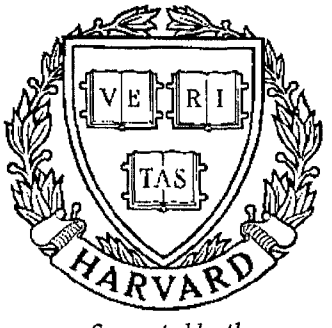
THESIS REPORT
Ph.D.

S Y S T E M S
R E S E A R C H
C E N T E R



Supported by the National Science Foundation Engineering Research Center Program (NSFD CD 8803012), the University of Maryland, Harvard University, and Industry

Burst Reduction Properties of the Leaky Bucket and the Calculus of Burstiness

by L. Kuang
Advisor: A.M. Makowski

Ph.D. 92-19

ABSTRACT

Title of Dissertation: Burst Reduction Properties Of The Leaky Bucket And The Calculus Of Burstiness

Lei Kuang, Doctor of Philosophy, 1992

Dissertation directed by: Armand M. Makowski, Professor
Department of Electrical Engineering

This dissertation considers an important issue in high-speed networks – the understanding of the interaction between the network and the bursty traffic it handles.

The first part of the dissertation analyzes the leaky bucket (LB), a proposed congestion control scheme for ATM networks. There are many preliminary results in the literature on the LB. Most of them, however, are numerical ones under some Markovian assumptions. In this dissertation, it is shown that under very mild assumptions, the LB is burst reducing. More importantly, this burst reduction property obeys some monotonicity properties in all the parameters of the LB. This property is then used to the design of the LB. Bounds and approximations for some other performance measures of the LB are also derived to facilitate the design. The mathematical tool developed in this part can also be used to study a large class of non-stationary stochastic processes.

In the second part of the dissertation, some general results on the calculus of burstiness are obtained for renewal processes. These results are useful in understanding the effects on the burstiness of the traffic in some systems through multiplexing, splitting, rate control, and scheduling. The burstiness is characterized by both the peakedness functional of the traffic and the squared coefficient of variation of its asymptotic version. As examples, problems in network scheduling and in the design of multiplexer with rate control mechanism are discussed.

Burst Reduction Properties Of The Leaky Bucket And The Calculus Of Burstiness

by

Lei Kuang

Dissertation submitted to the Faculty of the Graduate School of
the University of Maryland in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy
1992

Advisory Committee:

Professor Armand M. Makowski, Chairman/Advisor
Professor Evaggelos Geraniotis
Professor Mark Shayman
Professor Satish K. Tripathi
Assistant Professor Bernard L. Menezes

Dedication

TO MY BELOVED WIFE YEJUN

for her love, understanding, and support

Acknowledgements

I would like to first express my deep-felt gratitude to my advisor Prof. Armand M. Makowski for his ceaseless guidance, continuous encouragement, and unrivaled consideration. His persistence and serious attitude have been and will continuously be a reference model for me. I owe him very much for his providing me the freedom in selecting research topics and for his financial assistance.

I owe a significant debt of gratitude to Dr. P. Tsoucas for his guidance and financial assistance in my early graduate years. I am also grateful to Prof. E. Geraniotis, Prof. M. Shayman, Prof. B. L. Menezes, and Prof. S. K. Tripathi for their serving on my advisory committee and carefully reading this dissertation.

My special thanks are due to Prof. Anantharam, Prof. Shanthikumar, Dr. Eckberg, and Dr. Chang for providing me their most recent research results. I would like to thank Dr. J. Pan, Dr. L. Gün, Dr. Tedijianto, Dr. V. Subir, Dr. Z. Liu and Mr. Y. Kim for helpful discussions.

This work is conducted in the stimulating atmosphere of the Electrical Engineering Department and the Systems Research Center at the University of Maryland. This research was supported by the National Science Foundation under grants NFSD CDR-88-03012 and NCR-88-14566-02.

Lei Kuang

August 1992, College Park, Maryland

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

BACKGROUND AND SUMMARY

1.1 Background

Along with the advances in high-capacity optical fiber and VLSI techniques, the next generation of LANs (local area networks) and WANs (wide area networks) are expect to be able to handle data at Gbps (Gigabits per second) rates. Such high-speed networks will give rise to a number of new services in addition to the conventional voice service. These new services will include: high-speed data exchange, high-resolution still picture transfer, mixed-mode document exchange and retrieval, high-quality interactive videotex and, in particular, videophone, video conference, and distributed TV. Most of these services show very specific characteristics with respect to bit rate (constant or variable) or to the required quality of service (QOS) in terms of information loss, information delay, etc [51]. Many new QOS objectives exceed the capabilities of the recently introduced ISDN (integrated services digital network), which does not seem adequate to integrate all of those services [51]. Therefore, the CCITT has recommended that the new broadband ISDN (B-ISDN) be based on a new technique, the asynchronous transfer mode (ATM) [51].

The ATM is a fast packet switching technique based on virtual connection using small fixed-sized packets called cells. A cell consists of 53 octets, five of which are reserved for the cell header and 48 for user information. The ATM bearer service is capable of integrating any connection-oriented service. These include circuit-switched type services with constant bit rates, as well as those with highly variable bit rates, as are found in computer, video, and packetized voice communications.

This promising integration technique, however, raises a number of new problems due to the higher degree of resource sharing compared to that of the conventional

synchronous transfer mode (STM). Some of these problems are highly dependent on the characteristics of the user or traffic source. For example, the coincidence of peak bit rate of many connections sharing the same transmission, switching, or buffer resources may cause excessive delay or even information losses which must be kept within strict limits to meet the overall QOS objective in a B-ISDN.

Therefore it is of vital importance to have a good grasp of traffic behavior in high-speed networks. There have been many attempts in studying this traffic, although its characteristics are still not well known. The interaction between network operations and the traffic inside the network is even less understood. Due to the complexity of the problems studied, most of the existing results are of the numerical type under very restrictive assumptions which are not satisfied in many cases. The purpose of this thesis is to study the structural properties of the traffic and the effect of network operations (such as the leaky bucket input rate regulation) on these properties. The results we get are much less restrictive, and can be applied to a large range of traffic models.

1.2 Summary

In this thesis, we develop tools for studying the traffic in a high-speed network and the effect of some network operations (such as the leaky bucket, multiplexing, routing, and rate control) on the "burstiness" of the traffic.

The thesis contains two parts.

The first part of the thesis focuses on the study of an important network operation in high-speed networks – the leaky bucket (LB). Due to its importance, the LB has been studied by many researchers. However, most of these results are numerical ones under some Markovian assumptions. The contribution in this part of the thesis lies in new tools for studying the LB and for obtaining stochastic comparison results under very mild assumptions. This part has two chapters. In Chapter 2, we study the burst reduction property of the LB, which is a very important feature of this input rate regulation scheme. In particular, we show that under very mild assump-

tions, the output traffic from a LB will be less bursty than the input traffic. This property has been shown previously through numerical examples under Markovian assumptions. More importantly, we show that the burst reduction property of the LB obeys some monotonicity properties in all the parameters describing the LB. In Chapter 3, we study some other performance measures of the LB based on approximations. Together with the results obtained in Chapter 2, we discuss design issues of the LB.

In the second part of the thesis, we establish some general results on the calculus of burstiness. Via two measures of burstiness (the coefficient of variation and the peakedness functional), we study the effects of thinning and superposition on the burstiness of renewal processes. These results are useful in understanding certain network operations such as multiplexing, routing, and rate control. In Chapter 4, we discuss this issue with respect to the criterion of the squared coefficient of variation, while in Chapter 5, we discuss this issue again with respect to the peakedness functional. We conclude the second part by studying two examples in Chapter 6. The first example is to design a thinning scheme for a trunk formed by multiplexing a number of independent sources. The other example is to compare policies in traffic scheduling.

PART I

THE LEAKY BUCKET

CHAPTER 2

BURST REDUCTION PROPERTIES

2.1 Introduction

In order to guarantee a given QOS to the user, the network has to control the traffic flow according to the nature of the service. In high-speed networks like B-ISDN, a whole set of control schemes will be applied to the various levels of activity – connection, burst, and cell levels. Congestion control at burst and cell levels in high-speed networks presents a new challenge [3, 60]. In the first part of the thesis, we discuss a commonly recognized congestion control scheme at burst and cell levels in high-speed networks, the leaky bucket (LB).

There are two kinds of congestion control schemes in general: reactive congestion control (RCC) and preventive congestion control (PCC) [61]. In RCC, congestion is detected after it has occured and appropriate measures are taken in reaction to congestion conditions inside the network. PCC, on the other hand, seeks to prevent congestion from arising before it can adversely affect the QOS of the network connection. A survey of congestion controls is given by Jain [37].

In traditional X.25 packet switching networks, congestion control is based on end-to-end exchange of control messages called window flow control. After successfully receiving a packet, the destination node sends an ACK message back to the source node. A packet at the source node will enter the network only if the number of unacknowledged packets is less than the window size W. Otherwise it waits in a buffer outside the network. This scheme guarantees that at most W packets are in the network for a particular source-destination pair. Window flow control is an example of RCC which regulates the traffic flow at the access point based on current traffic levels within the network. A typical window flow control scheme is depicted

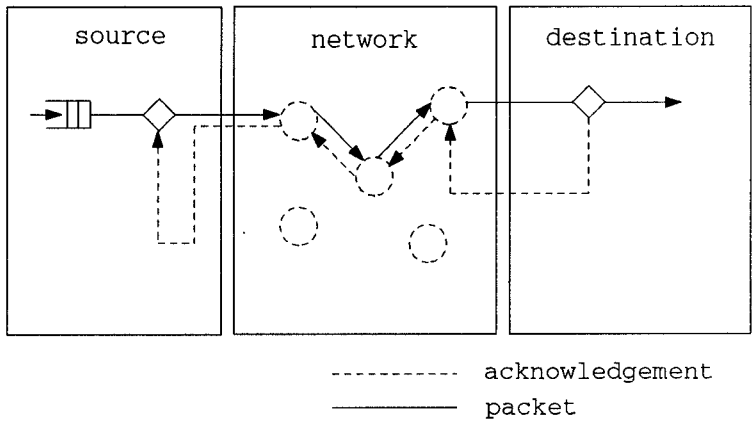


Figure 2.1: Window flow control in X.25 packet switching networks

in Fig. 2.1 [62].

This scheme, however, will not work in high-speed networks, where the propagation delay across the network typically dominates the switching and buffering delays. The reason for this is that while the transmission speeds in high-speed networks scale up, the propagation delay does not. Therefore the feedback from the destination is usually outdated, and the action that the source takes is too late to resolve congestion at the switching elements or at the buffers. For example [23], the transmission time for an ATM cell is approximately 2.8 μs at 150 Mbps, and the resulting time-constant in a cell queue will be in the order of 10 μs for moderate utilization, while end-to-end signal propagation delay though fiber for a 1000 km connection is approximately 5 ms. It can be seen that in order to make efficient use of network resources, the window size for high bit rate services would be in the order of thousands; this in turn requires a big processing burden and huge buffers, and results in large network delays. Furthermore, this scheme cannot guarantee user throughput. The nature of the traffic also affects the design of the congestion control. In high-speed networks, some real-time traffic cannot be slowed down to cope with network congestion, so that some level of bandwidth guarantee is required [60]. Finally, the high speed nature of the network requires that the control operation works at the speed of the communication link. Therefore, simple mechanisms which do not rely heavily on network feedback are desirable as they are suitable for

implementation in high speed hardware.

It is widely agreed that the most effective approach to congestion control in ATM networks is to use PCC rather than RCC [35, 60]. One feasible solution is to break the feedback loop of the window flow control scheme, which leads to an open-loop input rate regulation. A source policing control monitors the user data flow of each connection with respect to its negotiated traffic parameters. If a violation is detected, an appropriate action has to be taken, such as immediate dropping [11] or marking of excessive cells for loss in case of congestion [21, 26]. The best-known source policing function is the LB algorithm [57], which we study in this and in the next chapters. Other algorithms operate on cell counts within sliding or jumping time windows [17, 50], and are studied in [24, 25, 39, 47] among others.

The LB has been studied by many authors, existing results are first reviewed in the next section. For more information on LB, the reader can also refer to [4, 5, 18, 21, 23, 26, 27, 33, 41].

2.2 Literature Review

The performance measures which have been mostly discussed in the literature on the LB include the probability of cell loss, cell delay at the LB (or the queue length of the LB), throughput from the LB, and burst reduction properties of the LB.

Numerical results on cell delay and the probability of cell loss were obtained under some Markovian assumptions on the source, e.g., the two-state on/off arrival process [50], the Markovian arrival process (MAP [45]) [6], and the discrete-time Markovian arrival process (DMAP) [9]. Under similar assumptions, the queue length distribution of the input buffer at the access point was calculated in [1, 54] using the matrix-analytic technique [48, 49]. Numerical studies of the burst reduction properties of the LB were done for the Poisson arrival process [53], the two-state Markov modulated Poisson process (MMPP) [22], and the two-state Markov modulated Bernoulli process (MMBP) [34].

Although the LB appears to be a very simple mechanism (see the next section for detail), its analysis under general assumptions turns out to be quite difficult. The burst reduction property of the LB was established in a purely deterministic context in [44] by introducing the concept of a "burstiness curve". Independently of the work in this thesis, [2] proved recently the burst reduction property of the LB. The approach used in [2] combines a sample path construction with ideas from the theory of stationary processes. It is shown there that the burstiness (expressed as the steady-state queue lengths at a down-stream single-server queue with deterministic service time less than the token generation period of the LB) of the output traffic from the LB increases as the token buffer size increases. An LB with token pool size of one was studied with the help of majorization [13]. There the author showed that the throughput from the LB increases as the variability of the input traffic decreases. The throughput from LB is also shown to be increasing and stochastically concave in the token pool size [10]. Therefore, the cell loss at the LB is decreasing and stochastically convex in the token pool size. It was also shown that if the token pool size is larger and the token generation rate is larger for an un-buffered LB, then the

queue length of a down-stream single-server queue is stochastically larger. For a buffered LB, it was shown that the number of cells lost at any time is stochastically convex in the token pool size and in the input buffer size.

Another approach in the study of the LB is to use approximations. Fluid approximation was used by several authors [11, 28, 29, 31].

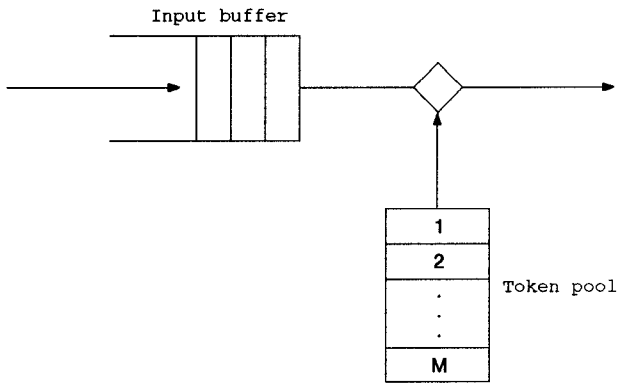


Figure 2.2: The leaky bucket

2.3 The Leaky Bucket

The original LB was proposed as a simple algorithm for monitoring and policing sources in an ATM environment [57]. It consists of a token pool of finite capacity and of a mechanism for generating tokens into the token pool at a constant rate. If the token pool happens to be full when a token is generated, the newly generated token is discarded. In the case where a token is generated exactly at an arrival epoch, we assume that the token is generated after the arrival of the cell. At each cell arrival epoch, the LB algorithm checks whether or not there are tokens in the token pool. If at least one token is found, the cell is passed to the network and a token is removed from the token pool; otherwise, the cell is lost. In this way the LB ensures that the departure rate from the LB will not exceed a certain value. However, cells may be lost from time to time due to lack of available tokens. In order to maintain a low probability of cell loss on conforming sources (i.e., sources which do not exceed the preset parameters such as average rate, peak rate, and burstiness), the LB must be so dimensioned that the token generation rate is relatively large, with the consequence that the (unbuffered) LB cannot effectively restrict non-conforming sources [34].

There are many variations of the original LB. The one discussed here is the buffered LB proposed in [14, 53]. In order to reduce the probability of cell loss, an input buffer is added to the original LB (Fig. 2.2). Cells which cannot obtain tokens upon their arrival will wait in the input buffer. Only when the input buffer is full,

is the arriving cell lost. Newly generated tokens are distributed to cells in the input buffer (if any) in a FCFS manner.

A LB of this kind has two basic functions: (i) reduce the rate of the traffic (i.e., police the traffic) and (ii) reduce the burstiness of the traffic (i.e., smooth the traffic). In this chapter, we focus on the burst reduction properties of the LB. The first function will be discussed in Chapter 3.

2.4 The Model

Since all cells are admitted into the LB except those which are policed, cells which enter the LB see a virtually infinite input buffer. Therefore, except where explicitly stated, we assume that the input buffer has infinite capacity. If we assume the input buffer to have infinite capacity, a LB can be characterized by two parameters, say M and D, where M is the size of the token pool, and D is the token generation period which is assumed constant. Tokens are assumed to be generated at times $\{kD, k = 0, 1, \cdots\}$. Cells are tagged upon arrival in the order of their arrival. We also assume that the input buffer operates as a FCFS queue. These assumptions are made only for notational convenience and do not affect the results we shall obtain. A LB with parameters D and M is denoted by $LB(D, M)$.

