_1, n_2)$ diverge to $+\infty$ when they are outside of $H_1 \cap H_2$ or $\overline{H}_1 \cap \overline{H}_2$. This indicates that $\eta_1^{n_1}\eta_2^{n_2}$ is greater than $\overline{\eta}_1^{n_1}\overline{\eta}_2^{n_2}$ in the region $H_1 \cap H_2$, and that $\overline{\eta}_1^{n_1}\overline{\eta}_2^{n_2}$ is greater than $\eta_1^{n_1}\eta_2^{n_2}$ in the region $\overline{H}_1 \cap \overline{H}_2$.

Hence we can guess that the bands B_1 and B_2 lie on the line determined by $\eta_1^{n_1}\eta_2^{n_2} = \overline{\eta}_1^{n_1}\overline{\eta}_2^{n_2}$, or equivalently,

$$n_2 = \tilde{a} n_1 \quad \text{with} \quad \tilde{a} = -\frac{\log \eta_1/\overline{\eta}_1}{\log \overline{\eta}_2/\eta_2}. \tag{2.8}$$

Each broken line in Figure 2.9 indicates this line for $\rho_2 = 0.75 \sim 0.90$. One can see that the line moves almost together with the bands.

2.5 Equations for η_k, $\overline{\eta}_k$ and $\tilde{\rho}_2$

The properties discussed in the previous section are for the specific model $E_2/H_2/1 \rightarrow /E_2/1$. We can see similar properties in most of cases.

To the author's knowledge, there are no papers which investigate such tail properties theoretically. However, through the study, they have a strong confidence that the values of η_k and $\overline{\eta}_k$ $(k = 1, 2)$ are given as solutions of some simultaneous equations written with LSTs of interarrival and service time distributions. In this section we present the equations and summarize some basic properties of η_k and $\overline{\eta}_k$. The assumptions to derive the equations and a brief explanation of the derivation process are given at the end of this

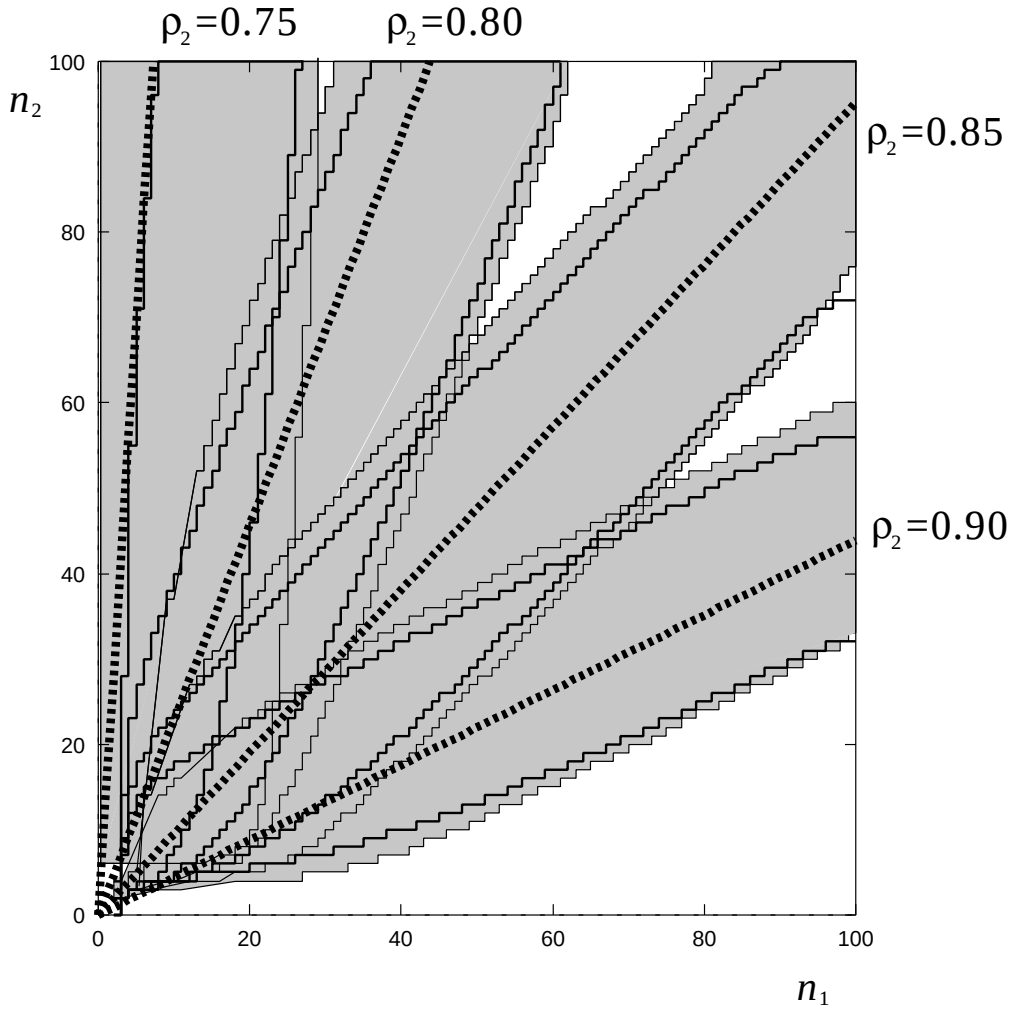


Figure 2.9: Movement of B_1 and B_2 in $E_2/H_2/1 \to /E_2/1$ with $\rho_1 = 0.6$

section.

Let $T^*(s)$ be the LST of the interarrival time distribution $PH(\boldsymbol{\alpha}, \boldsymbol{T})$. Denoting by $-\tau < 0$ an abscissa of convergence of T^*, the function $T^*(s)$ is then defined, positive and convex decreasing on the interval $(-\tau, \infty)$ [31]. Similarly, we denote by $S_k^*(s)$ the LST of the service time distribution $PH(\boldsymbol{\beta}, \boldsymbol{S}_k)$ for the k-th stage, and by $-\sigma_k < 0$ an abscissa of

convergence of S_k^*. Then $S_k^*(s)$ is defined, positive and convex decreasing on the interval $(-\sigma_k, \infty)$. When we mention about roots of an equation, we refer only real roots in the domain of the equation and count the number of roots by taking the multiplicity into account. For example, every double root is counted twice.

Consider the simultaneous equations

$$\begin{cases} T^*(s_0)S_1^*(-s_0) &= 1, \\ T^*(s_0)S_1^*(s_1)S_2^*(s_2) &= 1, \\ s_0 + s_1 + s_2 &= 0. \end{cases} \tag{2.9}$$

Let $f(s_0) = T^*(s_0)S_1^*(-s_0)$. Then $f(s_0)$ is a convex function of s_0 on $(-\tau, \sigma_1)$, and $f(0) = 1$. Hence the equation $f(s_0) = 1$ has two roots one of which is $s_0 = 0$. The derivative of $f(s_0)$ at $s_0 = 0$ is negative since $f'(0) = T^{*'}(0) - S_1^{*'}(0) = -1/\lambda + 1/\mu_1 = -(1-\rho_1)/\lambda < 0$. Hence the other root $s_0 = \omega_0$ is positive.

For the second equation of (2.9), eliminating s_1 by the third equation and inserting $s_0 = \omega_0$, we have

$$g(s_2) = T^*(\omega_0)S_1^*(-\omega_0 - s_2)S_2^*(s_2) = 1. \tag{2.10}$$

$g(s_2)$ is a convex function of s_2 on $(-\sigma_2, \sigma_1 - \omega_0)$. The equation (2.10) has a trivial root $s_2 = 0$, and hence it has one more root $s_2 = \omega_2$. We set $\omega_1 = -\omega_0 - \omega_2$. This triplet $(\omega_0, \omega_1, \omega_2)$ is our desired solution of (2.9), and we let

$$\eta_1 = T^*(\omega_0), \quad \eta_2 = \frac{1}{S_2^*(\omega_2)}. \tag{2.11}$$

Since ω_0 is positive, η_1 is strictly less than 1. Furthermore, η_1 is a monotone increasing function of ρ_1, and $\eta_1 \downarrow 0$ as $\rho_1 \downarrow 0$ while $\eta_1 \uparrow 1$ as $\rho_1 \uparrow 1$. On the other hand, ω_2 is negative

if ρ_2 is small but it may be positive if ρ_2 becomes large, and hence η_2 may exceed 1. η_2 can be regarded as a function of both ρ_1 and ρ_2. As a function of ρ_2, η_2 is monotone increasing and $\eta_2 \downarrow 0$ as $\rho_2 \downarrow 0$.

For $\overline{\eta}_k$'s, consider the simultaneous equations

$$\begin{cases} T^*(-\overline{s}_2)S_2^*(\overline{s}_2) &= 1, \\ T^*(\overline{s}_0)S_1^*(\overline{s}_1)S_2^*(\overline{s}_2) &= 1, \\ \overline{s}_0 + \overline{s}_1 + \overline{s}_2 &= 0. \end{cases} \tag{2.12}$$

A similar argument for the equations (2.9) can be applied to (2.12). The first equation defined on $(-\sigma_2, \tau)$ has two roots, zero and $\overline{\omega}_2$. Since $\rho_2 < 1$, $\overline{\omega}_2$ is strictly negative. By inserting $\overline{s}_2 = \overline{\omega}_2$ and $\overline{s}_1 = -\overline{s}_0 - \overline{\omega}_2$ into the left hand side of the second equation, we have an equation for $\overline{s}_2 = \overline{\omega}_2$. The left hand side of this equation is a convex function, and hence the equation has two roots. One is a trivial root $\overline{s}_0 = -\overline{\omega}_2$, and we denote the other as $\overline{\omega}_0$. We set $\overline{\omega}_1 = -\overline{\omega}_0 - \overline{\omega}_2$. This triplet $(\overline{\omega}_0, \overline{\omega}_1, \overline{\omega}_2)$ is the desired solution of (2.12). For this triplet, we set

$$\overline{\eta}_2 = \frac{1}{S_2^*(\overline{\omega}_2)}, \quad \overline{\eta}_1 = T^*(\overline{\omega}_0). \tag{2.13}$$

Since $\overline{\omega}_2$ is negative, $\overline{\eta}_2$ is less than 1. Moreover, $\overline{\eta}_2$ is a monotone increasing function of ρ_2, and $\overline{\eta}_2 \downarrow 0$ as $\rho_2 \downarrow 0$ while $\overline{\eta}_2 \uparrow 1$ as $\rho_2 \uparrow 1$. $\overline{\omega}_0$ may take positive or negative value, and hence $\overline{\eta}_1$ may be greater than 1. $\overline{\eta}_1$ is a function of both ρ_1 and ρ_2, and $\overline{\eta}_1 \to \eta_1$ as $\rho_1 \uparrow 1$.

Using η_k and $\overline{\eta}_k$ above, the slope $\tilde{a}$ of bands B_k is given by (2.8) in most of the models. In Section 2.4.4, we defined $\tilde{\rho}_2$ as ρ_2 at which $\eta_2 = \overline{\eta}_2$. However, this definition is not

suitable in a general situation since such $\tilde{\rho}_2$ may not be unique. We can show that, for fixed ρ_1, there exists a unique ρ_2 such that

$$\eta_1 = \eta_2 = \overline{\eta}_2. \tag{2.14}$$

This ρ_2 is suitable for $\tilde{\rho}_2$. When $\tilde{a}$ is well-defined for all $\rho_2 \in (0,1)$ except for the point $\rho_2 = \tilde{\rho}_2$, $\tilde{a}$ is strictly negative for $\rho_2 < \tilde{\rho}_2$ and is strictly positive for $\rho_2 > \tilde{\rho}_2$. Further, usually $\tilde{a} \uparrow +\infty$ as $\rho_2 \downarrow \tilde{\rho}_2$, and $\tilde{a} \downarrow 0$ as $\rho_2 \uparrow 1$.

Note that $\tilde{a}$ in (2.8) cannot be defined for $E_j/E_j/1 \to /E_j/1$ with $j = 1, 2, \ldots$, in which $\eta_1 = \overline{\eta}_1 = \rho_1^j$ and $\eta_2 = \overline{\eta}_2 = \rho_2^j$. In this case, a perturbation analysis indicates that

$$\tilde{\rho}_2 = \rho_2 \quad \text{and} \quad \tilde{a} = \frac{1-\rho_2}{\rho_1 - \rho_2} \tag{2.15}$$

are plausible definitions for these models.

For $c_k(i_k)$, we have the following expressions. Let $\boldsymbol{c}_k$ be the stochastic vector whose i_k-th element is $c_k(i_k)$.

$$\begin{aligned} \boldsymbol{c}_0 &= \frac{\omega_0}{1-\eta_1}\boldsymbol{\alpha}(\omega_0 \boldsymbol{I} - \boldsymbol{T})^{-1}, \\ \boldsymbol{c}_1 &= \frac{\omega_1}{1-\eta_2/\eta_1}\boldsymbol{\beta}_1(\omega_1 \boldsymbol{I} - \boldsymbol{S}_1)^{-1}, \\ \boldsymbol{c}_2 &= \frac{\omega_2}{1-1/\eta_2}\boldsymbol{\beta}_2(\omega_2 \boldsymbol{I} - \boldsymbol{S}_2)^{-1}. \end{aligned} \tag{2.16}$$

For $\overline{c}_k(i_k)$, we have similar expressions with $\overline{\omega}_k$ and $\overline{\eta}_k$ in places of ω_k and η_k, respectively.

The derivation process of the equations in (2.9) and (2.12) is rather complicated. Here we give basic assumptions for the equations and briefly outline the derivation process. The complete derivation process and related discussions will be given in Chapter 4.

For (2.9), we assume a geometric decay property for large n_1: There exists a constant η_1 such that

$$x(n_1, n_2; i_0, i_1, i_2) = \eta_1 x(n_1 - 1, n_2; i_0, i_1, i_2) \quad \text{for } \forall i_0, i_1, i_2 \text{ and } n_2 = 0, 1, 2, \ldots. \tag{2.17}$$

For fixed n_1 we consider the balance equation around the state $(n_1, n_2; i_0, i_1, i_2)$. We can use (2.17) to eliminate $x(n_1 - 1, *; *, *, *)$'s and $x(n_1 + 1, *; *, *, *)$'s in the equation so that the equation contains only $x(n_1, *; *, *, *)$'s. By considering such balance equations for all i_0, i_1, i_2 and $n_2 = 0, 1, 2, \ldots$, we can show that $x(n_1, n_2; i_0, i_1, i_2)$ takes a matrix-geometric form. If we use a similar technique to the one in [45], we obtain the relations in (2.9), (2.11) and (2.16). Thus, roughly speaking, if the stationary distribution has geometric tail in the direction of n_1-axis, the decay parameters are given by (2.11).

On the other hand, to derive the second and the third equations in (2.12) we assume for large n_2 that there exists a constant $\overline{\eta}_2$ such that

$$x(n_1, n_2; i_0, i_1, i_2) = \overline{\eta}_2 x(n_1, n_2 - 1; i_0, i_1, i_2) \quad \text{for } \forall i_0, i_1, i_2 \text{ and } n_1 = 0, 1, 2, \ldots.$$

Furthermore, to derive the first equation in (2.12), we have to assume that $x(n_1, n_2; i_0, i_1, i_2)$ is of the form $x_1(n_1; i_0, i_1)x_2(n_2, i_2)$. A similar process to the one above leads us to the relations in (2.12), (2.13) and $\overline{c}_k$ version of (2.16). In case of (2.9), the assumption of asymptotic product form is not necessary because instead we can use the fact that the behavior of the first stage is never affected by the second stage.

Thus, the above equations for the characteristic constants of the tail probabilities are derived from geometric decay assumptions and some product form assumptions. However,

the multiplicative constants G and $\overline{G}$ cannot be obtained in this line of derivation. To get the values of them, we have to execute some simulations or to solve the balance equations numerically. In the latter case, if we use the geometric decay property above, we can get the values of them with much smaller computational burden.

Chapter 3

Asymptotic Properties in a Quasi-Birth-and-Death Process with a Countable Number of Phases

3.1 Introduction

We consider a discrete-time Markov chain $\{X(n)\}$ on a state space $\mathcal{S} = \{(m, i) \,|\, m, i = 0, 1, 2, \cdots\}$ having a transition probability matrix of the block-tridiagonal form

$$\boldsymbol{P} = \begin{pmatrix} \boldsymbol{B}_0 & \boldsymbol{A}_0 & & & \\ \boldsymbol{C}_1 & \boldsymbol{B} & \boldsymbol{A} & & \\ & \boldsymbol{C} & \boldsymbol{B} & \boldsymbol{A} & \\ & & \boldsymbol{C} & \boldsymbol{B} & \ddots \\ & & & \ddots & \ddots \end{pmatrix} \tag{3.1}$$

after partitioning the state space into levels $\mathcal{L}_m = \{(m,i)\,|\,i = 0,1,2,\cdots\}$, $m = 0,1,2,\cdots$. Such a chain is called a quasi-birth-and-death (QBD) process with a countable number of phases in each level. We assume that the chain is irreducible and positive recurrent, and has the stationary probabilities $\pi(m,i),\ (m,i) \in \mathcal{S}$.

It is known [28] that the stationary probability vector $\boldsymbol{\pi} = (\,\boldsymbol{\pi}_m,\ m = 0,1,2,\dots\,)$ with subvectors $\boldsymbol{\pi}_m = (\,\pi(m,i),\ (m,i) \in \mathcal{L}_m\,)$ has a matrix-geometric form

$$\boldsymbol{\pi}_m = \boldsymbol{\pi}_1\, \boldsymbol{R}^{m-1}, \qquad m \geq 1, \tag{3.2}$$

where the *rate matrix* $\boldsymbol{R}$ is given as the minimal nonnegative solution of the matrix-quadratic equation

$$\boldsymbol{R} = \boldsymbol{A} + \boldsymbol{R}\,\boldsymbol{B} + \boldsymbol{R}^2\,\boldsymbol{C}. \tag{3.3}$$

If the number of phases in each level is finite, various properties on the rate matrix and stationary probabilities are known [30]. For example, the stationary distribution $\boldsymbol{\pi}$ has a geometric tail

$$\boldsymbol{\pi}_m \sim c\,\eta^m\,\boldsymbol{x} \qquad \text{as} \quad m \to \infty, \tag{3.4}$$

where η is the Perron-Frobenius eigenvalue of $\boldsymbol{R}$, $\boldsymbol{x}$ is a left eigenvector of $\boldsymbol{R}$ associated with η, and c is a multiplicative constant. This asymptotic property was used, for example, to investigate tail behaviors of the queue-length and waiting-time distributions in a $PH/PH/c$ queue [45] or more generally in a $GI/PH/c$ queue with versatile service distributions [33].

The purpose of this chapter is to present a sufficient condition for the asymptotic property (3.4) to hold in the case with a countable number of states in each level. This

result will be applied to a tandem queueing system $PH/PH/1 \to /PH/1$ to investigate tail properties of the joint queue-length distribution.

The remainder of this chapter is constructed as follows. We present our main theorems and prove them in the next section, and apply them to a continuous-time QBD process in Section 3.3.

3.2 Main Theorems

Before stating our main theorems, we introduce some notations. We denote by $\mathbf{0}$ and $\boldsymbol{O}$ a zero vector and a zero matrix, by $\boldsymbol{I}$ an identity matrix, and by $\boldsymbol{e}$ a column vector with elements all equal to one. Inequalities and limits are applied elementwise. Let $\mathcal{M}_k = \cup_{l=k}^{\infty} \mathcal{L}_l$, and

$$p_{i,j}^{L}(n) \;=\; P\{\, X(n) = (m,j),\, X(u) \in \mathcal{L}_m,\; 0 \le u \le n \mid X(0) = (m,i)\,\},$$

$$p_{i,j}^{M}(n) \;=\; P\{\, X(n) = (m,j),\, X(u) \in \mathcal{M}_m,\; 0 \le u \le n \mid X(0) = (m,i)\,\},$$

for $i, j \in \mathcal{L}_m,\ m \ge 2$ and $n = 0, 1, 2, \ldots$. From the QBD structure (3.1) these probabilities are independent of m. We set

$$v_{i,j} = \sum_{n=0}^{\infty} p_{i,j}^{L}(n) \quad \text{and} \quad u_{i,j} = \sum_{n=0}^{\infty} p_{i,j}^{M}(n).$$

These values are interpreted as follows: $v_{i,j}$ is the mean number of visits to state (m,j) of the chain starting from state (m,i) before it goes out of $\mathcal{L}_m$, and $u_{i,j}$ is the mean number

of visits to state (m, j) of the chain starting from state (m, i) before it goes down to level $m-1$ $(m > 1)$. Since the chain is positive recurrent, these values are all finite and non-negative. Let $\boldsymbol{V}$ and $\boldsymbol{U}$ be matrices with elements v_{ij} and u_{ij} respectively. Then we can easily see that $\boldsymbol{V}\boldsymbol{B} = \boldsymbol{B}\boldsymbol{V} = \boldsymbol{V} - \boldsymbol{I}$. Hence the inverse of the infinite matrix $\boldsymbol{I} - \boldsymbol{B}$ exists and given by $\boldsymbol{V}$. Hereafter we will write $(\boldsymbol{I} - \boldsymbol{B})^{-1}$ instead of $\boldsymbol{V}$.

The (i, j)th element of $\boldsymbol{R}$ is interpreted as the mean number that the chain which transits from state $(m-1, i)$ into somewhere in level m visits state (m, j) before it comes back to level $m-1$. A probabilistic argument shows that $\boldsymbol{R}$ is given by $\boldsymbol{A}\boldsymbol{U}$.

The so-called G-matrix $\boldsymbol{G}$ is defined by $\boldsymbol{U}\boldsymbol{C}$. Like $\boldsymbol{R}$, the matrix $\boldsymbol{G}$ is the minimal non-negative solution of the matrix-quadratic equation

$$\boldsymbol{G} = \boldsymbol{A}\,\boldsymbol{G}^2 + \boldsymbol{B}\,\boldsymbol{G} + \boldsymbol{C}. \tag{3.5}$$

We can see that $\boldsymbol{G}\boldsymbol{e} = \boldsymbol{U}\boldsymbol{C}\boldsymbol{e} = \boldsymbol{e}$ from the positive recurrence of the chain.

The followings are our main theorems.

Theorem 3.2.1. Assume that there exist a positive constant $\eta\,(< 1)$, a positive row vector $\boldsymbol{x}$ and a positive column vector $\boldsymbol{y}$ such that

(i) $$\boldsymbol{x}\,(\eta^{-1}\,\boldsymbol{A} + \boldsymbol{B} + \eta\,\boldsymbol{C}) = \boldsymbol{x}, \tag{3.6}$$

(ii) $$(\eta^{-1}\,\boldsymbol{A} + \boldsymbol{B} + \eta\,\boldsymbol{C})\,\boldsymbol{y} = \boldsymbol{y}, \tag{3.7}$$

(iii) $$\eta^{-1}\,\boldsymbol{x}\,\boldsymbol{A}\,\boldsymbol{y} \neq \eta\,\boldsymbol{x}\,\boldsymbol{C}\,\boldsymbol{y}, \tag{3.8}$$

(iv) $$\boldsymbol{x}\,\boldsymbol{e} < \infty. \text{ and } \boldsymbol{x}\,\boldsymbol{y} < \infty. \tag{3.9}$$

Then,

(a) $\boldsymbol{x}\,\boldsymbol{R} = \eta\,\boldsymbol{x}$, and

(b) $\boldsymbol{z} = \boldsymbol{A}\,(\eta^{-1}\,\boldsymbol{I} - \boldsymbol{G})\,\boldsymbol{y}$ is a non-zero, non-negative vector satisfying $\boldsymbol{R}\,\boldsymbol{z} = \eta\,\boldsymbol{z}$.

Let $\mathcal{L}_m(A)$ be the set of states in $\mathcal{L}_m$ corresponding to positive elements of the vector $\boldsymbol{A}\boldsymbol{e}$. Since $\boldsymbol{R} = \boldsymbol{A}\,\boldsymbol{U}$, the i-th row of $\boldsymbol{R}$ is a zero vector if $(m, i) \notin \mathcal{L}_m(A)$. In order to discuss asymptotic properties of $\boldsymbol{R}^m$, we assume that the submatrix $\boldsymbol{R}(A)$ of $\boldsymbol{R}$ corresponding to the index set $\mathcal{L}_m(A) \times \mathcal{L}_m(A)$ is irreducible. A sufficient condition for $\boldsymbol{R}(A)$ being irreducible is that the matrix

$$\begin{pmatrix} \boldsymbol{B} & \boldsymbol{A} & & \\ \boldsymbol{C} & \boldsymbol{B} & \boldsymbol{A} & \\ & \boldsymbol{C} & \boldsymbol{B} & \ddots \\ & & \ddots & \ddots \end{pmatrix} \tag{3.10}$$

formed from $\boldsymbol{P}$ by deleting rows and columns corresponding to states in $\mathcal{L}_0$ is irreducible. In fact, if the matrix (3.10) is irreducible, all the elements of $\boldsymbol{U}$ are positive, and hence the rows of $\boldsymbol{R}(A)$ are all positive. If $\boldsymbol{R}(A)$ is irreducible and if (a) and (b) in Theorem 3.2.1 hold, then $\boldsymbol{R}(A)$ is η-positive in the sense of Seneta [39] and we have

$$\boldsymbol{R}^m \sim \eta^m\,\boldsymbol{z}\,\boldsymbol{x} \qquad \text{as} \quad m \to \infty$$

assuming vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ are normalized so that $\boldsymbol{x}\,\boldsymbol{z} = 1$ (the finiteness of the product $\boldsymbol{x}\boldsymbol{z}$ is easily proved from (3.6) and (3.9)). Thus the following theorem is a direct consequence of Theorem 3.2.1 and the matrix-geometric form representation (3.2).

Theorem 3.2.2. Suppose that the conditions in Theorem 3.2.1 hold. Then, if $\boldsymbol{R}(A)$ is irreducible and $\boldsymbol{\pi}_1 \boldsymbol{z} < \infty$, the stationary probability vector $\boldsymbol{\pi}$ of the Markov chain $\{X(n)\}$ has a geometric tail:

$$\boldsymbol{\pi}_m \sim c\,\eta^m \boldsymbol{x} \qquad \text{as} \quad m \to \infty,$$

where c is a multiplicative constant. A sufficient condition for the irreducibility of $\boldsymbol{R}(A)$ is that the matrix (3.10) is irreducible, and a sufficient condition for $\boldsymbol{\pi}_1 \boldsymbol{z}$ to be finite is that $\boldsymbol{\pi}_1 \boldsymbol{y} < \infty$.

The statement (a) of Theorem 3.2.1 was first proved in Ramaswami & Taylor [36] as a special case of a QBD process with level dependent transition probabilities. For reader's convenience, here we give a direct proof of it together with a proof of the statement (b). In the following lemmas, we assume that the conditions of Theorem 3.2.1 hold.

Lemma 3.2.1. Products $\boldsymbol{x}\,\boldsymbol{R}$, $\boldsymbol{x}\,\boldsymbol{G}$, $\boldsymbol{R}\,\boldsymbol{y}$ and $\boldsymbol{G}\,\boldsymbol{y}$ are all finite, and

$$\boldsymbol{x}\,(\eta\,\boldsymbol{I} - \boldsymbol{R}) \;\geq\; \boldsymbol{0}, \tag{3.11}$$

$$(\eta^{-1}\,\boldsymbol{I} - \boldsymbol{G})\,\boldsymbol{y} \;\geq\; \boldsymbol{0}. \tag{3.12}$$

Proof The rate matrix $\boldsymbol{R}$ is given as the limit of an increasing sequence of matrices $\{\boldsymbol{R}^{(k)}\}$ defined by

$$\boldsymbol{R}^{(0)} = \boldsymbol{O}, \quad \boldsymbol{R}^{(k)} = \boldsymbol{A} + \boldsymbol{R}^{(k-1)}\boldsymbol{B} + \{\boldsymbol{R}^{(k-1)}\}^2\,\boldsymbol{C} \qquad \text{for} \quad k = 1, 2, \ldots.$$

Then an induction using Equation (3.3) proves that $\boldsymbol{x}\,\boldsymbol{R}^{(k)} \leq \eta\,\boldsymbol{x}$ for all k. Letting $k \to \infty$, we have (3.11). A similar argument using Equation (3.5) proves (3.12). From (3.11) and (3.9), $\boldsymbol{x}\,\boldsymbol{R}\,\boldsymbol{y} \leq \eta\,\boldsymbol{x}\,\boldsymbol{y} < \infty$. Hence products $\boldsymbol{x}\,\boldsymbol{R}$ and $\boldsymbol{R}\,\boldsymbol{y}$ are finite. Similarly we can see the finiteness of the products $\boldsymbol{x}\,\boldsymbol{G}$ and $\boldsymbol{G}\,\boldsymbol{y}$ from (3.12) and (3.9). □

Lemma 3.2.2.

$$\boldsymbol{x}\,(\eta\,\boldsymbol{I} - \boldsymbol{R})\,\boldsymbol{C}\,(\eta^{-1}\boldsymbol{I} - \boldsymbol{G}) \;=\; \boldsymbol{0}, \tag{3.13}$$

$$(\eta\,\boldsymbol{I} - \boldsymbol{R})\,\boldsymbol{z} \;=\; (\eta\,\boldsymbol{I} - \boldsymbol{R})\,\boldsymbol{A}\,(\eta^{-1}\boldsymbol{I} - \boldsymbol{G})\,\boldsymbol{y} \;=\; \boldsymbol{0} \tag{3.14}$$

Proof. From (3.6) we have

$$\boldsymbol{0} \;=\; \boldsymbol{x}\,(\eta^{-1}\boldsymbol{I} - \eta^{-2}\boldsymbol{A} - \eta^{-1}\boldsymbol{B} - \boldsymbol{C}), \tag{3.15}$$

and from (3.5)

$$\boldsymbol{0} \;=\; \boldsymbol{x}\,(\boldsymbol{G} - \boldsymbol{A}\,\boldsymbol{G}^2 - \boldsymbol{B}\,\boldsymbol{G} - \boldsymbol{C}). \tag{3.16}$$

Note that the production operation in (3.16) of the vector $\boldsymbol{x}$ and a matrix within parentheses is legitimate since $\boldsymbol{x}\,\boldsymbol{G}$ is finite from Lemma 3.2.1. Subtracting the right hand side of (3.16) from that of (3.15) we have

$$\begin{aligned}
\boldsymbol{0} \;&=\; \boldsymbol{x}\,(\eta^{-1}\boldsymbol{I} - \eta^{-2}\boldsymbol{A} - \eta^{-1}\boldsymbol{B} - \boldsymbol{C}) \;-\; \boldsymbol{x}\,(\boldsymbol{G} - \boldsymbol{A}\,\boldsymbol{G}^2 - \boldsymbol{B}\,\boldsymbol{G} - \boldsymbol{C}) \\
&=\; \boldsymbol{x}\,\{-\boldsymbol{A}\,(\eta^{-2}\boldsymbol{I} - \boldsymbol{G}^2) + (\boldsymbol{I} - \boldsymbol{B})\,(\eta^{-1}\boldsymbol{I} - \boldsymbol{G})\}
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{x}\,\{-\boldsymbol{A}(\eta^{-1}\boldsymbol{I}+\boldsymbol{G})(\eta^{-1}\boldsymbol{I}-\boldsymbol{G})+(\boldsymbol{I}-\boldsymbol{B})\,(\eta^{-1}\boldsymbol{I}-\boldsymbol{G})\}\\
&= \boldsymbol{x}\,(-\eta^{-1}\boldsymbol{A}-\boldsymbol{A}\,\boldsymbol{G}+\boldsymbol{I}-\boldsymbol{B})\,(\eta^{-1}\boldsymbol{I}-\boldsymbol{G})\\
&= \boldsymbol{x}(-\eta^{-1}\boldsymbol{A}-\boldsymbol{R}\,\boldsymbol{C}+\boldsymbol{I}-\boldsymbol{B})\,(\eta^{-1}\boldsymbol{I}-\boldsymbol{G}).
\end{aligned}
$$

Since $\boldsymbol{x}\,(-\eta^{-1}\boldsymbol{A}+\boldsymbol{I}-\boldsymbol{B})=\eta\,\boldsymbol{x}\,\boldsymbol{C}$ from (3.6), we have (3.13). Similarly, we obtain (3.14) from (3.5), (3.7) and Lemma 3.2.1.