Sometimes it is easier to describe the evolution of the LB through its states. To this end, we define

$P(t)$: the number of tokens available in the token pool at time $t \geq 0$

and

$Q(t)$: the number of cells waiting in the input buffer at time $t \geq 0$.

By convention, the processes P and Q are taken to be right continuous. By the assumptions made in Section 2.3, if a cell arrives exactly at token generation epoch t, $P(t)$ and $Q(t)$ remain unchanged. This convention removes the ambiguity when the token pool happens to be full when a cell arrives at a token generation epoch. Although different conventions lead to different sample path realizations, they will not affect the results of this chapter. From the definition of the LB, we see that for all $t \geq 0$, at least one of the quantities $Q(t)$ or $P(t)$ must be zero, (i.e., $Q(t)P(t) = 0$), and both quantities $Q(t)$ and $P(t)$ can be uniquely recovered from their difference $S(t)$ defined by

$$S(t) = Q(t) - P(t), \qquad t \geq 0. \tag{2.1}$$

If $S(t) \geq 0$, then $Q(t) = S(t)$ and $P(t) = 0$; and if $S(t) < 0$, then $Q(t) = 0$ and $P(t) = -S(t)$. We assume that the first cell arriving at time $t = 0$ finds the LB in state s. In order to emphasize the role of the initial state, we sometimes write $LB_s(D, M)$ to denote $LB(D, M)$ with initial state s.

The evolution of the LB can also be described by two sequences of $\mathbb{R}_+$-valued random variables (rvs) $\boldsymbol{\alpha} = \{\alpha_n, n = 1, 2, \cdots\}$ and $\boldsymbol{\delta} = \{\delta_n, n = 1, 2, \cdots\}$ with $\alpha_1 = \delta_1 = 0$. We interpret α_1 (resp. δ_1) as the first arrival (resp. departure) epoch as they should be under our assumptions. For $n = 1, 2, \cdots$, we interpret α_{n+1} (resp. δ_{n+1}) as the inter-arrival (resp. inter-departure) time between the n^{th} and the $(n+1)^{st}$ cells. The inter-arrival (resp. inter-departure) times may be zero to represent batch arrivals (resp. departures).

Our analysis will be done in both the transient and steady-state regimes. Although we impose no assumptions on $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ for the transient analysis, we need the following notion of convex stability for the steady-state analysis.

Definition 2.1 *For any $\mathbb{R}_+$-valued sequence of rvs $\boldsymbol{\zeta} = \{\zeta_n, n = 1, 2, \cdots\}$, we say that $\boldsymbol{\zeta}$ is convexly stable if there exists an integrable $\mathbb{R}_+$-valued rv ζ (i.e., $\mathbb{E}[\zeta] < \infty$) such that*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \phi(\zeta_i) = \mathbb{E}[\phi(\zeta)] \qquad a.s. \tag{2.2}$$

for any convex function $\phi : \mathbb{R}_+ \to \mathbb{R}$ provided the expectation exists, in which case we call ζ the asymptotic version of $\boldsymbol{\zeta}$.

The asymptotic version ζ of a convexly stable sequence $\boldsymbol{\zeta}$ is unique as stated in the following theorem.

Theorem 2.1 *The asymptotic version of a convexly stable sequence $\boldsymbol{\zeta}$ is unique, i.e., if ζ and ξ are two asymptotic versions of $\boldsymbol{\zeta}$, then*

$$\zeta \stackrel{d}{=} \xi \tag{2.3}$$

where $\stackrel{d}{=}$ denotes equality in distribution.

Proof. From (2.2) we have immediately that

$$\mathbb{E}[\phi(\zeta)] = \mathbb{E}[\phi(\xi)] \tag{2.4}$$

for any convex function $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that the expectations exist. For $s \geq 0$, taking $\phi(x) = e^{-sx}$, $x \geq 0$, we conclude that the Laplace-Stieltjes transforms of ζ and ξ coincide, and (2.3) follows from the fact that the Laplace-Stieltjes transform of a rv uniquely determines its distribution [8, Theorem 26.2]. ■

The convex stability of $\boldsymbol{\zeta}$ implies

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \zeta_i = \mathbb{E}[\zeta] \qquad a.s. . \tag{2.5}$$

In particular if the sequence of inter-arrival times $\boldsymbol{\alpha}$ is convexly stable, then the (long-run) cell arrival rate λ is well defined by

$$\lambda = \left(\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \alpha_i \right)^{-1} . \tag{2.6}$$

In this chapter, we always assume that the sequence of inter-arrival times $\boldsymbol{\alpha}$ is convexly stable, although the transient results hold without this assumption.. We shall say that the LB is stable for input traffic rate λ if

$$\lambda D < 1. \tag{2.7}$$

The convex stability condition (2.2) is not too restrictive; it holds for many arrival processes of interest as we now indicate.

Example 1 – Renewal processes. By the strong law of large numbers, the convex stability condition is satisfied for renewal processes with finite first moment.

Example 2 – Stationary processes. By [38, Theorem 5.6, p. 487], a stationary processes is convexly stable if it is ergodic.

The convex stability condition holds for many non-stationary processes as well. For example, deterministic periodic arrival processes with finite period are non-stationary in general, yet they are convexly stable.

It is not easy to show directly that a departure process from a LB is convexly stable. However, if the inter-departure times have a stationary version, then they are convexly stable, and the asymptotic version coincides with the stationary version.

2.5 Burst Reduction Properties

In this section, we present the burst reduction property of the LB. More importantly, we shall show that this property is monotone with respect to various parameters of the LB. The proofs of the theorems will be postponed to Section 2.8. In this section, we use the convex ordering of rvs to capture the burstiness of the inter-departure times from the LB. For the definitions and basic properties of stochastic orderings, the reader is referred to Appendix B and [52, 55].

From now on we assume that all the LBs start with the same initial condition (except in Theorem 2.6) and are fed by the same but arbitrary input sequence $\boldsymbol{\alpha}$ with rate λ. Since the output traffic from the LB tends to be more evenly spaced, it would be expected to be less "bursty" than the input traffic. Moreover, since the output traffic is governed by various parameters of the LB, it is desirable to find the relationship between these parameters and the burst reduction property of the LB. This leads to the following monotonicity properties with respect to various parameters of the LB.

2.5.1 Monotonicity in M

We first consider the monotonicity of the inter-departure times with respect to the token pool size M. Consider two LBs $LB(D, M)$ and $LB(D, \hat{M})$ with $D \geq 0$ and $1 \leq M \leq \hat{M}$, and let $\boldsymbol{\delta}$ and $\hat{\boldsymbol{\delta}}$ be sequences of inter-departure times from $LB(D, M)$ and $LB(D, \hat{M})$, respectively. Our first result says that the burstiness of the output traffic decreases as the size of the token pool decreases.

Theorem 2.2 *(Monotonicity in M) Assuming $\boldsymbol{\delta}$ and $\hat{\boldsymbol{\delta}}$ to be convexly stable and $\lambda D < 1$, we have*

$$\delta \leq_{cx} \hat{\delta}, \tag{2.8}$$

where δ and $\hat{\delta}$ are the asymptotic versions of $\boldsymbol{\delta}$ and $\hat{\boldsymbol{\delta}}$, respectively, and therefore, $c^2(\delta) \leq c^2(\hat{\delta})$.

An intuitive explanation of Theorem 2.2 is as follows. For any bursty sources, there will be periods in which only a few cells arrive (i.e., periods of silence) and periods in which cells arrive in large batches (i.e., periods of burst). During any silence period, the LB with a larger token pool is more likely to accumulate more tokens than the LB with a smaller token pool. During the next burst period, the LB with more available tokens will let more cells enter the network instantaneously, and the output traffic will thus tend to be more bursty. This intuition can be used in understanding other results in this section.

2.5.2 Monotonicity in D

The monotonicity of the inter-departure times with respect to D is considered next. We first assume $M = \infty$ so that no token is lost in this situation. Consider the two LBs $LB(D, \infty)$ and $LB(\tilde{D}, \infty)$ with $0 \leq \tilde{D} \leq D$. Let $\boldsymbol{\delta}$ and $\tilde{\boldsymbol{\delta}}$ be sequences of inter-departure times from $LB(D, \infty)$ and $LB(\tilde{D}, \infty)$, respectively. Paralleling the monotonicity result in M, we have the following result.

Theorem 2.3 *(Monotonicity in D) Assuming $\boldsymbol{\delta}$ and $\tilde{\boldsymbol{\delta}}$ to be convexly stable and $\lambda D < 1$, we have*

$$\delta \leq_{cx} \tilde{\delta}, \tag{2.9}$$

where δ and $\tilde{\delta}$ are the asymptotic versions of $\boldsymbol{\delta}$ and $\tilde{\boldsymbol{\delta}}$, respectively, and therefore, $c^2(\delta) \leq c^2(\tilde{\delta})$.

So in the case where no token will be lost, the slower the token generation rate is, the less bursty the output traffic from the LB will be.

The assumption $M = \infty$ is really not necessary. In fact, we only need to require that the token pool size of the faster LB be strictly larger than that of the slower one. This is shown in the next subsection.

2.5.3 Joint Monotonicity in M and D

Consider two LBs $LB(D, M)$ and $LB(\bar{D}, \bar{M})$ with $0 \leq \bar{D} \leq D$ and $1 \leq M < \bar{M}$, and let $\boldsymbol{\delta}$ and $\bar{\boldsymbol{\delta}}$ be the sequences of inter-departure times from $LB(D, M)$ and

$LB(\bar{D}, \bar{M})$, respectively.

Theorem 2.4 *(Monotonicity in D and M) Assuming $\boldsymbol{\delta}$ and $\bar{\boldsymbol{\delta}}$ to be convexly stable and $\lambda D < 1$, we have*

$$\delta \leq_{cx} \bar{\delta}, \tag{2.10}$$

where δ and $\bar{\delta}$ are the asymptotic versions of $\boldsymbol{\delta}$ and $\bar{\boldsymbol{\delta}}$, respectively, and therefore, $c^2(\delta) \leq c^2(\bar{\delta})$.

2.5.4 Burst Reduction Property

An important property of the LB is its ability of reducing the burstiness of the traffic. This can be derived directly from Theorem 2.4. Take $\bar{D} = 0$ in Theorem 2.4 so that no cell will be delayed. As a consequence $\bar{\boldsymbol{\delta}}$ is just the sequence of inter-arrival times $\boldsymbol{\alpha}$. Since in this case, the token pool does not play any role, we may choose any $\bar{M} > M$, and apply Theorem 2.4 to conclude

Theorem 2.5 *Assume $\boldsymbol{\alpha}$ to be convexly stable and $\lambda D < 1$. If $\boldsymbol{\delta}$ is also convexly stable, then we have*

$$\delta \leq_{cx} \alpha, \tag{2.11}$$

where α and δ are the asymptotic versions of $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, respectively, and therefore, $c^2(\delta) \leq c^2(\alpha)$.

2.5.5 Monotonicity in the Initial Condition

Another factor which affects the token availability and consequently the burstiness of the output traffic is the initial state of the LB. We now turn to the monotonicity with respect to the initial state of the LB.

Consider two LBs differing only in their initial states. Let $\boldsymbol{\delta}$ and $\check{\boldsymbol{\delta}}$ be the inter-departure times from $LB_s(D, M)$ and $LB_{\check{s}}(D, M)$, respectively, with $-M \leq \check{s} < s$. We have the following results.

Theorem 2.6 *(Monotonicity in the initial state) Assuming* $\boldsymbol{\delta}$ *and* $\check{\boldsymbol{\delta}}$ *to be convexly stable and* $\lambda D < 1$, *we have*

$$\delta \leq_{cx} \check{\delta}, \tag{2.12}$$

where δ *and* $\check{\delta}$ *are the asymptotic versions of* $\boldsymbol{\delta}$ *and* $\check{\boldsymbol{\delta}}$, *respectively, and therefore,* $c^2(\delta) \leq c^2(\check{\delta})$.

From Theorem 2.6 we conclude that the fewer tokens initially in the token pool (resp. the more cells initially waiting in the input buffer), the less bursty the output traffic of the LB. It may be asked whether Theorem 2.6 is really meaningful in that if both $\boldsymbol{\delta}$ and $\check{\boldsymbol{\delta}}$ are convexly stable, should we not get equality between their asymptotic versions. To answer this question, we consider the following example. Let $\boldsymbol{\alpha} = (0, 0.5, 1.5, 0.5, 1.5, \cdots)$ be a deterministic sequence. The output sequence from $LB_0(1,1)$ is then $\boldsymbol{\delta}_0 = (0, 1, 1, 1, 1, \cdots)$, while the output sequence from $LB_{-1}(1,1)$ is $\boldsymbol{\delta}_{-1} = (0, 0.5, 1.5, 0.5, 1.5, \cdots)$. It is plain that $\boldsymbol{\delta}_0$ has a stationary version which coincides with its asymptotic version and which is equal to 1 with probability 1, while $\boldsymbol{\delta}_{-1}$ does not have any stationary version but its asymptotic version is equal to 0.5 or 1.5 with probability 0.5. Therefore, if both sequences do not have stationary versions, strict inequality may hold in Theorem 2.6.

2.6 The Burst Structure

It will soon become clear that the results presented in Section 2.5 are direct consequences of the sample path behavior of the LB. To this end, we first explore the burst structure of the sample paths of the LB. For $n = 1, 2, \cdots$, the arrival epoch a_n (resp. departure epoch d_n) of the n^{th} cell can be expressed as

$$a_n = \sum_{i=1}^{n} \alpha_i \qquad \left(\text{ resp. } \; d_n = \sum_{i=1}^{n} \delta_i \right), \qquad n = 1, 2, \cdots. \tag{2.13}$$

With $l_1^{(1)} = 1$, we recursively define

$$l_k^{(2)} = \min\{i > l_k^{(1)} : S(a_i) > 0\}, \qquad k = 1, 2, \cdots \tag{2.14}$$

and

$$l_{k+1}^{(1)} = \min\{i > l_k^{(2)} : S(a_i) \leq 0\}, \qquad k = 1, 2, \cdots. \tag{2.15}$$

For $n = 1, 2, \cdots$, we say that the n^{th} cell is of the first kind if $l_k^{(1)} \leq n < l_k^{(2)}$, for some $k = 1, 2, \cdots$. A cell which is not of the first kind is said to be of the second kind. Cells of the first kind do not need to wait in the input buffer as they pass through the LB without any delay, so that their arrival epochs and departure epochs are the same, i.e.,

$$a_n = d_n, \qquad \text{if the } n^{th} \text{ cell is of the first kind.} \tag{2.16}$$

Cells of the second kind, however, have to wait for tokens to be generated in order to leave the LB, in which case we have

$$a_n < d_n, \qquad \text{if the } n^{th} \text{ cell is of the second kind.} \tag{2.17}$$

With the help of (2.13) – (2.15), we now define the burst cycles of the LB.

Definition 2.2 *We call*

$$B_k = [a_{l_k^{(1)}}, a_{l_{k+1}^{(1)}}), \qquad k = 1, 2, \cdots \tag{2.18}$$

the k^{th} burst cycle of the LB.

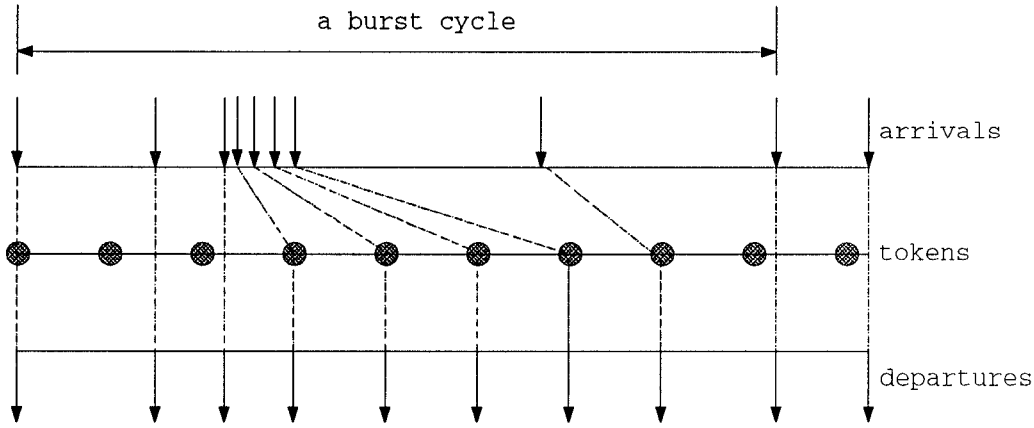


Figure 2.3: The burst cycles

For each $k = 1, 2, \cdots$ we see that $l_k^{(1)}$ is the tag of the first cell (of the first kind) in the k^{th} burst cycle, and $l_k^{(2)}$ is the tag of the first cell of the second kind in the k^{th} burst cycle. A typical realization of the burst cycles is shown in Fig. 2.3. Let $Z^{(1)}$ denote the collection of tags of all cells of the first kind, i.e.,

$$Z^{(1)} = \{n = 1, 2, \cdots : l_k^{(1)} \leq n < l_k^{(2)} \text{ for some } k = 1, 2, \cdots\}. \tag{2.19}$$

We see that the burst cycles of the LB are uniquely determined by $Z^{(1)}$. It is easily seen that each burst cycle begins with a group of cells of the first kind and is followed by a group of cells of the second kind. This pattern repeats itself upon forming successive burst cycles.