Lemma 3.2.3. $\boldsymbol{x}\,(\eta\,\boldsymbol{I}-\boldsymbol{R})\,\boldsymbol{C}\;=\;\boldsymbol{0}$.

Proof. Multiplying $\boldsymbol{e}$ to (3.13) from the right, then we have $\boldsymbol{x}\,(\eta\boldsymbol{I}-\boldsymbol{R})\,\boldsymbol{C}\,\boldsymbol{e}=\boldsymbol{0}$ since $\boldsymbol{x}\boldsymbol{e}<\infty$, $\boldsymbol{G}\boldsymbol{e}=\boldsymbol{e}$ and $\eta<1$. We know that $\boldsymbol{x}\,(\eta\boldsymbol{I}-\boldsymbol{R})\,\boldsymbol{C}$ is non-negative from (3.11) and $\boldsymbol{e}$ is positive. Hence, $\boldsymbol{x}\,(\eta\boldsymbol{I}-\boldsymbol{R})\,\boldsymbol{C}$ must be a zero vector.

Lemma 3.2.4. $\boldsymbol{x}\,(\eta\boldsymbol{I}-\boldsymbol{R})\;=\;\boldsymbol{0}$ or, equivalently, $\boldsymbol{x}\,\boldsymbol{R}\;=\;\eta\,\boldsymbol{x}$.

Proof. From (3.6) and (3.3) we have

$$
\eta\,\boldsymbol{x}\;=\;\boldsymbol{x}\,\boldsymbol{A}\,(\boldsymbol{I}-\boldsymbol{B})^{-1}+\eta^2\,\boldsymbol{x}\,\boldsymbol{C}\,(\boldsymbol{I}-\boldsymbol{B})^{-1}, \tag{3.17}
$$

$$
\boldsymbol{x}\,\boldsymbol{R}\;=\;\boldsymbol{x}\,\boldsymbol{A}\,(\boldsymbol{I}-\boldsymbol{B})^{-1}+\boldsymbol{x}\,\boldsymbol{R}^2\,\boldsymbol{C}\,(\boldsymbol{I}-\boldsymbol{B})^{-1}. \tag{3.18}
$$

Products in (3.17) are all finite from Lemma 3.2.1. Subtracting both sides of (3.18) from those of (3.17) we have

$$\begin{aligned}
\boldsymbol{x}\,(\eta\,\boldsymbol{I}-\boldsymbol{R}) &= \boldsymbol{x}\,(\eta^2\boldsymbol{I}-\boldsymbol{R}^2)\,\boldsymbol{C}\,(\boldsymbol{I}-\boldsymbol{B})^{-1} \\
&= \boldsymbol{x}\,(\eta\,\boldsymbol{I}-\boldsymbol{R})\,(\eta\,\boldsymbol{I}+\boldsymbol{R})\,\boldsymbol{C}\,(\boldsymbol{I}-\boldsymbol{B})^{-1} \\
&= \eta\,\boldsymbol{x}\,(\eta\,\boldsymbol{I}-\boldsymbol{R})\,\boldsymbol{C}\,(\boldsymbol{I}-\boldsymbol{B})^{-1}+\boldsymbol{x}\,(\eta\,\boldsymbol{I}-\boldsymbol{R})\,\boldsymbol{R}\boldsymbol{C}(\boldsymbol{I}-\boldsymbol{B})^{-1} \\
&= \boldsymbol{x}\,(\eta\,\boldsymbol{I}-\boldsymbol{R})\,\boldsymbol{A}\,\boldsymbol{U}\,\boldsymbol{C}\,(\boldsymbol{I}-\boldsymbol{B})^{-1} \qquad \text{(from Lemma 3.2.3)} \\
&= \boldsymbol{x}\,(\eta\,\boldsymbol{I}-\boldsymbol{R})\,\boldsymbol{A}\,\{\boldsymbol{U}\,\boldsymbol{C}\,(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{A}\}^{k-1}\,\boldsymbol{U}\,\boldsymbol{C}\,(\boldsymbol{I}-\boldsymbol{B})^{-1} \qquad (3.19)
\end{aligned}$$

for any $k \geq 1$. Hence

$$\boldsymbol{x}\,(\eta\,\boldsymbol{I}-\boldsymbol{R})\,\boldsymbol{A}\,\boldsymbol{e} = \boldsymbol{x}\,(\eta\,\boldsymbol{I}-\boldsymbol{R})\,\boldsymbol{A}\,\{\boldsymbol{U}\,\boldsymbol{C}\,(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{A}\}^{k}\boldsymbol{e}.$$

The (i,j)-th element of the matrix $\boldsymbol{U}\,\boldsymbol{C}\,(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{A}$ is interpreted as the probability that the chain starting from state (m,i) once reaches $\mathcal{L}_{m-1}$ and then returns to $\mathcal{L}_m$ at state (m,j) without going to $\mathcal{L}_{m-2}$. Therefore, from the positive recurrence of the chain, $\{\boldsymbol{U}\,\boldsymbol{C}\,(\boldsymbol{I}-\boldsymbol{B})^{-1}\boldsymbol{A}\}^{k}\,\boldsymbol{e}$ tends to a zero vector as $k \to \infty$. Hence $\boldsymbol{x}\,(\eta\,\boldsymbol{I}-\boldsymbol{R})\,\boldsymbol{A}\,\boldsymbol{e} = \boldsymbol{0}$. Since $\boldsymbol{x}\,(\eta\,\boldsymbol{I}-\boldsymbol{R})\,\boldsymbol{A}$ is nonnegative and $\boldsymbol{e}$ is positive, the former must be a zero vector. From (3.19) this implies that $\boldsymbol{x}\,(\eta\,\boldsymbol{I}-\boldsymbol{R}) = \boldsymbol{0}$.

Lemma 3.2.5. $\boldsymbol{z} = \boldsymbol{A}\,(\eta^{-1}\boldsymbol{I}-\boldsymbol{G})\,\boldsymbol{y}$ is non-zero and non-negative.

Proof. The non-negativity of $\boldsymbol{z}$ is easily seen from (3.12) in Lemma 3.2.1. The non-zero property of $\boldsymbol{z}$ is derived from (3.8) as

$$\boldsymbol{x}\,\boldsymbol{z} \;=\; \boldsymbol{x}\,\boldsymbol{A}\,(\eta^{-1}\boldsymbol{I}-\boldsymbol{G})\,\boldsymbol{y} \;=\; \eta^{-1}\boldsymbol{x}\,\boldsymbol{A}\,\boldsymbol{y}-\boldsymbol{x}\,\boldsymbol{R}\,\boldsymbol{C}\,\boldsymbol{y} \;=\; \eta^{-1}\boldsymbol{x}\,\boldsymbol{A}\,\boldsymbol{y}-\eta\,\boldsymbol{x}\,\boldsymbol{C}\,\boldsymbol{y} \neq 0.$$

Proof of Theorem 3.2.1. The relation (a) was proved in Lemma 3.2.4. The statement (b) comes from (3.14) in Lemma 3.2.2 and from Lemma 3.2.5. □

3.3 Continuous-Time Parameter Case

In this section we apply the theorems proved in the preceding section to a continuous-time Markov chain $\{\overline{X}(t),\, t \geq 0\}$ on the state space $\mathcal{S}$ having transition rate matrix

$$\overline{\boldsymbol{Q}} = \begin{pmatrix} \overline{\boldsymbol{B}}_0 & \overline{\boldsymbol{A}}_0 & & & \\ \overline{\boldsymbol{C}}_1 & \overline{\boldsymbol{B}} & \overline{\boldsymbol{A}} & & \\ & \overline{\boldsymbol{C}} & \overline{\boldsymbol{B}} & \overline{\boldsymbol{A}} & \\ & & \overline{\boldsymbol{C}} & \overline{\boldsymbol{B}} & \ddots \\ & & & \ddots & \ddots \end{pmatrix}. \tag{3.20}$$

This type of Markov chain is also called QBD process. We assume that the chain is positive recurrent and has stationary probabilities $\pi(m,i)$, $(m,i) \in \mathcal{S}_m$. The stationary probability vector $\overline{\boldsymbol{\pi}} = (\,\overline{\boldsymbol{\pi}}_m,\, m = 0,1,2,\dots\,)$ with subvectors $\overline{\boldsymbol{\pi}}_m = (\,\overline{\pi}(m,i),\, i \in \mathcal{L}_m\,)$ takes a matrix-geometric form with a rate matrix $\overline{\boldsymbol{R}}$. We also write the G-matrix as $\overline{\boldsymbol{G}}$. A submatrix $\overline{\boldsymbol{R}}(\overline{A})$ is similarly defined as $\boldsymbol{R}(A)$.

We may apply Theorems 3.2.1 and 3.2.2 directly to a discrete-time parameter Markov chain $\{\overline{X}(nh),\, n = 0, 1, 2, \ldots\}$ for some $h > 0$. However, this approach is not convenient since the conditions of the theorems are not directly related to the rate matrix (3.20).

Let $\overline{\boldsymbol{D}}$ and $\overline{\boldsymbol{D}}_0$ be diagonal matrices such that $\overline{\boldsymbol{D}} \geq -\overline{\boldsymbol{B}}$ and $\overline{\boldsymbol{D}}_0 \geq -\overline{\boldsymbol{B}}_0$. Then we consider a discrete-time Markov chain $\{X'(n),\, n = 0, 1, 2, \ldots\}$ with transition probability matrix

$$\boldsymbol{P}' = \begin{pmatrix} \boldsymbol{I} + \overline{\boldsymbol{D}}_0^{-1}\overline{\boldsymbol{B}}_0 & \overline{\boldsymbol{D}}_0^{-1}\overline{\boldsymbol{A}}_0 & & & \\ \overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{C}}_1 & \boldsymbol{I} + \overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{B}} & \overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{A}} & & \\ & \overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{C}} & \boldsymbol{I} + \overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{B}} & \overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{A}} & \\ & & \overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{C}} & \boldsymbol{I} + \overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{B}} & \ddots \\ & & & \ddots & \ddots \end{pmatrix}.$$

Vectors and matrices associated with $\{X'(n)\}$ are related to those for $\{\overline{X}(t)\}$ as

$$\boldsymbol{\pi}'_0 = \overline{\boldsymbol{\pi}}_0\,\overline{\boldsymbol{D}}_0, \qquad \boldsymbol{\pi}'_m = \overline{\boldsymbol{\pi}}_m\,\overline{\boldsymbol{D}}, \quad m = 1, 2, \ldots,$$

$$\boldsymbol{R}' = \overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{R}}\,\overline{\boldsymbol{D}}, \qquad \boldsymbol{G}' = \overline{\boldsymbol{G}},$$

$$\boldsymbol{x}' = \overline{\boldsymbol{x}}\,\overline{\boldsymbol{D}}, \quad \boldsymbol{y}' = \overline{\boldsymbol{y}}, \quad \boldsymbol{z}' = \overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{z}}.$$

Applying Theorems 3.2.1 and 3.2.2 to the chain $\{X'(n)\}$, we have the following theorems for the continuous-time Markov chain $\{\overline{X}(t)\}$.

Theorem 3.3.1. For the continuous-time Markov chain $\{\overline{X}(t)\}$, assume that the discrete-time Markov chain $\{X'(n)\}$ defined above is positive recurrent and that there exist a positive constant $\eta\,(< 1)$ and positive vectors $\overline{\boldsymbol{x}}$ and $\overline{\boldsymbol{y}}$ such that

(i) $\overline{\boldsymbol{x}}(\eta^{-1}\overline{\boldsymbol{A}} + \overline{\boldsymbol{B}} + \eta\overline{\boldsymbol{C}}) = \overline{\boldsymbol{0}}$,

(ii) $(\eta^{-1}\overline{\boldsymbol{A}} + \overline{\boldsymbol{B}} + \eta\overline{\boldsymbol{C}})\overline{\boldsymbol{y}} = \boldsymbol{0}$,

(iii) $\eta^{-1}\overline{\boldsymbol{x}}\,\overline{\boldsymbol{A}}\,\overline{\boldsymbol{y}} \neq \eta\overline{\boldsymbol{x}}\,\overline{\boldsymbol{C}}\,\overline{\boldsymbol{y}}$,

(iv) $\overline{\boldsymbol{x}}\,\overline{\boldsymbol{D}}\,\overline{\boldsymbol{e}} < \infty$ and $\overline{\boldsymbol{x}}\,\overline{\boldsymbol{D}}\,\overline{\boldsymbol{y}} < \infty$.

Then,

(a) $\overline{\boldsymbol{x}}\overline{\boldsymbol{R}} = \eta\overline{\boldsymbol{x}}$, and

(b) $\overline{\boldsymbol{z}} = \overline{\boldsymbol{A}}(\eta^{-1}\overline{\boldsymbol{I}} - \overline{\boldsymbol{G}})\overline{\boldsymbol{y}}$ is a non-zero, non-negative vector satisfying $\overline{\boldsymbol{R}}\overline{\boldsymbol{z}} = \eta\overline{\boldsymbol{z}}$.

Theorem 3.3.2. For the continuous-time Markov cahin $\{\overline{X}(t)\}$, suppose that the conditions in Theorem 3.3.1 hold. Then, if $\overline{\boldsymbol{R}}(\overline{A})$ is irreducible and $\overline{\boldsymbol{\pi}}_1\,\overline{\boldsymbol{D}}\,\overline{\boldsymbol{z}} < \infty$, then the stationary probability vector $\overline{\boldsymbol{\pi}}$ has a geometric tail:

$$\overline{\boldsymbol{\pi}}_m \sim c\,\eta^m\,\overline{\boldsymbol{x}} \qquad \text{as} \quad m \to \infty, \tag{3.21}$$

where c is a multiplicative constant. A sufficient condition for the irreducibility of $\overline{\boldsymbol{R}}(\overline{A})$ is that the matrix

$$\begin{pmatrix} \overline{\boldsymbol{B}} & \overline{\boldsymbol{A}} & & \\ \overline{\boldsymbol{C}} & \overline{\boldsymbol{B}} & \overline{\boldsymbol{A}} & \\ & \overline{\boldsymbol{C}} & \overline{\boldsymbol{B}} & \ddots \\ & & \ddots & \ddots \end{pmatrix} \tag{3.22}$$

formed from (3.20) by deleting rows and columns corresponding states in $\mathcal{L}_0$ is irreducible. And a sufficient condition for $\overline{\boldsymbol{\pi}}_1\,\overline{\boldsymbol{D}}\,\overline{\boldsymbol{z}}$ to be finite is that $\overline{\boldsymbol{\pi}}_1\,\overline{\boldsymbol{D}}\,\overline{\boldsymbol{y}} < \infty$.

If diagonal elements of $-\overline{\boldsymbol{B}}$ and $-\overline{\boldsymbol{B}}_0$ are bounded by $d\,(< \infty)$ from above, then the so-called uniformization technique can be applied, and the conditions of these theorems

become simpler. Let $\overline{\boldsymbol{D}}_0 = \overline{\boldsymbol{D}} = d\,\boldsymbol{I}$. Then the positive recurrence of $\{X'(n)\}$ is derived from that of the original chain $\{\overline{X}(t)\}$, and the theorems are reduced to the following corollaries.

Corollary 1. For the continuous-time Markov chain $\{\overline{X}(t)\}$, assume that diagonal elements of $-\overline{\boldsymbol{B}}$ and $-\overline{\boldsymbol{B}}_0$ are bounded by $d\,(<\infty)$ from above and that there exists a positive constant $\eta\,(<1)$ and positive vectors $\overline{\boldsymbol{x}}$ and $\overline{\boldsymbol{y}}$ such that

(i) $\overline{\boldsymbol{x}}(\eta^{-1}\overline{\boldsymbol{A}} + \overline{\boldsymbol{B}} + \eta\overline{\boldsymbol{C}}) = \overline{\boldsymbol{0}}$,

(ii) $(\eta^{-1}\overline{\boldsymbol{A}} + \overline{\boldsymbol{B}} + \eta\overline{\boldsymbol{C}})\overline{\boldsymbol{y}} = \boldsymbol{0}$,

(iii) $\eta^{-1}\overline{\boldsymbol{x}}\overline{\boldsymbol{A}}\overline{\boldsymbol{y}} \neq \eta\overline{\boldsymbol{x}}\overline{\boldsymbol{C}}\overline{\boldsymbol{y}}$,

(iv) $\overline{\boldsymbol{x}}\overline{\boldsymbol{e}} < \infty$ and $\overline{\boldsymbol{x}}\overline{\boldsymbol{y}} < \infty$.

Then,

(a) $\overline{\boldsymbol{x}}\overline{\boldsymbol{R}} = \eta\overline{\boldsymbol{x}}$, and

(b) $\overline{\boldsymbol{z}} = \overline{\boldsymbol{A}}(\eta^{-1}\overline{\boldsymbol{I}} - \overline{\boldsymbol{G}})\overline{\boldsymbol{y}}$ is a non-zero, non-negative vector satisfying $\overline{\boldsymbol{R}}\overline{\boldsymbol{z}} = \eta\overline{\boldsymbol{z}}$.

Corollary 2. For the continuous-time Markov cahin $\{\overline{X}(t)\}$, suppose that the assumption in Corollary 1 holds. Then, if $\overline{\boldsymbol{R}}(\overline{A})$ is irreducible and $\overline{\boldsymbol{\pi}}_1\overline{\boldsymbol{z}} < \infty$, then the stationary probability vector $\overline{\boldsymbol{\pi}}$ has a geometric tail (3.21). A sufficient condition for the irreducibility of $\overline{\boldsymbol{R}}(\overline{A})$ is that the matrix (3.22) is irreducible, and a sufficient condition for $\overline{\boldsymbol{\pi}}_1\overline{\boldsymbol{z}}$ to be finite is that $\overline{\boldsymbol{\pi}}_1\overline{\boldsymbol{y}} < \infty$.

Chapter 4

Asymptotic Properties of Stationary Distributions in Two-Stage Tandem Queueing Systems

4.1 Introduction

Tandem queueing systems are basic models in the theory of queues and have been studied for a long time. However, because of complexities of their stochastic structures, their properties are scarcely known except for cases with product form solutions. They are simplest models of queueing networks as well as direct extensions of single queueing systems. Hence the study of them are expected to connect the theory of single queueing systems with that of queueing networks. In this paper, we prove geometric decays of the stationary state

probability in a two-stage tandem queueing system $PH/PH/1 \to /PH/1$ with a buffer of infinite capacity.

In the ordinary one-stage queue $PH/PH/c$ with traffic intensity $\rho < 1$, it was shown that the stationary distribution has a geometric tail [45]. Let $\pi(n; i_0, i_1)$ be the stationary probability that there exist n customers in the system while the phases of the arrival and service processes are i_0 and i_1, respectively. Then

$$\pi(n; i_0, i_1) \sim GC_0(i_0)C_1(i_1)\eta^n, \qquad n \to \infty, \tag{4.1}$$

where $G, C_0(i_0), C_1(i_1)$ and η are some constants and $\sim$ indicates that the ratio of both sides tends to 1. These constants other than G can be easily obtained from the phase type representations of the interarrival and service time distributions. This kind of geometric decay property is very useful, for example, on the computation of the stationary state probabilities, or on the discussion of tail probabilities for estimating very small loss probabilities (e.g. less than 10^{-9}) of the corresponding finite queue.

Our main concern here is to prove a similar geometric tail property in the two-stage tandem queueing system $PH/PH/1 \to /PH/1$.

In Chapter 2, we have made a conjecture on the geometric decay of the stationary state probability in $PH/PH/1 \to /PH/1$ through an extensive numerical experiment. Let $\pi(n_1, n_2; i_0, i_1, i_2)$ be the stationary probability that there exist n_1 customers in the fist stage and n_2 customers in the second stage while the phases of the arrival process and the two service processes are i_0, i_1 and i_2, respectively. Then the conjecture asserts that the

stationary state probability decays geometrically as n_1 and/or n_2 become large but decay rates and multiplicative constants may be different according to the ratio of n_1 and n_2:

$$\pi(n_1, n_2; i_0, i_1, i_2) \sim \begin{cases} G\, C_0(i_0)\, C_1(i_1)\, C_2(i_2)\, \eta_1^{n_1} \eta_2^{n_2}, \\ \qquad \text{for large } n_1 \text{ and/or } n_2 \text{ such that } \ n_2 < \alpha\, n_1, \\ \overline{G}\, \overline{C}_0(i_0)\, \overline{C}_1(i_1)\, \overline{C}_2(i_2)\, \overline{\eta}_1^{n_1} \overline{\eta}_2^{n_2}, \\ \qquad \text{for large } n_1 \text{ and/or } n_2 \text{ such that } \ n_1 < \alpha^{-1}\, n_2, \end{cases} \tag{4.2}$$

where $\alpha = -\ln(\eta_1/\overline{\eta}_1)/\ln(\eta_2/\overline{\eta}_2)$ and constants η_k, $\overline{\eta}_k$ $(k = 1, 2)$ and $c_k(i_k)$, $\overline{c}_k(i_k)$ $(k = 0, 1, 2)$ are determined from the phase-type representations of the interarrival and service time distributions.

In this chapter, under a certain condition, we prove (4.2) in two special cases, the case where $n_1 \to \infty$ with n_2 being fixed and the case where $n_2 \to \infty$ with n_1 being fixed. The proof uses a result on the Matrix-geometric form solution of a quasi-birth-and-death (QBD) process with a *countable* number of phases in each level [48].

The reminder of the chapter is constructed as follows. In Section 4.2, we describe our two-stage tandem queueing model and state our main theorems in Section 4.2.2. Applying Theorem 3.3.1 in Chapter 3, we prove our theorems in Sections 4.3 and 4.4. In many places in this paper, we have to use various properties of solutions of four key systems of equations given in Section 4.2.2. These properties are proved in Appendix.

4.2 Geometric Decay Property for Two-Stage Tandem Queueing System

4.2.1 Model Description

We denote by $PH(\boldsymbol{a}, \boldsymbol{\Phi})$ a phase-type distribution represented by a continuous-time, finite-state, absorbing Markov chain with initial probability vector $\tilde{\boldsymbol{a}} = (\boldsymbol{a}, 0)$ and transition rate matrix $\tilde{\boldsymbol{\Phi}} = \begin{bmatrix} \boldsymbol{\Phi} & \boldsymbol{\gamma} \\ \mathbf{0} & 0 \end{bmatrix}$ (see [30]). We assume that there exist no redundant phases in any phase-type distributions. More definitely, the phase-type distribution $(\boldsymbol{a}, \boldsymbol{\Phi})$ is assumed to satisfy

$$-\boldsymbol{a}\boldsymbol{\Phi}^{-1} > \mathbf{0}. \tag{4.3}$$

We consider an open, two-stage tandem queueing system (Figure 4.1). Customers arrive at the first stage to be served there, move to the second to be served there again, and then go out of the system. The k-th stage ($k = 1, 2$) has a single server and a buffer of infinite capacity, so that neither loss nor blocking occurs. Interarrival times of customers are independent and identically distributed (i.i.d.) random variables subjecting to an irreducible phase-type distribution $PH(\boldsymbol{\alpha}, \boldsymbol{T})$. Service times at the server of the k-th stage are also i.i.d. variables subjecting to an irreducible phase-type distribution $PH(\boldsymbol{\beta}_k, \boldsymbol{S}_k)$. The interarrival times and the service times are assumed to be mutually independent. Customers are served according to the first-come first-served (FCFS) discipline at each stage.

The state of the system is represented by a quintuple $(n_1, n_2; i_0, i_1, i_2)$, where i_0 is the

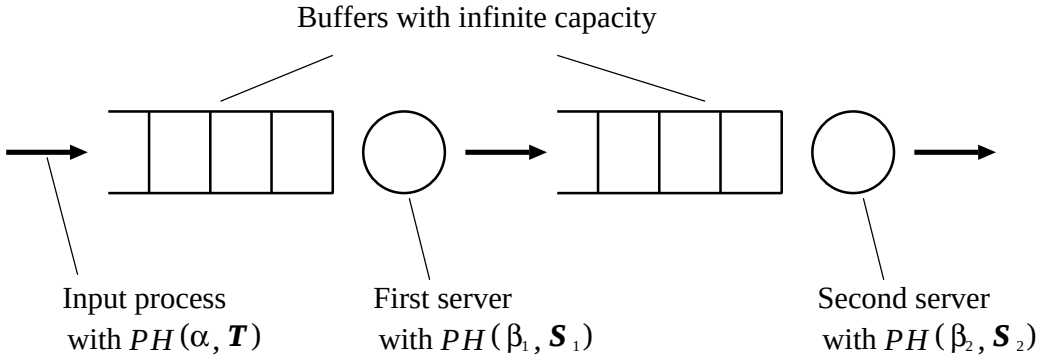


Figure 4.1: Two-stage tandem queueing system

phase of the arrival process, i_k is the phase of the service process at the k-th stage, and n_k is the number of customers in the k-th stage ($k = 1, 2$). Then the system behaves as a continuous-time Markov chain.

We denote the traffic intensity at the k-th stage by $\rho_k = \lambda/\mu_k$ where $1/\lambda$ is the mean interarrival time and $1/\mu_k$ is the mean service time at the k-th stage, and assume $\rho_1, \rho_2 < 1$ so that the chain is stable and has stationary probabilities $\pi(n_1, n_2; i_0, i_1, i_2)$.

We prepare some notations. Let $\boldsymbol{I}_0$ be the identity matrix with the same dimension as $\boldsymbol{T}$, and $\boldsymbol{I}_k$ be the identity matrix with the same dimension as $\boldsymbol{S}_k$. $\boldsymbol{e}_0$ denotes the column vector of the same dimension as $\boldsymbol{T}$ with all elements being equal to 1, and $\boldsymbol{e}_k$ the column vector of the same dimension as $\boldsymbol{S}_k$ with all elements being equal to 1 ($k = 1, 2$).

4.2.2 Geometric Decay Property of the Stationary Distribution

The marginal queue-length distribution of the first stage clearly has a geometric tail, since the behavior of the first stage is not affected by that of the second stage. Our concern is the tail property of the joint queue-length distribution of the first and the second stages or the asymptotic behavior of the stationary probabilities.

To describe the geometric decay, we introduce *decay parameters* $\eta_1, \eta_2, \overline{\eta}_1$ and $\overline{\eta}_2$ given as follows.

The Laplace-Stieltjes Transforms of the interarrival and service time distributions are given by

$$T^*(s) = \boldsymbol{\alpha}(s\boldsymbol{I}_0 - \boldsymbol{T})^{-1}\boldsymbol{\gamma}_0, \qquad S_k^*(s) = \boldsymbol{\beta}_k(s\boldsymbol{I}_k - \boldsymbol{S}_k)^{-1}\boldsymbol{\gamma}_k, \tag{4.4}$$

where

$$\boldsymbol{\gamma}_0 = -\boldsymbol{T}\boldsymbol{e}_0, \qquad \boldsymbol{\gamma}_k = -\boldsymbol{S}_k\boldsymbol{e}_k.$$

Note that, for any LSTs $F_1^*(s)$ and $F_2^*(s)$ for phase-type distributions, $f(s) = F_1^*(s)F_2^*(-s)$ is convex and hence $f(s) = C$ has at most two real roots for any constant C. This also implies that signs of derivatives $f'(s)$ at two roots are different.

Consider the equation

$$T^*(s)S_1^*(-s) = 1. \tag{4.5}$$

Since the function $f_1(s) = T^*(s)S_1^*(-s)$ is convex, there exist at most two real roots of (4.5). Since $s = 0$ is a trivial root, there exist another real root ω_0. We set

$$\eta_1 = T^*(\omega_0). \tag{4.6}$$

We also consider the equation

$$T^*(\omega_0)S_1^*(-s-\omega_0)S_2^*(s)=1. \tag{4.7}$$

The left-hand side is a convex function of s and zero is a trivial root. Let ω_2 be the other root, and set

$$\eta_2=\frac{1}{S_2^*(\omega_2)}. \tag{4.8}$$

In the same manner, we denote by $\overline{\omega}_2$ the root of the equation

$$T^*(-s)S_2^*(s)=1 \tag{4.9}$$

other than 0, and set

$$\eta_2=\frac{1}{S_2^*(\overline{\omega}_2)}. \tag{4.10}$$

Let $\overline{\omega}_0$ be the root of the equation

$$T^*(s)S_1^*(-s-\overline{\omega}_2)S_2^*(\overline{\omega}_2)=1 \tag{4.11}$$

other than $-\overline{\omega}_2$, and set

$$\eta_2=\{S_2^*(\omega_2)\}^{-1}. \tag{4.12}$$

Remarks.

1. Since ω_0 is positive, η_1 is strictly less than 1. This also implies that the derivative of $T^*(s)S_1^*(-s)$ at ω_0 must be positive, and that the derivative at origin must be negative. Furthermore, it is easily shown that η_1 is a monotone increasing function of ρ_1, and $\eta_1\downarrow 0$ as $\rho_1\downarrow 0$ while $\eta_1\uparrow 1$ as $\rho_1\uparrow 1$.

2. ω_2 is negative if ρ_2 is small but it may be positive if ρ_2 becomes large, and hence η_2 may exceed 1. η_2 can be regarded as a function of both ρ_1 and ρ_2. For a fixed ρ_1, η_2 is a monotone increasing function and $\eta_2 \downarrow 0$ as $\rho_2 \downarrow 0$.

3. Since $\overline{\omega}_2$ is negative, $\overline{\eta}_2$ is less than 1. It is a monotone increasing function of ρ_2, and $\overline{\eta}_2 \downarrow 0$ as $\rho_2 \downarrow 0$ while $\overline{\eta}_2 \uparrow 1$ as $\rho_2 \uparrow 1$.

4. As for $\overline{\eta}_1$, it may be greater than 1 since $\overline{\omega}_0$ may take positive or negative value. For a fixed ρ_2, $\overline{\eta}_1$ is a monotone increasing function and $\overline{\eta}_1 \to \eta_1$ as $\rho_1 \uparrow 1$.

In Chapter 2, we did extensive numerical experiments and calculated stationary probabilities for a variety of examples of two-stage tandem queueing systems $PH/PH/1 \to /PH/1$. They observed the results and conjectured some geometric decay properties of the tails of the stationary distributions as follows:

Conjecture 1. For fixed ρ_1, there exists a threshold $\tilde{\rho}_2$ and the behavior of $x(n_1, n_2; i_0, i_1, i_2)$ is different between the cases $\rho_2 < \tilde{\rho}_2$ and $\rho_2 > \tilde{\rho}_2$.

1. In the case $\rho_2 < \tilde{\rho}_2$, there exist constants $\eta_1, \eta_2, c_0(i_0), c_1(i_1), c_2(i_2)$ and G such that

$$\pi(n_1, n_2; i_0, i_1, i_2) \sim G\ c_0(i_0)c_1(i_1)c_2(i_2)\eta_1^{n_1}\eta_2^{n_2}$$

as $n_1, n_2 \to \infty$ on a line $n_2 = an_1 + b$ with rational $a > 0$ and b. This asymptotic representation is also valid when $n_1 \to \infty$ with fixed n_2 and when $n_2 \to \infty$ with fixed n_1.

2. In the case $\rho_2 > \tilde{\rho}_2$, there exists a positive constant $\tilde{a}$ such that the decay rates are different between the cases $0 < a < \tilde{a}$ and $a > \tilde{a}$ for the slope a of the line on which n_1 and n_2 increase. We denote the two sets of constants corresponding to these two cases as $\{\eta_1, \eta_2, c_0(i_0), c_1(i_1), c_2(i_2), G\}$ and $\{\overline{\eta}_1, \overline{\eta}_2, \overline{c}_0(i_0), \overline{c}_1(i_1), \overline{c}_2(i_2), \overline{G}\}$.

   (a) When $n_1, n_2 \to \infty$ on a line $n_2 = an_1 + b$ with rational a and b such that $0 < a < \tilde{a}$,

$$\pi(n_1, n_2; i_0, i_1, i_2) \sim G\ c_0(i_0)c_1(i_1)c_2(i_2)\eta_1^{n_1}\eta_2^{n_2}.$$

   This asymptotic representation is also valid when $n_1 \to \infty$ with fixed n_2.

   (b) When $n_1, n_2 \to \infty$ on a line $n_2 = an_1 + b$ with rational a and b such that $a > \tilde{a}$,

$$\pi(n_1, n_2; i_0, i_1, i_2) \sim \overline{G}\ \overline{c}_0(i_0)\overline{c}_1(i_1)\overline{c}_2(i_2)\overline{\eta}_1^{n_1}\overline{\eta}_2^{n_2}.$$

   This asymptotic representation is also valid when $n_2 \to \infty$ with fixed n_1.

Decay constants η_k and $\overline{\eta}_k$ $(k = 1, 2)$ is given by (4.6), (4.8), (4.10) and (4.12) below.

In this chapter we prove the following two theorems, which justify a part of the conjecture stated above.

Theorem 4.2.1. If $\eta_2 < 1$,

$$\pi(n_1, n_2; i_0, i_1, i_2) \sim G_1(n_2; i_0, i_1, i_2)\eta_1^{n_1} \quad (n_1 \to \infty), \tag{4.13}$$

where $G_1(n_2; i_0, i_1, i_2)$ is a constant. Furthermore,

$$G_1(n_2; i_0, i_1, i_2) \sim G_2\, C_0(i_0)C_1(i_1)C_2(i_2)\eta_2^{n_2} \quad (n_2 \to \infty), \tag{4.14}$$

where $C_0(i_0)$ is the i_0-th element of $\boldsymbol{\alpha}(\omega_0\boldsymbol{I}-\boldsymbol{T})^{-1}$, $C_1(i_1)$ is the i_1-th element of $\boldsymbol{\beta}_1\{-(\omega_0+\omega_2)\boldsymbol{I}_2-\boldsymbol{S}_1\}^{-1}$, and $C_2(i_2)$ are the i_2-th element of $\boldsymbol{\beta}_2\{\omega_2\boldsymbol{I}_2-\boldsymbol{S}_2\}^{-1}$. The constant G_2 does not depend on state of the system.

Theorem 4.2.2. If $\eta_1 < \overline{\eta}_2$ and $\overline{\eta}_1 < \overline{\eta}_2$, then for fixed n_1, i_0, i_1 and i_2,

$$\pi(n_1, n_2; i_0, i_1, i_2) \sim \overline{G}_2(n_1; i_0, i_1, i_2)\overline{\eta}_2^{n_2} \quad (n_2 \to \infty), \tag{4.15}$$

where $\overline{G}_2$ is a constant. In this case,

$$\overline{G}_2(n_1; i_0, i_1, i_2) \sim \overline{G}_1\,\overline{C}_0(i_0)\overline{C}_1(i_1)\overline{C}_2(i_2)\overline{\eta}_1^{n_1} \quad (n_1 \to \infty), \tag{4.16}$$

where $\overline{C}_0(i_0)$ is the i_0-th element of $\boldsymbol{\alpha}(\overline{\omega}_0\boldsymbol{I}_0-\boldsymbol{T})^{-1}$, $\overline{C}_1(i_1)$ is the i_1-th element of $\boldsymbol{\beta}_1\{-(\overline{\omega}_0-\overline{\omega}_2)\boldsymbol{I}_1-\boldsymbol{S}_1\}^{-1}$, and $\overline{C}_2(i_2)$ is the i_2-th element of $\boldsymbol{\beta}_2(\overline{\omega}_2\boldsymbol{I}_2-\boldsymbol{S}_2)^{-1}$. The constant $\overline{G}_1$ does not depend on state of the system.

4.2.3 Quasi-birth-and-death Process with Countable Number of Phases

To prove the theorems, we use Corollaries 1 and 2 proved in the preceding chapter. These corollaries are summarized in Proposition 1 below.

Consider a continuous time Markov chain $\{X(t)\}$ on the state space $\mathcal{S} = \{(m, i); m, i = 0, 1, 2, \ldots\}$. We partition $\mathcal{S}$ into subsets $\mathcal{L}_m = \{(m, i); i = 0, 1, 2, \ldots\}, m = 0, 1, 2, \ldots$. The subset $\mathcal{L}_m$ is called the m-th level. We assume that the transition rate matrix $\boldsymbol{Q}$ of $\{X(t)\}$

has a block-tridiagonal form

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{B}_0 & \boldsymbol{A}_0 & & & \\ \boldsymbol{C}_1 & \boldsymbol{B} & \boldsymbol{A} & & \\ & \boldsymbol{C} & \boldsymbol{B} & \boldsymbol{A} & \\ & & \boldsymbol{C} & \boldsymbol{B} & \ddots \\ & & & \ddots & \ddots \end{pmatrix}. \tag{4.17}$$

Such a chain is called a quasi-birth-and-death (QBD) process with a countable number of states in each level. Let $\boldsymbol{\pi} = (\boldsymbol{\pi}_0 \ \ \boldsymbol{\pi}_1 \ \ \cdots)$ be the stationary vector of the QBD process partitioned by $\mathcal{L}_m$'s.

Proposition 1. We assume that diagonal elements of $-\boldsymbol{B}$ and $-\boldsymbol{B}_0$ are bounded by $d(< \infty)$ from above. Suppose that there exist a positive constant $\eta(< 1)$ and positive vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ such that

$$\boldsymbol{x}\left(\frac{1}{\eta}\boldsymbol{A} + \boldsymbol{B} + \eta\boldsymbol{C}\right) = \boldsymbol{0}, \tag{4.18}$$

$$\left(\frac{1}{\eta}\boldsymbol{A} + \boldsymbol{B} + \eta\boldsymbol{C}\right)\boldsymbol{y} = \boldsymbol{0}, \tag{4.19}$$

$$\frac{1}{\eta}\boldsymbol{x}\boldsymbol{A}\boldsymbol{y} \neq \eta\boldsymbol{x}\boldsymbol{C}\boldsymbol{y}, \tag{4.20}$$

$$\boldsymbol{x}\boldsymbol{e} < \infty, \quad \textit{and} \quad \boldsymbol{x}\boldsymbol{y} < \infty. \tag{4.21}$$

If $\boldsymbol{\pi}_1\boldsymbol{y} < \infty$ and if the matrix

$$\boldsymbol{Q}' = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{A} & & \\ \boldsymbol{C} & \boldsymbol{B} & \boldsymbol{A} & \\ & \boldsymbol{C} & \boldsymbol{B} & \ddots \\ & & \ddots & \ddots \end{pmatrix} \tag{4.22}$$

is irreducible, then $\boldsymbol{\pi}$ has a geometric tail:

$$\boldsymbol{\pi}_m \sim C\eta^m\boldsymbol{x}. \tag{4.23}$$

In the following sections, we show that this proposition holds for two cases of our two-stage tandem queueing systems. In these cases, the state space is divided into levels by the number of customers in the first or the second stages.

4.3 Rate Matrix and Its Invariant Vectors for n_1-based Decomposition

First, we arrange the states $(n_1, n_2; i_0, i_1, i_2)$ of the Markov chain $\{X(t)\}$ derived from $PH/PH/1 \to /PH/1$ in lexicographic order. We partition the state space according to n_1, i.e., we let

$$\mathcal{L}_m = \{(n_1, n_2; i_0, i_1, i_2) | n_1 = m\}, \quad m = 0, 1, 2, \ldots. \tag{4.24}$$

We denote by $\boldsymbol{Q}$ the transition rate matrix of the chain corresponding to the arrangement above, and by $\boldsymbol{\pi} = (\boldsymbol{\pi}_0 \ \ \boldsymbol{\pi}_1 \ \ \cdots)$ the stationary vector partitioned according to $\mathcal{L}_m$'s. Then $\boldsymbol{Q}$ is of the block-tridiagonal form as (4.17) with

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{\gamma}_0\boldsymbol{\alpha} \otimes \boldsymbol{I}_1 & & & \\ & \boldsymbol{\gamma}_0\boldsymbol{\alpha} \otimes \boldsymbol{I}_1 \otimes \boldsymbol{I}_2 & & \boldsymbol{O} \\ & & \boldsymbol{\gamma}_0\boldsymbol{\alpha} \otimes \boldsymbol{I}_1 \otimes \boldsymbol{I}_2 & \\ & \boldsymbol{O} & & \boldsymbol{\gamma}_0\boldsymbol{\alpha} \otimes \boldsymbol{I}_1 \otimes \boldsymbol{I}_2 \\ & & & & \ddots \end{pmatrix}, \tag{4.25}$$

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{T} \oplus \boldsymbol{S}_1 & & & & & \\ \boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \otimes \boldsymbol{\gamma}_2 & \boldsymbol{T} \oplus \boldsymbol{S}_1 \oplus \boldsymbol{S}_2 & & & & \\ & \boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \otimes \boldsymbol{\gamma}_2\boldsymbol{\beta}_2 & \boldsymbol{T} \oplus \boldsymbol{S}_1 \oplus \boldsymbol{S}_2 & & \\ & & \boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \otimes \boldsymbol{\gamma}_2\boldsymbol{\beta}_2 & \boldsymbol{T} \oplus \boldsymbol{S}_1 \oplus \boldsymbol{S}_2 & \\ & & & \ddots & \ddots \end{pmatrix}, \tag{4.26}$$

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1\boldsymbol{\beta}_1 \otimes \boldsymbol{\beta}_2 & & & \\ & \boldsymbol{O} & \boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1\boldsymbol{\beta}_1 \otimes \boldsymbol{I}_2 & & \\ & & \boldsymbol{O} & \boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1\boldsymbol{\beta}_1 \otimes \boldsymbol{I}_2 & \\ & & & \boldsymbol{O} & \ddots \\ & & & & \ddots \end{pmatrix}. \tag{4.27}$$

where we write the Kronecker product of two matrices or vectors as $\otimes$ and the Kronecker sum of them as $\oplus$, respectively. Then $\{X(t)\}$ is regarded as a QBD process with a countable number of phases in each level.

We define some vectors. $\boldsymbol{v}_0$ and $\boldsymbol{v}_1$ are column vectors given by

$$\boldsymbol{v}_0 = (\omega_0\boldsymbol{I}_0 - \boldsymbol{T})^{-1}\boldsymbol{\gamma}_0, \quad \boldsymbol{v}_1 = (-\omega_0\boldsymbol{I}_1 - \boldsymbol{S}_1)^{-1}\boldsymbol{\gamma}_1, \tag{4.28}$$

and $\boldsymbol{u}_0$ and $\boldsymbol{u}_1$ are row vectors defined by

$$\boldsymbol{u}_0 = \boldsymbol{\alpha}(\omega_0\boldsymbol{I}_0 - \boldsymbol{T})^{-1}, \quad \boldsymbol{u}_1 = \boldsymbol{\beta}_1(-\omega_0\boldsymbol{I}_1 - \boldsymbol{S}_1)^{-1}. \tag{4.29}$$

From (4.3), all elements of these vectors are positive.

We prove Theorem 4.2.1 through a series of lemmas.

Lemma 4.3.1.

$$\boldsymbol{u}_0\boldsymbol{\gamma}_0 = \boldsymbol{\alpha}\boldsymbol{v}_0 = T^*(\omega_0) = \eta_1, \qquad \boldsymbol{u}_1\boldsymbol{\gamma}_1 = \boldsymbol{\beta}_1\boldsymbol{v}_1 = S_1^*(-\omega_0) = \frac{1}{\eta_1}, \tag{4.30}$$

$$\boldsymbol{u}_0\left(\frac{1}{\eta_1}\boldsymbol{\gamma}_0\boldsymbol{\alpha}+\boldsymbol{T}\right)=\omega_0\boldsymbol{u}_0, \qquad \boldsymbol{u}_1\left(\eta_1\boldsymbol{\gamma}_1\boldsymbol{\beta}_1+\boldsymbol{S}_1\right)=-\omega_0\boldsymbol{u}_1, \tag{4.31}$$

$$\left(\frac{1}{\eta_1}\boldsymbol{\gamma}_0\boldsymbol{\alpha}+\boldsymbol{T}\right)\boldsymbol{v}_0=\omega_0\boldsymbol{v}_0, \qquad \left(\eta_1\boldsymbol{\gamma}_1\boldsymbol{\beta}_1+\boldsymbol{S}_1\right)\boldsymbol{v}_1=-\omega_0\boldsymbol{v}_1. \tag{4.32}$$

Proof. The first two equations are trivial from definitions. From (4.29), we have

$$\begin{aligned}
\boldsymbol{u}_0\left(\omega_0\boldsymbol{I}_0-\boldsymbol{T}\right) &= \boldsymbol{\alpha}, \\
\omega_0\boldsymbol{u}_0 &= \boldsymbol{\alpha}+\boldsymbol{u}_0\boldsymbol{T} \\
&= \frac{1}{\eta_1}\boldsymbol{u}_0\boldsymbol{\gamma}_0\boldsymbol{\alpha}+\boldsymbol{u}_0\boldsymbol{T} \\
&= \boldsymbol{u}_0\left(\frac{1}{\eta_1}\gamma_0\boldsymbol{\alpha}+\boldsymbol{T}\right).
\end{aligned}$$

The reminders can be proved in a similar manner,

Note: Conversely, the vector $\boldsymbol{u}_0$ which satisfies Equation (4.31) in Lemma 4.3.1 must have a form given by (4.29) up to a multiplicative constant.

Let

$$\boldsymbol{K}\equiv\frac{1}{\eta_1}\boldsymbol{A}+\boldsymbol{B}+\eta_1\boldsymbol{C}=\begin{pmatrix}
\tilde{\boldsymbol{B}}_0 & \tilde{\boldsymbol{A}}_0 & & & \\
\tilde{\boldsymbol{C}}_1 & \tilde{\boldsymbol{B}} & \tilde{\boldsymbol{A}} & & \\
 & \tilde{\boldsymbol{C}} & \tilde{\boldsymbol{B}} & \tilde{\boldsymbol{A}} & \\
 & & \tilde{\boldsymbol{C}} & \tilde{\boldsymbol{B}} & \ddots \\
 & & & \ddots & \ddots
\end{pmatrix}.$$

Then by definitions,

$$\tilde{\boldsymbol{A}}_0 = \eta_1\boldsymbol{I}_0\otimes\boldsymbol{\gamma}_1\boldsymbol{\beta}_1\otimes\boldsymbol{\beta}_2,$$

$$
\begin{aligned}
\tilde{\boldsymbol{B}}_0 &= \frac{1}{\eta_1}\boldsymbol{\gamma}_0\boldsymbol{\alpha}\otimes\boldsymbol{I}_1+\boldsymbol{T}\oplus\boldsymbol{S}_1,\\
\tilde{\boldsymbol{C}}_1 &= \boldsymbol{I}_0\otimes\boldsymbol{I}_1\otimes\boldsymbol{\gamma}_2,\\
\tilde{\boldsymbol{A}} &= \eta_1\boldsymbol{I}_0\otimes\boldsymbol{\gamma}_1\boldsymbol{\beta}_1\otimes\boldsymbol{I}_2,\\
\tilde{\boldsymbol{B}} &= \frac{1}{\eta_1}\boldsymbol{\gamma}_0\boldsymbol{\alpha}\otimes\boldsymbol{I}_1\otimes\boldsymbol{I}_2+\boldsymbol{T}\oplus\boldsymbol{S}_1\oplus\boldsymbol{S}_2,\\
\tilde{\boldsymbol{C}} &= \boldsymbol{I}_0\otimes\boldsymbol{I}_1\otimes\boldsymbol{\gamma}_2\boldsymbol{\beta}_2.
\end{aligned}
\tag{4.33}
$$

We introduce a column vector $\boldsymbol{y}_{\eta_1}$ by

$$
\boldsymbol{y}_{\eta_1}=\begin{pmatrix}
\boldsymbol{v}_0\otimes\boldsymbol{v}_1\\
\boldsymbol{v}_0\otimes\boldsymbol{v}_1\otimes\boldsymbol{e}_2\\
\boldsymbol{v}_0\otimes\boldsymbol{v}_1\otimes\boldsymbol{e}_2\\
\vdots
\end{pmatrix}.
\tag{4.34}
$$

Lemma 4.3.2. $\boldsymbol{y}_{\eta_1}$ is positive, and

$$
\boldsymbol{K}\boldsymbol{y}_{\eta_1}=\boldsymbol{0}.
$$

Proof. The positivity of $\boldsymbol{y}_{\eta_1}$ is clear from the definition. The vector $\boldsymbol{K}\boldsymbol{y}_{\eta_1}$ is written as

$$
\begin{pmatrix}
\tilde{\boldsymbol{B}}_0(\boldsymbol{v}_0\otimes\boldsymbol{v}_1)+\tilde{\boldsymbol{A}}_0(\boldsymbol{v}_0\otimes\boldsymbol{v}_1\otimes\boldsymbol{e}_2)\\
\tilde{\boldsymbol{C}}_1(\boldsymbol{v}_0\otimes\boldsymbol{v}_1)+(\tilde{\boldsymbol{B}}+\tilde{\boldsymbol{A}})(\boldsymbol{v}_0\otimes\boldsymbol{v}_1\otimes\boldsymbol{e}_2)\\
(\tilde{\boldsymbol{C}}+\tilde{\boldsymbol{B}}+\tilde{\boldsymbol{A}})(\boldsymbol{v}_0\otimes\boldsymbol{v}_1\otimes\boldsymbol{e}_2)\\
(\tilde{\boldsymbol{C}}+\tilde{\boldsymbol{B}}+\tilde{\boldsymbol{A}})(\boldsymbol{v}_0\otimes\boldsymbol{v}_1\otimes\boldsymbol{e}_2)\\
\vdots
\end{pmatrix}.
$$

The j-th subvector is equal to the zero vector since from Lemma 4.3.1 and (4.33)

$$\begin{aligned}
&(\tilde{\boldsymbol{C}} + \tilde{\boldsymbol{B}} + \tilde{\boldsymbol{A}})(\boldsymbol{v}_0 \otimes \boldsymbol{v}_1 \otimes \boldsymbol{e}_2) \\
&= \left\{ \left(\frac{1}{\eta_1} \boldsymbol{\gamma}_0 \boldsymbol{\alpha} + \boldsymbol{T} \right) \oplus (\eta_1 \boldsymbol{\gamma}_1 \boldsymbol{\beta}_1 + \boldsymbol{S}_1) \oplus (\boldsymbol{\gamma}_2 \boldsymbol{\beta}_2 + \boldsymbol{S}_2) \right\} (\boldsymbol{v}_0 \otimes \boldsymbol{v}_1 \otimes \boldsymbol{e}_2) \\
&= \left\{ \left(\frac{1}{\eta_1} \boldsymbol{\gamma}_0 \boldsymbol{\alpha} + \boldsymbol{T} \right) \boldsymbol{v}_0 \right\} \otimes \boldsymbol{v}_1 \otimes \boldsymbol{e}_2 + \boldsymbol{v}_0 \otimes \{ (\eta_1 \boldsymbol{\gamma}_1 \boldsymbol{\beta}_1 + \boldsymbol{S}_1) \boldsymbol{v}_1 \} \otimes \boldsymbol{e}_2 + \\
&\qquad \boldsymbol{v}_0 \otimes \boldsymbol{v}_1 \otimes \{ (\boldsymbol{\gamma}_2 \boldsymbol{\beta}_2 + \boldsymbol{S}_2) \boldsymbol{e}_2 \} \\
&= \omega_0 \boldsymbol{v}_0 \otimes \boldsymbol{v}_1 \otimes \boldsymbol{e}_2 + -\boldsymbol{v}_0 \otimes (-\omega_0 \boldsymbol{v}_1) \otimes \boldsymbol{e}_2 + \boldsymbol{0} = \boldsymbol{0}.
\end{aligned}$$

The first two subvectors are also equal to the zero vectors. Thus $\boldsymbol{K}\boldsymbol{y}_{\eta_1} = 0$.

Lemma 4.3.3. If $\eta_2 < 1$, there exists a positive vector $\boldsymbol{x}_{\eta_1}$ such that

$$\boldsymbol{x}_{\eta_1} \boldsymbol{K} = \boldsymbol{0}, \quad \boldsymbol{x}_{\eta_1} \boldsymbol{e} < \infty, \quad \boldsymbol{x}_{\eta_1} \boldsymbol{y}_{\eta_1} < \infty.$$

For the proof of this lemma, we define the matrix $\mathrm{diag}(\boldsymbol{\phi})$ for an arbitrary column vector $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots)^t$ by

$$\mathrm{diag}(\boldsymbol{\phi}) = \begin{pmatrix} \phi_1 & & & \\ & \phi_2 & & \\ & & \phi_3 & \\ & & & \ddots \end{pmatrix}.$$

The operator $\mathrm{diag}(\cdot)$ satisfies the following equalities for column vectors $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$:

$$\mathrm{diag}(\boldsymbol{\phi} \otimes \boldsymbol{\psi}) = \mathrm{diag}(\boldsymbol{\phi}) \otimes \mathrm{diag}(\boldsymbol{\psi}), \qquad \mathrm{diag}(\boldsymbol{\phi} \otimes \boldsymbol{\psi})^{-1} = \mathrm{diag}(\boldsymbol{\phi})^{-1} \otimes \mathrm{diag}(\boldsymbol{\psi})^{-1}.$$

Proof. In order to exploit known results for Markov chains, we transform $\boldsymbol{K}$ so that it becomes a transition rate matrix. Consider a transformation $\boldsymbol{K}_D$ of $\boldsymbol{K}$ by the matrix $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{y}_{\eta_1})$,

$$\boldsymbol{K}_D \equiv \boldsymbol{D}^{-1}\boldsymbol{K}\boldsymbol{D} = \begin{pmatrix} \tilde{\boldsymbol{D}}_0^{-1}\tilde{\boldsymbol{B}}_0\tilde{\boldsymbol{D}}_0 & \tilde{\boldsymbol{D}}_0^{-1}\tilde{\boldsymbol{A}}_0\tilde{\boldsymbol{D}} & & & \\ \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{C}}_1\tilde{\boldsymbol{D}}_0 & \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{D}} & \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}} & & \\ & \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{D}} & \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{D}} & \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}} & \\ & & \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{D}} & \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{D}} & \ddots \\ & & & \ddots & \ddots \end{pmatrix},$$

where $\tilde{\boldsymbol{D}}_0 = \mathrm{diag}(\boldsymbol{v}_0 \otimes \boldsymbol{v}_1)$ and $\tilde{\boldsymbol{D}} = \mathrm{diag}(\boldsymbol{v}_0 \otimes \boldsymbol{v}_1 \otimes \boldsymbol{e}_2)$. In this case, $\boldsymbol{K}_D$ becomes a transition rate matrix of a QBD process with finite number of phases in each level.

We shall show that the transition rate matrix $\boldsymbol{K}_D$ is ergodic if $\eta_2 < 1$, by using Theorem 3.1.1 in [30]. Since the matrix

$$\begin{aligned} &\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}} + \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{D}} + \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{D}} \\ &= \left\{\mathrm{diag}(\boldsymbol{v}_0)^{-1}\left(\frac{1}{\eta_1}\boldsymbol{\gamma}_0\boldsymbol{\alpha} + \boldsymbol{T}\right)\mathrm{diag}(\boldsymbol{v}_0)\right\} \oplus \\ &\qquad \{\mathrm{diag}(\boldsymbol{v}_1)^{-1}(\eta_1\boldsymbol{\gamma}_1\boldsymbol{\beta}_1 + \boldsymbol{S}_1)\,\mathrm{diag}(\boldsymbol{v}_1)\} \oplus (\boldsymbol{\gamma}_2\boldsymbol{\beta}_2 + \boldsymbol{S}_2) \end{aligned} \tag{4.35}$$

is a transition rate matrix of a finite dimension, and also it is irreducible from the assumption (4.3), it has a non-negative invariant vector $\tilde{\boldsymbol{\pi}}$. Theorem 3.1.1 in [30] says that $\boldsymbol{K}_D$ *is ergodic if and only if*

$$\tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}})\tilde{\boldsymbol{e}} < \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{D}})\tilde{\boldsymbol{e}}, \tag{4.36}$$

where $\tilde{\boldsymbol{e}}$ is the column vector of the same dimension as $\tilde{\boldsymbol{D}}$ with all elements being equal to 1.

Since the matrix in (4.35) is the Kronecker sum of three smaller matrices, $\tilde{\boldsymbol{\pi}}$ is represented as the Kronecker product of the eigenvectors of these smaller matrices. If we let

$$\tilde{\boldsymbol{\pi}} = \{\boldsymbol{u}_0 \operatorname{diag}(\boldsymbol{v}_0)\} \otimes \{\boldsymbol{u}_1 \operatorname{diag}(v v_1)\} \otimes \{\boldsymbol{\beta}_2(-\boldsymbol{S}_2)^{-1}\},$$

then from (4.35) and Lemma 4.3.1,

$$\begin{aligned}
&\tilde{\boldsymbol{\pi}}\left(\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}} + \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{D}} + \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{D}}\right) \\
&= \left\{\boldsymbol{u}_0\left(\frac{1}{\eta_1}\boldsymbol{\gamma}_0\boldsymbol{\alpha} + \boldsymbol{T}\right)\operatorname{diag}(\boldsymbol{v}_0)\right\} \otimes \{\boldsymbol{u}_1 \operatorname{diag}(\boldsymbol{v}_1)\} \otimes \{\boldsymbol{\beta}_2(-\boldsymbol{S}_2)^{-1}\} + \\
&\qquad \{\boldsymbol{u}_0 \operatorname{diag}(\boldsymbol{v}_0)\} \otimes \{\boldsymbol{u}_1 (\eta_1\boldsymbol{\gamma}_1\boldsymbol{\beta}_1 + \boldsymbol{S}_1) \operatorname{diag}(\boldsymbol{v}_1)\} \otimes \{\boldsymbol{\beta}_2(-\boldsymbol{S}_2)^{-1}\} + \\
&\qquad \{\boldsymbol{u}_0 \operatorname{diag}(\boldsymbol{v}_0)\} \otimes \{\boldsymbol{u}_1 \operatorname{diag}(\boldsymbol{v}_1)\} \otimes \{\boldsymbol{\beta}_2(-\boldsymbol{S}_2)^{-1}(\boldsymbol{\gamma}_2\boldsymbol{\beta}_2 + \boldsymbol{S}_2)\} \\
&= \{\omega_0\boldsymbol{u}_0 \operatorname{diag}(\boldsymbol{v}_0)\} \otimes \{\boldsymbol{u}_1 \operatorname{diag}(\boldsymbol{v}_1)\} \otimes \{\boldsymbol{\beta}_2(-\boldsymbol{S}_2)^{-1}\} + \\
&\qquad \{\boldsymbol{u}_0 \operatorname{diag}(\boldsymbol{v}_0)\} \otimes \{-\omega_0\boldsymbol{u}_1 \operatorname{diag}(\boldsymbol{v}_1)\} \otimes \{\boldsymbol{\beta}_2(-\boldsymbol{S}_2)^{-1}\} + \\
&\qquad \{\boldsymbol{u}_0 \operatorname{diag}(\boldsymbol{v}_0)\} \otimes \{\boldsymbol{u}_1 \operatorname{diag}(\boldsymbol{v}_1)\} \otimes \{-\boldsymbol{\beta}_2 + \boldsymbol{\beta}_2\} = \boldsymbol{0}.
\end{aligned}$$

Hence $\tilde{\boldsymbol{\pi}}$ is the invariant vector of the matrix (4.35).