The concept of burst cycles reveals the simple structure of the sample paths of the LB. From Fig. 2.3 we see that the inter-departure time between two consecutive arrivals of the second kind is constant and equal to the token generation period D. The inter-departure times between consecutive arrivals of the first kind are complicated. However, as we shall see later, this contributes no difficulty to the analysis. It is this simple structure that makes our transient analysis possible.

We now conclude this section by proving two technical lemmas. Let LB and $\widetilde{LB}$ be two LBs with the same input sequence $\boldsymbol{\alpha}$ and output sequences $\boldsymbol{\delta}$ and $\tilde{\boldsymbol{\delta}}$, respectively.

Lemma 2.1 *If the departure epochs of LB and* $\widetilde{LB}$ *satisfy*

$$d_n \geq \tilde{d}_n, \qquad n = 1, 2, \cdots, \tag{2.20}$$

then we have

$$Z^{(1)} \subset \tilde{Z}^{(1)}. \tag{2.21}$$

Proof. Suppose (2.21) not to hold. Then there is an arrival epoch, say a_n, with n in $Z^{(1)}$ but not in $\tilde{Z}^{(1)}$, in which case $d_n = a_n < \tilde{d}_n$ by (2.16) and (2.17), thus contradicting (2.20). ■

Sometimes the condition (2.20) can be more easily verified by a condition on the states of the LBs as it is now shown in the following lemma.

Lemma 2.2 *If the states of LB and* $\widetilde{LB}$ *satisfy*

$$0 \geq S(0) \geq \tilde{S}(0) \tag{2.22}$$

and

$$S(t) \geq \tilde{S}(t), \qquad t > 0, \tag{2.23}$$

then (2.20) holds.

Proof. Conditions (2.22) and (2.23) readily imply

$$S(d_n) \geq \tilde{S}(d_n), \qquad n = 1, 2, \cdots. \tag{2.24}$$

Since the arrival processes to the two LBs are the same, (2.24) in turn implies that fewer arrivals are waiting in the input buffer of $\widetilde{LB}$ at these epochs. Since the input buffer operates according to the FCFS discipline, (2.24) implies that the n^{th} arrival in $\widetilde{LB}$ has also left by time d_n, $n = 1, 2, \cdots$, and (2.20) therefore holds. ■

The condition $S(0) \leq 0$ in (2.22) is necessary, as can be seen from the following counter-example. Suppose $S(0) = 5$ and $\tilde{S}(0) = 0$. So (2.22) holds without $S(0) \leq 0$.

Suppose there are K arrivals during $[0, d_n)$ with $S(d_n) = 1$ and $\tilde{S}(d_n) = 0$. A closer look reveals that there were $K + 4$ departures from LB by time d_n, and only K departures from $\tilde{LB}$ by the same time. Therefore, $\tilde{d}_{K+4} > d_{K+4}$ which contradicts (2.20).

2.7 Sample Paths Comparison Via Majorization

For the notion of majorization and its properties, the reader is referred to Appendix A and [46].

Now we state and prove a general result which gives a sufficient condition for the comparability of two LBs in some suitable sense. To simplify the notation, we use the following convention. For any sequence $\boldsymbol{\xi} = \{\xi_n, n = 1, 2, \cdots\}$ of $\mathbb{R}$-valued rvs, we define the $\mathbb{R}^{m-n+1}$-valued rvs $\boldsymbol{\xi}_{m,n}$, $m \leq n$, $n = 1, 2, \cdots$ by

$$\boldsymbol{\xi}_{m,n} = (\xi_m, \cdots, \xi_n). \tag{2.25}$$

Theorem 2.7 *Let $LB_s(D, M)$ and $LB_{\tilde{s}}(\tilde{D}, \tilde{M})$, be any two LBs with the same input sequence $\boldsymbol{\alpha}$ and output sequences $\boldsymbol{\delta}$ and $\tilde{\boldsymbol{\delta}}$, respectively. If their departure epochs satisfy*

$$d_n \geq \tilde{d}_n, \qquad n = 1, 2, \cdots, \tag{2.26}$$

then we have

$$\boldsymbol{\delta}_{1,n} \prec^w \tilde{\boldsymbol{\delta}}_{1,n}, \qquad n = 1, 2, \cdots \tag{2.27}$$

and

$$\boldsymbol{\delta}_{1,n} \prec \tilde{\boldsymbol{\delta}}_{1,n}, \qquad n \in Z^{(1)}. \tag{2.28}$$

Proof. By Lemma 2.1, (2.26) implies

$$Z^{(1)} \subset \tilde{Z}^{(1)}. \tag{2.29}$$

Consider the burst cycles determined by $Z^{(1)}$. For any burst cycle B_k, $k = 1, 2, \cdots$, we have from (2.16) that

$$d_n = \tilde{d}_n = a_n, \qquad l_k^{(1)} \leq n \leq l_k^{(2)} \tag{2.30}$$

and therefore,

$$\boldsymbol{\delta}_{l_k^{(1)},n} = \tilde{\boldsymbol{\delta}}_{l_k^{(1)},n}, \qquad l_k^{(1)} \leq n \leq l_k^{(2)}. \tag{2.31}$$

If for all $k = 1, 2, \cdots$, we can prove

$$\boldsymbol{\delta}_{l_k^{(2)},n} \prec^w \tilde{\boldsymbol{\delta}}_{l_k^{(2)},n}, \qquad l_k^{(2)} \le n < l_{k+1}^{(1)} \tag{2.32}$$

and

$$\boldsymbol{\delta}_{l_k^{(2)},l_{k+1}^{(1)}} \prec \tilde{\boldsymbol{\delta}}_{l_k^{(2)},l_{k+1}^{(1)}}, \tag{2.33}$$

then Theorem 2.7 follows from the closure property of majorization under concatenation (Lemma A.2). Upon examming Fig. 2.3, we conclude that without loss of generality, it suffices to prove (2.32) and (2.33) in the first burst cycle, as is done in the following lemma. ■

Lemma 2.3 *Under the assumptions of Theorem 2.7, we have*

$$\boldsymbol{\delta}_{l_1^{(2)},n} \prec^w \tilde{\boldsymbol{\delta}}_{l_1^{(2)},n}, \qquad l_1^{(2)} \le n < l_2^{(1)} \tag{2.34}$$

and

$$\boldsymbol{\delta}_{l_1^{(2)},l_2^{(1)}} \prec \tilde{\boldsymbol{\delta}}_{l_1^{(2)},l_2^{(1)}}. \tag{2.35}$$

Proof. From Fig. 2.3 we observe that

$$\tilde{\delta}_{l_1^{(2)}} < \delta_{l_1^{(2)}} < D \le \delta_{l_2^{(1)}} < \tilde{\delta}_{l_2^{(1)}}, \tag{2.36}$$

$$\delta_i = D, \qquad l_1^{(2)} < i < l_2^{(1)}, \tag{2.37}$$

and

$$\sum_{i=l_1^{(2)}}^{l_2^{(1)}} \delta_i = \sum_{i=l_1^{(2)}}^{l_2^{(1)}} \tilde{\delta}_i. \tag{2.38}$$

Condition (2.26) also implies

$$\sum_{i=l_1^{(2)}}^{k} \tilde{\delta}_i \le \sum_{i=l_1^{(2)}}^{k} \delta_i, \qquad l_1^{(2)} \le k < l_2^{(1)}. \tag{2.39}$$

Now (2.35) follows from Lemma 2.4 below which mimics the structure of the burst cycles of the two LBs.

To prove (2.34), fix n with $l_1^{(2)} \leq n < l_2^{(1)}$ and define an auxiliary sequence of rvs $\boldsymbol{\zeta}$ by

$$\zeta_i = \begin{cases} \tilde{\delta}_i, & i = 1, 2, \cdots, \text{ and } i \neq n; \\ \tilde{\delta}_n + d_n - a_n, & i = n. \end{cases} \tag{2.40}$$

We can interprate $\boldsymbol{\zeta}$ as a modification of $\tilde{\boldsymbol{\delta}}$ which coincides with $\tilde{\boldsymbol{\delta}}$ except that the n^{th} departure is forced to delay an amount of time $d_n - a_n$, so that the second burst cycle begins earlier at the n^{th} departure epoch. Using (2.35), we get

$$\boldsymbol{\delta}_{l_1^{(2)},n} \prec \boldsymbol{\zeta}_{l_1^{(2)},n}. \tag{2.41}$$

It is also clear that

$$\boldsymbol{\zeta}_{l_1^{(2)},n} \geq \tilde{\boldsymbol{\delta}}_{l_1^{(2)},n} \tag{2.42}$$

and (2.34) now follows from Lemma A.1. ■

Lemma 2.4 *If vectors* $\boldsymbol{x} = (x_1, \ldots, x_n)$ *and* $\boldsymbol{y} = (y_1, \ldots, y_n)$ *in* $\mathbb{R}_+^n$ *satisfy the conditions*

$$\text{(a)} \qquad x_1 \leq y_1 \leq D, \tag{2.43}$$

$$\text{(b)} \qquad y_2 = y_3 = \cdots = y_{n-1} = D, \tag{2.44}$$

$$\text{(c)} \qquad D \leq y_n \leq x_n, \tag{2.45}$$

$$\text{(d)} \qquad \sum_{i=1}^{k} x_i \leq \sum_{i=1}^{k} y_i, \qquad k = 1, \cdots, n-1, \tag{2.46}$$

$$\text{(e)} \qquad \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \tag{2.47}$$

then the following ordering

$$\boldsymbol{y} \prec \boldsymbol{x} \tag{2.48}$$

is true.

Proof. From the definition of majorization, by taking into account (e), we only need to show that

$$\sum_{i=k}^{n} y_{[i]} \geq \sum_{i=k}^{n} x_{[i]}, \qquad k = 2, \cdots, n. \tag{2.49}$$

From (a)–(c), we have

$$y_1 \leq y_2 \leq \cdots \leq y_{n-1} \leq y_n, \tag{2.50}$$

so that

$$\sum_{i=k}^{n} y_{[i]} = \sum_{i=1}^{n-k+1} y_i, \qquad k = 1, \cdots, n. \tag{2.51}$$

From (d), we conclude

$$\sum_{i=1}^{n-k+1} y_i \geq \sum_{i=1}^{n-k+1} x_i \geq \sum_{i=k}^{n} x_{[i]} \qquad k = 2, \cdots, n \tag{2.52}$$

where the last inequality results from (A.1), and (2.48) follows. ■

Combining Theorem 2.7 and Lemma 2.2, we have the following corollary.

Corollary 2.1 *Let $LB_s(D, M)$ and $LB_{\tilde{s}}(\tilde{D}, \tilde{M})$, be any two LBs with the same input sequence $\boldsymbol{\alpha}$ and output sequences $\boldsymbol{\delta}$ and $\tilde{\boldsymbol{\delta}}$, respectively. If their states satisfy*

$$0 \geq s \geq \tilde{s} \tag{2.53}$$

and

$$S(t) \geq \tilde{S}(t), \qquad t > 0, \tag{2.54}$$

then we have

$$\boldsymbol{\delta}_{1,n} \prec^w \tilde{\boldsymbol{\delta}}_{1,n}, \qquad n = 1, 2, \cdots \tag{2.55}$$

and

$$\boldsymbol{\delta}_{1,n} \prec \tilde{\boldsymbol{\delta}}_{1,n}, \qquad n \in Z^{(1)}. \tag{2.56}$$

2.8 Proofs of the Main Results

In this section, we first establish relations between the transient result (Theorem 2.7 or Corollary 2.1) and our main results (the steady-state results of Section 2.4). Then we verify that the conditions in Corollary 2.1 are satisfied in each of the Theorems 2.2 – 2.6. The notion of convex stability introduced in Section 2.4 serves as a bridge between the transient result and the steady-state results. The following lemma formalizes this relationship.

Lemma 2.5 *Let $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ be two sequences of $\mathbb{R}_+$-valued rvs, and suppose that there exists a sequence of (possibly random) integers $\{n_k : k = 1, 2, \cdots\}$ such that $\lim_{k\to\infty} n_k = \infty$ and*

$$\boldsymbol{\zeta}_{1,n_k} \prec \boldsymbol{\eta}_{1,n_k}, \qquad k = 1, 2, \cdots. \tag{2.57}$$

If $\boldsymbol{\zeta}$ and $\boldsymbol{\eta}$ are both convexly stable, then their asymptotic versions ζ and η are ordered as

$$\zeta \leq_{cx} \eta, \tag{2.58}$$

and therefore, $c^2(\zeta) \leq c^2(\eta)$.

Proof. From Lemma A.3, (2.57) implies

$$\frac{1}{n_k}\sum_{i=1}^{n_k}\phi(\zeta_i) \leq \frac{1}{n_k}\sum_{i=1}^{n_k}\phi(\eta_i), \qquad k = 1, 2, \cdots \tag{2.59}$$

for any convex function $\phi : \mathbb{R}_+ \to \mathbb{R}$. Taking limit as $k \to \infty$, the convex stability assumption implies

$$\mathbb{E}[\phi(\zeta)] \leq \mathbb{E}[\phi(\eta)], \tag{2.60}$$

provided the expectations in (2.60) exist, whence (2.58) follows from the definition of convex ordering. ■

Now we begin to prove Theorems 2.2 – 2.4, and Theorem 2.6. The stability condition $\lambda D < 1$ ensures that all LBs in the theorems are stable for the input rate

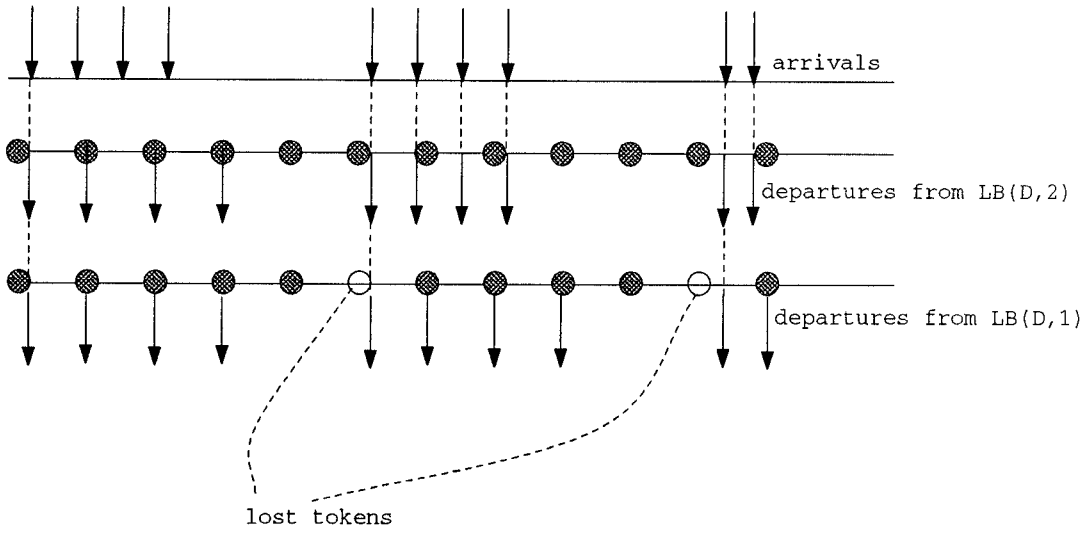


Figure 2.4: Output traffic from LBs of different token pool sizes

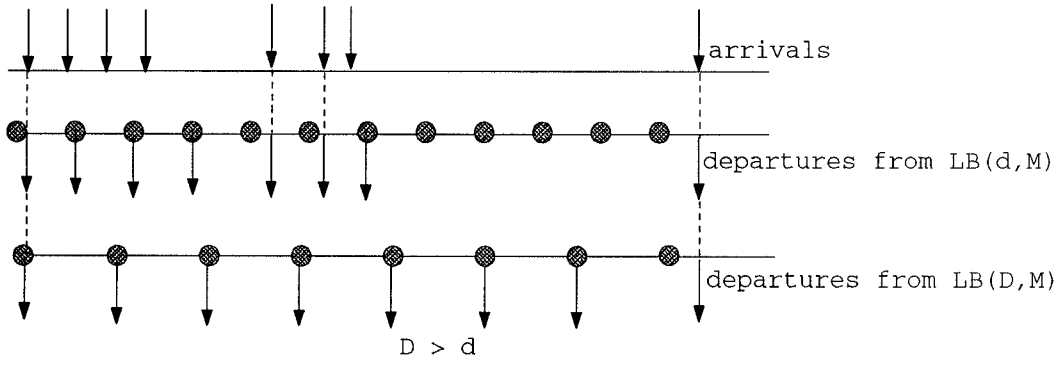


Figure 2.5: Output traffic from LBs of different token generation periods

λ, which in turn guarantees that the set $Z^{(1)}$ of indices of arrivals of the first kind has infinite cardinality. In the proofs of Theorems 2.2 – 2.4, we follow the convention made in Section 2.3 for simplicity. For the more general case, one observes that the only difficulty arises in the first burst cycle. However, by the stability condition, the first burst cycle is necessarily finite, and therefore contributes none to the limit. From Corollary 2.1 and Lemma 2.5, we only need to show that the states of the LBs in each theorem are properly ordered. As illustrations, the sample paths of the LBs with different token generation periods or LBs with different token pool size are shown in Figures 2.4 – 2.5.

Proof of Theorem 2.2. We shall show that

$$S(t) \geq \hat{S}(t), \qquad t \geq 0. \tag{2.61}$$

Let $\{t_n,\ n = 0, 1, \cdots\}$ be the sequence of arrival and token generation epochs (the first arrival and the first token generation epochs are by convention zero, i.e., $t_0 = 0$). We observe that $\{S(t),\ t \geq 0\}$ and $\{\hat{S}(t),\ t \geq 0\}$ are (by convention right-continuous) piecewise constant with jumps at $t_i,\ i = 0, 1, \cdots$, so that (2.61) need to be verified only at these epochs. We proceed by induction. Clearly we have $S(t_0) = \hat{S}(t_0) = 0$. Assume that $S(t_i) \geq \hat{S}(t_i)$ for some $i = 0, 1, \cdots$. It is easy to see that

$$S(t_{i+1}) = \max\{S(t_i) + \Delta, -M\} \tag{2.62}$$

and

$$\hat{S}(t_{i+1}) = \max\{\hat{S}(t_i) + \Delta, -\hat{M}\} \tag{2.63}$$

where

$$\Delta = \begin{cases} 1 & \text{if } t_{i+1} \text{ is an arrival epoch,} \\ -1 & \text{if } t_{i+1} \text{ is a token generation epoch.} \end{cases} \tag{2.64}$$

Since $-M \geq -\hat{M}$, we have $S(t_{i+1}) \geq \hat{S}(t_{i+1})$ by the induction hypothesis. ■

Proof of Theorem 2.3. As in the proof of Theorem 2.2, we need to show that

$$S(t) \geq \tilde{S}(t), \qquad t \geq 0. \tag{2.65}$$

Since we have assumed $M = \infty$, no token can be lost in either systems. The numbers of tokens generated up to time $t > 0$ in both systems are therefore given by $\left\lfloor \frac{t}{D} \right\rfloor$ and $\left\lfloor \frac{t}{\tilde{D}} \right\rfloor$, respectively, where$\lfloor x \rfloor$ is the largest integer less than or equal to x. If $N(t),\ t \geq 0$, denotes the number of arrivals in $(0, t]$, then the states of the two LBs are given by

$$S(t) = N(t) - \left\lfloor \frac{t}{D} \right\rfloor, \qquad t \geq 0 \tag{2.66}$$

and

$$\tilde{S}(t) = N(t) - \left\lfloor \frac{t}{\tilde{D}} \right\rfloor, \qquad t \geq 0, \tag{2.67}$$

respectively. Since $\tilde{D} \leq D$, (2.65) follows from (2.66) and (2.67). ■

Proof of Theorem 2.4. As in the proof of Theorem 2.2, we need to show

$$S(t) \geq \bar{S}(t), \qquad t \geq 0. \tag{2.68}$$

Let $\bar{L}(t)$ denote the number of tokens discarded during the interval $(0, t]$ due to a full token pool at $LB(\bar{D}, \bar{M})$, and define

$$\tau_k = \inf\{t \geq 0 : \bar{L}(t) = k + 1\}, \qquad k = 0, 1, \cdots. \tag{2.69}$$

We observe that

$$\bar{S}(\tau_k) = -\bar{M} \text{ and } S(\tau_k) \geq -M, \qquad k = 0, 1, \cdots, \tag{2.70}$$

so that

$$S(\tau_k) - \bar{S}(\tau_k) \geq -M + \bar{M} \geq 1, \qquad k = 0, 1, \cdots. \tag{2.71}$$

For t in (τ_k, τ_{k+1}), $k = 0, 1, \cdots$, there are exactly $\left\lfloor \frac{t - \tau_k}{\bar{D}} \right\rfloor$ tokens generated by $LB(\bar{D}, \bar{M})$ and at most $\left\lfloor \frac{t - \tau_k}{D} \right\rfloor + 1$ tokens generated by $LB(D, M)$ during the interval $(\tau_k, t]$. Furthermore, no token is discarded from $LB(\bar{D}, \bar{M})$ during the interval (τ_k, τ_{k+1}), $k = 0, 1, \cdots$. If $N(s, t)$ denotes the number of arrivals during the interval $(s, t]$, then we have

$$\begin{cases} S(t) \geq N(0, t) - \left\lfloor \frac{t}{D} \right\rfloor, \\ \\ \bar{S}(t) = N(0, t) - \left\lfloor \frac{t}{\bar{D}} \right\rfloor, \end{cases} \qquad t \in [0, \tau_0) \tag{2.72}$$

and

$$\begin{cases} S(t) \geq S(\tau_k) + N(\tau_k, t) - \left(\left\lfloor \frac{t - \tau_k}{D} \right\rfloor + 1 \right), \\ \\ \bar{S}(t) = \bar{S}(\tau_k) + N(\tau_k, t) - \left\lfloor \frac{t - \tau_k}{\bar{D}} \right\rfloor, \end{cases} \qquad t \in (\tau_k, \tau_{k+1}), \quad k = 0, 1, \cdots. \tag{2.73}$$

As a result, we get

$$S(t) - \bar{S}(t) \geq \left\lfloor \frac{t}{\bar{D}} \right\rfloor - \left\lfloor \frac{t}{D} \right\rfloor \geq 0, \qquad t \in [0, \tau_0) \tag{2.74}$$

and

$$S(t)-\bar{S}(t) \geq S(\tau_k)-\bar{S}(\tau_k)+\left\lfloor \frac{t-\tau_k}{\bar{D}} \right\rfloor - \left\lfloor \frac{t-\tau_k}{D} \right\rfloor -1 \geq 0, \qquad t \in (\tau_k, \tau_{k+1}). \quad (2.75)$$

Thus (2.68) holds for all $t \geq 0$. ■

Proof of Theorem 2.6. In order to use Corollary 2.1, we first assume that $0 \geq s \geq \tilde{s}$, and define $\{t_n,\ n = 0, 1, \cdots\}$ as in the proof of Theorem 2.2. We want to show that

$$S(t_i) \geq \check{S}(t_i), \qquad i = 0, 1, \cdots. \quad (2.76)$$

By assumption, (2.76) holds for $i = 0$. With Δ defined by (2.64), we have

$$S(t_{i+1}) = \max\{S(t_i) + \Delta, -M\}, \qquad i = 1, 2, \cdots \quad (2.77)$$

and

$$\check{S}(t_{i+1}) = \max\{\check{S}(t_i) + \Delta, -M\}, \qquad i = 1, 2, \cdots, \quad (2.78)$$

so that (2.76) follows by induction.

In general, if $s > 0$, then by the stability assumption, there exists $T = kD$ for some $k < \infty$ such that $S(T) = 0$ for the first time. Since the arrivals up to time T are all delayed at LB, while the same arrivals may or may not be delayed at $\check{LB}$, we have $0 = S(T) \geq \check{S}(T)$. Taking T as the new time origin, and since the finite time interval $[0, T)$ has no contribution to the limits, we are done. ■

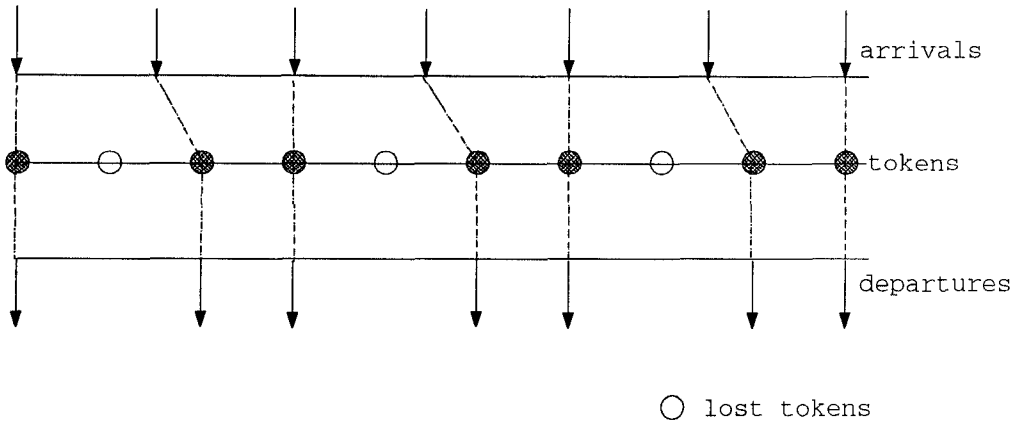


Figure 2.6: LB without token buffering increases burstiness

2.9 Comments and Extensions

2.9.1 No Token Buffering

The results in this chapter cannot be applied to LBs without buffering for tokens. In the case of no token buffering, the burst cycles become infinite so the transient analysis may fail. Fig. 2.6 illustrates a case where the output traffic is more bursty than the input traffic for a LB without token buffering.

2.9.2 Monotonicity in D

The monotonicity in D is not true in general. The following is a counterexample for two LBs with token pool size one and different token generation rates ($D_1 < D_2$ and $M_1 = M_2 = 1$). We see that the output traffic from the LB with the smaller token generation rate is in fact more bursty.

2.9.3 The Unstable Case

If the stability condition (2.7) does not hold, it is readily seen that the input queue will eventually build up, and the inter-departure times will become deterministic of length D; a formal proof of this fact can be found in [2]. In that case, the departure process will be the least bursty, but the cell delay at the LB will become infinite which is not acceptable in practice.

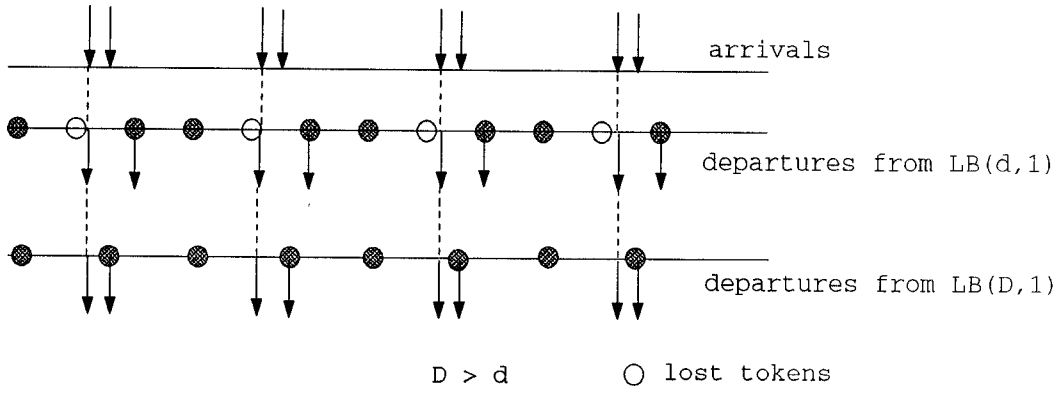


Figure 2.7: Two LBs with the same token pool size and different token generation rates

In the unstable case, we may no longer be able to take limits along the subsequence $Z^{(1)}$ as we did in the proof of Theorem 2.7, since $Z^{(1)}$ may contain only finitely many elements. So without the stability assumption Theorems 2.2–2.6 may not hold. Weaker results, however, can be obtained by taking advantage of the notion of weak majorization. From the first part of Theorem 2.7 and Lemma A.4, we conclude to the following result.

Theorem 2.8 *Without the stability assumption of (2.7), Theorems 2.2–2.6 still hold by replacing* $\leq_{cx}$ *with* $\leq_{dcx}$.

It is worth pointing out that $\delta \leq_{dcx} \alpha$ does not imply $c^2(\delta) \leq c^2(\alpha)$.

CHAPTER 3

BOUNDS, APPROXIMATIONS, AND DESIGN ISSUES

In Chapter 2, we studied the burst reduction properties of the LB. In order to make the output traffic from the LB as smooth as possible, it is desirable to choose the token generation period D as large as possible and the token pool size M as small as possible. However, the side-effect of doing so is to impose extra cell delay at the LB which must also be taken into consideration. In this chapter, we shall establish bounds and approximations for the cell delay at the LB and for some other parameters of interest, e.g., the mean and peak cell rates of the output traffic. Based on these results, we shall discuss the design issues of the LB.

We shall derive the bounds and approximations in Section 3.2. The bounds for the mean and peak cell rates are obtained trivially. Our basic approach to approximate the cell delay is to first bound it by the waiting time in a suitable G/D/1 queue, and then to derive approximations of the waiting time of the G/D/1 queue. To justify the approximation, we first show in the next section that in heavy traffic the LB and the G/D/1 queue become the same in that they have the same heavy traffic diffusion limits.

3.1 Heavy Traffic Diffusion Limits

In this section, we shall prove some heavy traffic limit theorems for the LB. These results are useful in justifying the approximation result for the cell delay we shall derive in Section 3.2; they are also of independent interest.

We observe that whenever the input buffer of a LB is non-empty, the LB behaves just like a G/D/1 queue with service time equal to the token generation period. Therefore, we expect that the two systems will have the same heavy traffic diffusion

limits. This will be made rigorous in the following subsections.

References [15], [43], and [58] provide good surveys of heavy traffic limit theorems. In Appendix C we collect some of the important results needed in our proof, and in Appendix D, we summarize some heavy traffic results for GI/GI/1 queues.

We first present a lemma which serves as a bridge between the LB and the corresponding G/D/1 queue. Let $\{N^G(t),\ t \geq 0\}$ be the queue size process of a G/D/1 queue with service time D; and let $\{N^{LB}(t),\ t \geq 0\}$ the queue size process of a LB with token pool size M and token generation period D. Comparing the definition of $\{S(t),\ t \geq 0\}$ in Section 2.4, we see that

$$N^{LB}(t) = (S(t))^+, \qquad t \geq 0. \tag{3.1}$$

Lemma 3.1 *If the LB and the G/D/1 queue have the same input sequence, then*

$$N^G(0) = N^{LB}(0) \tag{3.2}$$

implies

$$0 \leq N^G(t) - N^{LB}(t) \leq M + 1, \qquad t \geq 0. \tag{3.3}$$

Proof. First we show that

$$N^G(t) - N^{LB}(t) \geq 0, \qquad t \geq 0. \tag{3.4}$$

Consider the LB as a G/G/1 queue with service being the generation of a token for each arrival. Since the G/G/1 queue can provide service even before the arrival of the customer (generate tokens and store them in the token pool), and it has the same service time as the G/D/1 queue, the real service rate in the G/G/1 queue is faster than the service rate of the G/D/1 queue. Therefore the G/G/1 queue (i.e., the LB) has a shorter queue length than at the G/D/1 queue. Next we use induction to show that

$$N^G(t) - S(t) \leq M + 1, \qquad t \geq 0. \tag{3.5}$$

By assumption (3.2), we have

$$N^G(0) - S(0) \leq M, \tag{3.6}$$

and therefore, (3.5) holds for $t = 0$. Let $\{s_i, \ i = 0, 1, \cdots\}$ be the token generation epochs (i.e., $s_i = iD$, $i = 0, 1, \cdots$), we now show that

$$N^G(s_i^-) - S(s_i^-) \leq M, \qquad i = 0, 1, \cdots. \tag{3.7}$$

and

$$N^G(t) - S(t) \leq M + 1, \qquad s_i \leq t < s_{i+1}, \ i = 0, 1, \cdots. \tag{3.8}$$

For $0 \leq s \leq t$, we denote by $N(s,t)$ the number of arrivals in the interval $[s,t)$. Fixing $i = 0, 1, \cdots$, we have

$$S(s_{i+1}^-) = \max\{-M, S(s_i^-) - 1\} + N(s_i, s_{i+1}) \tag{3.9}$$

and

$$N^G(s_{i+1}^-) = \max\{0, N^G(s_i^-) - 1\} + N(s_i, s_{i+1}). \tag{3.10}$$

From (3.9) and (3.10) we get

$$N^G(s_{i+1}^-) - S(s_{i+1}^-) = \max\{0, N^G(s_i^-) - 1\} - \max\{-M, S(s_i^-) - 1\}. \tag{3.11}$$

If $N^G(s_i^-) > 0$, then by the induction hypothesis we have,

$$\begin{aligned} N^G(s_{i+1}^-) - S(s_{i+1}^-) &= N^G(s_i^-) - 1 - \max\{-M, S(s_i^-) - 1\} \\ &\leq N^G(s_i^-) - 1 - (S(s_i^-) - 1) \\ &\leq M. \end{aligned} \tag{3.12}$$

If $N^G(s_i^-) = 0$, then

$$\begin{aligned} N^G(s_{i+1}^-) - S(s_{i+1}^-) &= 0 - \max\{-M, S(s_i^-) - 1\} \\ &\leq 0 - (-M) = M. \end{aligned} \tag{3.13}$$

Since there is at most one departure from both systems during the interval $[s_i, s_{i+1})$, we have

$$N^G(t) - S(t) \leq M + 1, \qquad s_i \leq t < s_{i+1}. \tag{3.14}$$

Combining (3.1) and (3.5), we finally get

$$N^G(t) - N^{LB}(t) \leq N^G(t) - S(t) \leq M + 1, \qquad t \geq 0. \tag{3.15}$$

■

To study the heavy traffic diffusion limits, we consider a sequence of arrival processes $\{\alpha_i^r,\ i = 1, 2, \cdots\}$ with intensity λ_r indexed by $r = 1, 2, \cdots$. We assume that this sequence satisfies Assumption (A1) defined in Appendix D so that as $r \to \infty$, we have

$$Var\,(\alpha_1^r) \to \sigma \tag{3.16}$$

with $0 < \sigma < \infty$.