Now we evaluate the both sides of (4.36).

$$\begin{aligned}
\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{e}} &= \left[\boldsymbol{u}_0 \otimes \boldsymbol{u}_1 \otimes \left\{\boldsymbol{\beta}_2(-\boldsymbol{S}_2)^{-1}\right\}\right] \times (\eta_1\boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1\boldsymbol{\beta}_1 \otimes \boldsymbol{I}_2) \times (\boldsymbol{v}_0 \otimes \boldsymbol{v}_1 \otimes \boldsymbol{e}_2) \\
&= \{\boldsymbol{\alpha}(\omega_0\boldsymbol{I}_0 - \boldsymbol{T})^{-1}\boldsymbol{v}_0\} \otimes \frac{1}{\eta_1} \otimes \{\boldsymbol{\beta}_2(-\boldsymbol{S}_2)^{-1}\boldsymbol{e}_2\} \\
&= \frac{1}{\eta_1}\, T^{*\prime}(s_0)\Big|_{s_0=\omega_0}\; S_2^{*\prime}(s_2)\Big|_{s_2=0},
\end{aligned}$$

where a prime represents a derivative, and

$$\tilde{\boldsymbol{\pi}}\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{e}} = \left[\left\{\boldsymbol{\alpha}(\omega_0\boldsymbol{I}_0 - \boldsymbol{T})^{-1}\right\} \otimes \left\{\boldsymbol{\beta}_1(-\omega_0\boldsymbol{I}_1 - \boldsymbol{S}_1)^{-1}\right\} \otimes \left\{\boldsymbol{\beta}_2(-\boldsymbol{S}_2)^{-1}\right\}\right] \times$$

$$(\boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \otimes \boldsymbol{\gamma}_2\boldsymbol{\beta}_2) \times (\boldsymbol{v}_0 \otimes \boldsymbol{v}_1 \otimes \boldsymbol{e}_2)$$

$$\begin{aligned} &= \left\{\boldsymbol{\alpha}\left(\omega_0\boldsymbol{I}_0 - \boldsymbol{T}\right)^{-1}\boldsymbol{v}_0\right\} \times \left\{\boldsymbol{\beta}_1\left(-\omega_0\boldsymbol{I}_1 - \boldsymbol{S}_1\right)^{-1}\boldsymbol{v}_1\right\} \\ &= T^{*'}(s_0)\Big|_{s_0=\omega_0}\ S_1^{*'}(s_1)\Big|_{s_1=-\omega_0} = T^{*'}(s_0)\Big|_{s_0=\omega_0}\ S_1^{*'}(-\omega_0 - s_2)\Big|_{s_2=0}. \end{aligned}$$

Therefore, the difference of both sides in inequality (4.36) is

$$\begin{aligned} &\tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}})\tilde{\boldsymbol{e}} - \tilde{\boldsymbol{\pi}}(\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{D}}) \\ &= T^{*'}(s_0)\Big|_{s_0=\omega_0}\left\{\frac{1}{\eta_1}\,S_2^{*'}(s_2)\Big|_{s_2=0} - S_1^{*'}(-\omega_0 - s_2)\Big|_{s_2=0}\right\} \\ &= T^{*'}(s_0)\Big|_{s_0=\omega_0}\left\{S_1^*(-\omega_0)\,S_2^{*'}(s_2)\Big|_{s_2=0} - S_1^{*'}(-\omega_0 - s_2)\Big|_{s_2=0}\,S_2^*(0)\right\} \qquad (4.37) \\ &= T^{*'}(s_0)\Big|_{s_0=\omega_0}\ \{S_1^*(-\omega_0 - s_2)S_2^*(s_2)\}'|_{s_2=0}. \end{aligned}$$

Since the derivative of $T^*(s_0)$ is negative by the monotonicity of LSTs, the inequality (4.36) holds iff the derivative of the function

$$f_2(s_2) = S_1^*(-\omega_0 - s_2)S_2^*(s_2)$$

at the origin is positive. Remember that $f_2(s_2)$ is convex, and the equation $\eta_1 f_2(s_2) = 1$ has two roots, zero and ω_2. This means that the inequality $f_2'(0) > 0$ holds iff $f_2'(\omega_2) < 0$, or equivalently, $\eta_2 < 1$. Thus we have shown that $\boldsymbol{K}_D$ is ergodic iff $\eta_2 < 1$.

Under this condition, we know that there exists a positive vector $\boldsymbol{x}_D$ such that

$$\boldsymbol{x}_D\boldsymbol{K}_D = \boldsymbol{0}, \quad \boldsymbol{x}_D\boldsymbol{e} = 1.$$

Then, the positive vector $\boldsymbol{x}_{\eta_1} = \boldsymbol{x}_D\boldsymbol{D}$ satisfies $\boldsymbol{x}_{\eta_1}\boldsymbol{K} = \boldsymbol{0}$ and $\boldsymbol{x}_{\eta_1}\boldsymbol{y}_{\eta_1} = 1 < \infty$. From the

definition of $\boldsymbol{y}_{\eta_1}$, there exists a positive number d_1 such that $\boldsymbol{y}_{\eta_1} < d_1\boldsymbol{e}$. Hence

$$\boldsymbol{x}_{\eta_1}\boldsymbol{e} < d_1\boldsymbol{x}_{\eta_1}\boldsymbol{y}_{\eta_1} < d_1 < \infty.$$

Lemma 4.3.4.

$$\eta_1^{-1}\boldsymbol{x}_{\eta_1}\boldsymbol{A}\boldsymbol{y}_{\eta_1} \neq \eta_1\boldsymbol{x}_{\eta_1}\boldsymbol{C}\boldsymbol{y}_{\eta_1}.$$

Proof. We define a matrix $\boldsymbol{E}$ as

$$\boldsymbol{E} = \begin{pmatrix} \boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \\ \boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \otimes \boldsymbol{e}_2 \\ \boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \otimes \boldsymbol{e}_2 \\ \vdots \end{pmatrix}.$$

It is clear that

$$\begin{aligned} \boldsymbol{y}_{\eta_1} &= \boldsymbol{E}(\boldsymbol{v}_0 \otimes \boldsymbol{v}_1), \\ \boldsymbol{K}\boldsymbol{E} &= \boldsymbol{E}\left(\frac{1}{\eta_1}\boldsymbol{\gamma}_0\boldsymbol{\alpha} + \boldsymbol{T}\right) \oplus \left(\eta_1\boldsymbol{\gamma}_1\boldsymbol{\beta}_1 + \boldsymbol{S}_1\right), \qquad (4.38) \\ \boldsymbol{A}\boldsymbol{E} = \boldsymbol{E}(\boldsymbol{\gamma}_0\boldsymbol{\alpha} \otimes \boldsymbol{I}_1), &\qquad \boldsymbol{C}\boldsymbol{E} = \boldsymbol{E}(\boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1\boldsymbol{\beta}_1), \end{aligned}$$

and hence

$$\frac{1}{\eta_1}\boldsymbol{A}\boldsymbol{y}_{\eta_1} = \boldsymbol{E}(\boldsymbol{\gamma}_0 \otimes \boldsymbol{v}_1), \quad \eta_1\boldsymbol{C}\boldsymbol{y}_{\eta_1} = \boldsymbol{E}(\boldsymbol{v}_0 \otimes \boldsymbol{\gamma}_1).$$

Postmultiplying $\boldsymbol{E}$ to the equality $\boldsymbol{x}_{\eta_1}\boldsymbol{K} = \boldsymbol{0}$ and applying (4.38), we have

$$\boldsymbol{x}_{\eta_1}\boldsymbol{E}\left\{\left(\frac{1}{\eta_1}\boldsymbol{\gamma}_0\boldsymbol{\alpha} + \boldsymbol{T}\right) \oplus (\eta_1\boldsymbol{\gamma}_1\boldsymbol{\beta}_1 + \boldsymbol{S}_1)\right\} = \boldsymbol{0}.$$

From this equation and Lemma 4.3.1, we see that $\boldsymbol{x}_{\eta_1}\boldsymbol{E} = \boldsymbol{u}_0 \otimes \boldsymbol{u}_1$ up to a multiplicative constant. Again from Lemma 4.3.1, we have

$$\begin{aligned}
\frac{1}{\eta_1}\boldsymbol{x}_{\eta_1}\boldsymbol{A}\boldsymbol{y}_{\eta_1} - \eta_1\boldsymbol{x}_{\eta_1}\boldsymbol{C}\boldsymbol{y}_{\eta_1} &= \boldsymbol{x}_{\eta_1}\boldsymbol{E}(\boldsymbol{\gamma}_0 \otimes \boldsymbol{v}_1) - \boldsymbol{x}_{\eta_1}\boldsymbol{E}(\boldsymbol{v}_0 \otimes \boldsymbol{\gamma}_1) \\
&= \boldsymbol{u}_0\boldsymbol{\gamma}_0 \cdot \boldsymbol{u}_1\boldsymbol{v}_1 - \boldsymbol{u}_0\boldsymbol{v}_0 \cdot \boldsymbol{u}_1\boldsymbol{\gamma}_1 \\
&= \eta_1\boldsymbol{u}_1\boldsymbol{v}_1 - \frac{1}{\eta_1}\boldsymbol{u}_0\boldsymbol{v}_0 \\
&= \boldsymbol{\alpha}(\omega_0\boldsymbol{I}_0 - \boldsymbol{T})^{-1}\boldsymbol{\gamma}_0\boldsymbol{\beta}_1(-\omega_0\boldsymbol{I}_1 - \boldsymbol{S}_1)^{-2}\boldsymbol{\gamma}_1 \\
&\quad - \boldsymbol{\alpha}(\omega_0\boldsymbol{I}_0 - \boldsymbol{T})^{-2}\boldsymbol{\gamma}_0\boldsymbol{\beta}_1(-\omega_0\boldsymbol{I}_1 - \boldsymbol{S}_1)^{-1}\boldsymbol{\gamma}_1 \\
&= \{T^*(s_0)S_1^*(-s_0)\}'\Big|_{s_0=\omega_0}.
\end{aligned}$$

The function $f_1(s_0) = T^*(s_0)S_1^*(-s_0)$ is convex and $f_1(0) = f_1(\omega_0) = 1$. However, ω_0 must be positive since $\eta_1 = T^*(\omega_0) < 1$. It results that $f_1'(\omega_0) > 0$ and the inequality of the lemma holds.

Lemma 4.3.5.

$$\boldsymbol{\pi}_1\boldsymbol{y}_{\eta_1} < \infty.$$

Proof. Since $\boldsymbol{y}_{\eta_1} < d_1\boldsymbol{e}$,

$$\boldsymbol{\pi}_1\boldsymbol{y}_{\eta_1} < d_1\boldsymbol{\pi}_1\boldsymbol{e} < \infty.$$

To prove the asymptotic form of $\boldsymbol{x}_{\eta_1}$, we introduce a partition of each $\mathcal{L}_m$:

$$l_k = \{(n_2; i_0, i_1, i_2) | n_2 = k\}, \quad k = 0, 1, 2, \ldots,$$

and denote by $\boldsymbol{x}_{\eta_1} = (\boldsymbol{x}_{\eta_1}(0) \ \ \boldsymbol{x}_{\eta_1}(1) \ \ \cdots)$ the row vector $\boldsymbol{x}_{\eta_1}$ partitioned according to l_k's.

Lemma 4.3.6. If $\eta_2 < 1$, then

$$\boldsymbol{x}_{\eta_1}(n_2) \sim G_2\boldsymbol{x}_0 \otimes \boldsymbol{x}_1 \otimes \boldsymbol{x}_2,$$

where G_2 is a certain constant and

$$\begin{aligned}
\boldsymbol{x}_0 &= \boldsymbol{\alpha}(\omega_0\boldsymbol{I} - \boldsymbol{T})^{-1}, \\
\boldsymbol{x}_1 &= \boldsymbol{\beta}_1\{-(\omega_0 + \omega_2)\boldsymbol{I} - \boldsymbol{S}_1\}^{-1}, \\
\boldsymbol{x}_2 &= \boldsymbol{\beta}_2(\omega_2\boldsymbol{I} - \boldsymbol{S}_1)^{-1}.
\end{aligned}$$

Proof. The ordinary matrix-geometric theory by Neuts [30] can be applied to a Markov chain having the transition rate matrix $\boldsymbol{K}_D$, since it is a QBD process with finite number of phases in each level.

Let $\tilde{\boldsymbol{R}}_D$ be the rate matrix of $\boldsymbol{K}_D$. Then $\tilde{\boldsymbol{R}}_D$ is the minimal non-negative solution to the matrix equation

$$\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}} + \tilde{\boldsymbol{R}}_D\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{D}} + \tilde{\boldsymbol{R}}_D^2\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{D}} = \boldsymbol{O}.$$

We denote by $\tilde{\eta}$ the Perron-Frobenius eigenvalue of $\tilde{\boldsymbol{R}}_D$ and $\tilde{\boldsymbol{x}}_D$ be the corresponding left eigenvector: $\tilde{\boldsymbol{x}}_D\tilde{\boldsymbol{R}}_D = \tilde{\eta}\tilde{\boldsymbol{x}}_D$. Then we have

$$\begin{aligned}
\boldsymbol{0} &= \tilde{\boldsymbol{x}}_D\left(\frac{1}{\tilde{\eta}}\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{D}} + \tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{D}} + \tilde{\eta}\tilde{\boldsymbol{D}}^{-1}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{D}}\right) \\
&= \tilde{\boldsymbol{x}}_D\left[\left\{\mathrm{diag}(\boldsymbol{v}_0)^{-1}\left(\frac{1}{\eta_1}\boldsymbol{\gamma}_0\boldsymbol{\alpha} + \boldsymbol{T}\right)\mathrm{diag}(\boldsymbol{v}_0)\right\}\oplus\right. \\
&\qquad \left.\left\{\mathrm{diag}(\boldsymbol{v}_1)^{-1}\left(\frac{\eta_1}{\tilde{\eta}}\boldsymbol{\gamma}_1\boldsymbol{\beta}_1 + \boldsymbol{S}_1\right)\mathrm{diag}(\boldsymbol{v}_1)\right\}\oplus(\tilde{\eta}\boldsymbol{\gamma}_2\boldsymbol{\beta}_2 + \boldsymbol{S}_2)\right].
\end{aligned}$$

Since the matrix in brackets is the Kronecker sum of three smaller matrices, $\tilde{\boldsymbol{x}}_D$ can be represented as the Kronecker products of three eigenvectors of these three matrices. That is, $\tilde{\boldsymbol{x}}_D = \tilde{\boldsymbol{x}}_0 \otimes \tilde{\boldsymbol{x}}_1 \otimes \tilde{\boldsymbol{x}}_2$ where $\tilde{\boldsymbol{x}}_0$, $\tilde{\boldsymbol{x}}_1$ and $\tilde{\boldsymbol{x}}_2$ are row vectors satisfying

$$\begin{aligned}
\tilde{\boldsymbol{x}}_0\left\{\mathrm{diag}(\boldsymbol{v}_0)^{-1}\left(\frac{1}{\eta_1}\boldsymbol{\gamma}_0\boldsymbol{\alpha} + \boldsymbol{T}\right)\mathrm{diag}(\boldsymbol{v}_0)\right\} &= \xi_0\tilde{\boldsymbol{x}}_0, \\
\tilde{\boldsymbol{x}}_1\left\{diag(\boldsymbol{v}_1)^{-1}\left(\frac{\eta_1}{\tilde{\eta}}\boldsymbol{\gamma}_1\boldsymbol{\beta}_1 + \boldsymbol{S}_1\right)\mathrm{diag}(\boldsymbol{v}_1)\right\} &= \xi_1\tilde{\boldsymbol{x}}_1, \\
\tilde{\boldsymbol{x}}_2\left\{(\tilde{\eta}\boldsymbol{\gamma}_2\boldsymbol{\beta}_2 + \boldsymbol{S}_2)\right\} &= \xi_2\tilde{\boldsymbol{x}}_2, \\
\xi_0 + \xi_1 + \xi_2 &= 0.
\end{aligned}$$

Then, as in the proof of Lemma 4.3.1 we can show that

$$\eta_1 = \boldsymbol{\alpha}(\xi_0\boldsymbol{I}_0 - \boldsymbol{T})^{-1}\boldsymbol{\gamma}_0 = T^*(\xi_0), \tag{4.39}$$

$$\frac{\tilde{\eta}}{\eta_1} = \boldsymbol{\beta}_1(\xi_1\boldsymbol{I}_1 - \boldsymbol{S}_1)^{-1}\boldsymbol{\gamma}_1 = S_1^*(\xi_1), \tag{4.40}$$

$$\frac{1}{\tilde{\eta}} = \boldsymbol{\beta}_2(\xi_2\boldsymbol{I}_2 - \boldsymbol{S}_2)^{-1}\boldsymbol{\gamma}_2 = S_2^*(\xi_2). \tag{4.41}$$

From the monotonicity of LSTs, we have $\xi_0 = \omega_0$. Eliminating $\tilde{\eta}$ and ξ_1 from (4.40) and (4.41), we obtain the equation (4.7) and hence ξ_2 must be zero or ω_2. Since $\boldsymbol{K}_D$ is ergodic, $\tilde{\eta}$ should be less than 1. Then we have $\xi_2 = \omega_2$ and $\tilde{\eta} = \eta_2$.

We can also derive explicit forms of $\tilde{\boldsymbol{x}}_k$'s as

$$\begin{aligned}
\tilde{\boldsymbol{x}}_0 &= \boldsymbol{\alpha}(\omega_0\boldsymbol{I} - \boldsymbol{T})^{-1}\,\mathrm{diag}(\boldsymbol{v}_0), \\
\tilde{\boldsymbol{x}}_1 &= \boldsymbol{\beta}_1\{-(\omega_0+\omega_2)\boldsymbol{I} - \boldsymbol{S}_1\}^{-1}\,\mathrm{diag}(\boldsymbol{v}_1), \\
\tilde{\boldsymbol{x}}_2 &= \boldsymbol{\beta}_2(\omega_2\boldsymbol{I} - \boldsymbol{S}_1)^{-1},
\end{aligned}$$

up to multiplicative constants. Therefore, $\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{x}}_D\boldsymbol{D}^{-1}$ is an eigenvector of the rate matrix $\tilde{\boldsymbol{R}}$ of $\boldsymbol{K}$ corresponding to the eigenvalue $\tilde{\eta} = \eta_2$.

It results that, if $\eta_2 < 1$, the vector $\boldsymbol{x}_{\eta_1} = (\boldsymbol{x}_{\eta_1}(0) \;\; \boldsymbol{x}_{\eta_1}(1) \;\; \cdots)$ has a geometric tail

$$\begin{aligned}
\boldsymbol{x}_{\eta_1}(n_2) &\sim G_2\tilde{\eta}^{n_2}\tilde{\boldsymbol{x}} \\
&= G_2\eta_2^{n_2}\boldsymbol{\alpha}(\omega_0\boldsymbol{I} - \boldsymbol{T})^{-1} \otimes \boldsymbol{\beta}_1\{-(\omega_0+\omega_2)\boldsymbol{I} - \boldsymbol{S}_1\}^{-1} \otimes \boldsymbol{\beta}_2(\omega_2\boldsymbol{I} - \boldsymbol{S}_2)^{-1}, \qquad n_2 \to \infty.
\end{aligned}$$

This is the vector representation of (4.14).

Proof of Theorem 4.2.1 Since there are no redundant phases in $PH(\boldsymbol{\alpha}, \boldsymbol{T})$ and $PH(\boldsymbol{\beta}_k, \boldsymbol{S}_k)$ $(k = 1, 2)$, it is clear that $\boldsymbol{Q}'$ is irreducible. Then the above lemmas prove the whole of the theorem. □

4.4 Rate Matrix and Its Invariant Vectors for n_2-based Decomposition

Next, we rearrange the states $(n_1, n_2; i_0, i_1, i_2)$ of $\{X(t)\}$ first in order of n_2, and then for fixed n_2 they are arranged lexicographically. We define a new partition of the state space by

$$\overline{\mathcal{L}}_m = \{(n_1, n_2; i_0, i_1, i_2) | n_2 = m\}, \quad m = 0, 1, 2, \ldots. \tag{4.42}$$

We denote by $\overline{\boldsymbol{Q}}$ the transition rate matrix of the chain corresponding to the arrangement above, and by $\overline{\boldsymbol{\pi}} = (\overline{\boldsymbol{\pi}}_0 \ \ \overline{\boldsymbol{\pi}}_1 \ \ \cdots)$ the stationary vector of $\{X(t)\}$ partitioned according to $\overline{\mathcal{L}}_m$'s. Then $\overline{\boldsymbol{Q}}$ is of a block-tridiagonal form

$$\overline{\boldsymbol{Q}} = \begin{pmatrix} \overline{\boldsymbol{B}}_0 & \overline{\boldsymbol{A}}_0 & & & \\ \overline{\boldsymbol{C}}_1 & \overline{\boldsymbol{B}} & \overline{\boldsymbol{A}} & & \\ & \overline{\boldsymbol{C}} & \overline{\boldsymbol{B}} & \overline{\boldsymbol{A}} & \\ & & \overline{\boldsymbol{C}} & \overline{\boldsymbol{B}} & \ddots \\ & & & \ddots & \ddots \end{pmatrix}, \tag{4.43}$$

where

$$\overline{\boldsymbol{A}} = \begin{pmatrix} \boldsymbol{O} & & & & \\ \boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1 \otimes \boldsymbol{I}_2 & \boldsymbol{O} & & & \\ & \boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1 \boldsymbol{\beta}_1 \otimes \boldsymbol{I}_2 & \boldsymbol{O} & & \\ & & \boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1 \boldsymbol{\beta}_1 \otimes \boldsymbol{I}_2 & \boldsymbol{O} & \\ & & & \ddots & \ddots \end{pmatrix}, \tag{4.44}$$

$$\overline{\boldsymbol{B}} = \begin{pmatrix} \boldsymbol{T} \oplus \boldsymbol{S}_2 & \boldsymbol{\gamma}_0 \boldsymbol{\alpha} \otimes \boldsymbol{\beta}_1 \otimes \boldsymbol{I}_2 & & & \\ & \boldsymbol{T} \oplus \boldsymbol{S}_1 \oplus \boldsymbol{S}_2 & \boldsymbol{\gamma}_0 \boldsymbol{\alpha} \otimes \boldsymbol{I}_1 \otimes \boldsymbol{I}_2 & & \\ & & \boldsymbol{T} \oplus \boldsymbol{S}_1 \oplus \boldsymbol{S}_2 & \boldsymbol{\gamma}_0 \boldsymbol{\alpha} \otimes \boldsymbol{I}_1 \otimes \boldsymbol{I}_2 & \\ & & & \boldsymbol{T} \oplus \boldsymbol{S}_1 \oplus \boldsymbol{S}_2 & \ddots \\ & & & & \ddots \end{pmatrix}, \tag{4.45}$$

$$\overline{\boldsymbol{C}} = \begin{pmatrix} \boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_2\boldsymbol{\beta}_2 & & & \\ & \boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \otimes \boldsymbol{\gamma}_2\boldsymbol{\beta}_2 & & \boldsymbol{O} \\ & & \boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \otimes \boldsymbol{\gamma}_2\boldsymbol{\beta}_2 & \\ & \boldsymbol{O} & & \boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \otimes \boldsymbol{\gamma}_2\boldsymbol{\beta}_2 \\ & & & & \ddots \end{pmatrix}. \quad (4.46)$$

We define the following vectors:

$$\overline{\boldsymbol{u}}_0 = \boldsymbol{\alpha}(-\overline{\omega}_2\boldsymbol{I}_0 - \boldsymbol{T})^{-1}, \qquad \overline{\boldsymbol{u}}_2 = \boldsymbol{\beta}_2(\overline{\omega}_2\boldsymbol{I}_2 - \boldsymbol{S}_2)^{-1}, \quad (4.47)$$

$$\overline{\boldsymbol{v}}_0 = (-\overline{\omega}_2\boldsymbol{I}_0 - \boldsymbol{T})^{-1}\boldsymbol{\gamma}_0, \qquad \overline{\boldsymbol{v}}_2 = (\overline{\omega}_2\boldsymbol{I}_2 - \boldsymbol{S}_2)^{-1}\boldsymbol{\gamma}_2. \quad (4.48)$$

Lemma 4.4.1.

$$\overline{\boldsymbol{u}}_0\boldsymbol{\gamma}_0 = \boldsymbol{\alpha}\overline{\boldsymbol{v}}_0 = T^*(-\overline{\omega}_2) = \overline{\eta}_2, \qquad \overline{\boldsymbol{u}}_2\boldsymbol{\gamma}_2 = \boldsymbol{\beta}_2\overline{\boldsymbol{v}}_2 = S_2^*(\overline{\omega}_2) = \frac{1}{\overline{\eta}_2}, \quad (4.49)$$

$$\overline{\boldsymbol{u}}_0\left(\frac{1}{\overline{\eta}_2}\boldsymbol{\gamma}_0\boldsymbol{\alpha} + \boldsymbol{T}\right) = -\overline{\omega}_2\overline{\boldsymbol{u}}_0, \qquad \overline{\boldsymbol{u}}_2\left(\overline{\eta}_2\boldsymbol{\gamma}_2\boldsymbol{\beta}_2 + \boldsymbol{S}_2\right) = \overline{\omega}_2\overline{\boldsymbol{u}}_2, \quad (4.50)$$

$$\left(\frac{1}{\overline{\eta}_2}\boldsymbol{\gamma}_0\boldsymbol{\alpha} + \boldsymbol{T}\right)\overline{\boldsymbol{v}}_0 = -\overline{\omega}_2\overline{\boldsymbol{v}}_0, \qquad \left(\overline{\eta}_2\boldsymbol{\gamma}_2\boldsymbol{\beta}_2 + \boldsymbol{S}_2\right)\overline{\boldsymbol{v}}_2 = \overline{\omega}_2\overline{\boldsymbol{v}}_2. \quad (4.51)$$

Proof. All these equations are easily proved as in Lemma 4.3.1.

First, we define $\overline{\boldsymbol{y}}_{\overline{\eta}_2}$ to satisfy (4.19) in Proposition 1 for $\eta = \overline{\eta}_2$. Let

$$\overline{\boldsymbol{K}} \equiv \left(\frac{1}{\overline{\eta}_2}\overline{\boldsymbol{A}} + \overline{\boldsymbol{B}} + \overline{\eta}_2\overline{\boldsymbol{C}}\right) = \begin{pmatrix} \widehat{\boldsymbol{B}}_0 & \widehat{\boldsymbol{A}}_0 & & \\ \widehat{\boldsymbol{C}}_1 & \widehat{\boldsymbol{B}} & \widehat{\boldsymbol{A}} & \\ & \widehat{\boldsymbol{C}} & \widehat{\boldsymbol{B}} & \widehat{\boldsymbol{A}} \\ & & \widehat{\boldsymbol{C}} & \widehat{\boldsymbol{B}} & \ddots \\ & & & \ddots & \ddots \end{pmatrix},$$

where

$$\begin{aligned}
\widehat{\boldsymbol{A}}_0 &= \boldsymbol{\gamma}_0\boldsymbol{\alpha} \otimes \boldsymbol{\beta}_1 \otimes \boldsymbol{I}_2, \\
\widehat{\boldsymbol{B}}_0 &= \boldsymbol{T} \oplus \boldsymbol{S}_2 + \overline{\eta}_2 \boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_2\boldsymbol{\beta}_2, \\
\widehat{\boldsymbol{C}}_1 &= \frac{1}{\overline{\eta}_2}\boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1 \otimes \boldsymbol{I}_2, \\
\widehat{\boldsymbol{A}} &= \boldsymbol{\gamma}_0\boldsymbol{\alpha} \otimes \boldsymbol{I}_1 \otimes \boldsymbol{I}_2, \\
\widehat{\boldsymbol{B}} &= \boldsymbol{T} \oplus \boldsymbol{S}_1 \oplus \boldsymbol{S}_2 + \overline{\eta}_2 \boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \otimes \boldsymbol{\gamma}_2\boldsymbol{\beta}_2, \\
\widehat{\boldsymbol{C}} &= \frac{1}{\overline{\eta}_2}\boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1\boldsymbol{\beta}_1 \otimes \boldsymbol{I}_2.
\end{aligned} \tag{4.52}$$

We define $\overline{\boldsymbol{y}}_{\overline{\eta}_2}$ as

$$\overline{\boldsymbol{y}}_{\overline{\eta}_2} = \begin{pmatrix} \overline{\boldsymbol{v}}_0 \otimes \overline{\boldsymbol{v}}_2 \\ \overline{\eta}_2^{-1}\overline{\boldsymbol{v}}_0 \otimes \boldsymbol{e}_1 \otimes \overline{\boldsymbol{v}}_2 \\ \overline{\eta}_2^{-2}\overline{\boldsymbol{v}}_0 \otimes \boldsymbol{e}_1 \otimes \overline{\boldsymbol{v}}_2 \\ \vdots \end{pmatrix}.$$

Lemma 4.4.2.

$$\overline{\boldsymbol{K}}\overline{\boldsymbol{y}}_{\overline{\eta}_2} = \boldsymbol{0}.$$

Proof. From Lemma 4.4.1, it is easily checked that $\overline{\boldsymbol{K}}\overline{\boldsymbol{y}}_{\overline{\eta}_2} = \boldsymbol{0}$ in a similar manner as in Lemma 4.3.2.

Lemma 4.4.3. If $\overline{\eta}_1 < \overline{\eta}_2$, then there exists a positive vector $\overline{\boldsymbol{x}}_{\overline{\eta}_2}$ such that

$$\overline{\boldsymbol{x}}_{\overline{\eta}_2}\overline{\boldsymbol{K}} = \boldsymbol{0}, \quad \overline{\boldsymbol{x}}_{\overline{\eta}_2}\boldsymbol{e} < \infty, \quad \overline{\boldsymbol{x}}_{\overline{\eta}_2}\overline{\boldsymbol{y}}_{\overline{\eta}_2} < \infty.$$

Proof. We consider the transformation $\overline{\boldsymbol{K}}_{\overline{D}}$ of $\overline{\boldsymbol{K}}$ by $\overline{\boldsymbol{D}} = diag(\overline{\boldsymbol{y}}_{\overline{\eta}_2})$. That is, we define $\overline{\boldsymbol{K}}_{\overline{D}}$ as

$$\overline{\boldsymbol{K}}_{\overline{D}} = \begin{pmatrix} \widehat{\boldsymbol{D}}_0^{-1}\widehat{\boldsymbol{B}}_0\widehat{\boldsymbol{D}}_0 & \overline{\eta}_2^{-1}\widehat{\boldsymbol{D}}_0^{-1}\widehat{\boldsymbol{A}}_0\widehat{\boldsymbol{D}} & & \\ \overline{\eta}_2\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{C}}_1\widehat{\boldsymbol{D}}_0 & \widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{D}} & \overline{\eta}_2^{-1}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}} & \\ & \overline{\eta}_2\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{C}}\widehat{\boldsymbol{D}} & \widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{D}} & \overline{\eta}_2^{-1}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}} \\ & & \overline{\eta}_2\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{C}}\widehat{\boldsymbol{D}} & \widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{D}} \\ & & & \ddots \end{pmatrix}.$$

where $\widehat{\boldsymbol{D}}_0 = diag(\overline{\boldsymbol{v}}_0 \otimes \overline{\boldsymbol{v}}_2)$ and $\widehat{\boldsymbol{D}} = diag(\overline{\boldsymbol{v}}_0 \otimes \boldsymbol{e}_1 \otimes \overline{\boldsymbol{v}}_2)$. Note that $\overline{\boldsymbol{K}}_{\overline{D}}$ is a transition rate matrix of a Markov chain which is a QBD process with finitely many phases in each level. We shall show that the Markov chain is ergodic using the matrix-geometric theory by Neuts.