For $r = 1, 2, \cdots$, we denote by LB_r the LB with token pool size M_r and token generation period D_r. Since we expect the heavy traffic limit theorems for the LB to be the same as those of a G/D/1 queue, we define a sequence of G/D/1 queues G_r with service time D_r, $r = 1, 2, \cdots$, accordingly. Quantities which are related to the LB (resp. the GI/D/1 queue) are superscribed by LB (resp. G).

To get the appropriate heavy traffic results, we impose some additional assumptions.

Assumption (A2): The quantities D_r and λ_r change in r in such a way that as $r \to \infty$,

$$D_r \to D \tag{3.17}$$

with

$$\sqrt{r}(D_r - \lambda_r^{-1}) \to C,\ -\infty < C < 0. \tag{3.18}$$

It is easy to see that Assumption (A2) implies $\lambda_r^{-1} \to D$.

Assumption (A3): The token pool size does not grow too fast, i.e., we assume as $r \to \infty$,

$$\frac{M_r}{\sqrt{r}} \to 0. \tag{3.19}$$

In the following subsections, $f : D[0,1] \to D[0,1]$ is a reflection mapping and $\mathcal{W}$ is the Wiener process on $[0,1]$ as defined in Appendix D.

3.1.1 The Queue Size Process

For $r = 1, 2, \cdots$, we define the $D[0,1]$-valued rv η^{LB_r}, the normalized queue size process of the LB, by

$$\eta^{LB_r}(t) \triangleq \frac{N^{LB_r}(rt)}{\sqrt{r}}, \qquad 0 \le t \le 1. \tag{3.20}$$

Theorem 3.1 *Under Assumptions (A1), (A2), and (A3), we have*

$$\eta^{LB_r} \Longrightarrow f(\frac{\sigma}{D^{3/2}}\mathcal{W} + d) \tag{3.21}$$

where

$$d(t) = Ct, \qquad 0 \le t \le 1 \tag{3.22}$$

with σ and C being defined by (3.16) and (3.18), respectively.

Proof. Let G^r, $r = 1, 2, \cdots$, be a sequence of GI/D/1 queues. It is well known (see Appendix D) that under Assumptions (A1) and (A2), we have

$$\eta^{G_r} \Longrightarrow f(\frac{\sigma}{D^{3/2}}\mathcal{W} + d), \tag{3.23}$$

where

$$\eta^{G_r}(t) \triangleq \frac{N^{G_r}(rt)}{\sqrt{r}}, \qquad 0 \le t \le 1, \quad r = 1, 2, \cdots \tag{3.24}$$

is the normalized queue size process of the GI/D/1 queues. By the converging together theorem (Theorem C.1), it suffices to show that as $r \to \infty$,

$$m(\eta^{LB_r}, \eta^{G_r}) \xrightarrow{P} 0, \tag{3.25}$$

and this follows from the next lemma. ■

Lemma 3.2 *Under Assumption (A3), we have*

$$m(\eta^{G_r}, \eta^{LB_r}) \to 0 \qquad a.s. \tag{3.26}$$

as $r \to \infty$.

Proof. From [36], for $r = 1, 2, \cdots$, we have

$$\begin{aligned} m(\eta^{G_r}, \eta^{LB_r}) &\leq \sup_{0 \leq t \leq 1} \left| \eta^{G_r}(t) - \eta^{LB_r}(t) \right| \\ &= \sup_{0 \leq t \leq 1} \frac{\left| N^{G_r}(rt) - N^{LB_r}(rt) \right|}{\sqrt{r}} \\ &\leq \frac{(M_r + 1)}{\sqrt{r}} \qquad a.s.. \end{aligned} \tag{3.27}$$

The second inequality follows from Lemma 3.1. Under Assumption (A3), (3.26) then follows from (3.27). ■

3.1.2 The Waiting Time Process

The normalized waiting time process is defined as follows.

$$\omega^{LB_r}(t) \triangleq \frac{W^{LB_r}_{\lfloor rt \rfloor}}{\sqrt{r}}, \qquad 0 \leq t \leq 1, \quad r = 1, 2, \cdots \tag{3.28}$$

where for each $r = 1, 2, \cdots$, $W_i^{LB_r}$ is the waiting time of the i^{th} $(i = 1, 2, \cdots)$ arrival at LB_r. Then we have the following theorem.

Theorem 3.2 *Under Assumptions (A1), (A2), and (A3), as $r \to \infty$, we have*

$$\omega^{LB_r} \Longrightarrow f(\sigma \mathcal{W} + d), \tag{3.29}$$

where σ and d are defined by (3.16) and (3.22), respectively.

Proof. Define the corresponding normalized waiting time process for the GI/D/1 queue by

$$\omega^{G_r}(t) \triangleq \frac{W^{G_r}_{\lfloor rt \rfloor}}{\sqrt{r}} \qquad 0 \leq t \leq 1, \quad r = 1, 2, \cdots. \tag{3.30}$$

It is well known (see Appendix D) that under Assumptions (A1) and (A2),

$$\omega^{G_r} \Longrightarrow f(\sigma \mathcal{W} + d). \tag{3.31}$$

By the converging together theorem (Theorem C.1), it suffices to show that

$$m(\omega^{G_r}, \omega^{LB_r}) \xrightarrow{P} 0 \tag{3.32}$$

as $r \to \infty$. This follows from Lemma 3.3. ■

Lemma 3.3 *Under Assumption (A3), if $D_r \to D$ as $r \to \infty$, then we have*

$$m(\omega^{G_r}, \omega^{LB_r}) \to 0 \qquad a.s. \tag{3.33}$$

as $r \to \infty$.

Proof. Observe that if the difference between the queue lengths of the LB and of the GI/D/1 queue is m, then the difference between the two waiting times is at most $(m+1)D_r$. The rest of the proof is very similar to that of Lemma 3.2. ■

3.2 Bounds and Approximations

3.2.1 The Mean Cell Waiting Time

In Section 3.1, we showed that the LB and the corresponding GI/D/1 queue have the same heavy traffic diffusion limits. Therefore, by a well-known result [40], the steady-state mean waiting time $\bar{W}$ of cells in the input buffer is approximately

$$\bar{W} \approx \frac{\lambda\sigma^2}{2(1-\lambda D)} \tag{3.34}$$

where σ^2 is the variance of the inter-arrival time, λ is the arrival rate, and D is the token generation period. This approximation becomes accurate, in fact exact, in heavy traffic.

From Lemma 3.1 we see that the number of cells in the LB is bounded above by the number of cells in the corresponding GI/D/1 queue, so that the waiting time at the LB is bounded above by the waiting time at the GI/D/1 queue. It is also well known [40] that the quantity at the right-hand side of (3.34) is actually an upper bound of the mean waiting time in the GI/D/1 queue. Therefore, we have

$$\bar{W} \leq \frac{\lambda\sigma^2}{2(1-\lambda D)}, \tag{3.35}$$

and the bound is tight in heavy traffic.

However, the bound (3.35) is very loose in light traffic. In fact, when D goes to zero, this bound does not go to zero. So tighter approximations are needed. Consider an M/D/1 queue. It is well known [40] that in this case the mean waiting time is given by

$$\bar{W} = \frac{\lambda D^2}{2(1-\lambda D)}. \tag{3.36}$$

Motivated by the form of (3.35) and (3.36), we propose the following approximation.

$$\bar{W} \approx \frac{\lambda^{1+h}\sigma^2 D^h}{2(1-\lambda D)}, \qquad h \geq 0. \tag{3.37}$$

We see that the right-hand side of (3.37) is an decreasing function of $h \geq 0$. If $h = 0$, then (3.37) reduces to the upper bound (3.34), while since $\lambda^2\sigma^2 = 1$ for Poisson arrivals, if $h = 2$, then the right-hand side of (3.37) is just the mean waiting time of

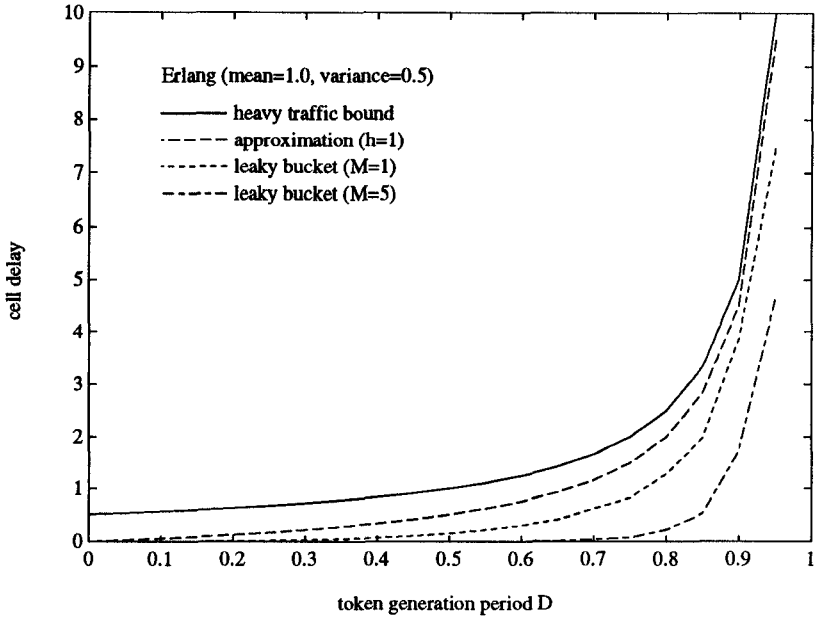


Figure 3.1: Mean cell delay (Erlang distribution)

an M/D/1 queue given by (3.36). So (3.37) provides an reasonable approximation to the mean waiting time, the accuracy of which can be adjusted by proper choice of the parameter h. For a conservative estimation, the value of h should be chosen relatively small ($h \ll 1$, as is in the case of heavy traffic) to make an over-estimate rather than an under-estimate. Figs. 3.1–3.3 show the approximate cell delays for some typical distributions of the inter-arrival times and the corresponding simulation results. If we choose $h = 1$, then we get the following approximation

$$\bar{W} \approx \frac{\lambda^2 \sigma^2 D}{2(1 - \lambda D)}. \tag{3.38}$$

This approximation has a very simple form, and from Figs. 3.1–3.3, we see that it is very good. Therefore, we shall use (3.38) as our approximation of cell delay in the design.

3.2.2 The Mean and Peak Cell Rates

As recommended by the CCITT, the traffic for variable bit rate services can be characterized by the mean cell rate (R_m) and the peak cell rate (R_p) [12]. The mean cell rate is measured over a long period T_l, and the peak cell rate is measured over

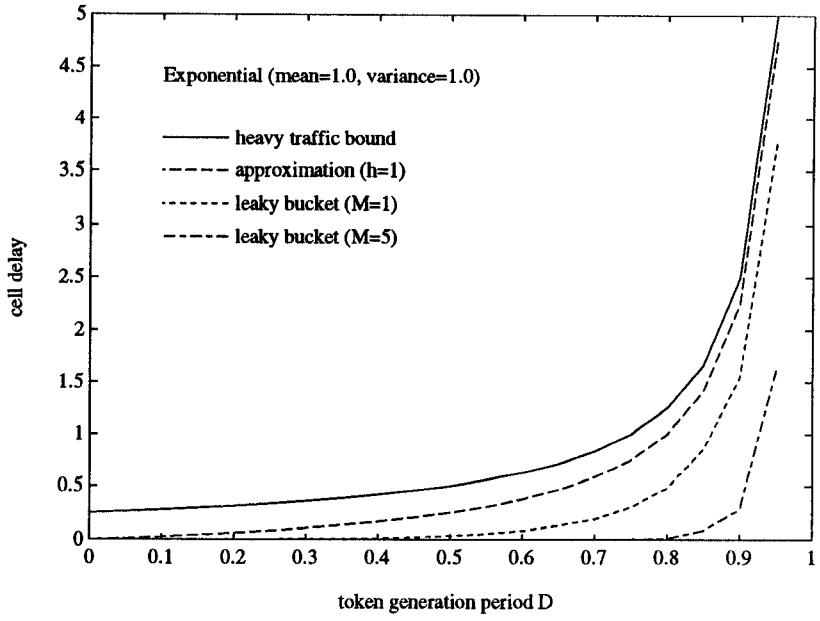


Figure 3.2: Mean cell delay (exponential distribution)

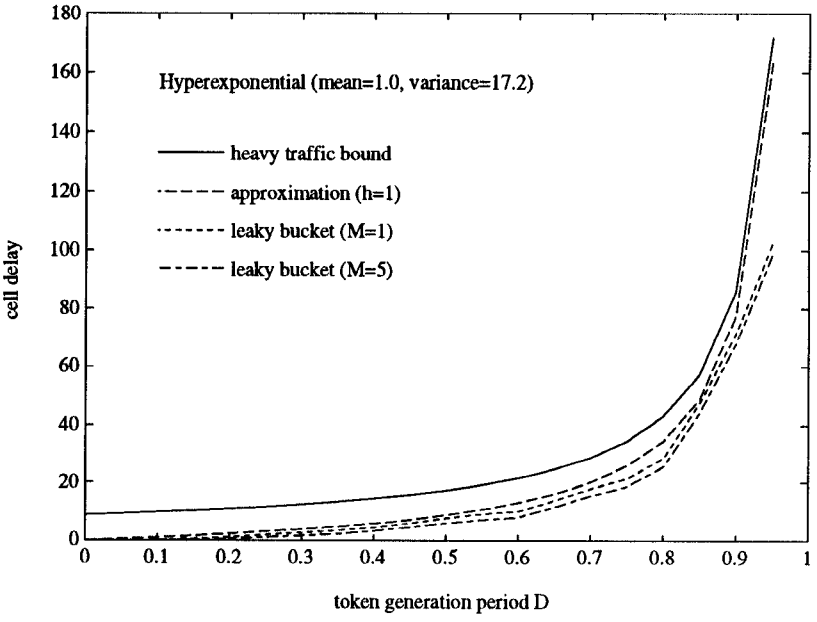


Figure 3.3: Mean cell delay (Hyper-exponential distribution)

a short period T_s. Since for any interval of length T, the maximum number of cells that can leave the LB is bounded by $M + 1 + \frac{T}{D}$, the mean cell rate is bounded by

$$R_m \leq D^{-1} + \frac{M+1}{T_l} \tag{3.39}$$

while the peak cell rate is bounded by

$$R_p \leq D^{-1} + \frac{M+1}{T_s}. \tag{3.40}$$

3.3 Design Issues

3.3.1 Parameter Design

We now discuss some design issues of the LB from an engineering point of view. Our objective is to choose the parameters of the LB (the size of the token pool (M), the token generation period (D), and the size of the input buffer (B)) so as to meet some design criteria. These criteria usually include the mean cell rate (R_m), the peak cell rate (R_p), the cell delay at the LB (D_{cell}), and the probability of cell loss at the LB (P_{loss}). Here we add another criterion, the burstiness of the output traffic (measured by the squared coefficient of variation of the inter-departure times in steady-state, and denoted by $c(D, M, B)$ to emphasize its dependency on the parameters of the LB). Although this last criterion somewhat overlaps with the mean and peak rates, and is difficult to specify, it has some very good monotonicity properties as we have already seen in Chapter 2. By choosing this criterion as our objective function, it is easy to formulate the design problem as an optimization problem. With $0 < r_m < r_p$, $w > 0$, and $0 < p < 1$, this optimization problem (P1) can be stated as follows:

$$
\begin{aligned}
(P1) \qquad & \textit{minimize} && c(D, M, B) && (3.41)\\
& \textit{subject to} && R_m \leq r_m, && (3.42)\\
& && R_p \leq r_p, && (3.43)\\
& && D_{cell} \leq w, && (3.44)\\
& && P_{loss} \leq p, && (3.45)\\
& && (D, M, B) \in \mathbb{R}_+^2 \times \mathbb{N}\,. && (3.46)
\end{aligned}
$$

Let $\Psi \subset \mathbb{R}_+^2 \times \mathbb{N}$ denote the feasible set of (P1), i.e., Ψ consists of all triples (D, M, B) in $\mathbb{R}_+^2 \times \mathbb{N}$ such that (3.42)–(3.45) hold. To solve this constrained problem, we first need to determine the feasible set Ψ. To determine Ψ exactly is very difficult. In the following, we shall derive a heuristic approach to estimate this set and to find a sub-optimal solution. Our method consists of three steps.