Since the matrix $(\overline{\eta}_2^{-1}\widehat{\boldsymbol{A}} + \widehat{\boldsymbol{B}} + \overline{\eta}_2\widehat{\boldsymbol{C}})$ can be rewritten as

$$\{(\overline{\eta}_2^{-1}\boldsymbol{\gamma}_0\boldsymbol{\alpha} + \boldsymbol{T}) \oplus (\boldsymbol{\gamma}_1\boldsymbol{\beta}_1 + \boldsymbol{S}_1) \oplus (\overline{\eta}_2\boldsymbol{\gamma}_2\boldsymbol{\beta}_2 + \boldsymbol{S}_2)\},$$

we let

$$\widehat{\boldsymbol{\pi}} = \overline{\boldsymbol{u}}_0 \otimes \{\boldsymbol{\beta}_1(-\boldsymbol{S}_1)^{-1}\} \otimes \overline{\boldsymbol{u}}_2. \tag{4.53}$$

Then $\widehat{\boldsymbol{\pi}}$ satisfies the equation

$$\widehat{\boldsymbol{\pi}}(\overline{\eta}_2^{-1}\widehat{\boldsymbol{A}} + \widehat{\boldsymbol{B}} + \overline{\eta}_2\widehat{\boldsymbol{C}}) = \boldsymbol{0}.$$

Thus the vector $\widehat{\boldsymbol{\pi}}\widehat{\boldsymbol{D}}$ satisfies the equation

$$\widehat{\boldsymbol{\pi}}\widehat{\boldsymbol{D}}(\overline{\eta}_2^{-1}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}} + \widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{D}} + \overline{\eta}_2\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{C}}\widehat{\boldsymbol{D}}) = \boldsymbol{0},$$

Employing Theorem 3.1.1 in [30], $\overline{\boldsymbol{K}}_{\overline{D}}$ is ergodic iff

$$\widehat{\boldsymbol{\pi}}\widehat{\boldsymbol{D}}(\overline{\eta}_2^{-1}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}})\widehat{\boldsymbol{e}} < \widehat{\boldsymbol{\pi}}\widehat{\boldsymbol{D}}(\overline{\eta}_2^{-1}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{C}}\widehat{\boldsymbol{D}})\widehat{\boldsymbol{e}}, \tag{4.54}$$

where $\widehat{\boldsymbol{e}}$ is the column vector of the same dimension as $\widehat{\boldsymbol{D}}$ with all elements being equal to 1. The difference of both sides of the above inequality is rewritten as

$$\begin{aligned}
&\widehat{\boldsymbol{\pi}}\widehat{\boldsymbol{D}}(\overline{\eta}_2^{-1}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}})\widehat{\boldsymbol{e}} - \widehat{\boldsymbol{\pi}}\widehat{\boldsymbol{D}}(\overline{\eta}_2^{-1}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{C}}\widehat{\boldsymbol{D}})\widehat{\boldsymbol{e}} \\
&= \overline{\eta}_2^{-1}\widehat{\boldsymbol{\pi}}\widehat{\boldsymbol{A}}(\overline{\boldsymbol{v}}_0 \otimes \boldsymbol{e}_1 \otimes \overline{\boldsymbol{v}}_2) - \overline{\eta}_2\widehat{\boldsymbol{\pi}}\widehat{\boldsymbol{C}}(\overline{\boldsymbol{v}}_0 \otimes \boldsymbol{e}_1 \otimes \overline{\boldsymbol{v}}_2) \\
&= \left\{\boldsymbol{\alpha}(-\overline{\omega}_2\boldsymbol{I}_0 - \boldsymbol{T})^{-1}\boldsymbol{\gamma}_0\boldsymbol{\alpha}_0\overline{\boldsymbol{v}}_0\right\} \otimes \left\{\boldsymbol{\beta}_1(-\boldsymbol{S}_1)^{-1}\boldsymbol{e}_1\right\} \otimes \left\{\boldsymbol{\beta}_2(\overline{\omega}_2\boldsymbol{I}_2 - \boldsymbol{S}_2)^{-1}\overline{\boldsymbol{v}}_2\right\} - \\
&\qquad \overline{\eta}_2\left\{\boldsymbol{\alpha}(-\overline{\omega}_2\boldsymbol{I}_0 - \boldsymbol{T})^{-1}\overline{\boldsymbol{v}}_0\right\} \otimes \left\{\boldsymbol{\beta}_1(-\boldsymbol{S}_1)^{-1}\boldsymbol{\gamma}_1\boldsymbol{\beta}_1\right\} \otimes \left\{\boldsymbol{\beta}_2(\overline{\omega}_2\boldsymbol{I}_2 - \boldsymbol{S}_2)^{-1}\overline{\boldsymbol{v}}_2\right\} \\
&= \overline{\eta}_2 T^*(-\overline{\omega}_2)\ S_1^{*'}(-s_0 - \overline{\omega}_2)\Big|_{s0=-\overline{\omega}_2} \cdot \{-S_2^{*'}(s_2)\}\Big|_{s_2=\overline{\omega}_2} - \\
&\qquad \overline{\eta}_2\ T^{*'}(s_0)\Big|_{s_0=-\overline{\omega}_2}\ S_1^*(0)\ \{-S_2^{*'}(s_2)\}\Big|_{s_2=\overline{\omega}_2} \\
&= \overline{\eta}_2\ \{T^*(s_0)S_1^*(-s_0 - \overline{\omega}_2)\}'|_{s_0=-\overline{\omega}_2}\ \{-S_2^{*'}(s_2)\}\Big|_{s_2=\overline{\omega}_2}. \qquad (4.55)
\end{aligned}$$

It is easily shown that the function

$$g(s_0) = T^*(s_0)S_1^*(-s_0 - \overline{\omega}_2)$$

is convex, and hence the equation $g(s_0) = \overline{\eta}_2$ has two roots, $-\overline{\omega}_2$ and $\overline{\omega}_0$. From the assumption $\overline{\eta}_1 < \overline{\eta}_2$, we have $\overline{\omega}_0 > -\overline{\omega}_2$ since $\overline{\eta}_1 = T^*(\overline{\omega}_0)$ and $\overline{\eta}_2 = T^*(-\overline{\omega}_2)$. Hence

$$g'(s_0)|_{s_0=-\overline{\omega}_2} = \{T^*(s_0)S_1^*(-s_0 - \overline{\omega}_2)\}'|_{s_0=-\overline{\omega}_2} < 0,$$

and this implies the difference (4.55) is negative. Thus the inequality (4.54) holds, and the Markov chain with transition rate matrix $\overline{\boldsymbol{K}}_{\overline{D}}$ is ergodic. Then there exists a positive vector $\overline{\boldsymbol{x}}_{\overline{D}}$ such that $\overline{\boldsymbol{x}}_{\overline{D}}\overline{\boldsymbol{K}}_{\overline{D}} = \boldsymbol{0}$ and $\overline{\boldsymbol{x}}_{\overline{D}}\boldsymbol{e} = 1$. If we define $\overline{\boldsymbol{x}}_{\overline{\eta}_2} = \overline{\boldsymbol{x}}_{\overline{D}}\overline{\boldsymbol{D}}^{-1}$,

$$\overline{\boldsymbol{x}}_{\overline{\eta}_2}(\widehat{\boldsymbol{A}} + \overline{\eta}_2\widehat{\boldsymbol{B}} + \overline{\eta}_2^2\widehat{\boldsymbol{C}}) = \boldsymbol{0}, \quad \overline{\boldsymbol{x}}_{\overline{\eta}_2}\overline{\boldsymbol{y}}_{\overline{\eta}_2} = 1 < \infty.$$

From the definition of $\overline{\boldsymbol{y}}_{\overline{\eta}_2}$, there exists a positive constant d_2 such that $\boldsymbol{e} < d_2\overline{\boldsymbol{y}}_{\overline{\eta}_2}$. Hence

$$\overline{\boldsymbol{x}}_{\overline{\eta}_2}\boldsymbol{e} < d_2\overline{\boldsymbol{x}}_{\overline{\eta}_2}\overline{\boldsymbol{y}}_{\overline{\eta}_2} < \infty.$$

Lemma 4.4.4.

$$\overline{\eta}_2^{-1}\overline{\boldsymbol{x}}_{\overline{\eta}_2}\overline{\boldsymbol{A}}\overline{\boldsymbol{y}}_{\overline{\eta}_2} < \overline{\eta}_2\overline{\boldsymbol{x}}_{\overline{\eta}_2}\overline{\boldsymbol{C}}\overline{\boldsymbol{y}}_{\overline{\eta}_2}.$$

Proof. By the definition, we can rewrite $\overline{\boldsymbol{K}}_{\overline{D}}$ as

$$\overline{\boldsymbol{K}}_{\overline{D}} = \overline{\boldsymbol{K}}_{01} \oplus (\widehat{\boldsymbol{S}}_2 + \overline{\eta}_2\widehat{\gamma}_2\widehat{\boldsymbol{\beta}}_2 - \overline{\omega}_2\boldsymbol{I}_2), \tag{4.56}$$

where

$$\overline{\boldsymbol{K}}_{01} = \begin{pmatrix} \widehat{\boldsymbol{T}} + \overline{\omega}_2\boldsymbol{I}_2 & \overline{\eta}_2^{-1}\widehat{\gamma}_0\widehat{\boldsymbol{\alpha}} \otimes \beta_1 & & & \\ \boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1 & \widehat{\boldsymbol{T}} \oplus \boldsymbol{S}_1 + \overline{\omega}_2\boldsymbol{I}_0 \otimes \boldsymbol{I}_1 & \overline{\eta}_2^{-1}\widehat{\gamma}_0\widehat{\boldsymbol{\alpha}} \otimes \boldsymbol{I}_1 & & \\ & \boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1\boldsymbol{\beta}_1 & \widehat{\boldsymbol{T}} \oplus \boldsymbol{S}_1 + \overline{\omega}_2\boldsymbol{I}_0 \otimes \boldsymbol{I}_1 & \overline{\eta}_2^{-1}\widehat{\gamma}_0\widehat{\boldsymbol{\alpha}} \otimes \boldsymbol{I}_1 & \\ & & \boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1\boldsymbol{\beta}_1 & \widehat{\boldsymbol{T}} \oplus \boldsymbol{S}_1 + \overline{\omega}_2\boldsymbol{I}_0 \otimes \boldsymbol{I}_1 & \ddots \\ & & & \ddots & \ddots \end{pmatrix},$$

and

$$\widehat{\boldsymbol{T}} = \mathrm{diag}(\overline{\boldsymbol{v}}_0)^{-1}\boldsymbol{T}\,\mathrm{diag}(\overline{\boldsymbol{v}}_0), \qquad \widehat{\boldsymbol{S}}_2 = \mathrm{diag}(\overline{\boldsymbol{v}}_2)^{-1}\boldsymbol{S}_2\,\mathrm{diag}(\overline{\boldsymbol{v}}_2),$$

$$\widehat{\boldsymbol{\alpha}} = \boldsymbol{\alpha}\,\mathrm{diag}(\overline{\boldsymbol{v}}_0), \qquad \widehat{\boldsymbol{\beta}}_2 = \boldsymbol{\beta}_2\,\mathrm{diag}(\overline{\boldsymbol{v}}_2),$$

$$\widehat{\boldsymbol{\gamma}}_0 = \mathrm{diag}(\overline{\boldsymbol{v}}_0)^{-1}\boldsymbol{\gamma}_0, \qquad \widehat{\boldsymbol{\gamma}}_2 = \mathrm{diag}(\overline{\boldsymbol{v}}_2)^{-1}\boldsymbol{\gamma}_2.$$

It is easily checked that $\overline{\boldsymbol{K}}_{01}$ is a transition rate matrix of a Markov chain which is a QBD process with finitely many states in each level. The ergodicity of $\overline{\boldsymbol{K}}_{01}$ is shown by virtue of the Theorem 3.1.1 of Neuts [30] again, and hence there exists the stationary probability vector $\overline{\boldsymbol{x}}_{01}$ of the chain such that

$$\overline{\boldsymbol{x}}_{01}\overline{\boldsymbol{K}}_{01} = \boldsymbol{0}, \quad \overline{\boldsymbol{x}}_{01}\boldsymbol{e} = 1.$$

From (4.56), $\overline{\boldsymbol{x}}_{\overline{D}}$ is written as

$$\overline{\boldsymbol{x}}_{\overline{D}} = (\overline{\boldsymbol{u}}_2\overline{\boldsymbol{v}}_2)^{-1}\overline{\boldsymbol{x}}_{01} \otimes \overline{\boldsymbol{u}}_2\,\mathrm{diag}(\overline{\boldsymbol{v}}_2).$$

Then we have

$$\begin{aligned}
\overline{\eta}_2\overline{\boldsymbol{x}}_{\overline{\eta}_2}\overline{\boldsymbol{C}}\overline{\boldsymbol{y}}_{\overline{\eta}_2} &= \overline{\eta}_2\overline{\boldsymbol{x}}_{\overline{D}}\overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{C}}\overline{\boldsymbol{y}}_{\overline{\eta}_2} \\
&= [\overline{\boldsymbol{x}}_{01} \otimes \{\overline{\boldsymbol{u}}_2\,\mathrm{diag}(\overline{\boldsymbol{v}}_2)\}]
\begin{pmatrix}
\boldsymbol{e}_0 \otimes \mathrm{diag}(\overline{\boldsymbol{v}}_2)^{-1}\boldsymbol{\gamma}_2 \\
\boldsymbol{e}_0 \otimes \boldsymbol{e}_1 \otimes \mathrm{diag}(\overline{\boldsymbol{v}}_2)^{-1}\boldsymbol{\gamma}_2 \\
\boldsymbol{e}_0 \otimes \boldsymbol{e}_1 \otimes \mathrm{diag}(\overline{\boldsymbol{v}}_2)^{-1}\boldsymbol{\gamma}_2 \\
\vdots
\end{pmatrix} \\
&= \frac{1}{\overline{\eta}_2}.
\end{aligned}$$

On the other hand, since

$$\frac{1}{\overline{\eta}_2}\overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{A}}\overline{\boldsymbol{y}}_{\overline{\eta}_2} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{e}_0 \otimes \boldsymbol{\gamma}_1 \otimes \boldsymbol{e}_2 \\ \boldsymbol{e}_0 \otimes \boldsymbol{\gamma}_1 \otimes \boldsymbol{e}_2 \\ \vdots \end{pmatrix},$$

we have

$$\begin{aligned} \frac{1}{\overline{\eta}_2}\overline{\boldsymbol{x}}_{\overline{\eta}_2}\overline{\boldsymbol{A}}\overline{\boldsymbol{y}}_{\overline{\eta}_2} &= \frac{1}{\overline{\eta}_2}\overline{\boldsymbol{x}}_{\overline{D}}\overline{\boldsymbol{D}}^{-1}\overline{\boldsymbol{A}}\overline{\boldsymbol{y}}_{\overline{\eta}_2} \\ &= [\overline{\boldsymbol{x}}_{01} \otimes \{\overline{\boldsymbol{u}}_2 \operatorname{diag}(\overline{\boldsymbol{v}}_2)\}] \begin{pmatrix} 0 \\ \boldsymbol{e}_0 \otimes \boldsymbol{\gamma}_1 \otimes \boldsymbol{e}_2 \\ \boldsymbol{e}_0 \otimes \boldsymbol{\gamma}_1 \otimes \boldsymbol{e}_2 \\ \vdots \end{pmatrix} \\ &= \overline{\boldsymbol{x}}_{01} \begin{pmatrix} 0 \\ \boldsymbol{e}_0 \otimes \boldsymbol{\gamma}_1 \\ \boldsymbol{e}_0 \otimes \boldsymbol{\gamma}_1 \\ \vdots \end{pmatrix}. \end{aligned} \tag{4.57}$$

Note that $\boldsymbol{e}_0 \otimes \boldsymbol{\gamma}_1 = (\boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1\boldsymbol{\beta}_1)(\boldsymbol{e}_0 \otimes \boldsymbol{e}_1)$. This means that the i-th element of $\boldsymbol{e}_0 \otimes \boldsymbol{\gamma}_1$ is the rate that the Markov chain $\overline{\boldsymbol{K}}_{01}$ at state (n, i) goes down to level $(n-1)$. Hence the quantity (4.57) is the rate that the chain $\overline{\boldsymbol{K}}_{01}$ goes one level down. From the balance equation, this rate is equal to the rate that the chain goes one level up:

$$\frac{1}{\overline{\eta}_2}\overline{\boldsymbol{x}}_{\overline{\eta}_2}\overline{\boldsymbol{A}}\overline{\boldsymbol{y}}_{\overline{\eta}_2} = \overline{\boldsymbol{x}}_{01} \begin{pmatrix} \widehat{\boldsymbol{\gamma}}_0 \\ \widehat{\boldsymbol{\gamma}}_0 \otimes \boldsymbol{e}_1 \\ \widehat{\boldsymbol{\gamma}}_0 \otimes \boldsymbol{e}_1 \\ \vdots \end{pmatrix} = \begin{pmatrix} \boldsymbol{I}_0 \\ \boldsymbol{I}_0 \otimes \boldsymbol{e}_1 \\ \boldsymbol{I}_0 \otimes \boldsymbol{e}_1 \\ \vdots \end{pmatrix} \widehat{\boldsymbol{\gamma}}_0. \tag{4.58}$$

Let

$$\boldsymbol{E}_{01} = \begin{pmatrix} \boldsymbol{I}_0 \\ \boldsymbol{I}_0 \otimes \boldsymbol{e}_1 \\ \boldsymbol{I}_0 \otimes \boldsymbol{e}_1 \\ \vdots \end{pmatrix},$$

and postmultiply it to $\overline{\boldsymbol{x}}_{01}\overline{\boldsymbol{K}}_{01} = \boldsymbol{0}$. Then we have

$$\begin{aligned}\boldsymbol{0} &= \overline{\boldsymbol{x}}_{01}\overline{\boldsymbol{K}}_{01}\boldsymbol{E}_{01} = \overline{\boldsymbol{x}}_{01}\begin{pmatrix} \widehat{\boldsymbol{T}} + \overline{\omega}_2\boldsymbol{I}_2 + \overline{\eta}_2^{-1}\widehat{\boldsymbol{\gamma}}_0\boldsymbol{\alpha} \\ (\widehat{\boldsymbol{T}} + \overline{\omega}_2\boldsymbol{I}_2 + \overline{\eta}_2^{-1}\widehat{\boldsymbol{\gamma}}_0\boldsymbol{\alpha}) \otimes \boldsymbol{e}_1 \\ (\widehat{\boldsymbol{T}} + \overline{\omega}_2\boldsymbol{I}_2 + \overline{\eta}_2^{-1}\widehat{\boldsymbol{\gamma}}_0\boldsymbol{\alpha}) \otimes \boldsymbol{e}_1 \\ \vdots \end{pmatrix} \\ &= \overline{\boldsymbol{x}}_{01}\boldsymbol{E}_{01}(\widehat{\boldsymbol{T}} + \overline{\omega}_2\boldsymbol{I}_2 + \overline{\eta}_2^{-1}\widehat{\boldsymbol{\gamma}}_0\boldsymbol{\alpha}).\end{aligned}$$

This indicates that $\overline{\boldsymbol{x}}_{01}\boldsymbol{E}_{01}$ can be regarded as the stationary probability vector of a Markov chain with transition rate matrix $\widehat{\boldsymbol{T}} + \overline{\omega}_2\boldsymbol{I}_2 + \overline{\eta}_2^{-1}\widehat{\boldsymbol{\gamma}}_0\boldsymbol{\alpha}$, and is given by

$$\overline{\boldsymbol{x}}_{01}\boldsymbol{E}_{01} = (\overline{\boldsymbol{u}}_0\overline{\boldsymbol{v}}_0)^{-1}\overline{\boldsymbol{u}}_0\,\mathrm{diag}(\overline{\boldsymbol{v}}_0),$$

where $\overline{\boldsymbol{u}}_0$ is defined in (4.48). Thus, from (4.58), we have

$$\frac{1}{\overline{\eta}_2}\overline{\boldsymbol{x}}_{\overline{\eta}_2}\overline{\boldsymbol{A}}\overline{\boldsymbol{y}}_{\overline{\eta}_2} = (\overline{\boldsymbol{u}}_0\overline{\boldsymbol{v}}_0)^{-1}\overline{\boldsymbol{u}}_0\,\mathrm{diag}(\overline{\boldsymbol{v}}_0)\widehat{\boldsymbol{\gamma}}_0 = (\overline{\boldsymbol{u}}_0\overline{\boldsymbol{v}}_0)^{-1}\cdot\overline{\eta}_2.$$

Therefore,

$$\begin{aligned}\overline{\eta}_2\overline{\boldsymbol{x}}_{\overline{\eta}_2}\overline{\boldsymbol{C}}\overline{\boldsymbol{y}}_{\overline{\eta}_2} - \frac{1}{\overline{\eta}_2}\overline{\boldsymbol{x}}_{\overline{\eta}_2}\overline{\boldsymbol{A}}\overline{\boldsymbol{y}}_{\overline{\eta}_2} &= (\overline{\boldsymbol{u}}_2\overline{\boldsymbol{v}}_2)^{-1}\frac{1}{\overline{\eta}_2} - \frac{\overline{\eta}_2}{\overline{\boldsymbol{u}}_0\overline{\boldsymbol{v}}_0} \\ &= \frac{(\overline{\boldsymbol{u}}_2\overline{\boldsymbol{v}}_2)^{-1}}{\overline{\boldsymbol{u}}_0\overline{\boldsymbol{v}}_0}\left\{\frac{1}{\overline{\eta}_2}\overline{\boldsymbol{u}}_0\overline{\boldsymbol{v}}_0 - \overline{\eta}_2\overline{\boldsymbol{u}}_2\overline{\boldsymbol{v}}_2\right\} \\ &= \frac{1}{(\overline{\boldsymbol{u}}_2\overline{\boldsymbol{v}}_2)(\overline{\boldsymbol{u}}_0\overline{\boldsymbol{v}}_0)}\left\{T^*(-s)S_2^*(s)\right\}'|_{s=\overline{\omega}_2} < 0.\end{aligned}$$

Lemma 4.4.5. If $\eta_1 < \overline{\eta}_2$, then $\overline{\boldsymbol{\pi}}_1\overline{\boldsymbol{y}}_{\overline{\eta}_2} < \infty$.

Proof. Since $\overline{\boldsymbol{\pi}}_1 \leq \sum_{m=0}^{\infty} \overline{\boldsymbol{\pi}}_m$, we have

$$\overline{\boldsymbol{\pi}}_1 \overline{\boldsymbol{y}}_{\overline{\eta}_2} \leq \sum_{m=0}^{\infty} \overline{\boldsymbol{\pi}}_m \overline{\boldsymbol{y}}_{\overline{\eta}_2}. \tag{4.59}$$

Since the behavior of the first stage is not affected by the second one, the stationary probability vector $\sum_{m=0}^{\infty} \overline{\boldsymbol{\pi}}_m$ asymptotically behaves similar to the marginal stationary probability vector of the Markov chain derived from the first stage. That is, it decays geometrically with rate η_1 as $n_1 \to \infty$. Since $\overline{\boldsymbol{y}}_{\overline{\eta}_2}$ decays with rate $\overline{\eta}_2^{-1}$, the inner product $\sum_{m=0}^{\infty} \overline{\boldsymbol{\pi}}_m \overline{\boldsymbol{y}}_{\overline{\eta}_2}$ is finite if $\eta_1/\overline{\eta}_2 < 1$.

To prove the asymptotic form of $\overline{\boldsymbol{x}}_{\overline{\eta}_2}$, we consider the partition of each $\overline{\mathcal{L}}_m$. That is, we let

$$\overline{l}_k = \{(n_1; i_0, i_1, i_2) | n_1 = k\}, \quad k = 0, 1, 2, \ldots,$$

and denote by $\overline{\boldsymbol{x}}_{\overline{\eta}_2} = (\overline{\boldsymbol{x}}_{\overline{\eta}_1}(0) \;\; \overline{\boldsymbol{x}}_{\overline{\eta}_1}(1) \;\; cdots)$ the row vector $\overline{\boldsymbol{x}}_{\overline{\eta}_2}$ partitioned according to $\overline{l}_k$'s.

Lemma 4.4.6. If $\overline{\eta}_1 < 1$, then

$$\overline{\boldsymbol{x}}_{\overline{\eta}_2}(n_1) \sim \overline{G}_1 \overline{\eta}_1^{n_1} \overline{\boldsymbol{x}}_0 \otimes \overline{\boldsymbol{x}}_1 \otimes \overline{\boldsymbol{x}}_2, \quad (n_1 \to \infty),$$

where $\overline{G}_1$ is a certain constant and

$$\begin{aligned}
\overline{\boldsymbol{x}}_0 &= \boldsymbol{\alpha}(\overline{\omega}_0 \boldsymbol{I} - \boldsymbol{T})^{-1}, \\
\overline{\boldsymbol{x}}_1 &= \boldsymbol{\beta}_1 \{-(\overline{\omega}_0 + \overline{\omega}_2)\boldsymbol{I} - \boldsymbol{S}_1\}^{-1}, \\
\overline{\boldsymbol{x}}_2 &= \boldsymbol{\beta}_2 (\overline{\omega}_2 \boldsymbol{I} - \boldsymbol{S}_1)^{-1}.
\end{aligned}$$

Proof. The ordinary matrix-geometric theory by Neuts [30] can be applied as well as the proof of Lemma 4.4.6. Let $\widehat{\boldsymbol{R}}_{\overline{D}}$ be the rate matrix of $\overline{\boldsymbol{K}}_{\overline{D}}$. Then $\widehat{\boldsymbol{R}}_{\overline{D}}$ is the minimal non-negative solution to the matrix equation

$$\frac{1}{\overline{\eta}_2}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}} + \widehat{\boldsymbol{R}}_{\overline{D}}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{D}} + \overline{\eta}_2\widehat{\boldsymbol{R}}_{\overline{D}}^2\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{C}}\widehat{\boldsymbol{D}} = \boldsymbol{O}.$$

We denote by $\widehat{\eta}$ an Perron-Frobenius eigenvalue of $\widehat{\boldsymbol{R}}_{\overline{D}}$ and $\widehat{\boldsymbol{x}}_{\overline{D}}$ be the corresponding left eigenvector, that is, $\widehat{\boldsymbol{x}}_{\overline{D}}\widehat{\boldsymbol{R}}_{\overline{D}} = \widehat{\eta}\widehat{\boldsymbol{x}}_{\overline{D}}$. Then we have

$$\widehat{\boldsymbol{x}}_{\overline{D}}\left(\frac{1}{\widehat{\eta}\overline{\eta}_2}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}} + \widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{D}} + \widehat{\eta}\overline{\eta}_2\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{C}}\widehat{\boldsymbol{D}}\right) = \boldsymbol{0}.$$

and hence the vector $\widehat{\boldsymbol{x}}_{\overline{D}}$ is the invariant vector of the matrix

$$\begin{aligned}
&\left(\frac{1}{\widehat{\eta}\overline{\eta}_2}\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{D}} + \widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{B}}\widehat{\boldsymbol{D}} + \widehat{\eta}\overline{\eta}_2\widehat{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{C}}\widehat{\boldsymbol{D}}\right) \\
&= \left\{\mathrm{diag}(\overline{\boldsymbol{v}}_0)^{-1}\left(\frac{1}{\widehat{\eta}\overline{\eta}_2}\boldsymbol{\gamma}_0\boldsymbol{\alpha} + \boldsymbol{T}\right)\mathrm{diag}(\overline{\boldsymbol{v}}_0)\right\} \oplus \\
&\qquad \left\{diag(\boldsymbol{v}_1)^{-1}\left(\widehat{\eta}\boldsymbol{\gamma}_1\boldsymbol{\beta}_1 + \boldsymbol{S}_1\right)\right\} \oplus \left\{\left(\overline{\eta}_2\boldsymbol{\gamma}_2\boldsymbol{\beta}_2 + \boldsymbol{S}_2\right)\mathrm{diag}(\overline{\boldsymbol{v}}_2)\right\}.
\end{aligned}$$

Then $\widehat{\boldsymbol{x}}_{\overline{D}}$ can be represented as $\widehat{\boldsymbol{x}}_0 \otimes \widehat{\boldsymbol{x}}_1 \otimes \widehat{\boldsymbol{x}}_2$ such that

$$\begin{aligned}
\widehat{\boldsymbol{x}}_0\left\{\mathrm{diag}(\overline{\boldsymbol{v}}_0)^{-1}\left(\frac{1}{\widehat{\eta}\overline{\eta}_2}\boldsymbol{\gamma}_0\boldsymbol{\alpha} + \boldsymbol{T}\right)\mathrm{diag}(\overline{\boldsymbol{v}}_0)\right\} &= \widehat{\xi}_0\tilde{\boldsymbol{x}}_0, \\
\tilde{\boldsymbol{x}}_1\left\{\left(\widehat{\eta}\boldsymbol{\gamma}_1\boldsymbol{\beta}_1 + \boldsymbol{S}_1\right)\right\} &= \widehat{\xi}_1\tilde{\boldsymbol{x}}_1, \\
\tilde{\boldsymbol{x}}_2\left\{\mathrm{diag}(\overline{\boldsymbol{v}}_2)^{-1}\left(\overline{\eta}_2\boldsymbol{\gamma}_2\boldsymbol{\beta}_2 + \boldsymbol{S}_2\right)\mathrm{diag}(\overline{\boldsymbol{v}}_2)\right\} &= \widehat{\xi}_2\tilde{\boldsymbol{x}}_2, \\
\widehat{\xi}_0 + \widehat{\xi}_1 + \widehat{\xi}_2 &= 0.
\end{aligned}$$

Then we have

$$\widehat{\eta}\overline{\eta}_2 = \boldsymbol{\alpha}(\widehat{\xi}_0\boldsymbol{I}_0 - \boldsymbol{T})^{-1}\boldsymbol{\gamma}_0 = T^*(\widehat{\xi}_0), \tag{4.60}$$

$$\frac{1}{\widehat{\eta}} = \boldsymbol{\beta}_1(\widehat{\xi}_1\boldsymbol{I}_1 - \boldsymbol{S}_1)^{-1}\boldsymbol{\gamma}_1 = S_1^*(\widehat{\xi}_1), \tag{4.61}$$

$$\frac{1}{\overline{\eta}_2} = \boldsymbol{\beta}_2(\widehat{\xi}_2\boldsymbol{I}_2 - \boldsymbol{S}_2)^{-1}\boldsymbol{\gamma}_2 = S_2^*(\widehat{\xi}_2). \tag{4.62}$$

From the monotonicity of LSTs, $\widehat{\xi}_2 = \overline{\omega}_2$. Eliminating $\widehat{\eta}$ and $\widehat{\xi}_1$ from (4.60) and (4.61), we obtain the equation (4.11) and $\widehat{\xi}_0$ must be $-\overline{\omega}_2$ or $\overline{\omega}_0$. Since $\overline{\boldsymbol{K}}_{\overline{D}}$ is ergodic under the condition of the theorem, however, $\widehat{\eta}$ must be less than 1. If $\widehat{\xi}_0 = -\overline{\omega}_2$ then $\widehat{\eta} = \overline{\eta}_2^{-1}$ exceeds 1 by definition, and hence $\widehat{\xi}_0 = \overline{\omega}_0$. and $\widehat{\eta} = \overline{\eta}_1$.

We can also derive explicit forms of $\widehat{\boldsymbol{x}}_k$'s as

$$\begin{aligned}
\widehat{\boldsymbol{x}}_0 &= \boldsymbol{\alpha}(\overline{\omega}_0\boldsymbol{I} - \boldsymbol{T})^{-1}\,\mathrm{diag}(\overline{\boldsymbol{v}}_0),\\
\widehat{\boldsymbol{x}}_1 &= \boldsymbol{\beta}_1\{-(\overline{\omega}_0 + \overline{\omega}_2)\boldsymbol{I} - \boldsymbol{S}_1\}^{-1},\\
\widehat{\boldsymbol{x}}_2 &= \boldsymbol{\beta}_2(\overline{\omega}_2\boldsymbol{I} - \boldsymbol{S}_2)^{-1}\,\mathrm{diag}(\overline{\boldsymbol{v}}_2),
\end{aligned}$$

up to multiplicative constants. Therefore,

$$\widehat{\boldsymbol{x}} = \widehat{\boldsymbol{x}}_{\overline{D}}\widehat{\boldsymbol{D}}^{-1} = \{\boldsymbol{\alpha}(\overline{\omega}_0\boldsymbol{I} - \boldsymbol{T})^{-1}\} \otimes [\boldsymbol{\beta}_1\{-(\overline{\omega}_0 + \overline{\omega}_2)\boldsymbol{I} - \boldsymbol{S}_1\}^{-1}] \otimes \{\boldsymbol{\beta}_2(\overline{\omega}_2\boldsymbol{I} - \boldsymbol{S}_2)^{-1}\} \tag{4.63}$$

is an invariant vector of the rate matrix $\widehat{\boldsymbol{R}}$ of $\overline{\boldsymbol{K}}$ corresponding to the eigenvalue $\widehat{\eta} = \overline{\eta}_1$.

Proof of Theorem 4.2.2 Since there are no redundant phases in $PH(\boldsymbol{\alpha}, \boldsymbol{T})$ and $PH(\boldsymbol{\beta}_k, \boldsymbol{S}_k)$ $(k = 1, 2)$, it is clear that

$$\overline{\boldsymbol{Q}}' = \begin{pmatrix} \overline{\boldsymbol{B}} & \overline{\boldsymbol{A}} & & \\ \overline{\boldsymbol{C}} & \overline{\boldsymbol{B}} & \overline{\boldsymbol{A}} & \\ & \overline{\boldsymbol{C}} & \overline{\boldsymbol{B}} & \ddots \\ & & \ddots & \ddots \end{pmatrix}.$$

is irreducible. Since $\overline{\eta}_2 < 1$ by definition, the condition $\overline{\eta}_1 < \overline{\eta}_2$ is the sufficient condition of $\overline{\eta}_1 < 1$ in Lemma 4.4.6. Then the above lemmas prove the whole of the theorem. □

Chapter 5

Numerical Computation for Tandem Queueing Systems on a Parallel Computer

5.1 Introduction

The purpose of this chapter is to discuss problems arising when we use the aggregation/disaggregationmethod, a popular numerical method for solving linear equations, on a parallel computer.

The numerical analysis has been an essential tool not only for applications, but also for progressing of novel theories. For example, it is often employed to make conjectures on unresolved problems. We use numerical analyses, in this thesis, for making some conjec-

tures on the tail behaviors of the stationary distributions in two- and three-stage tandem queueing systems.

The numerical analysis in the wide sense includes simulation, numerical solutions for differential equations, and computational approaches by using Markov chain models. We focus here the last ones. To compute the characteristic values of various stochastic models, the Markov chain is one of the most popular and strong tools. Once we derive a Markov chain from the model we are interested, then most of the characteristics of the system can be derived from the stationary distribution $\boldsymbol{x}$ of the Markov chain which is obtained by solving balance equations:

$$\boldsymbol{x}\boldsymbol{Q} = \boldsymbol{0}, \qquad \boldsymbol{x}\boldsymbol{e} = 1. \tag{5.1}$$

Since the balance equations form a system of linear equations $\boldsymbol{x}\boldsymbol{A} = \boldsymbol{b}$, we can apply the ordinary numerical method for linear equations to the computation of the stationary distribution $\boldsymbol{x}$. There are a great number of numerical methods for linear equations, some of which are conscious of the structure of matrices while others are for general matrices.

The LU decomposition method, a variation of the Gauss elimination method, is considered as the best among *direct methods* for general linear equations. It can get the solutions with relatively less computational burdens. However, direct methods sometimes generate serious errors, especially in the case of very large matrices. Thus the LU decomposition method, as well as other direct methods, is usually considered to be inappropriate for the balance equations of large-scale Markov chains.

Many of practical systems need a great amount of states to describe by Markov chains,

and hence we have to treat vectors and matrices with tremendously large dimensions. To compute these large-scale problem, *iterative methods* are commonly used. For example, we usually use the Jacobi method or the Gauss-Seidel iterative method for non-structured Markov chains. These methods are sometimes referred as the power methods. With these iterative methods, however, it often takes a large number of iterations to converge, since these methods are unconscious of the structure of the Markov chain. In many cases, the transition rate matrix $\boldsymbol{Q}$ is structured. For example, $\boldsymbol{Q}$ has a block-tridiagonal form in the case of embedded Markov chains derived from queueing systems such as $GI/PH/c$ or $PH/G/1$. It is empirically known that the power method takes much iteration for this type of matrices.

The block Gauss-Seidel method, a variant of the Gauss-Seidel method, uses the block structure of matrices to accelerate the speed of convergence. Also the aggregation/disaggregationmethod [37, 44, 49] (often abbreviated to the A/D method) uses the block structure as well as the property of stochastic vectors and matrices. Nowadays, the A/D method is considered as one of the most effective methods for computations of stationary distributions, and there are a varieties of algorithms (See [38] and [42] for example).

Nowadays, the architecture of computer used for computations is coming to a very important factor to determine which algorithm is faster. A good example is the partitioning problem, which is the main issue of this chapter, when we implement a numerical method to a massive parallel computer. A massive parallel computer has a number of (from 128 to

over 1000) processor elements, called *cells*, and provide high-speed computing. However, data transfer called message passing from one cell to another takes a considerable time, and this might cause a bottleneck of the computation. Therefore, we should organize the program so that the amount of data transfer becomes as small as possible. The main issue of this chapter is how to decrease the data transfer in calculation of the stationary probabilities of tandem queueing systems by the A/D method.

In Section 5.2, we describe the concept of the A/D method for general Markov chains. The application of the A/D method to the three-stage tandem queueing systems are discussed in Section 5.3. We introduce Fujitsu AP1000 in Section 5.4 as a typical example of massive parallel computers. Finally, in Section 5.5, we consider to divide and allocate the states of the three-stage tandem queueing system to cells in AP1000.

5.2 Aggregation/Disaggregation Method

There are two keywords to express the A/D method; *state decomposition* and *conditional probabilities*.

In a Markov chain derived from a practical problem, transitions from a state are usually restricted to a small number of states compared with the whole state space. In such a case, it is convenient to decompose the state space into disjoint subspaces called *lumps* according to relations among states.

The state space $\mathcal{S}$ of the original Markov chain with N states us decomposed into K

lumps $\mathcal{L}_0, \mathcal{L}_1, \ldots, \mathcal{L}_{K-1}$. We denote the number of states in $\mathcal{L}_i$ by N_i. Then

$$\begin{aligned} \mathcal{S} &= \mathcal{L}_0 \cup \mathcal{L}_1 \cup \cdots \cup \mathcal{L}_{K-1}, \\ N &= \sum_{i=0}^{K-1} N_i, \\ \mathcal{L}_i \cap \mathcal{L}_j &= \phi \quad \text{for all} \quad 0 \le i, j < K, \quad i \ne j. \end{aligned}$$

We call subscript of the lump $\mathcal{L}_i$ as a *macro state*, and each state in $\mathcal{S}$ as a *micro state*.

We divide the stationary probability vector $\boldsymbol{x}$ to K subvectors and the transition rate matrix $\boldsymbol{Q}$ into $K \times K$ submatrices according to the state decomposition above:

$$\begin{aligned} \boldsymbol{Q} &= \begin{pmatrix} \boldsymbol{Q}_{0,0} & \boldsymbol{Q}_{0,1} & \cdots & \boldsymbol{Q}_{0,K-1} \\ \boldsymbol{Q}_{1,0} & \boldsymbol{Q}_{1,1} & \cdots & \boldsymbol{Q}_{1,K-1} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Q}_{K-1,0} & \boldsymbol{Q}_{K-1,1} & \cdots & \boldsymbol{Q}_{K-1,K-1} \end{pmatrix}, \\ \boldsymbol{x} &= (\boldsymbol{x}_0 \ \boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_{K-1}), \end{aligned} \tag{5.2}$$

where $\boldsymbol{Q}_{i,j}$ is an $N_i \times N_j$-submatrix, and $\boldsymbol{x}_i$ is a row vector with N_i elements $(0 \le i, j < K)$.

For this decomposition, we introduce two new chains for macro states and micro states. We denote by $\tilde{\boldsymbol{x}}$ the row vector with

$$\tilde{x}_i = \boldsymbol{x}_i \boldsymbol{e}_i \quad (i = 0, 1, \ldots, K-1), \tag{5.3}$$

in its ith element, where $\boldsymbol{e}_i$ is the N_i-dimensional column vector with all elements equal to 1. Here $\tilde{\boldsymbol{x}}$ can be regarded as the stationary probability vector for macro states. We also define $\boldsymbol{b}_i$ as

$$\boldsymbol{b}_i = \frac{1}{\tilde{x}_i} \boldsymbol{x}_i \quad (i = 0, 1, \ldots, K-1). \tag{5.4}$$

$\boldsymbol{b}_i$ is considered as the "conditional" stationary probability vector for micro states in the lump $\mathcal{L}_i$.

We form $K \times N$-matrix $\boldsymbol{B}$ and $N \times K$-matrix $\boldsymbol{E}$ as follows:

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{b}_0 & \mathbf{0}_1 & \cdots & \mathbf{0}_{K-1} \\ \mathbf{0}_0 & \boldsymbol{b}_1 & \cdots & \mathbf{0}_{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_0 & \mathbf{0}_1 & \cdots & \boldsymbol{b}_{K-1} \end{pmatrix}, \tag{5.5}$$

$$\boldsymbol{E} = \begin{pmatrix} \boldsymbol{e}_0 & \mathbf{0}_0^T & \cdots & \mathbf{0}_0^T \\ \mathbf{0}_1^T & \boldsymbol{e}_1 & \cdots & \mathbf{0}_1^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{K-1}^T & \mathbf{0}_{K-1}^T & \cdots & \boldsymbol{e}_{K-1} \end{pmatrix}, \tag{5.6}$$

where $\mathbf{0}_i$ is the N_i-dimensional row vector with all elements equal to zero ($i = 0, 1, \ldots, K-1$), and the superscript ^{T} represents a transpose.

By the definition of $\boldsymbol{B}$ and $\boldsymbol{E}$, we have

$$\boldsymbol{E}\tilde{\boldsymbol{e}} = \boldsymbol{e}, \quad \boldsymbol{B}\boldsymbol{E} = \tilde{\boldsymbol{I}}, \tag{5.7}$$

where $\tilde{\boldsymbol{e}}$ is the K-dimensional column vector with all elements equal to 1, and $\tilde{\boldsymbol{I}}$ is the K-dimensional identity matrix. We also have by the definition of $\tilde{\boldsymbol{x}}$ and $\boldsymbol{b}_i$'s

$$\tilde{\boldsymbol{x}} = \boldsymbol{x}\boldsymbol{E}, \quad \tilde{\boldsymbol{x}}\boldsymbol{B} = \boldsymbol{x}. \tag{5.8}$$

From (5.1), (5.7) and (5.8), we have

$$\mathbf{0} = \boldsymbol{x}\boldsymbol{Q}\boldsymbol{E} = \tilde{\boldsymbol{x}}\boldsymbol{B}\boldsymbol{Q}\boldsymbol{E}. \tag{5.9}$$

Hence, if we denote $\tilde{\boldsymbol{Q}} = \boldsymbol{B}\boldsymbol{Q}\boldsymbol{E}$, then $\tilde{\boldsymbol{x}}$ and $\tilde{\boldsymbol{Q}}$ satisfy

$$\tilde{\boldsymbol{x}}\tilde{\boldsymbol{Q}} = \tilde{\mathbf{0}}, \tag{5.10}$$

$$\tilde{\boldsymbol{x}}\tilde{\boldsymbol{e}} \;=\; 1,$$

where $\tilde{\mathbf{0}}$ is the K-dimensional row vector with all elements equal to zero. If we denote the (i,j)-element of $\tilde{\boldsymbol{Q}}$ by $\tilde{q}_{i,j}$, then

$$\tilde{q}_{i,j} = \boldsymbol{b}_i \boldsymbol{Q}_{i,j} \boldsymbol{e}_j \quad (0 \le i,j < K).$$

It is easily shown that $\tilde{q}_{i,j}$ is the transition rate from lump i to j with a given conditional stationary probability vector for lump i. Then $\tilde{\boldsymbol{Q}}$ represents the transition rates of the process on macro states, though it is not a Markov chain.

Note that, if all of $\boldsymbol{b}_i$'s are known, we can calculate the stationary probability vector $\tilde{\boldsymbol{x}}$ for macro states as a solution of much smaller chain than the original Markov chain. Since the ith subvector $\boldsymbol{x}_i$ of the original stationary probability vector $\boldsymbol{x}$ is given by $\boldsymbol{x}_i = \tilde{x}_i \boldsymbol{b}_i (i = 0, 1, \ldots, K-1)$, we can obtain $\boldsymbol{x}$ from $\tilde{\boldsymbol{x}}$. This is the interpretation of the phase in which $\tilde{\boldsymbol{x}}$ is obtained from $\boldsymbol{b}_i$'s. This phase will be referred as the *aggregation phase*.

We proceed to the introduction of the phase where $\boldsymbol{b}_i$ for a given i is obtained from $\tilde{\boldsymbol{x}}$ and other $\boldsymbol{b}_j$'s. This phase will be referred as the *disaggregation phase*. From (5.1) and (5.4),

$$\mathbf{0}_i \;=\; \sum_{j=0}^{K-1} \boldsymbol{x}_j \boldsymbol{Q}_{j,i} \tag{5.11}$$

$$\;=\; \sum_{j=0}^{K-1} \tilde{x}_j \boldsymbol{b}_j \boldsymbol{Q}_{j,i}. \tag{5.12}$$

Hence

$$\tilde{x}_i \boldsymbol{b}_i \boldsymbol{Q}_{i,i} \;=\; -\sum_{j \ne i} \tilde{x}_j \boldsymbol{b}_j \boldsymbol{Q}_{j,i}, \tag{5.13}$$

$$\boldsymbol{b}_i \boldsymbol{e}_i \;=\; 1.$$

Equation (5.13) can be regarded as the balance equation in lump i. If $\tilde{\boldsymbol{x}}$ and all $\boldsymbol{b}_j$'s ($j \neq i$) are given, then we can obtain $\boldsymbol{b}_i$ by solving the equations (5.13).

5.2.1 Algorithm of the A/D method

The A/D method applies (5.10) and (5.13) alternately to generate converging sequences of $\tilde{\boldsymbol{x}}$ and $\boldsymbol{b}_i$'s converged.

Let $\tilde{\boldsymbol{x}}^{(n)}$ be the vector corresponding to $\tilde{\boldsymbol{x}}$ in the nth iteration, and $\boldsymbol{b}_i^{(n)}$ the vector corresponding to $\boldsymbol{b}_i$ in the nth iteration ($i = 0, 1, \ldots, K-1$). We also define $\boldsymbol{B}^{(n)}$ the matrix with $\boldsymbol{b}_i^{(n)}$ in place of $\boldsymbol{b}_i$ in $\boldsymbol{B}$ defined in (5.5).

The algorithm is as follows:

Step 1: Take an appropriate initial values of $\boldsymbol{b}_0^{(0)}, \boldsymbol{b}_1^{(0)}, \ldots, \boldsymbol{b}_{K-1}^{(0)}$, and let $n = 1$.

Step 2: Calculate $\tilde{\boldsymbol{Q}}^{(n)} = \boldsymbol{B}^{(n-1)}\boldsymbol{Q}\boldsymbol{E}$.

Step 3: Solve the balance equation for $\tilde{\boldsymbol{x}}^{(n)}$

$$\tilde{\boldsymbol{x}}^{(n)}\tilde{\boldsymbol{Q}}^{(n)} \;=\; \tilde{\boldsymbol{0}}, \tag{5.14}$$

$$\tilde{\boldsymbol{x}}^{(n)}\tilde{\boldsymbol{e}} \;=\; 1. \tag{5.15}$$

Step 4: For each $i = 0, 1, \ldots, K-1$, solve the systems of equations

$$\boldsymbol{z}_i^{(n)}\boldsymbol{Q}_{i,i} = -\sum_{j \neq i} \tilde{x}_j^{(n)} \boldsymbol{b}_j^{(n-1)} \boldsymbol{Q}_{j,i}, \tag{5.16}$$

for $\boldsymbol{z}_i^{(n)}$, and calculate $\boldsymbol{b}_i^{(n)}$ by

$$\boldsymbol{b}_i^{(n)} = \frac{1}{\boldsymbol{z}_i^{(n)} \boldsymbol{e}_i} \boldsymbol{z}_i^{(n)} \quad (i = 0, 1, \ldots, K-1).$$

Step 5: Check the convergence: If $\tilde{\boldsymbol{x}}^{(n)}$ and $\boldsymbol{b}_i^{(n)}$'s are sufficiently close to $\tilde{\boldsymbol{x}}^{(n-1)}$ and $\boldsymbol{b}_i^{(n-1)}$'s, then stop the iteration. Otherwise, set $n = n + 1$ and go back to Step 2.

The equations in Step 4 can be replaced with

$$\boldsymbol{z}_i^{(n)} \boldsymbol{Q}_{i,i} = -\sum_{j<i} \tilde{x}_j^{(n)} \boldsymbol{b}_j^{(n)} \boldsymbol{Q}_{j,i} - \sum_{j>i} \tilde{x}_j^{(n)} \boldsymbol{b}_j^{(n-1)} \boldsymbol{Q}_{j,i}.$$

to apply the latest values of $\boldsymbol{b}_j^{(n)}$'s $(j < i)$ for the calculation of $\boldsymbol{z}_i^{(n)}$.

Note that, in Steps 2 and 3, we have to solve systems of equations. The A/D method does not offer any specific method to solve these equations. However, since these systems of equations are usually much smaller than the balance equation of the original Markov chain, we can easily calculate them with the power method or the LU decomposition method.

Moreover, if $\tilde{\boldsymbol{Q}}^{(n)}$ and $\boldsymbol{Q}_{i,i}$'s are upper row-triangle, tridiagonal, or of other simple form, then the calculation of $\tilde{\boldsymbol{x}}^{(n)}$ and $\boldsymbol{b}_i^{(n)}$'s may be done with simple substitutions. For example, a quasi-birth-and-death process has a transition rate matrix of the block-tridiagonal form, and hence $\tilde{\boldsymbol{Q}}^{(n)}$ is of a tridiagonal form.

Also note that there is no proof which guarantees the convergence of the A/D method so far. Empirically, however, we can see the convergences with the A/D method in most of practical cases.

5.3 Application of the A/D Method to a Three-stage Tandem Queueing System

As an example of a Markov chain analyzed by the A/D method on a parallel computer, we take a continuous-time Markov chain derived from a three-stage tandem queueing system with phase-type interarrival and service time distributions and with buffers of infinite capacity.

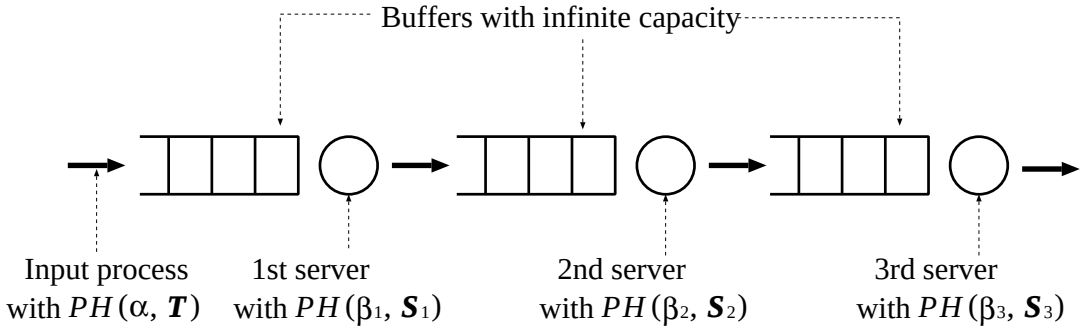


Figure 5.1: Three-stage tandem queueing system

Customers arrive at the first stage to be served there, move to the second and then the third to be served there again, and eventually go out of the system. Customers are served according to first-come first-served (FCFS) discipline at each stage. The kth stage ($k = 1, 2, 3$) has a single server and a buffer of infinite capacity, so that no loss or blocking occurs. Interarrival times of customers are independent and identically distributed (i.i.d.) random variables subjecting to a phase-type distribution $PH(\boldsymbol{\alpha}, \boldsymbol{T})$. Service times at the kth stage are also i.i.d. variables subjecting to a phase-type distribution $PH(\boldsymbol{\beta}_k, \boldsymbol{S}_k)$. The

interarrival times and the service times are assumed to be mutually independent.

The state of the system is represented by a 7-tuple $(n_1, n_2, n_3; i_0, i_1, i_2, i_3)$, where i_0 is the phase of the arrival process, i_k is the phase of the service process at the kth stage, and n_k is the number of customers in the kth stage $(k = 1, 2, 3)$. Then the system behaves as a continuous-time Markov chain. Let $D(n_1, n_2, n_3)$ be the set of states with n_1, n_2 and n_3 in the first three places of the 7-tuple representation.

The transition rate matrix $\boldsymbol{Q}$ can be represented by using matrices and vectors above. Let $\boldsymbol{Q}(n_1, n_2, n_3; n_1', n_2', n_3')$ be the submatrix of $\boldsymbol{Q}$ consisting of transition rates from states in $D(n_1, n_2, n_3)$ to states in $D(n_1', n_2', n_3')$. Using Kronecker sum and product operations, it is written as

$$\begin{aligned}
\boldsymbol{Q}(n_1, n_2, n_3; n_1, n_2, n_3) &= \boldsymbol{T} \oplus \boldsymbol{S}_1 \oplus \boldsymbol{S}_2 \oplus \boldsymbol{S}_3, \\
\boldsymbol{Q}(n_1, n_2, n_3; n_1 + 1, n_2, n_3) &= \boldsymbol{\gamma}_0 \boldsymbol{\alpha} \otimes \boldsymbol{I}_1 \otimes \boldsymbol{I}_2 \otimes \boldsymbol{I}_3, \\
\boldsymbol{Q}(n_1, n_2, n_3; n_1 - 1, n_2 + 1, n_3) &= \boldsymbol{I}_0 \otimes \boldsymbol{\gamma}_1 \boldsymbol{\beta}_1 \otimes \boldsymbol{I}_2 \otimes \boldsymbol{I}_3, \\
\boldsymbol{Q}(n_1, n_2, n_3; n_1, n_2 - 1, n_3 + 1) &= \boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \otimes \boldsymbol{\gamma}_2 \boldsymbol{\beta}_2 \otimes \boldsymbol{I}_3, \\
\boldsymbol{Q}(n_1, n_2, n_3; n_1, n_2, n_3 - 1) &= \boldsymbol{I}_0 \otimes \boldsymbol{I}_1 \otimes \boldsymbol{I}_2 \otimes \boldsymbol{\gamma}_3 \boldsymbol{\beta}_3, \\
\boldsymbol{Q}(n_1, n_2, n_3; n_1', n_2', n_3') &= \boldsymbol{O}, \qquad \text{otherwise}
\end{aligned}$$

for $n_1, n_2, n_3 \geq 2$. $\boldsymbol{Q}(n_1, n_2, n_3; n_1', n_2', n_3')$ may take a slightly different form if one of n_k's or n_k''s is equal to 0.

This Markov chain has some special properties which are useful for the application of the A/D method.

Property 1: State transitions into $D(n_1, n_2, n_3)$ from outside is possible only from neighboring sets $D(n_1-1, n_2, n_3)$, $D(n_1+1, n_2-1, n_3)$, $D(n_1, n_2+1, n_3-1)$, and $D(n_1, n_2, n_3+1)$.

Property 2: A non-zero submatrix $\boldsymbol{Q}(n_1, n_2, n_3; n_1', n_2', n_3')$ such that $(n_1, n_2, n_3) \neq (n_1', n_2', n_3')$ is given by a Kronecker product of a matrix of dyadic form $\boldsymbol{\gamma}_k \boldsymbol{\beta}_k$ and identity matrices.

Furthermore, from the structure of the model itself, marginal probabilities of the first two stages can be obtained by solving smaller models:

Property 3: The marginal stationary state probabilities $\pi_1(n_1; i_0, i_1)$ of the first stage can be obtained by solving a one-stage model consisting of the first stage of the original model.

Property 4: The marginal stationary state probabilities $\pi_2(n_1, n_2; i_0, i_1, i_2)$ of the first two stages can be obtained by solving a two-stage model consisting of the first and the second stages of the original model.

As stated in section 1, the aim of the authors for analyzing the tandem queueing system is to know the tail behavior of the stationary state probabilities. The stationary state probabilities $\pi(n_1, n_2, n_3; i_0, i_1, i_2, i_3)$ have to be calculated over a wide range, for example in the range $0 \leq n_k \leq N_k - 1$ with sufficiently large N_k's, say $N_k = 100$. If all s_k's are equal to 3, then the total number of states $M = s_0 s_1 s_2 s_3 N_1 N_2 N_3$ to be calculated is about

81,000,000. Such a huge number of states can be treated only by using the A/D method on a parallel computer.

In order to analyze the Markov chain introduced above, it seems reasonable to apply the A/D method successively to the one-stage model, to the two-stage model and then to the three-stage model.

Let $L_2(n_1, n_2; i_0, i_1, i_2)$ be the set of states having n_1 and n_2 in the first and the second places and i_0, i_1 and i_2 in the fourth, fifth and sixth places of the 7-tuple representation, and let $L_1(n_1; i_0, i_1) = \cup_{n_2,i_2} L_2(n_1, n_2; i_0, i_1, i_2)$. Clearly $L_1(n_1; i_0, i_1)$'s form a decomposition by the local state of the first stage and $L_2(n_1, n_2; i_0, i_1, i_2)$'s form a decomposition by the local states of the first two stages. From Property 3 above, the stationary probabilities $\pi_1(n_1; i_0, i_1)$ for the lumps $L_1(n_1; i_0, i_1)$ are obtained by analyzing the one-stage queue. This can be effectively done by using the algorithm proposed in [6], a variation of the A/D method. The stationary probabilities $\pi_2(n_1, n_2; i_0, i_1, i_2)$ for the sets $L_2(n_1, n_2; i_0, i_1, i_2)$ are obtained by solving the two-stage model, as stated in Property 4 above, using the A/D method by considering $L_1(n_1; i_0, i_1)$'s are lumps there. In this case, we don't need to perform the aggregation phase since values of $\pi_1(n_1; i_0, i_1)$'s have already been known. Note that the two-stage model has $M_2 = s_0 s_1 s_2 N_1 N_2 = 270,000$ states. Hence it may be solvable by using a computer with a single processor, but faster if one uses a parallel computer.

The stationary probabilities $\pi(n_1, n_2, n_3; i_0, i_1, i_2, i_3)$ have to be then calculated by the

A/D method with lumps $L_2(n_1, n_2; i_0, i_1, i_2)$. Again, in this case, we don't need to do the aggregation phase. The problem here is the allocation of the calculation in the disaggregation phase for these lumps. Since the data transfer between cells of the parallel computer requires a considerable time, we have to allocate lumps so that the amount of data transfer becomes as small as possible. We will discuss this problem in the next section.

5.4 Architecture of Parallel Computer

The parallel machine which we implement the A/D method on is the Fujitsu AP1000 highly parallel computer [15] at Fujitsu Laboratories, Ltd. In this section, we give an overview of the architecture of the AP1000, and mention its difference from traditional non-parallel computers in case of the implementation.

5.4.1 System Configuration of AP1000

The AP1000 machine which we use has one host processor and 256 or 512 processor elements called *cells* [2]. Each cell consists of a Sparc processor, memory, and a data channel controller (see Figure 5.2). The main characteristics of the AP1000 are as follows.

MIMD Design MIMD (Multiple Instruction-stream, Multiple Data-stream) means that each cell can execute its own tasks and has its own data. With this architecture, we can execute in parallel not only single type of tasks, but also different types of tasks.

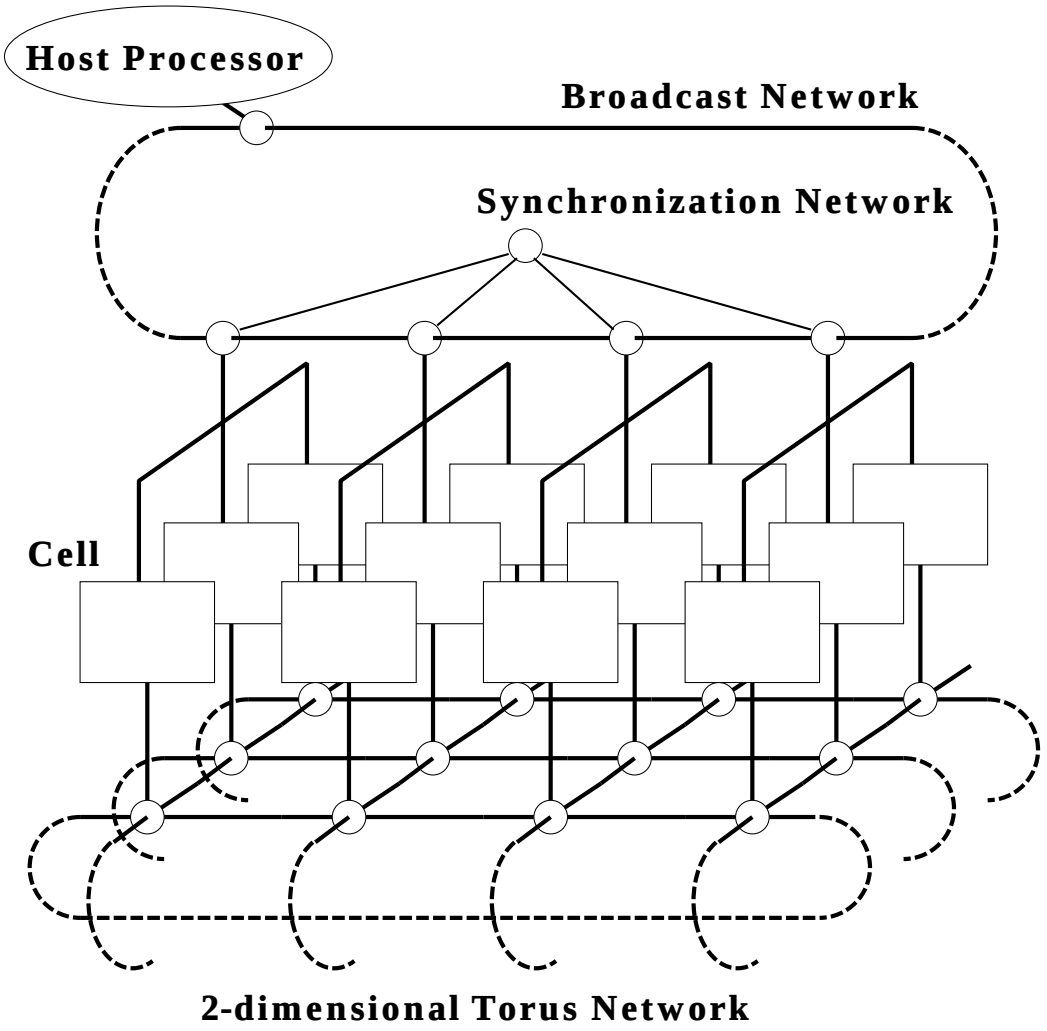


Figure 5.2: AP1000 configuration

Two Dimensional Torus Network Cells are connected each other via high-speed data transfer channels. These channels construct a two dimensional torus topology network, and each cell is considered as a grid point of the network. Then, the data transfer between two cells q steps away from each other takes q times as long as the data transfer between neighboring cells. Each cell-to-cell channel has a transfer rate of 25Mbytes/sec.