Step 1: To approximate the conditions.

We approximate R_m and R_p by their bounds derived in Section 3.2.2, so that (3.42) and (3.43) become

$$D^{-1} + \frac{M+1}{T_l} \leq r_m \tag{3.47}$$

and

$$D^{-1} + \frac{M+1}{T_s} \leq r_p, \tag{3.48}$$

respectively. We next approximate D_{cell} by (3.38) to write (3.44) as

$$\frac{\lambda^2 \sigma^2 D}{2(1-\lambda D)} \leq w, \tag{3.49}$$

or equivalently,

$$D \leq \frac{2w}{\lambda(2w + \lambda\sigma^2)}. \tag{3.50}$$

We denote by $\tilde{\Psi}$ the set of points in $\mathbb{R}_+^2 \times \mathbb{N}$ which satisfy (3.45), (3.47), (3.48), and (3.50). It can be seen that $\tilde{\Psi}$ is an approximation of the feasible set, and it is usually smaller than the feasible set Ψ. So, using $\tilde{\Psi}$ instead of the feasible set, we shall get a sub-optimal solution. We then go on to the estimation of the set $\tilde{\Psi}$.

Step 2: To separate the problem.

The idea is to separate the choices of B from D and M. It is clear that $B = \infty$ satisfies (3.45). In fact, this choice of B is very conservative. The quantities D and M can then be determined from (3.47), (3.48), and (3.50). We denote the set of D and M which satisfies (3.47), (3.48), and (3.50) by Φ (Fig. 3.4).

Step 3: To find the "optimal" value of D, M, and B.

If Φ is non-empty, then we choose, invoking Theorem 2.4, the smallest possible M and the largest possible D in Φ to minimize $c(D, M, \infty)$, in which case, they are given by

$$D = \frac{2w}{\lambda(2w + \lambda\sigma^2)} \tag{3.51}$$

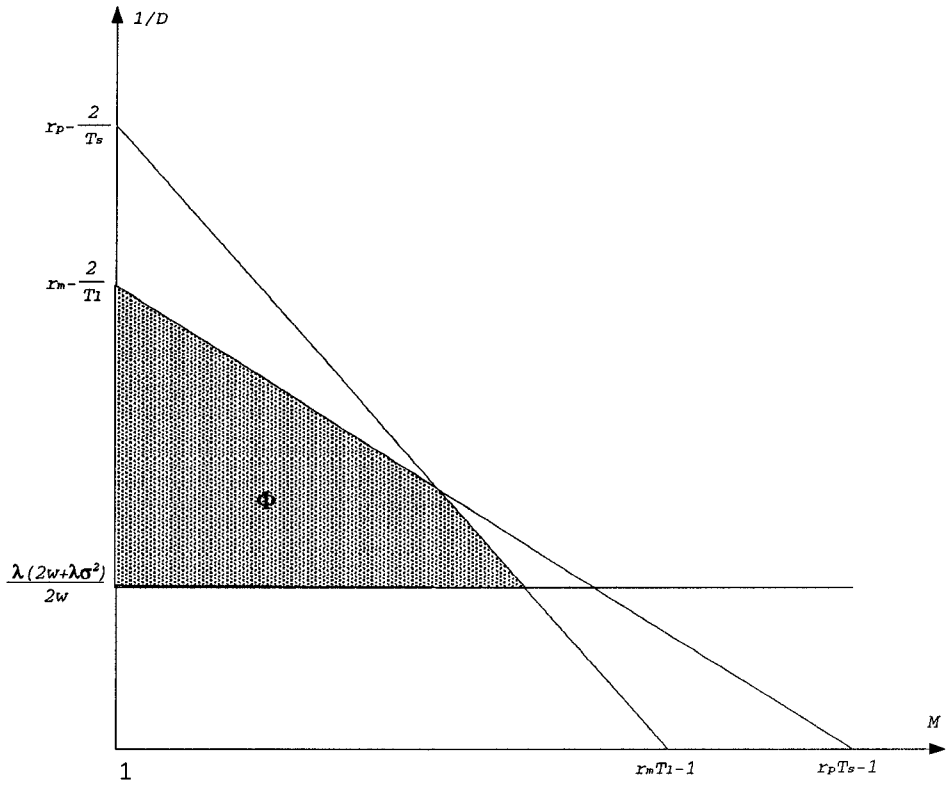


Figure 3.4: The set Φ

and

$$M = 1. \tag{3.52}$$

It is clear that with $B = \infty$, and with D and M being chosen according to (3.51) and (3.52), we get a feasible solution to (P1) although it may not be the optimal solution. It is also clear that if we replace B by any other finite value that satisfies (3.46), we get another feasible solution. By doing so, the probability of cell loss will no longer be zero, however, the cell delay and the mean and peak cell rates will all decrease. The choice of B can be done by means of dimensioning [34], and we shall not discuss it here.

From the discussion in this subsection, we make the following observations.

First, the set $\tilde{\Psi}$ may be empty. In this case, it does not necessarily mean that the feasible set is empty, because most of the conditions (3.47), (3.48), and (3.50) are derived from upper bounds of the corresponding quantities. Therefore we may relax some of the conditions and try again.

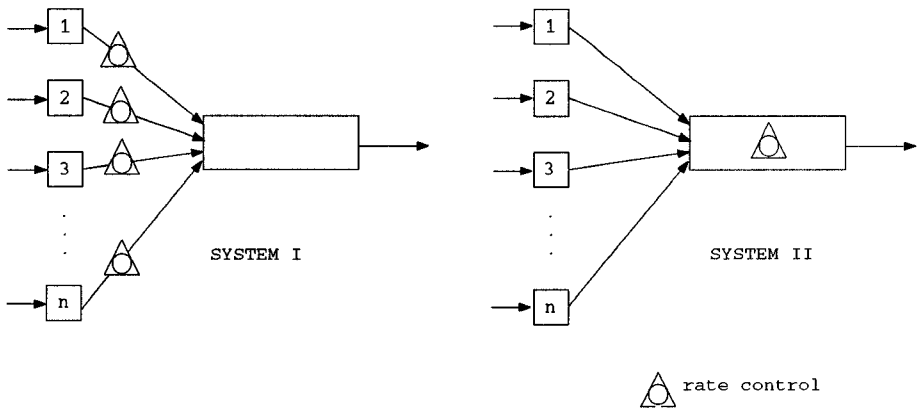


Figure 3.5: Multiplexer with LBs.

Secondly, if T_l is chosen to be very large and T_s to be very small compared to the token generation period D, then we see from (3.47) and (3.48) that the mean cell rate r_m will be determined mainly by the token generation period D and the peak cell rate r_p will be determined mainly by the token pool size M.

3.3.2 Deployment of LBs in the Network

In this subsection, we consider the use of LBs to reduce the burstiness of the traffic in a network. Suppose we have n sources which are multiplexed in a trunk. Our task is to apply LB technique to reduce the burstiness of the traffic on the trunk. We consider the two configurations shown in Fig. 3.5. The first configuration (SYSTEM I) is to apply a single LB at the trunk, and we call it the centralized implementation. The second configuration is to apply to each source a LB and we call it the distributed implementation. In the distributed implementation, we further consider two cases (SYSTEM II and III). In SYSTEM II all LBs generate tokens at the same time, while in SYSTEM III the token generation epochs among all LBs are equally spaced. To compare the two configurations, we assume that the total token generation rate in the second configuration is equal to the token generation rate of the single LB in the first configuration. However, we assume all LBs have token pool size of one. Our analysis is based on simulation, the results of which are depicted in Fig. 3.6.

From Fig. 3.6 we see that for smooth arrival processes (Erlang for example),

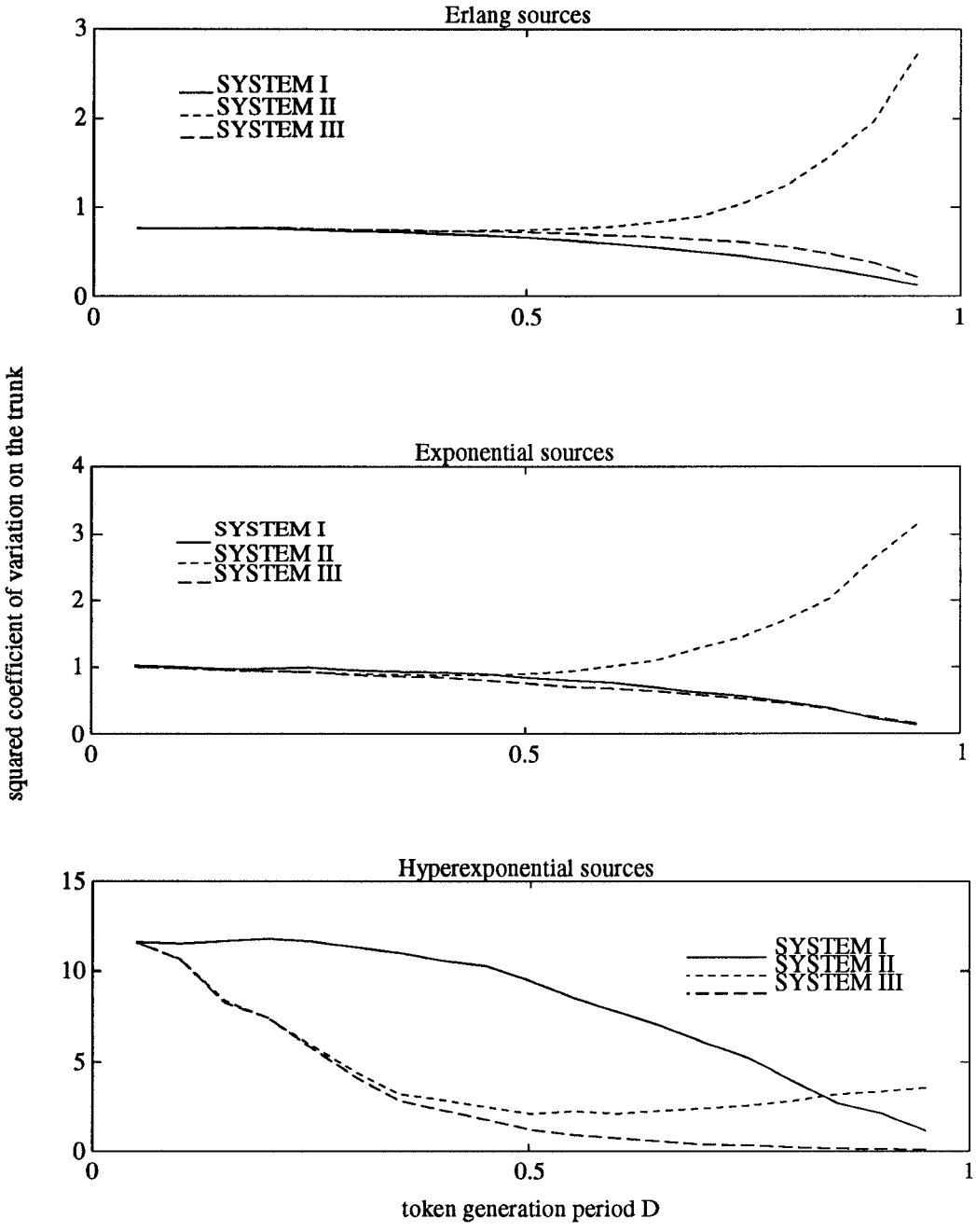


Figure 3.6: Comparison of systems I, II and III with LB rate control

centralized implementation (SYSTEM I) is better than distributed implementation. However, as the arrival process becomes more bursty (Hyperexponential for example), SYSTEM III is the best.

The reason SYSTEM II is not as good as system III is obvious, since it has a tendency to produce batch arrivals at the token generation epochs (especially when the LBs are designed to work in heavy traffic).

To see that SYSTEM III is preferred to SYSTEM I when the sources are bursty, we provide the following intuitive explanation. We can view the centralized LB in SYSTEM I as a variation of the distributed LBs by sharing a common token pool. Thus in the centralized implementation, each source can use tokens generated by all the n LBs. Therefore, during any burst, the sources are less regulated, resulting in more bursty traffic.

Since the traffic in high-speed networks are bursty, our conclusion is that the distributed implementation is preferred. Another advantage of distributed implementation is that it guarantees the fairness among sources.

PART II

CALCULUS OF BURSTINESS

CHAPTER 4

THE SQUARED COEFFICIENT OF VARIATION

4.1 Thinning of Point Processes

In Chapter 2, we have succeeded in using the squared coefficient of variation to measure burstiness. In this chapter, we shall continue to explore the use of this measure in characterizing the burst reduction properties of thinning. Roughly speaking, to thin a process is to delete a portion of its arrivals to form a new process. The rate control of an arrival process (such as the un-buffered LB), or the splitting of a process are examples of thinning. In the following, we formally define two kinds of thinning, random thinning and its deterministic realizations, which we shall study in this chapter.

Let $\boldsymbol{A} = \{a_n, n = 1, 2, \cdots\}$ be a point process with arrival epochs $\{a_n, n = 1, 2, \cdots\}$. With $a_0 = 0$, define the inter-arrival times of $\boldsymbol{A}$ by

$$X_n = a_n - a_{n-1}, \qquad n = 1, 2, \cdots. \tag{4.1}$$

Throughout this chapter, we assume that the rvs $\{X_n, n = 1, 2, \cdots\}$ are i.i.d. rvs which have the same distribution as the $\mathbb{R}_+$-valued rv X, i.e., we assume that this process is a renewal process. We also assume that X is integrable, i.e., $\mathbb{E}[X] < \infty$.

4.1.1 Random Thinning

For $0 < p \leq 1$, the p-thinning of a point process $\boldsymbol{A}$ is another point process, denoted by $\boldsymbol{A}_p$, resulting from independently deleting each arrival from $\boldsymbol{A}$ with probability $1 - p$. Since the original process is a renewal process, the resulting process is also a renewal process, and the inter-arrival times $\{Y_n^p,\ n = 1, 2, \cdots\}$ of $\boldsymbol{A}_p$ are i.i.d. rvs

distributed according to some generic $\mathbb{R}_+$-valued rv Y_p defined by

$$Y_p \stackrel{d}{=} \sum_{n=1}^{M} X_n \tag{4.2}$$

where M is a geometric rv with parameter p independent of $\boldsymbol{A}$. It is well known that

$$\mathbb{P}\{M = m\} = p(1-p)^{m-1}, \qquad m = 1, 2, \cdots. \tag{4.3}$$

4.1.2 Deterministic Realizations of Random Thinning

If p is a rational, then it can be expressed as

$$p = \frac{v}{u+v} \tag{4.4}$$

with u and v being mutually prime nonnegative integers. In this case, the p-thinning defined in the subsection 4.1.1 can be implemented deterministically. A deterministic realization of the p-thinning with p defined by (4.4) is to periodically delete u arrivals out of every $u+v$ arrivals in a deterministic fashion. By choosing different arrivals to delete, we have $\dfrac{(u+v)!}{u!v!}$ ways to deterministically realize a p-thinning. To analyze the process that results from a deterministic realization of the p-thinning, we first seek a good characterization of such process. To this end, for each pair of positive integers (u, v) we define the set

$$\Phi_{u,v} := \left\{ \boldsymbol{w} = (w_1, \cdots, w_v) \in \mathbb{N}^v : w_i \geq 1, \;\; i = 1, \cdots, v; \;\; \sum_{i=1}^{v} w_i = u + v \right\}. \tag{4.5}$$

The elements of $\Phi_{u,v}$ are in one-to-one correspondence with each of the deterministic realizations discussed above. The inter-arrival times $\{Y_k^{\boldsymbol{w}}, \;\; k = 1, 2, \cdots\}$ of the process that corresponds to $\boldsymbol{w}$ in $\Phi_{u,v}$ are independent rvs distributed according to

$$Y_k^{\boldsymbol{w}} \stackrel{d}{=} \sum_{j=1}^{w_i} X_j, \quad \text{if} \quad k = lv + i, \; i = 1, \cdots, v, \; l = 0, 1, \cdots. \tag{4.6}$$

We call such a realization the $\boldsymbol{w}$-thinning of $\boldsymbol{A}$ with resulting process denoted by $\boldsymbol{A}\boldsymbol{w}$. It is plain that $\{Y_k^{\boldsymbol{w}}, k = 1, 2, \cdots\}$ is no longer a renewal process unless w_n is constant for all $n = 1, \cdots, v$. However, the following lemma tells us that $\{Y_k^{\boldsymbol{w}}, k = 1, 2, \cdots\}$ is convexly stable.

Lemma 4.1 *The inter-arrival times* $\{Y_k^{\boldsymbol{w}}, k = 1,2,\cdots\}$ *of the* $\boldsymbol{w}$*-thinning of a renewal process* $\boldsymbol{A}$ *is convexly stable. Furthermore, the asymptotic version* $Y^{\boldsymbol{w}}$ *of* $\{Y_k^{\boldsymbol{w}}, k = 1,2,\cdots\}$ *is given by*

$$Y^{\boldsymbol{w}} \stackrel{d}{=} \sum_{j=1}^{w_i} X_j, \quad w.p. \ \frac{1}{v}, \qquad i = 1,\cdots,v. \tag{4.7}$$

Note that the sequence $\{Y_k^{\boldsymbol{w}}, k = 1,2,\cdots\}$ is a periodic i.i.d. sequence with period v as will be defined later. Therefore the convex stability of $\{Y_k^{\boldsymbol{w}}, k = 1,2,\cdots\}$ follows from the strong law of large numbers for such sequences (Lemma 4.2), while the form of the asymptotic version can be obtained by inspection via the uniqueness theorem of Theorem 2.1.

Let $\{X_{i,j},\ i = 0,1,\cdots;\ j = 1,2,\cdots\}$ be a double sequence of $\mathbb{R}$-valued rvs such that for each $j = 1,2,\cdots$, the sequence $\{X_{i,j},\ i = 0,1,\cdots\}$ is a sequence of i.i.d. rvs. Fix $k = 1,2,\cdots$, and define the sequence of $\mathbb{R}$-valued rvs $\{Y_n, n = 1,2,\cdots\}$ by

$$Y_{ik+j} = X_{i,j}, \qquad i = 0,1,\cdots;\ j = 1,\cdots,k. \tag{4.8}$$

We say that the rvs $\{Y_n, n = 1,2,\cdots\}$ form a periodic i.i.d. sequence with period k ($k = 1,2,\cdots$). We see that periodic i.i.d. sequences form a natural generalization of the notion of i.i.d. sequence. Now we establish a strong law of large numbers for periodic i.i.d. sequences in the following lemma.

Lemma 4.2 *(Strong Law of Large Numbers for Periodic I.I.D. Sequences)*

If $\{Y_n, n = 1,2,\cdots\}$ *is an integrable periodic i.i.d. sequence of* $\mathbb{R}$*-valued rvs with period* k*, then we have*

$$\lim_{m\to\infty} \frac{1}{m}\sum_{n=1}^{m} Y_n = \frac{1}{k}\sum_{n=1}^{k} \mathbb{E}[Y_n] \quad a.s. . \tag{4.9}$$

Proof. We prove the case where $\{Y_n, n = 1,2,\cdots\}$ are $\mathbb{R}_+$-valued rvs. A sharper argument shows that rvs can be taken to be $\mathbb{R}$-valued, too. Let

$$Z_i = \sum_{j=1}^{k} X_{i,j}, \qquad i = 0,1,\cdots. \tag{4.10}$$

Then the rvs $\{Z_i,\ i=0,1,\cdots\}$ form an i.i.d. sequence of $\mathbb{R}_+$-valued rvs, and by the strong law of large numbers, we have

$$\begin{aligned}\lim_{l\to\infty}\frac{1}{l}\sum_{i=0}^{l-1}Z_i &= \mathbb{E}[Z_0] \qquad a.s.\\ &= \sum_{n=1}^{k}\mathbb{E}[X_{0,n}] \qquad a.s.\\ &= \sum_{n=1}^{k}\mathbb{E}[Y_n] \qquad a.s.\ . \end{aligned} \tag{4.11}$$

From (4.10), we have

$$\frac{1}{m}\sum_{n=1}^{m}Y_n = \frac{1}{m}\sum_{i=0}^{k(m)-1}Z_i + \frac{1}{m}\sum_{n=k(m)k+1}^{m}Y_n, \qquad m=1,2,\cdots \tag{4.12}$$

where $k(m)=\left\lfloor\frac{m}{k}\right\rfloor$, $m=1,2,\cdots$, is the number of periods in the first m rvs. Since

$$0\le m-k(m)k<k, \qquad m=1,2,\cdots, \tag{4.13}$$

the second term in the right-hand side of (4.12) satisfies

$$0\le\sum_{n=k(m)k+1}^{m}Y_n\le Z_{k(m)}, \qquad m=1,2,\cdots. \tag{4.14}$$

The rvs $\{Y_n, n=1,2,\cdots\}$ being nonnegative, we see from (4.12) and (4.14) that

$$\frac{k(m)}{m}\frac{1}{k(m)}\sum_{i=0}^{k(m)-1}Z_i\le\frac{1}{m}\sum_{n=1}^{m}Y_n\le\frac{k(m)+1}{m}\frac{1}{k(m)+1}\sum_{i=0}^{k(m)}Z_i, \qquad m=1,2,\cdots, \tag{4.15}$$

and (4.9) now follows from (4.11) and (4.15) by noting that

$$\lim_{m\to\infty}\frac{k(m)}{m}=\frac{1}{k}. \tag{4.16}$$

■

As a special case of p-thinning with $v=1$ and $u=m$ for some $m=0,1,\cdots$, the deterministic realization always delete the first m arrivals and keep the $(m+1)^{st}$ one. Under these assumptions, the resulting process is a renewal process with inter-arrival times being distributed according to the rv Y^m defined by

$$Y^m\stackrel{d}{=}\sum_{i=1}^{m+1}X_i. \tag{4.17}$$

We shall discuss this case in the next chapter.

With the notation introduced so far, we shall proceed to the discussion of the burst reduction properties of random thinning and of its deterministic realizations via the squared coefficient of variation.

4.2 p-Thinning

Let $\boldsymbol{A}$ be a renewal process and let $\boldsymbol{A}_p$ be its p-thinning for $0 < p \leq 1$ as defined in Section 4.1.1. Recall that the inter-arrival times of $\boldsymbol{A}_p$ are i.i.d. rvs $\{Y_n^p,\ n = 1, 2, \cdots\}$ distributed as the rv Y_p defined in (4.2). By conditioning on M, we get

$$\mathbb{E}[Y_p] = \mathbb{E}[M]\,\mathbb{E}[X] = \frac{1}{p}\,\mathbb{E}[X] \tag{4.18}$$

and

$$\begin{aligned} Var(Y_p) &= \mathbb{E}[M]\ Var(X) + Var(M)\,(\mathbb{E}[X])^2 && (4.19)\\ &= \frac{1}{p}\,Var(X) + \frac{1-p}{p^2}\,(\mathbb{E}[X])^2\,. && (4.20) \end{aligned}$$

The following is an easy consequence of these facts.

Lemma 4.3 *The squared coefficient of variation of the p-thinning of a renewal process is*

$$C^2(Y_p) = \frac{C^2(X)}{\mathbb{E}[M]} + C^2(M) = pC^2(X) + 1 - p \tag{4.21}$$

where M is a geometric rv with parameter p.

From (4.21) we see that the mapping $p \to C^2(Y_p)$ is a linear mapping. It is strictly increasing if $C^2(X) < 1$ or strictly decreasing if $C^2(X) > 1$. Recall that the squared coefficient of variation of a Poisson process is one and is independent of the rate of the process. Combining this last comment with (4.21), we get

Theorem 4.1 *The p-thinning of a renewal process is burst reducing if and only if the original process is more bursty than a Poisson process, i.e.,*

$$C^2(Y_p) \leq C^2(X) \tag{4.22}$$

if and only if

$$C^2(X) \geq 1 \tag{4.23}$$

with equality holding if and only if $C^2(X) = 1$.

The effects of thinning are two-folds. On the one hand, thinning reduces burstiness through merging adjacent intervals. On the other hand, the irregularity of the thinning scheme introduces extra burstiness. From Theorem 4.1, we see that if the original process is Poisson, the two opposite effects cancel out. If the original process is smooth enough (less bursty than the Poisson process), then the latter effect dominates the former, and the burstiness therefore increases. If the original process is more bursty than the Poisson process, then the former effect dominates the latter, and the burstiness therefore decreases.

4.3 w-Thinning

Let u and v be fixed (not necessarily mutually prime) integers such that $u \geq 0$ and $v > 0$. For $\boldsymbol{w}$ in $\Phi_{u,v}$ (defined in (4.5)), let $\boldsymbol{A_w}$ be the $\boldsymbol{w}$-thinning of a renewal process $\boldsymbol{A}$ as defined in Section 4.1.2. From Lemma 4.1 we know that the inter-arrival times $\{Y_k^{\boldsymbol{w}},\ k = 1, 2, \cdots\}$ of $\boldsymbol{A_w}$ are convexly stable and the asymptotic version is given by (4.7). As different choices of $\boldsymbol{w}$ in $\Phi_{u,v}$ lead to different realizations of the p-thinning with p given by (4.4), we provide in the following lemma a comparison of these realizations.

Lemma 4.4 *For $\boldsymbol{w}$ and $\boldsymbol{w}'$ in $\Phi_{u,v}$, if $\boldsymbol{w} \prec \boldsymbol{w}'$, then*

$$\mathbb{E}\left[\sum_{i=1}^{v}\phi(Y_i^{\boldsymbol{w}})\right] \leq \mathbb{E}\left[\sum_{i=1}^{v}\phi(Y_i^{\boldsymbol{w}'})\right] \tag{4.24}$$

for any convex function $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that the expectations are well defined.

This lemma is a direct consequence of the following lemma.

Lemma 4.5 *Let $\boldsymbol{m} = (m_1, \cdots, m_K)$ and $\boldsymbol{n} = (n_1, \cdots, n_K)$ be two elements of $\mathbb{N}^K$. Let $\{X_{ij}, i, j = 1, 2, \cdots\}$ be a collection of i.i.d. integrable $\mathbb{R}_+$-valued rvs, and define the $\mathbb{R}_+^K$-valued rvs $\boldsymbol{Y} = (Y_1, \cdots, Y_K)$ and $\boldsymbol{Z} = (Z_1, \cdots, Z_K)$ by*

$$Y_i = \sum_{j=1}^{m_i} X_{ij} \quad \text{and} \quad Z_i = \sum_{j=1}^{n_i} X_{ij}, \qquad i = 1, \cdots, K. \tag{4.25}$$

Then $\boldsymbol{m} \prec \boldsymbol{n}$ implies

$$\mathbb{E}\left[\sum_{i=1}^{K}\phi(Y_i)\right] \leq \mathbb{E}\left[\sum_{i=1}^{K}\phi(Z_i)\right] \tag{4.26}$$

for all convex function $\phi : \mathbb{R}_+ \to \mathbb{R}$.

Proof. Without loss of generality, we may assume that $n_1 < m_1 \leq m_2 < n_2$ and $n_i = m_i, i = 3, \cdots, K$ [46, Lemma 2.B.1, p.21]. For any convex function $\phi : \mathbb{R}_+ \to \mathbb{R}$, we find

$$\mathbb{E}\left[\sum_{i=1}^{K}\phi(Z_i)\right] - \mathbb{E}\left[\sum_{i=1}^{K}\phi(Y_i)\right] = \mathbb{E}\left[\phi(Z_1) + \phi(Z_2) - \phi(Y_1) - \phi(Y_2)\right]. \tag{4.27}$$

Denoting

$$U := \sum_{j=m_2+1}^{n_2} X_{2j}, \tag{4.28}$$

and observing that

$$m_1 - n_1 = n_2 - m_2, \tag{4.29}$$

we see that

$$Y_1 \stackrel{d}{=} Z_1 + U \quad \text{and} \quad Z_2 = Y_2 + U. \tag{4.30}$$

Thus (4.27) yields

$$\begin{aligned} & \mathbb{E}\left[\sum_{i=1}^{K}\phi(Z_i)\right] - \mathbb{E}\left[\sum_{i=1}^{K}\phi(Y_i)\right] \\ = \; & \mathbb{E}\left[\phi(Y_2+U) - \phi(Y_2) + \phi(Z_1) - \phi(Z_1+U)\right] \\ \geq \; & 0 \end{aligned} \tag{4.31}$$

by the convexity of ϕ. ■

By the closure property of majorization under concatenation (Lemma A.2), Lemma 4.4 can be easily extended to any number of periods of the realization. Invoking Lemma 4.1, we conclude to the following theorem.

Theorem 4.2 *For $\boldsymbol{w}$ and $\boldsymbol{w}'$ in $\Phi_{u,v}$, let $Y_{\boldsymbol{w}}$ and $Y_{\boldsymbol{w}'}$ be the asymptotic versions of the inter-arrival times of $\boldsymbol{A}_{\boldsymbol{w}}$ and $\boldsymbol{A}_{\boldsymbol{w}'}$, respectively. If $\boldsymbol{w} \prec \boldsymbol{w}'$, then*

$$Y_{\boldsymbol{w}} \leq_{cx} Y_{\boldsymbol{w}'}, \tag{4.32}$$

and therefore,

$$C^2(Y_{\boldsymbol{w}}) \leq C^2(Y_{\boldsymbol{w}'}). \tag{4.33}$$

We next calculate the squared coefficient of variation of $Y_{\boldsymbol{w}}$ for an arbitrary $\boldsymbol{w}$ in $\Phi_{u,v}$. With

$$\overline{w} = \frac{1}{v}\sum_{i=1}^{v} w_i \quad \text{and} \quad \overline{w^2} = \frac{1}{v}\sum_{i=1}^{v} w_i^2, \qquad \boldsymbol{w} \in \Phi_{u,v}, \tag{4.34}$$

we observe from (4.7) that

$$\mathbb{E}\left[Y_{\boldsymbol{w}}\right] = \frac{1}{v}\sum_{i=1}^{v} \mathbb{E}\left[\sum_{j=1}^{w_i} X_j\right] = \frac{1}{v}\sum_{i=1}^{v} w_i\, \mathbb{E}\left[X\right] = \overline{w}\, \mathbb{E}\left[X\right] \tag{4.35}$$

and

$$\begin{aligned}\mathbb{E}\left[Y_{\boldsymbol{w}}^2\right] &= \frac{1}{v}\sum_{i=1}^{v}\mathbb{E}\left[\left(\sum_{j=1}^{w_i}X_j\right)^2\right] \\ &= \overline{w}\,Var\,(X) + \overline{w^2}(\mathbb{E}\,[X])^2. \end{aligned} \tag{4.36}$$

Combining (4.35) and (4.36), we finally see that

$$\begin{aligned} C^2(Y_{\boldsymbol{w}}) &= \frac{\mathbb{E}\,[Y_{\boldsymbol{w}}^2]}{(\mathbb{E}\,[Y_{\boldsymbol{w}}])^2} - 1 \\ &= \frac{Var\,(X)}{\overline{w}(\mathbb{E}\,[X])^2} + \frac{\overline{w^2}}{(\overline{w})^2} - 1 \\ &= \frac{1}{\overline{w}}C^2(X) + \frac{\overline{w^2} - (\overline{w})^2}{(\overline{w})^2}. \end{aligned} \tag{4.37}$$

Lemma 4.6 *For each $\boldsymbol{w} \in \Phi_{u,v}$, the squared coefficient of variation of the asymptotic version of the inter-arrival times of the $\boldsymbol{w}$-thinning is given by*

$$C^2(Y_{\boldsymbol{w}}) = \frac{C^2(X)}{\overline{w}} + c^2(\boldsymbol{w}) \tag{4.38}$$

with

$$c^2(\boldsymbol{w}) = \frac{\overline{w^2} - \overline{w}^2}{\overline{w}^2} \tag{4.39}$$

denoting the squared coefficient of variation of $\boldsymbol{w} = (w_1, \cdots, w_v)$.

By a routine calculation, we conclude to the following theorem.

Theorem 4.3 *For any $\boldsymbol{w}$ in $\Phi_{u,v}$, $\boldsymbol{w}$-thinning of a renewal process is burst reducing, i.e.,*

$$C^2(Y_{\boldsymbol{w}}) \leq C^2(X) \tag{4.40}$$

if and only if

$$c^2(\boldsymbol{w}) \leq (1 - \frac{1}{\overline{w}})C^2(X) \tag{4.41}$$

with equality holding if and only if

$$c^2(\boldsymbol{w}) = (1 - \frac{1}{\overline{w}})C^2(X). \tag{4.42}$$

From Theorem 4.3, we see that like p-thinning, $\boldsymbol{w}$-thinning is not always burst reducing. It is the burstiness introduced by the variation of the vector $\boldsymbol{w}$ that prevents it from reducing the burstiness of the original process. This issue will be revisited in the next section.

The question is now whether or not there is an element $\boldsymbol{w}$ in $\Phi_{u,v}$ such that the condition (4.41) is satisfied. Before answering this question, we first show that the quantity $C^2(Y_{\boldsymbol{w}})$ can be minimized by properly choosing $\boldsymbol{w}$ in $\Phi_{u,v}$. Define the element $\boldsymbol{w}^*$ in $\mathbb{N}^v$ by

$$w_i^* = \begin{cases} a+1, & i = 1, \cdots, m, \\ a, & i = m+1, \cdots, v \end{cases} \tag{4.43}$$

where

$$a = \left\lfloor \frac{u+v}{v} \right\rfloor \quad \text{and} \quad m = u + v - av. \tag{4.44}$$

It is clear that $0 \leq m < v$, so that $\boldsymbol{w}^*$ is indeed well defined. Moreover we have $a \geq 1$ and

$$\sum_{i=1}^{v} w_i^* = m(a+1) + (v-m)a = m + av = u + v. \tag{4.45}$$

In other words, $\boldsymbol{w}^*$ is indeed an element of $\Phi_{u,v}$. We claim that $\boldsymbol{w}^*$ minimizes $C^2(Y_{\boldsymbol{w}})$ over $\Phi_{u,v}$.