The number of rows and the number of cells on each row can be changed by the user, and it is realized logically on the two dimensional torus network.

There are two other channels to connect all cells and the host processor; one is for the one-to-N data transfer referred as *broadcast*, and another is for the synchronization among cells.

Distributed Memory Each cell has local memory of 16Mbytes, which is sufficient to store a million of stationary probabilities as double-precision floating-point numbers.

Message Passing The AP1000 has no shared memory for cells, and a cell cannot access to the local memory on other cells. When the task on a cell intends to transfer the data needed by another cell, it sends the data asynchronously through the torus network. This scheme is known as *message passing*.

With message passing, cells don't receive any interrupts for data transfer and hence all tasks can be executed asynchronously. This also implies that the host processor and cells cannot explicitly request data transfer to other cells, while cells can passively wait and receive data.

5.4.2 Differences from Non-Parallel Computers

Because the architecture of parallel computers is quite different from that of non-parallel ones, we must be careful of how we implement the algorithm on a parallel machine.

1. The speed of the data transfer through the torus network is much less than the speed of the data reference on the local memory. We then have to partition the problem into tasks so that the number of data transfer between tasks can be as small as possible.

2. If a specific task takes much longer time to finish execution than other tasks, parallelization is of no effect. To equalize the load of cells, we should carefully partition the problem or even dynamically allocate and partition the tasks.

3. In message passing, each task has to know a priori who requires the data transfer.

When we implement the A/D method on a parallel computer, we have to take into account of these points.

5.5 Allocation of Lumps to Cells

To make our discussion clearer, hereafter we will consider the case where we calculate the stationary probabilities of the three-stage model with $s_l = 3$ and $N_l = 100$ for all l by a parallel computer with 512 cells. We also assume that, for every l, $\boldsymbol{T}_l$ is an upper triangular matrix with positive numbers in all upper off-diagonal entries, and $\boldsymbol{\alpha}_l$ and $\boldsymbol{t}_l$ have positive numbers in all their entries. These assumptions are necessary for estimating the amount of data transfer later.

n \ i	1	2	3	4	5
0	$L_1(0,1,0)$		$L_1(0,2,0)$		$L_1(0,3,0)$
1	$L_1(1,1,1)$ $L_1(1,1,2)$	$L_1(1,1,3)$ $L_1(1,2,1)$	$L_1(1,2,2)$ $L_1(1,2,3)$	$L_1(1,3,1)$ $L_1(1,3,2)$	$L_1(1,3,3)$
	⋮	⋮	⋮	⋮	⋮
$n-1$					
n	$L_1(n,1,1)$ $L_1(n,1,2)$	$L_1(n,1,3)$ $L_1(n,2,1)$	$L_1(n,2,2)$ $L_1(n,2,3)$	$L_1(n,3,1)$ $L_1(n,3,2)$	$L_1(n,3,3)$
$n+1$					

Figure 5.3: A naive allocation of lumps to cells

5.5.1 A Naive Allocation

First we consider a naive allocation of lumps to cells. Since there are 512 cells and 900 L_1 lumps, it may seem natural to allocate two $L_1(n_1; j_0, j_1)$'s to one cell (see Figure 5.3), namely six hundreds L_2 lumps to one cell. More definitely, we arrange cells $C(n,i)$, $n = 0, 1, \ldots, 99$ and $i = 1, 2, \ldots, 5$, on a two-dimensional plane, and to cell $C(n,i)$ we allocate lumps $L_2(n_1, n_2; j_0, j_1, j_2)$ such that $n_1 = n$ and $3\,(j_0 - 1) + j_1 = 2i - 1$ or $2i$.

We shall evaluate the amount of data transfer among cells required for one iteration of the disaggregation phase. We denote by $\boldsymbol{\pi}_1(n_1; j_0, j_1)$ the row vector with entries $\pi(n_1, n_2, n_3; j_0, j_1, j_2, j_3)$ with n_1, j_0, j_1 in the first, fourth and fifth places of the 7-tuple representation. The order of the vector is $s_2 s_3 N_2 N_3 = 90,000$. To perform the disag-

gregation phase for lumps in cells $C(n,1) \sim C(n,5)$, we need information about vectors $\boldsymbol{\pi}_1(n-1;j_0,j_1)$'s and $\boldsymbol{\pi}_1(n+1;j_0,j_1)$'s.

First let us consider data transfer related to transitions associated with arrivals. Let j_1 be fixed to 1. We may send each vector $\boldsymbol{\pi}_1(n-1;j_0,1)$ to every cell in which the information of the vector is needed. Then, in our naive allocation, we send $\boldsymbol{\pi}_1(n-1;1,1)$ from $C(n-1,1)$ to three cells $C(n,1), C(n,2)$ and $C(n,4)$, and also vectors $\boldsymbol{\pi}_1(n-1;2,1)$ and $\boldsymbol{\pi}_1(n-1;3,1)$ from $C(n-1,2)$ and $C(n-1,4)$ to the same three destination cells. This requires totally nine transfers of data of size 90,000. This number of transfers can be decreased by exploiting property 2 in section 4.

We note that for lumps in cells $C(n,1), C(n,2)$ and $C(n,4)$ with $j_1 = 1$ it is sufficient to receive information on a single vector $\boldsymbol{t} = \sum_j t_{0j}\, \boldsymbol{\pi}_1(n-1;j,1)$ instead of three vectors $\boldsymbol{\pi}_1(n-1;1,1), \boldsymbol{\pi}_1(n-1;2,1)$ and $\boldsymbol{\pi}_1(n-1;3,1)$. By taking a Kronecker product of $\boldsymbol{t}$ and $\boldsymbol{\alpha}_0$, they can construct necessary vectors. This can be done in our allocation as follows.

- Send $\boldsymbol{\pi}_1(n-1;2,1)$ from $C(n-1,2)$ to $C(n-1,1)$.
- Send $\boldsymbol{\pi}_1(n-1;3,1)$ from $C(n-1,4)$ to $C(n-1,1)$.
- After calculating the sum $\boldsymbol{t}$ at $C(n-1,1)$, send $\boldsymbol{t}$ to $C(n,1), C(n,2)$ and $C(n,4)$ for the computation of the next disaggregation phase of $\boldsymbol{\pi}_1(n;1,1)$, $\boldsymbol{\pi}_1(n;2,1)$ and $\boldsymbol{\pi}_1(n;3,1)$.

The above process requires data transfer of size 90,000 five times. This amount is for fixed $j_1 = 1$. For all lumps in $C(n,1) \sim C(n,5)$, we have to transfer data of size 90,000 fifteen

n \ i	1	2	3
0	$L_1(0,1,0)$	$L_1(0,2,0)$	$L_1(0,3,0)$
1	$L_1(1,1,1)$ $L_1(1,1,3)$ $L_1(1,1,2)$	$L_1(1,2,1)$ $L_1(1,2,3)$ $L_1(1,2,2)$	$L_1(1,3,1)$ $L_1(1,3,3)$ $L_1(1,3,2)$
	$\vdots$	$\vdots$	$\vdots$
$n-1$			
n	$L_1(n,1,1)$ $L_1(n,1,3)$ $L_1(n,1,2)$	$L_1(n,2,1)$ $L_1(n,2,3)$ $L_1(n,2,2)$	$L_1(n,3,1)$ $L_1(n,3,3)$ $L_1(n,3,2)$
$n+1$			

Figure 5.4: Another naive allocation of lumps to cells

$(= 5 \times 3$ (number of possible j_1's)) times.

Similarly for transitions associated with service completions at the first stage, we have to transfer data of size 90,000 nine times, and for phase transitions in the arrival process and the service process at the first stage, we have to transfer data of size 90,000 twelve times.

Totally we need to transfer 36 times for one calculation of 5 cells $C(n,1) \sim C(n,5)$ in the disaggregation phase. Hence on the whole we need to transfer data of size $90,000 \times 36 \times 100$ (number of possible n's) =324,000,000 per one iteration of the disaggregation phase.

If we allocate three $L_1(n_1; j_0, j_1)$'s to one cell as indicated in Figure 5.4, data transfer is somewhat decreased on the part associated with service completions at the fist stage

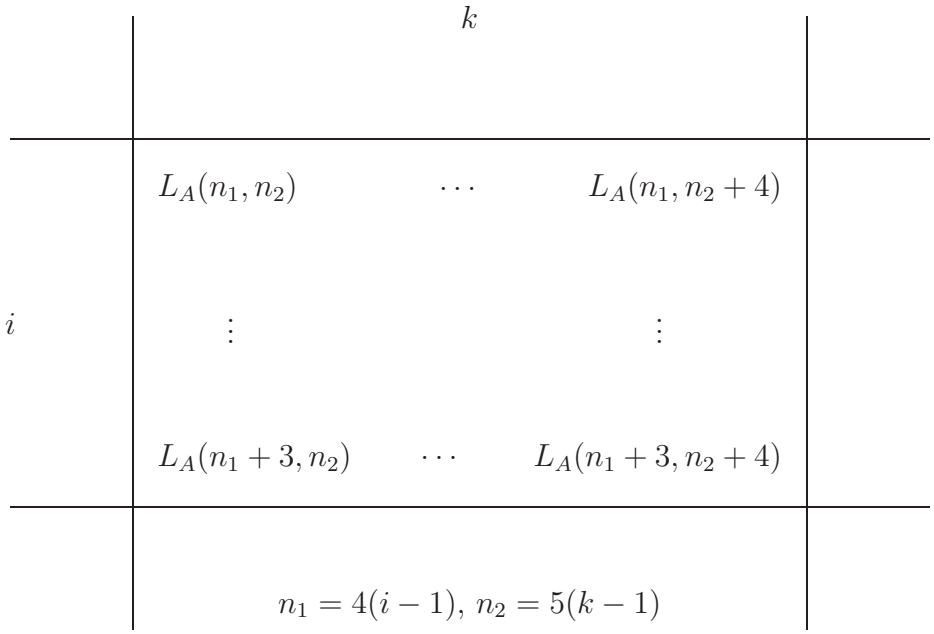


Figure 5.5: An efficient allocation of lumps to cells

and phase transitions in the arrival process and the service process at the first stage. The total amount of data transfer between cells becomes to $90,000 \times 27 \times 100 = 243,000,000$. By increasing the number of lumps allocated to a cell into 3/2 times that of the case of Figure 5.3 the amount of data transfer decreases into 3/4.

In these cases some data transfer occurs between non-neighboring cells. This may require additional transfer time. So, we should look for more efficient allocation of lumps.

5.5.2 An Efficient Allocation

We show a more efficient allocation of lumps to cells. In the naive allocations in the preceding subsection, we allocate $L_1(n_1; 1, j_1)$, $L_1(n_1; 2, j_1)$ and $L_1(n_1; 3, j_1)$ into different

cells. This requires some data transfer between corresponding cells. This kind of data transfer can be removed by allocating lumps with the same n_1 and n_2 into a cell.

Let $L_A(n_1, n_2)$ be the set of states with n_1 and n_2 in the first and the second places of the 7-tuple representation. We arrange cells in two dimensional as 25×20, and to cell $C(i, k)$ we allocate lumps contained in the sets $L_A(n_1, n_2)$ such that $4(i-1) \leq n_1 \leq 4i-1$ and $5(k-1) \leq n_2 \leq 5k-1$, see Figure 5.5. Then for transitions associated with arrivals, 5 vectors of size 8,100 are sent from $C(i-1, k)$ to $C(i, k)$. For transitions associated with service completions at first stage, 4 vectors of the same size are sent from $C(i+1, k)$, 3 vectors are sent from $C(i, k-1)$, and one vector is sent from $C(i+1, k-1)$. For transitions associated with service completions at the second stage, 4 vectors of the same size are sent from $C(i, k+1)$. Hence for the calculation of the disaggregation phase in $C(i, k)$, totally 17 vectors of size 8,100 should be sent from neighboring cells. On the whole the amount of data transfer for the next calculation of the disaggregation phase is evaluated as $8,100 \times 17 \times 500 = 68,850,000$. This is only one fifth of those of naive allocations.

5.5.3 An Allocation with Least Amount of Data Transfer

The allocation proposed in the preceding subsection is not the one with the least amount of data transfer. Let us consider a set surrounded by the curved line in Figure 5.6 on the (n_1, n_2)-plane. If we allocate the lumps contained in the set to one cell, the amount of data transfer required for the calculation of the disaggregation phase of the cell is evaluated as follows.

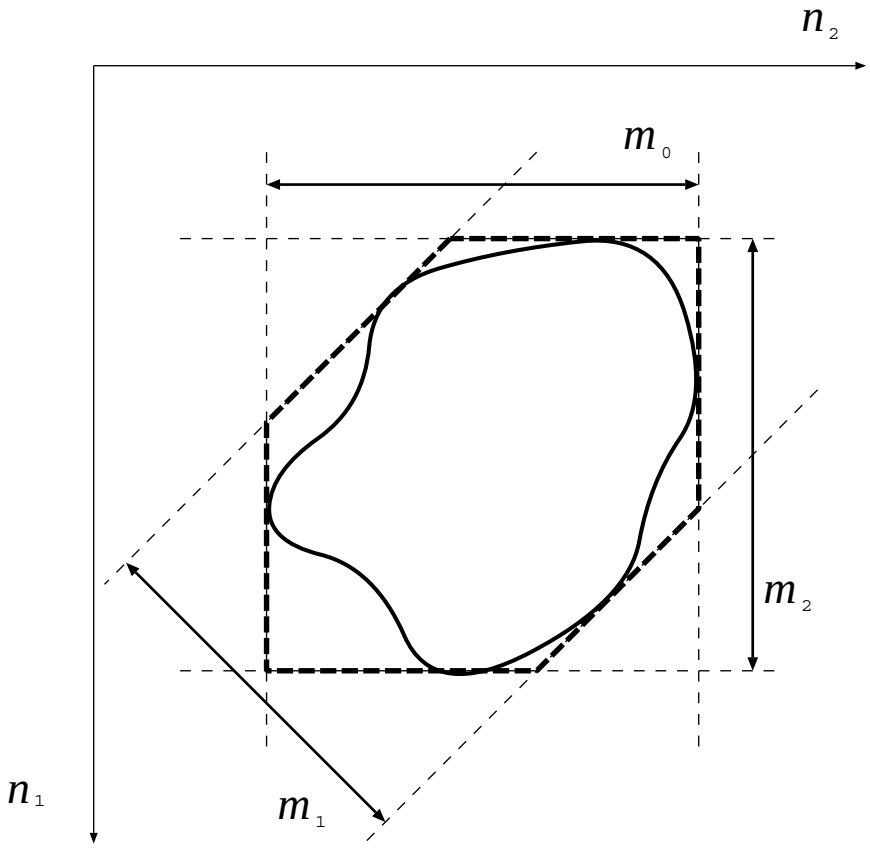


Figure 5.6: Diameters of a set

For transitions associated with arrivals, the number of vectors to be sent from the upper cells is equal to m_0, the diameter of the set from above. Similarly, for transitions associated with service completions at the second stage, the number of vectors to be sent is equal to m_2, the diameter of the set from side. For transitions associated with service completions at the first stage, we have to see the set from the angle of 45°. Then the number of lumps at peripheral is about $\sqrt{2}$ times the diameter m_1. Thus the number of vectors to be sent

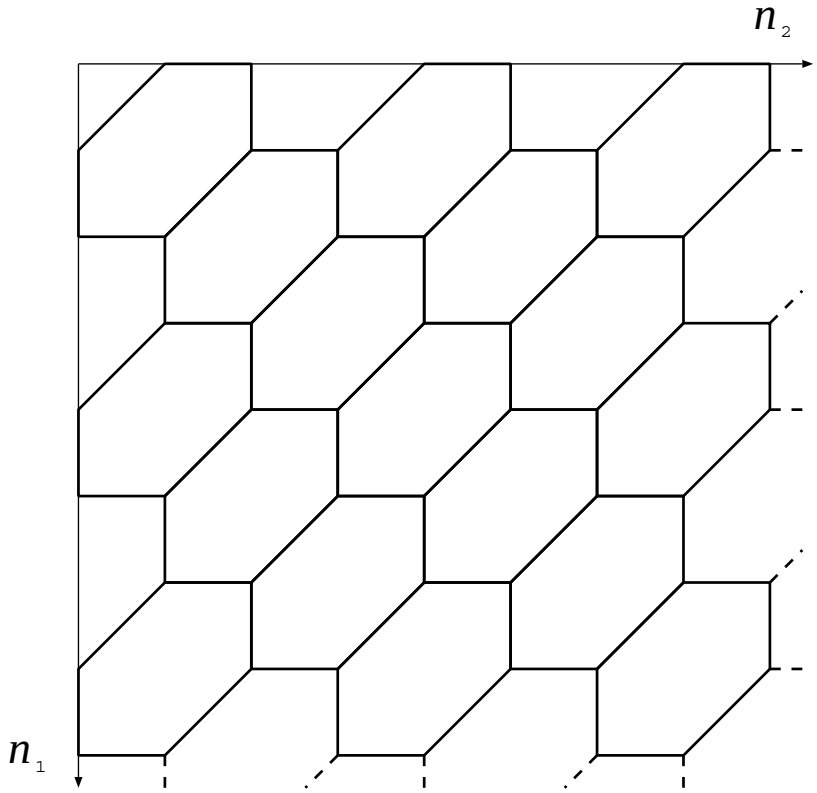


Figure 5.7: The partition for the most efficient allocation

for the calculation of the cell is equal to

$$m \;=\; m_0 \,+\, \sqrt{2}\, m_1 \,+\, m_2 \,.$$

Note that the amount of data transfer depends only on diameters m_0, m_1 and m_2, and does not depend on the area of the set. It is easily seen that for given diameters m_0, m_1 and m_2, the set having maximum area is the one designated by the hexagon whose edges are parallel to the two axes and the 45° line.

A trite calculation shows that the ratio of the value m to the square root of the area

of the hexagon is maximum when $m_0 = m_2 = \sqrt{2}m_1$. Figure 5.7 shows a covering of the whole (n_1, n_2)-plane by such hexagons. This is the most efficient allocation of lumps to cells from the view point of data transfer. The ratio of m to the square root of the area of the hexagon is $2\sqrt{3}$. However, this allocation has a weak point that the partition in Figure 5.7 does not fit for the two dimensional array structure of the cells. Hence, if we use this allocation we have to send data to non-neighboring cells and this requires additional overheads for data transfer.

One may consider that it is more natural to cover the plane by squares. Let the square has edges of length a. Then diameters defined in Figure 5.6. are given as $m_0 = m_2 = a$ and $m_1 = \sqrt{2}a$. Hence $m = 4a$ and the area is a^2. The ratio of m to the square root of the area is equal to 4. This is not so bad compared with the least data transfer case $2\sqrt{3}$ stated above. Almost all of the data transfer in this allocation is done between neighboring cells except one vector to the lump at a lower left corner.

Considering difficulties of allocation of lumps and of treatment of peripheral sets in the least data transfer allocation, the author thinks that the covering by squares is the best choice in practice. The efficient allocation in the preceding subsection is almost the best.

5.6 Concluding Remarks

The aggregation/disaggregation method (the A/D method) is a suitable numerical method for analyzing large scale Markov chains using a parallel computer.

In this chapter we have discussed application of the A/D method with an example of a three-stage tandem queueing system, and shown that we have to be careful on the allocation of lumps in the A/D method to the cells of the parallel computer since data transfer between cells takes a considerable time. At effective allocations the amount of data transfer is less than a tenth of that at naive allocations. This great improvement is realized by matching the structural relationship among lumps to the physical structure of cells on the computer. Thus without an adequate choice of allocation the A/D method could not become a powerful tool for calculating the stationary distributions of gigantic Markov chains.

The author has just started the calculation on a parallel computer. The data on computing times and data transfer times and so on will be reported in an earliest occasion.

Chapter 6

Numerical Experiments on Tail Behavior of Stationary Distributions in Three-stage Tandem Queueing Systems

6.1 Introduction

In this chapter, we observe geometric decay properties of the joint queue-length probability $p(n_1, n_2, n_3)$ in the three-stage tandem queueing system $PH/PH/1 \to /PH/1 \to /PH/1$ from numerical experiments using a massive parallel computer.

In Chapter 2, from the results of extensive numerical experiments for the two-stage

tandem queueing system $PH/PH/1 \to /PH/1$, we observed the joint probability $p(n_1, n_2)$ that there are n_k customers in the kth stage ($k = 1, 2$) has two types of asymptotic product form: For large n_1 and n_2 we have

$$p(n_1, n_2) \sim \begin{cases} G\eta_1^{n_1}\eta_2^{n_2} & \text{if } n_1 > an_2 \\ \overline{G}\overline{\eta}_1^{n_1}\overline{\eta}_2^{n_2} & \text{if } n_1 < an_2 \end{cases}$$

where $a = -\log\overline{\eta}_2/\eta_2/\log\eta_1/\overline{\eta}_1$ and decay parameters $\eta_1, \eta_2, \overline{\eta}_1$ and $\overline{\eta}_2$ are given as solutions of (2.9) and (2.12). However, the numerical experiments for only the "two"-stage model itself seems to be insufficient to investigate the theoretical aspect of queueing network models. To proceed one more step, we make experiments for three-stage models.

Here we scrutinize numerical results for the three-stage tandem queueing system $PH/PH/1 \to /PH/1 \to /PH/1$ as a typical extension of two-stage tandem queueing system. We see that $p(n_1, n_2, n_3)$ decays geometrically and find two (not three) types of geometric decay depending on the traffic intensities of the first, second and third stages: If we define decay rates

$$r_1(n_1, n_2, n_3) = \frac{p(n_1+1, n_2, n_3)}{p(n_1, n_2, n_3)}, \tag{6.1}$$

$$r_2(n_1, n_2, n_3) = \frac{p(n_1, n_2+1, n_3)}{p(n_1, n_2, n_3)}, \tag{6.2}$$

$$r_3(n_1, n_2, n_3) = \frac{p(n_1, n_2, n_3+1)}{p(n_1, n_2, n_3)}, \tag{6.3}$$

$$\tag{6.4}$$

then they approximately coincide with constants $\{\eta_1, \eta_2, \eta_3\}$ or $\{\overline{\eta}_1, \overline{\eta}_2, \overline{\eta}_3\}$ given by the

systems of equations

$$\begin{cases} T^*(\omega_0)S_1^*(-\omega_0) = 1, & \eta_1 = T^*(\omega_0), \\ T^*(\omega_0)S_1^*(\omega_1)S_2^*(-\omega_0-\omega_1) = 1, & \eta_2 = \eta_1 S_1^*(\omega_1), \\ T^*(\omega_0)S_1^*(\omega_1)S_2^*(\omega_2)S_3^*(-\omega_0-\omega_1-\omega_2) = 1, & \eta_3 = \eta_2 S_2^*(\omega_2), \end{cases} \tag{6.5}$$

$$\begin{cases} T^*(-\overline{\omega}_3)S_3^*(\overline{\omega}_3) = 1, & \overline{\eta}_3 = [S_3^*(\overline{\omega}_3)]^{-1}, \\ T^*(-\overline{\omega}_2-\omega_3)S_2^*(\overline{\omega}_2)S_3^*(\overline{\omega}_3) = 1, & \overline{\eta}_2 = \overline{\eta}_3[S_2^*(\overline{\omega}_2)]^{-1}, \\ T^*(-\overline{\omega}_1-\overline{\omega}_2-\overline{\omega}_3)S_1^*(\overline{\omega}_1)S_2^*(\overline{\omega}_2)S_3^*(\overline{\omega}_3) = 1, & \overline{\eta}_1 = \overline{\eta}_2[S_1^*(\overline{\omega}_1)]^{-1}. \end{cases} \tag{6.6}$$

For the numerical experiments, we use a massive parallel computer. Even with this ultramodern equipment, the types of models are rather restricted than the experiment for the $PH/PH/1 \rightarrow /PH/1$ models because of the limitation of the sizes of computable models.

6.2 Three-stage Tandem Queueing System

Here we consider an open, three-stage tandem queueing system (Figure 6.1).

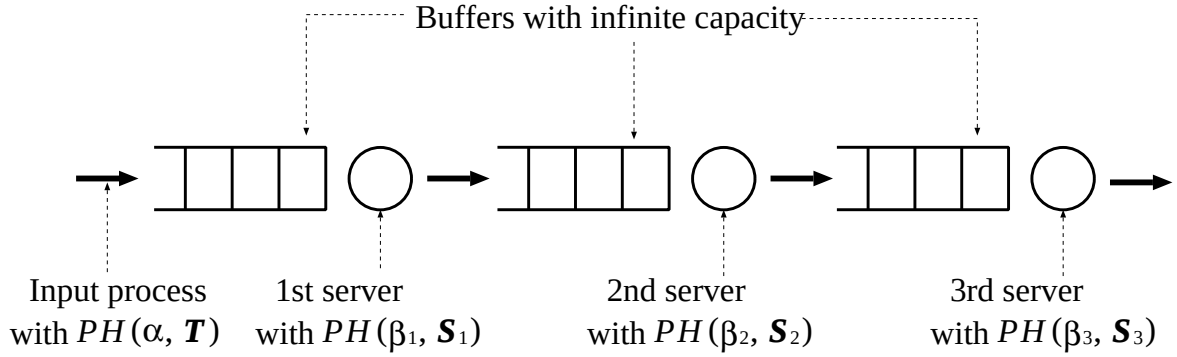


Figure 6.1: Three-stage tandem queueing system

Customers arrive at the first stage to be served there, move to the second and then

the third to be served there again, and eventually go out of the system. Customers are served according to first-come first-served (FCFS) discipline at each stage. The kth stage ($k = 1, 2, 3$) has a single server and a buffer of infinite capacity, so that no loss or blocking occurs. Interarrival times of customers are independent and identically distributed (i.i.d.) random variables subjecting to a phase-type distribution $PH(\boldsymbol{\alpha}, \boldsymbol{T})$. Service times at the kth stage are also i.i.d. variables subjecting to a phase-type distribution $PH(\boldsymbol{\beta}_k, \boldsymbol{S}_k)$. The interarrival times and the service times are assumed to be mutually independent.

The state of the system is represented by a 7-tuple $(n_1, n_2, n_3; i_0, i_1, i_2, i_3)$, where i_0 is the phase of the arrival process, i_k is the phase of the service process at the kth stage, and n_k is the number of customers in the kth stage ($k = 1, 2, 3$). Then the system behaves as a continuous-time Markov chain.

We denote the traffic intensity at the kth stage by $\rho_k = \lambda/\mu_k$ where $1/\lambda$ is the mean interarrival time and $1/\mu_k$ is the mean service time at the kth stage ($k = 1, 2, 3$). We assume $\rho_1, \rho_2, \rho_3 < 1$ so that the chain is stable and has stationary probabilities $x(n_1, n_2; i_0, i_1, i_2)$.

6.3 Numerical Experiments

To see the tail behavior of the joint queue-length distribution, we made numerical experiments for several models. The number of models we calculated are rather restricted than that in the case of two-stage models in Chapter 2, because the size of the three-stage models are around 100 times as large as that of the two-stage models in order to get the

comparable precision for the numerical results.

For models $M/E_2/1 \rightarrow /H_2/1 \rightarrow /E_2/1$ and $E_2/H_2/1 \rightarrow /E_2/1 \rightarrow /H_2/1$, we tested systematically with $\rho_k = .2, .4, .6, .8 (k = 1, 2, 3)$, and observed changes of the tail behaviors by the traffic intensities in detail. For the two-phase hyperexponential distribution (H_2), we used the one with the density function of the form

$$s(x) = 0.2e^{-4\mu x} + 0.8e^{-\mu x}, \ x > 0.$$

For the calculation of the stationary probabilities, we employed the aggregation/disaggregation method described in Chapter 5. Since our model has infinite number of states, we have to truncate the state space for all of $n_k (k = 1, 2, 3)$ in the calculation. However, in an iteration of the aggregation/disaggregation method, a new value of $x(n_1, n_2, n_3; *, *, *, *)$ is calculated from current values of neighboring states $x(n_1 - 1, n_2, n_3; *, *, *, *)$, $x(n_1 + 1, n_2 - 1, n_3; *, *, *, *)$, $x(n_1, n_2 + 1, n_3 - 1; *, *, *, *)$, and $x(n_1, n_2, n_3 + 1; *, *, *, *)$. Therefore, if we truncate the state space at $n_k = \nu_k$, we have to estimate the values of

$$\begin{aligned}
&x(\nu_1 + 1, n_2, n_3; *, *, *, *) \quad \text{for} \quad 0 \leq n_2 \leq \nu_2, 0 \leq n_3 \leq \nu_3, \\
&x(n_1, \nu_2 + 1, n_3; *, *, *, *) \quad \text{for} \quad 0 \leq n_1 \leq \nu_1, 0 \leq n_3 \leq \nu_3, \\
&x(n_1, \nu_2, n_3 + 1; *, *, *, *) \quad \text{for} \quad 0 \leq n_1 \leq \nu_1, 0 \leq n_2 \leq \nu_2.
\end{aligned}$$

In our experiments, we estimated those values by assuming geometric decay for these variables, namely, e.g., $x(\nu_1 + 1, n_2, n_3; *, *, *, *)$ was estimated as $\{x(\nu_1, n_2, n_3; *, *, *, *)\}^2 / x(\nu_1 - 1, n_2, n_3; *, *, *, *)$.

The truncation points $\nu_k (k = 1, 2, 3)$ were set to 64 in all the cases. So the number of states to be calculated was $2^3 \times 64^3 \approx 2 \times 10^6$ for $M/E_2/1 \to /H_2/1 \to /E_2/1$, and was $2^4 \times 64^3 \approx 4 \times 10^6$ for $E_2/H_2/1 \to /E_2/1 \to /H_2/1$. By the author's experiences, it is almost impossible to solve the balance equations with several millions of states by contemporary workstations, and this time we used a Fujitsu AP-1000 massive parallel computer with 512 processors and 16 megabytes of memory for each processor, thanks to Fujitsu laboratories Inc. The program was written in C with parallel processing extension.

The computational burden is practically $\mathcal{O}(N^3)$ with $N = \nu_1 \times \nu_2 \times \nu_3$, and it increases rapidly as $\rho_k \to 1$. Table 6.1 tabulates the CPU time for the computation of $E_2/H_2/1 \to /E_2/1 \to /H_2/1$ with $\rho_1 = .6, \rho_2 = .8$ and $\rho_2 = .2, .4$, and .6.

Table 6.1: The CPU time for $E_2/H_2/1 \to /E_2/1 \to /H_2/1$ with $\rho_1 = 0.6, \rho_2 = 0.8$, and $\nu_1 = \nu_2 = \nu_3 = 64$

ρ_2	0.2	0.4	0.6
CPU time [sec.]	610	1105	2660

6.4 Observation of the Numerical Results

In this section, we observe the results of the numerical experiments explained in the previous section in order to show some conjectures on the asymptotic behavior of the joint queue-length probabilities in the three-stage tandem queueing system.

Here we take the case of $E_2/H_2/1 \rightarrow /E_2/1 \rightarrow /H_2/1$ with $\rho_1 = 0.6, \rho_2 = 0.8, \rho_3 = 0.4$ as a typical example, and see its tail properties in detail. We start with observing ratios of two neighboring joint probabilities for the numbers of customers in the steady state.

6.4.1 Decay Rates of the Joint Queue-length Probability

Let $p(n_1, n_2, n_3)$ be the joint probability that there exist n_k customers in the kth stage ($k = 1, 2, 3$) in the steady state. Namely, $p(n_1, n_2, n_3) = \sum_{i_0} \sum_{i_1} \sum_{i_2} \sum_{i_3} x(n_1, n_2, n_3; i_0, i_1, i_2, i_3)$. We are interested in the decay rates of the joint queue-length probability, namely the ratios of neighboring $p(n_1, n_2, n_3)$'s:

$$\begin{aligned} r_1(n_1, n_2, n_3) &= \frac{p(n_1+1, n_2, n_3)}{p(n_1, n_2, n_3)}, \\ r_2(n_1, n_2, n_3) &= \frac{p(n_1, n_2+1, n_3)}{p(n_1, n_2, n_3)}, \\ r_3(n_1, n_2, n_3) &= \frac{p(n_1, n_2, n_3+1)}{p(n_1, n_2, n_3)}. \end{aligned}$$

First we scrutinize the behavior of $r_1(n_1, n_2, n_3)$. Figures 6.2a and 6.2b show graphs of $r_1(n_1, n_2, 10)$ and $r_1(n_1, n_2, 60)$, respectively. Figure 6.2a is similar to Figure 2.2a in Chapter 2, while Figure 6.2b is not.

In Figure 6.2a, we see that $r_1(n_1, n_2, 10)$ is relatively large very near the n_1 axis but it is in between η_1 and $\overline{\eta}_1$ in most of the region of (n_1, n_2). Especially $r_1(n_1, n_2, 10)$ is close to η_1 in a region in which n_1 is relatively larger than n_2 and it is close to $\overline{\eta}_1$ in a region in which n_2 is relatively larger than n_1.

In Figure 6.2b, $r_1(n_1, n_2, 60)$ shows the same quality as $r_1(n_1, n_2, 10)$ though the inter-

mediate region is shifted to right.

Figures 6.2c and 6.2d show the nature similar to Figures 6.2a and 6.2b, respectively. Also, the difference between Figures 6.2c and Figures 6.2d is almost same as that between Figures 6.2a and Figures 6.2a.

Figure 6.3a describes regions of (n_1, n_2) in which $r_1(n_1, n_2, 10)$ is close to η_1 or $\overline{\eta}_1$. In the dark gray region, labeled H_1, $r_1(n_1, n_2, 10)$ is close to η_1, namely $|r_1(n_1, n_2, 10) - \eta_1| < \varepsilon_1$ with $\varepsilon_1 = 0.2 \times |\eta_1 - \overline{\eta}_1|$, and in the light gray region, labeled $\overline{\mathrm{H}}_1$, $r_1(n_1, n_2, 10)$ is close to $\overline{\eta}_1$, namely $|r_1(n_1, n_2, 10) - \overline{\eta}_1| < \varepsilon_1$. The band B_1 between H_1 and $\overline{\mathrm{H}}_1$ represents the region where $r_1(n_1, n_2, 10)$ smoothly changes from $\eta_1 - \varepsilon_1$ to $\overline{\eta}_1 + \varepsilon_1$. Figure 6.3b describes regions of (n_1, n_2) in which $r_1(n_1, n_2, 10)$ is close to η_1 or $\overline{\eta}_1$ in the same manner as Figure 6.3b.

The band B_1 in Figure 6.3b is parallel with that in Figure 6.3a, though it is shifted to right. The shift will be due to the the sample values of n_3. The broken lines in Figure 6.3a and Figure 6.3b represent $\eta_1^{n_1}\eta_2^{n_2}\eta_3^{n_3} = \overline{\eta}_1^{n_1}\overline{\eta}_2^{n_2}\overline{\eta}_3^{n_3}$ where $n_3 = 10$ and 60, respectively.

Next we observe the behavior of $r_2(n_1, n_2, n_3)$. Figure 6.4 shows the graphs of $r_2(n_1, n_2, 10)$ and $r_2(n_1, n_2, 60)$. In Figure 6.4a, $r_2(n_1, n_2, 10)$ behaves just as $r_2(n_1, n_2)$ in Figure 2.2b. Figure 6.4b does not resembles Figure 6.4a, but the relationship between these two graphs are analogous to that between Figures 6.2a and 6.2b.

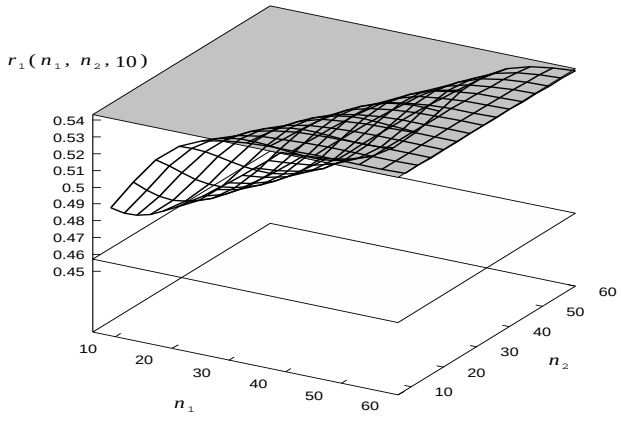
Figure 6.5 describes regions of (n_1, n_2) in which $r_1(n_1, n_2, N_3)$ ($N_3 = 10$ or 60) are close to η_2 or $\overline{\eta}_2$.

Figures 6.4c and 6.4d appear to be quite different from Figure 6.4, but the difference can be reasonable. In Figure 6.4a, both $r_2(n_1, n_2, 10)$ and $r_2(n_1, n_2, 10)$ are close to $\overline{\eta}_2$ in

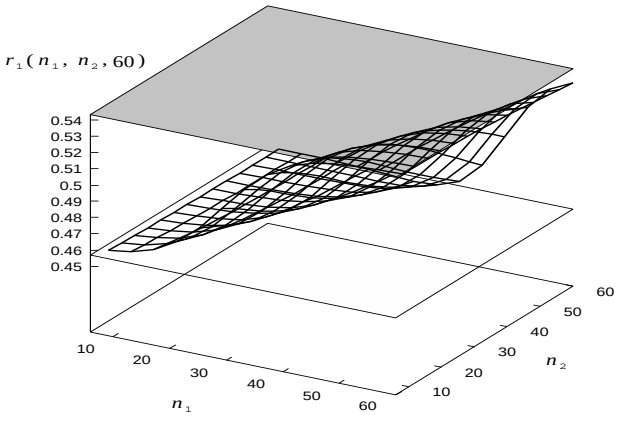
which $n_1 \approx 10$, while they are close to η_2 in which $n_1 \approx 60$. These agree with the behaviors of $r_2(10, n_2, n_3)$ and $r_2(60, n_2, n_3)$.

$r_3(n_1, n_2, n_3)$ is the last to be observed. Figures 6.6a and 6.6b shows the graphs of $r_3(n_1, 10, n_3)$ and $r_3(n_1, 60, n_3)$, respectively. Though we don't have any graphs in Chapter 2 which correspond to Figure 6.6, we can see that r_3 shows the same behavior as $r_1(n_1, n_2, n_3)$ and $r_2(n_1, n_2, n_3)$. That is, both $r_3(n_1, 10, n_3)$ and $r_3(n_1, 60, n_3)$ are near to η_3 where n_3 is relatively larger than n_3, while they are near to $\overline{\eta}_3$ with n_3 being relatively smaller than n_2. The difference between Figures 6.6a and 6.6b is similar to that between Figures 6.4a and Figures 6.4b.

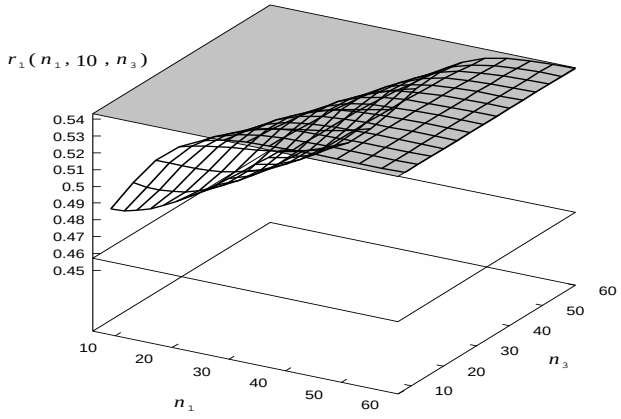
Figures 6.6c and 6.6d shows the graphs of $r_3(10, n_2, n_3)$ and $r_3(60, n_1, n_3)$, respectively. The behaviors of these graphs are similarly to those in Figures 6.4c and 6.4d respectively.



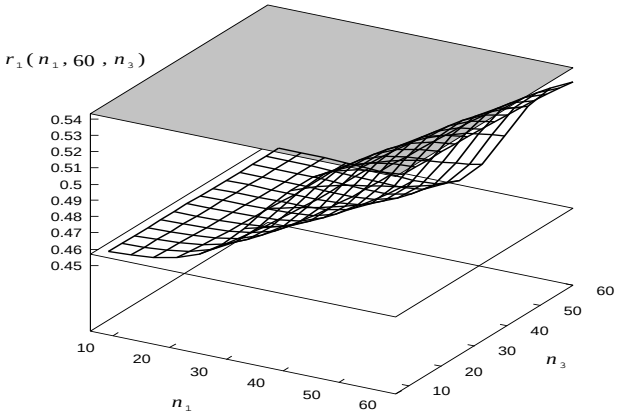
a: $N_3 = 10$



b: $N_3 = 60$

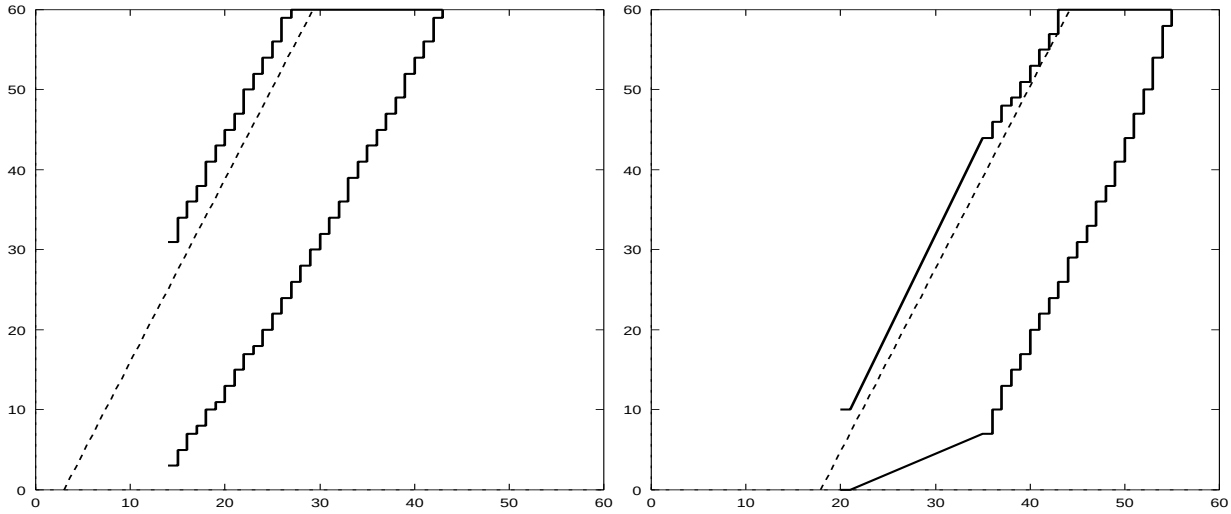


c: $N_2 = 10$



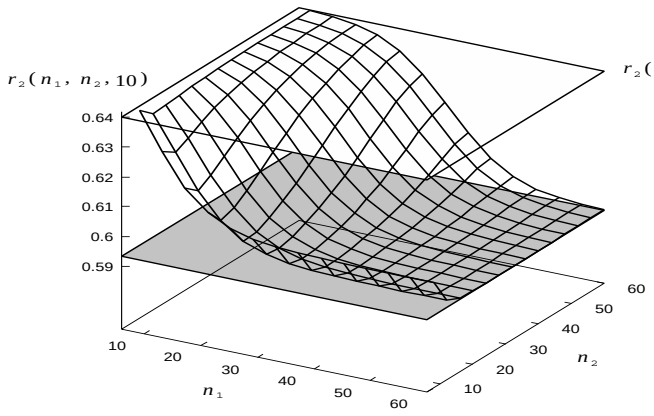
d: $N_2 = 60$

Figure 6.2: $r_1(n_1, N_2, n_3)$ behavior in $E_2/H_2/1 \rightarrow /E_2/1 \rightarrow /H_2/1$ $(\rho_1 = 0.6, \rho_2 = 0.8, \rho_3 = 0.4)$

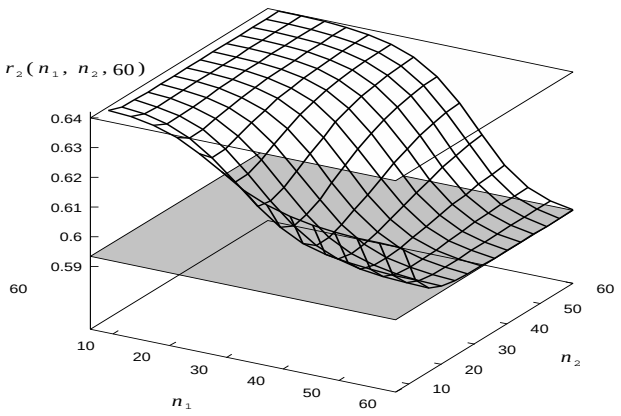


a: $r_1(n_1, n_2, 10)$ b: $r_1(n_1, n_2, 60)$

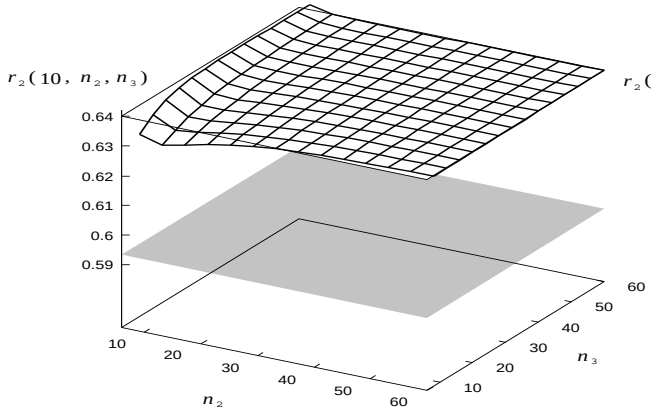
Figure 6.3: Characterization of $r_1(n_1, n_2, N_3)$ surface



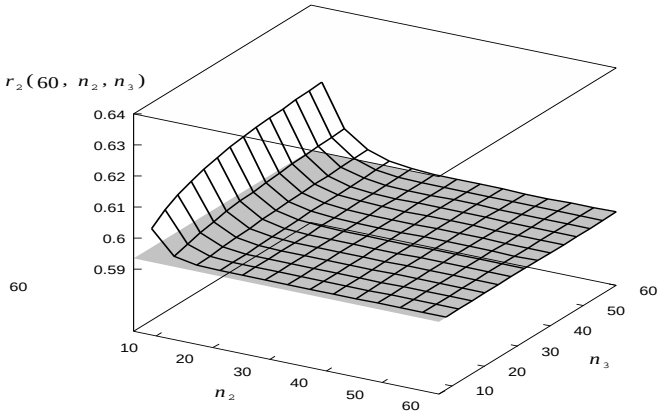
a: $N_3 = 10$



b: $N_3 = 60$

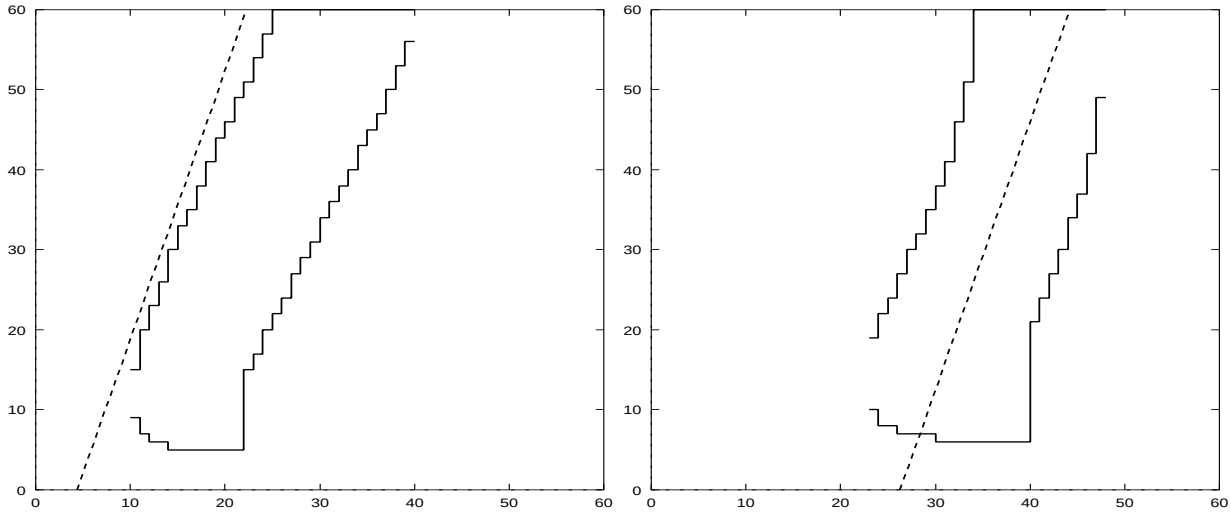


c: $r_2(10, n_2, n_3)$



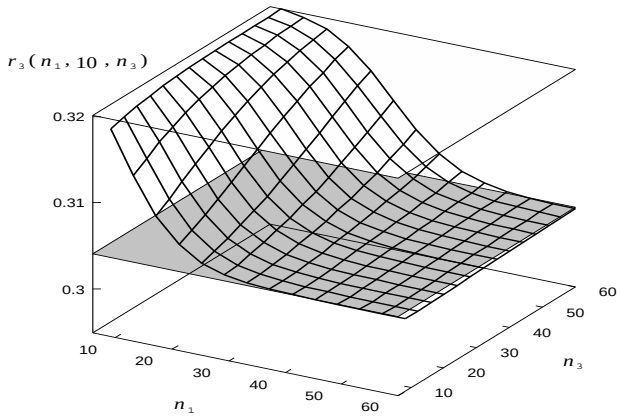
d: $r_2(60, n_2, n_3)$

Figure 6.4: r_2 behavior in $E_2/H_2/1 \rightarrow /E_2/1 \rightarrow /H_2/1$ $(\rho_1 = 0.6, \rho_2 = 0.8, \rho_3 = 0.4)$

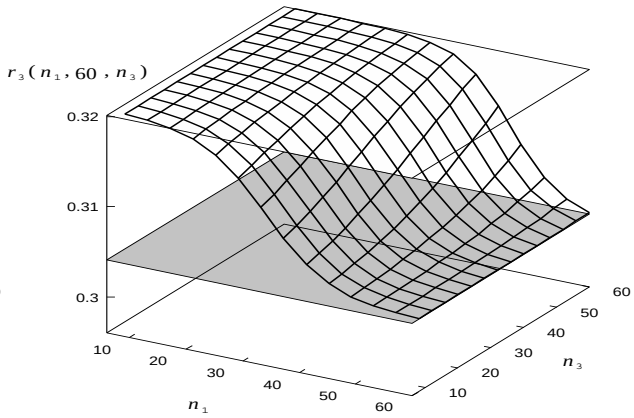


a: $r_2(n_1, n_2, 10)$ b: $r_2(n_1, n_2, 60)$

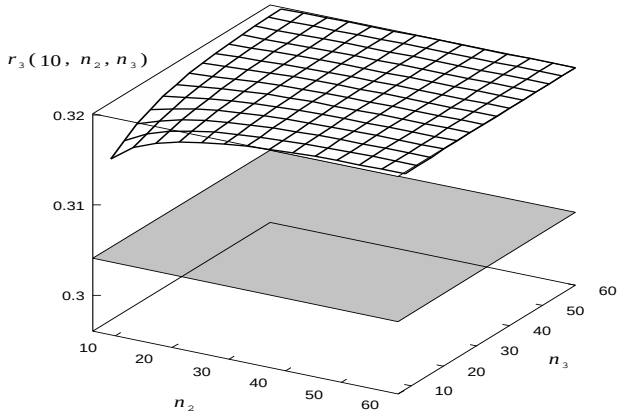
Figure 6.5: Characterization of $r_2(n_1, n_2, N_3)$ surface



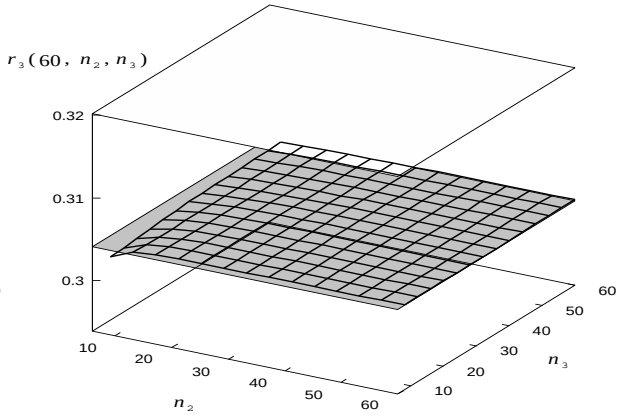
a: $r_3(n_1, 10, n_3)$



b: $r_3(n_1, 60, n_3)$



c: $r_3(10, n_2, n_3)$



d: $r_3(60, n_2, n_3)$

Figure 6.6: r_3 behavior in $E_2/H_2/1 \rightarrow /E_2/1 \rightarrow /H_2/1$ $(\rho_1 = 0.6, \rho_2 = 0.8, \rho_3 = 0.4)$

Bibliography

[1] Tayfur Altiok. Approximate analysis of queues in series with phase-type service times and blocking. *Operations Research*, **37**(4): pp.601–610, 1989.

[2] F. Baskett, K. M. Chandy, R. R. Muntz, and F. Palacios. Open, closed and mixed networks of queues with different classes of customers. *Journal of ACM*, **22**: pp.248–260, 1975.

[3] Nicholas Beaumont. Some results pertaining to lumping of a Markov chain. *Asia-Pacific Journal of Operational Research*, **9**: pp.1–8, 1992.

[4] O. J. Boxma. On a tandem queueing model with identical service times at both counters, I. *Advances in Applied Probability*, **11**: pp.616–643, 1979.

[5] O. J. Boxma. On a tandem queueing model with identical service times at both counters, II. *Advances in Applied Probability*, **11**: pp.644–659, 1979.

[6] Thomas M. Chen. On the independence of sojourn times in tandem queues. *Advances in Applied Probability*, **21**: pp.488–489, 1989.

[7] Adrian E. Conway and Nicolas D. Georganas. *Queueing Networks — Exact Computational Algorithms: A Unified Theory Based on Decomposition and Aggregation*. The MIT Press, 1989.

[8] Robert B. Cooper and Donald Gross. On the convergence of Jacobi and Gauss-Seidel iteration for steady-state probabilities of finite-state continuous-time Markov chains. *Communications in Statistics, Stochastic Models*, **7**(1): pp.185–189, 1991.

[9] P. J. Courtois and G. Louchard. Approximation of eigencharacteristics in nearly-completely decomposable stochastic systems. *Stochastic Processes and Their Applications*, **4**: pp.283–296, 1976.

[10] William Feller. *An Introduction to Probability Theory and Its Applications, Volume 2*. John Wiley & Sons, 1966.

[11] Kou Fujimoto and Yukio Takahashi. Numerical experiments for the tail probabilities in two-stage tandem queueing systems. In *The Proceedings of Performance Models for Information Communication Networks*, 1993.

[12] Kou Fujimoto and Yukio Takahashi. Tail behavior of the steady-state distribution in two-stage tandem queues: Numerical experiment and conjecture. *Journal of the Operations Research Society of Japan*, **39**(4): pp.525–540, 1996.

[13] A. Ganesh and V. Anantharam. Stationary tail probabilities in exponential server tandem queues with renewal arrivals. In Frank P. Kelly and Ruth J. Williams, editors,

Stochastic Networks: The IMA Volumes in Mathematics and Its Applications volume **71**, pages pp.367–385. Springer-Verlag, 1995.

[14] Guang-Hui Hsu and Uwe Jensen. The matched queueing network PH/M/c $\rightarrow$ $\circ$PH/PH/1. *Queueing Systems*, **13**: pp.315–333, 1993.

[15] Hiroaki Ishihara, Takeshi Horie, Satoshi Inao, Toshiyuki Shimizu, Sadayuki Kato, and Satoshi Ikesaka. Third generation message passing computer AP1000. In *the Proceedings of International Symposium on Supercomputing*, pages 44–65, 1991.

[16] F. I. Karpelevitch and A. Ya. Kreinin. Joint distribution in Poissonian tandem queues. *Queueing Systems*, **12**: pp.273–286, 1992.

[17] Tsuyoshi Katayama. Mean sojourn times in a multi-stage tandem queue served by a single server. *Journal of the Operations Research Society of Japan*, **31**(2): pp.233–251, 1988.

[18] Leonard Kleinrock. *Queueing Systems, Volume 1: Theory.* John Wiley & Sons, 1975.

[19] Dieter König and Masakiyo Miyazawa. Relationships and decomposition in the delayed bernoulli feedback queueing system. *Journal of Applied Probability*, **25**:169–183, 1988.

[20] Dieter König and Volker Schmidt. Relationships between time/customer stationary characteristics of tandem queues attended by a single server. *Journal of the Operations Research Society of Japan*, **27**(3): pp.191–204, 1984.

[21] Guy Latouche and Marcel F. Neuts. Efficient algorithmic solutions to exponential tandem queues with blocking. *SIAM Journal on Algebraic and Discrete Methods*, **1**: pp.93–106, 1980.

[22] Pierre Le Gall. The overall sojourn time in tandem queues with identical successive service times and renewal input. *Stochastic Processes and their Applications*, **52**: pp.165–178, 1994.

[23] Xiao-Gao Liu and John A. Buzacott. A decomposition-related throughput property of tandem queueing networks with blocking. *Queueing Systems*, **13**: pp.361–383, 1993.

[24] Guy Louchard and Guy Latouche. Geometric bounds on iterative approximations for nearly completely decomposable Markov chains. *Journal of Applied Probability*, **27**: pp.521–529, 1990.

[25] Naoki Makimoto. Quasi-stationary distributions in a PH/PH/c queue. *Commun. Statist. — Stochastic Models*, **9**((2)): pp.195–212, 1993.

[26] Naoki Makimoto and Yukio Takahashi. Asymptotic behavior of the stationary distribution in phase-type tandem queues. In *the Proceedings of Performance Models for Information Communication Networks*, pages 347–358. Special Interest Group: Queueing Theory and Its Application, The Operations Research Society of Japan, 1993.

[27] Ivo Marek and Daniel B. Szyld. Local convergence of the (exact and inexact) iterative aggragetion method for linear systems and Markov operators. *Numerische Mathematik*, **69**: pp.61–82, 1994.

[28] Douglas R. Miller. Computation of steady-state probabilities for $M/M/1$ priority queues. *Operations Research*, **29**(5): pp.945–958, 1981.

[29] Masakiyo Miyazawa. The derivation of invariant relations in complex queueing systems with stationary inputs. *Advances in Applied Probability*, **15**: pp.874–885, 1983.

[30] Marcel F. Neuts. *Matrix-Geometric Solutions in Stochastic Models*. The John Hopkins University Press, 1981.

[31] Marcel F. Neuts. The abscissa of convergence of the Laplace-Stieltjes transform of a PH-distribution. *Commun. Statist. — Simulation and Computation*, **13**: pp.367–373, 1984.

[32] Marcel F. Neuts. The caudal characteristic curve of queues. *Advances in Applied Probability*, **18**: pp.221–254, 1986.

[33] Marcel F. Neuts and Yukio Takahashi. Asymptotic behavior of the stationary distributions in the GI/PH/c queue with heterogeneous servers. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **57**: pp.441–452, 1981.

[34] H. T. Papadopoulos and C. Heavey. Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines. *European Journal of Operational Research*, **92**: pp.1–27, 1996.

[35] V. Ramaswami. From the matrix-geometric to the matrix-exponential. *Queueing Systems*, **6**: pp.229–260, 1990.

[36] V. Ramaswami and Peter G. Taylor. Some properties of the rate operators in level dependent quasi-birth-and-death processes with a countable number of phases. *Commun. Statist.—Stochastic Models*, **12**(5): pp.143–164, 1996.

[37] Paul J. Schweitzer. Aggregation methods for large Markov chains. In G. Iazeolla, P. J. Courtois, and A. Hordijk, editors, *Mathematical Computer Performance and Reliability*, pages 275–286. North-Holland, Amsterdam, 1984.

[38] L. P. Seelen. An algorithm for Ph/Ph/c queues. *European Journal of Operational Research*, **23**: pp.118–127, 1986.

[39] E. Seneta. *Non-negative Matrices and Markov Chains, 2nd ed.* Springer-Verlag, 1980.

[40] Bhaskar Sengupta. Markov processes whose steady state distribution is matrix-exponential with an application to the $GI/PH/1$ queue. *Advances in Applied Probability*, **21**: pp.159–180, 1989.

[41] Yu Song and Yukio Takahashi. Aggregate approximation for tandem queueing system with production blocking. *Journal of the Operations Research Society of Japan*, **34**: pp.329–353, 1991.

[42] Ushio Sumita and Maria Rieders. Application of the replacement process approach for computing the ergodic probability vector of large scale row-continuous Markov chains. *Journal of the Operations Research*, **33**: pp.279–306, 1990.

[43] W. Szczotka and F. P. Kelly. Asymptotic stationarity of queues in series and the heavy traffic approximation. *The Annals of Probability*, **18**(3): pp.1232–1248, 1990.

[44] Yukio Takahashi. A lumping method for numerical calculations of stationary distributions of Markov chains. Research Reports on Information Sciences B-18, Tokyo Institute of Technology, June 1975.

[45] Yukio Takahashi. Asymptotic exponentiality of the tail of the waiting time distribution in a PH/PH/c queue. *Advances in Applied Probability*, **13**: pp.619–630, 1981.

[46] Yukio Takahashi. A new type aggregation method for large Markov chains and its application to queueing networks. In M. Akiyama, editor, *Proceedings of ITC 11*, pages 3–4 A.1.1–4. Elsevier, 1985.

[47] Yukio Takahashi. Aggregate approximation for acyclic queueing networks with communication blocking. In H. G. Perros and T. Altiok, editors, *Queueing Networks with Blocking*, pages 33–46. Elsevier Science Publishers B.V., 1989.

[48] Yukio Takahashi, Naoki Makimoto, and Kou Fujimoto. Asymptotic properties in a quasi-birth-and-death process with a countable number of phases. Research Report on Mathematical and Computing Sciences B-321, Tokyo Institute of Technology, 1996.

[49] Yukio Takahashi and Yoshinori Takami. A numerical method for the steady-state probabilities of a GI/G/c queueing system in a general class. *Journal of the Operations Research*, **19**: pp.147–157, 1976.