Theorem 4.4 *The vector $\boldsymbol{w}^*$ defined by (4.43)–(4.44) minimizes $C^2(Y_{\boldsymbol{w}})$ over $\Phi_{u,v}$. Furthermore, $\boldsymbol{w}^*$ is unique in the sense that any $\boldsymbol{w}$ in $\Phi_{u,v}$ that minimizes $C^2(Y_{\boldsymbol{w}})$ is a permutation of $\boldsymbol{w}^*$.*

Proof. For $\boldsymbol{w}$ in $\Phi_{u,v}$, we see that $\overline{w} = \dfrac{u+v}{v}$, and minimizing $C^2(Y_{\boldsymbol{w}})$ over $\Phi_{u,v}$ is thus equivalent to minimizing $c^2(\boldsymbol{w})$ over $\Phi_{u,v}$. It is well known that $c^2(\boldsymbol{w})$ is (strictly) Schur-convex [46, 3.D.2.], therefore,

$$c^2(\boldsymbol{w}^*) \leq c^2(\boldsymbol{w}) \quad \text{if} \quad \boldsymbol{w}^* \prec \boldsymbol{w}, \tag{4.46}$$

with equality in (4.46) only if $\boldsymbol{w}$ is obtained from $\boldsymbol{w}^*$ by permutation. This can also be seen from Theorem 4.2.

Now it suffices to show that $\boldsymbol{w}^* \prec \boldsymbol{w}$ for all $\boldsymbol{w}$ in $\Phi_{u,v}$. From the definition of majorization, it suffices to show that for each $\boldsymbol{w}$ in $\Phi_{u,v}$, the inequalities

$$\sum_{i=1}^{k} w_{[i]} \geq k(a+1), \qquad k = 1, \cdots, m \tag{4.47}$$

and

$$\sum_{i=1}^{k} w_{[i]} \geq m(a+1) + (k-m)a, \qquad k = m+1, \cdots, v-1. \tag{4.48}$$

hold. We show this by contradiction upon making use of the non-increasing property of $\{w_{[i]},\ i = 1, \cdots, v\}$. If (4.47) fails, then we have

$$\sum_{i=1}^{k_0} w_{[i]} < k_0(a+1) \ \text{ for some } \ k_0 = 1, \cdots, m, \tag{4.49}$$

in which case, we necessarily have $w_{[k_0]} < a+1$ for otherwise (4.49) cannot be true. Therefore

$$w_{[i]} \leq a, \qquad i = k_0, \cdots, v, \tag{4.50}$$

and inequalities (4.49) and (4.50) lead to the following contradiction

$$v\overline{w} = \sum_{i=1}^{v} w_i < k_0(a+1) + (v-k_0)a = k_0 + va \leq m + va = v\overline{w}. \tag{4.51}$$

If (4.48) fails, then we have

$$\sum_{i=1}^{k_1} w_{[i]} < m(a+1) + (k_1 - m)a \ \text{ for some } \ k_1 = m+1, \cdots, v. \tag{4.52}$$

And this in turn implies (as shown below) that $w_{[k_1]} \leq a$, whence

$$w_{[i]} \leq a, \qquad i = k_1, \cdots, v. \tag{4.53}$$

From (4.52) and (4.53) we then get the following contradiction

$$v\overline{w} = \sum_{i=1}^{v} w_i < m(a+1) + (k_1 - m)a + (v-k_1)a = m + va = v\overline{w}. \tag{4.54}$$

To see that (4.52) implies (4.53), we assume that (4.53) is not true. Therefore we have $w_{[i]} \geq a+1$ for all $i = 1, \cdots, k_1$, whence,

$$\sum_{i=1}^{k_1} w_{[i]} \geq k_1(a+1) = m(a+1) + (k_1 - m)(a+1) > m(a+1) + (k_1 - m)a, \tag{4.55}$$

in contradiction with (4.52).

Finally we show the uniqueness of $\boldsymbol{w}^*$ (up to a permutation). We have already shown that $\boldsymbol{w}^* \prec \boldsymbol{w}$ for any $\boldsymbol{w}$ in $\Phi_{u,v}$. If $\boldsymbol{w}$ is not a permutation of $\boldsymbol{w}^*$, then we have $c^2(\boldsymbol{w}^*) < c^2(\boldsymbol{w})$ by the strict Schur-convexity of $c^2(\boldsymbol{w})$ [46, 3.A.1], and $\boldsymbol{w}$ cannot be a minimizer of $c^2(\boldsymbol{w})$. ■

The squared coefficient of variation of $\boldsymbol{w}^*$ can be obtained through some algebra, and is given by

$$c^2(\boldsymbol{w}^*) = \frac{m(v-m)}{(u+v)^2} \tag{4.56}$$

with m being defined in (4.44). Therefore, we have the following result.

Lemma 4.7 *The smallest squared coefficient of variation which can be achieved through deterministic realization of the p-thinning is*

$$C^2(Y_{\boldsymbol{w}^*}) = \frac{v}{u+v} C^2(X) + \frac{m(v-m)}{(u+v)^2}. \tag{4.57}$$

In general, it is very difficult to express (4.57) as a function of p. However, we can bound the value of $C^2(Y_{\boldsymbol{w}^*})$ as

$$pC^2(X) \leq C^2(Y_{\boldsymbol{w}^*}) \leq pC^2(X) + \frac{p^2}{4}. \tag{4.58}$$

This bound is tight in the sense that the lower bound is achieved when $m = 0$ (i.e., u is a multiple of v) and the upper bound is achieved when $m = \frac{v}{2}$ (i.e., the remainder of u/v is $v/2$). The lower bound is obvious while the upper bound can be obtained as follows: Fix u and v with $p = \frac{u}{u+v}$. The function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \frac{x(v-x)}{(u+v)^2}, \qquad x \in \mathbb{R} \tag{4.59}$$

is concave with maximum value at $x = \frac{v}{2}$ given by

$$f\left(\frac{v}{2}\right) = \frac{v^2}{4(u+v)^2} = \frac{p^2}{4}. \tag{4.60}$$

From (4.58) we see that we can make the squared coefficient of variation of the asymptotic version of the inter-arrival times of $\boldsymbol{w}$-thinning arbitrarily small by letting p goes to zero. This is in sharp contrast to the case of p-thinning, where the

squared coefficient of variation of the asymptotic version of the inter-arrival times goes to one as p goes to zero.

Now we come back to our earlier question – whether there exists $\boldsymbol{w}$ in $\Phi_{u,v}$ such that the $\boldsymbol{w}$-thinning is burst reducing (or that condition (4.41) is satisfied). The following theorem gives a sufficient condition for this to happen.

Theorem 4.5 *A sufficient condition for the existence of $\boldsymbol{w}$ in $\Phi_{u,v}$ such that the $\boldsymbol{w}$-thinning is burst reducing is*

$$C^2(X) > \frac{p^2}{4(1-p)}. \tag{4.61}$$

Proof. If (4.61) is satisfied, then from (4.58) we have

$$C^2(Y_{\boldsymbol{w}*}) \leq pC^2(X) + \frac{p^2}{4} < C^2(X). \tag{4.62}$$

■

Observe that if $p < 2\sqrt{2} - 2 \approx 0.828$, then $\frac{p^2}{4(1-p)} < 1$. Therefore, $\boldsymbol{w}$-thinning may still be burst reducing even if the original process is less bursty than a Poisson process.

Our discussion so far focuses on the assumption that p has the form of (4.4) with u and v mutually prime. One natural extension is to allow p to be expressed as

$$p = \frac{kv}{k(u+v)}, \qquad k = 1, 2, \cdots. \tag{4.63}$$

The question is whether the optimal deterministic realization over $\Phi_{ku,kv}$ has some monotonicity properties in k. This is explored in the following theorem.

Theorem 4.6 *For p defined by (4.63) with u and v mutually prime, we have*

$$\min_{\boldsymbol{w} \in \Phi_{ku,kv}} C^2(Y_{\boldsymbol{w}}) = \min_{\boldsymbol{w} \in \Phi_{u,v}} C^2(Y_{\boldsymbol{w}}) \qquad k = 1, 2, \cdots. \tag{4.64}$$

Proof. Our discussion so far has made no use of the fact that u and v are mutually prime. Therefore all the results we obtained can be directly used for $u' = ku$ and

$v' = kv$ for some fixed $k = 1, 2, \cdots$. By Theorem 4.4, $\boldsymbol{w}'$ which minimizes $C^2(Y_{\boldsymbol{w}})$ over $\Phi_{u',v'}$ is given by

$$w_i' = \begin{cases} a' + 1, & i = 1, \cdots, m', \\ a', & i = m' + 1, \cdots, v' \end{cases} \tag{4.65}$$

where

$$a' = \left\lfloor \frac{u' + v'}{v'} \right\rfloor \quad \text{and} \quad m' = u' + v' - a'v'. \tag{4.66}$$

Simple calculation now yields

$$a' = a \quad \text{and} \quad m' = km, \tag{4.67}$$

so that

$$\overline{w'} = \overline{w^*} \quad \text{and} \quad \overline{w'^2} = \overline{w^{*2}}, \tag{4.68}$$

and (4.64) thus holds by virtue of Lemma 4.6. ■

From Theorem 4.6, it is sufficient to consider only the simplest case with u and v mutually prime.

4.4 Comparison of p-Thinning and w-Thinning

Using the notation of Sections 4.2 and 4.3, we note that

$$\overline{w} = \mathbb{E}[M] = \frac{1}{p}, \qquad \boldsymbol{w} \in \Phi_{u,v} \tag{4.69}$$

and this leads via (4.21) and (4.38) to the following theorem.

Theorem 4.7 *For any* $p = \dfrac{v}{u+v}$ *and* $\boldsymbol{w}$ *in* $\Phi_{u,v}$ *with* u *and* v *mutually prime, we have*

$$C^2(Y_{\boldsymbol{w}}) \leq C^2(Y_p) \quad \textit{iff} \quad c^2(\boldsymbol{w}) \leq C^2(M) \tag{4.70}$$

with equality holding if and only if $c^2(\boldsymbol{w}) = C^2(M)$, *where* M *is a geometric rv with parameter* p.

Theorem 4.7 tells us that deterministic realization may not always be better than random thinning. It is better only when the variability, measured by the squared coefficient of variation, of $\boldsymbol{w}$ is less than that of a geometric rv with the same mean. For example, let $u = 2$ and $v = 5$, i.e., $p = 5/7$. For $\boldsymbol{w} = (1,1,1,1,3)$ in $\Phi_{2,5}$, we get $\overline{w} = 1.4$, $\overline{w^2} = 2.6$, and therefore $c^2(\boldsymbol{w}) = 0.3265 > 0.2857 = C^2(M)$. Thus by Theorem 4.7, we have $C^2(Y_{\boldsymbol{w}}) > C^2(Y_p)$.

However, the following theorem says that the deterministic realization defined via (4.43)–(4.44) is always better than random thinning.

Theorem 4.8 *If* $\boldsymbol{w}^*$ *is defined by (4.43)–(4.44), then we have*

$$C^2(Y_{\boldsymbol{w}^*}) < C^2(Y_p). \tag{4.71}$$

Proof. By Theorem 4.7, we need to show that $c^2(\boldsymbol{w}^*) < 1 - p$. Upon using (4.56) this is equivalent to showing

$$c^2(\boldsymbol{w}^*) - (1-p) = \frac{m(v-m) - u(u+v)}{(u+v)^2} < 0, \tag{4.72}$$

where m is defined by (4.44). With $g : \mathbb{R} \to \mathbb{R}$ defined by

$$g(x) = -x^2 + vx - u(u+v), \qquad x \in \mathbb{R}, \tag{4.73}$$

we see that (4.72) is equivalent to

$$g(m) < 0. \tag{4.74}$$

Case 1: If $v^2 - 4u(u+v) < 0$, then $g(x) = 0$ has no real solution. Since g is a concave function, we have $g(x) < 0$ for all x in $\mathbb{R}$, and in particularly, $g(m) < 0$.

Case 2: If $v^2 - 4u(u+v) \geq 0$, then it is easy to see that $v \geq (2+2\sqrt{2})u$, in which case

$$1 \leq a = \left\lfloor \frac{u+v}{v} \right\rfloor \leq \left\lfloor \frac{u+(2+\sqrt{2})u}{(2+\sqrt{2})u} \right\rfloor = 1. \tag{4.75}$$

Therefore we have $a = 1$, whence $m = u$. And we finally get

$$g(m) = -u^2 + vu - u(u+v) = -2u^2 < 0. \tag{4.76}$$

The proof is now completed. ■

4.5 Minimum Variance Thinning

It is not difficult to see that the asymptotic versions of the inter-arrival times of both random thinning and its deterministic realizations are special cases of the general rv Y defined as follows: Let Z be an $\mathbb{N}$-valued rv with probability mass distribution (pmd)

$$\mathbb{P}\{Z = n\} = p_n, \qquad n = 1, 2, \cdots, \tag{4.77}$$

and let $\{X_n, n = 1, 2, \cdots\}$ be a sequence of i.i.d. $\mathbb{R}_+$-valued rvs independent of Z. The $\mathbb{R}_+$-valued rv Y is then defined by

$$Y = \sum_{n=1}^{Z} X_n. \tag{4.78}$$

It is readily seen that if Z is a geometric rv with parameter p, then Y is the asymptotic version of the inter-arrival times under p-thinning; and if Z is uniformly distributed among $(w_1, \cdots, w_v)$, then Y is the asymptotic version of the inter-arrival times under $\boldsymbol{w}$-thinning. In particular, if $Z = a + 1$ w.p. $\frac{m}{v}$ and $Z = a$ w.p. $\frac{v - m}{v}$ as defined in (4.44), then Y is the asymptotic version of the inter-arrival times under $\boldsymbol{w}^*$-thinning.

As a generalization of Lemma 4.3, we get

$$C^2(Y) = \frac{1}{\mathbb{E}[Z]} C^2(X) + C^2(Z) \tag{4.79}$$

where X is a $\mathbb{R}_+$-valued rv with the same distribution as the rvs $\{X_n, n = 1, 2, \cdots\}$.

Thus we can view Y as the asymptotic version of the inter-arrival times of some sort of thinning. We see from (4.79) that summation of i.i.d. rvs reduces the squared coefficient of variation by a factor of $\mathbb{E}[Z]$, but on the other hand, the randomness of Z introduces additional variability to the squared coefficient of variation by an amount of $C^2(Z)$. Based on this observation, we ask whether there is a pmd which minimizes the variability of Z while achieving the same amount of thinning.

For a given thinning requirement, the problem of minimum variance thinning is thus to find an optimal pmd so as to minimize $C^2(Y)$ over all possible pmds subject to $\mathbb{E}[Z]$ being constant. From (4.77) and (4.79), this is equivalent to solving the

following minimization problem (P) where.

$$
\begin{aligned}
& \textit{minimize} && \sum_{n=1}^{\infty} n^2 p_n && (4.80)\\
(P) \qquad & \textit{subject to} && \sum_{n=1}^{\infty} n p_n = c \geq 1, && (4.81)\\
& && \sum_{n=1}^{\infty} p_n = 1, && (4.82)\\
& && p_n \geq 0, \qquad n = 1, 2, \cdots. && (4.83)
\end{aligned}
$$

The following lemma gives a necessary condition to the solution to (P).

Lemma 4.8 *Let* $\{q_n, n = 1, 2, \cdots\}$ *be any feasible solution to (P). If there exists* $i < j < k$ *such that* $q_i > 0$ *and* $q_k > 0$, *then* $\{q_n, n = 1, 2, \cdots\}$ *is not optimal.*

Proof. Let

$$r = \frac{k-j}{k-i}. \tag{4.84}$$

It is easy to see that $0 < r < 1$. With Δ being given by

$$0 < \Delta \leq \min\left\{\frac{q_i}{r}, \frac{q_k}{1-r}\right\}, \tag{4.85}$$

we define the pmd $\{p_n,\ n = 1, 2, \cdots\}$ by

$$
\begin{cases}
p_i = q_i - r\Delta \\
p_j = q_j + \Delta \\
p_k = q_k - (1-r)\Delta \\
p_n = q_n, \qquad n \neq i, j, k.
\end{cases} \tag{4.86}
$$

It is easy to check that $\{p_n, n = 1, 2, \cdots\}$ is indeed a valid pmd satisfying (4.81). Now simple calculation yields

$$
\begin{aligned}
& \sum_{n=1}^{\infty} n^2 p_n - \sum_{n=1}^{\infty} n^2 q_n \\
&= \Delta[j^2 - ri^2 - (1-r)k^2] \\
&= \Delta \frac{j^2(k-i) - i^2(k-j) - k^2(j-i)}{k-i} \\
&= -\Delta(k-j)(j-i) \\
&< 0.
\end{aligned} \tag{4.87}
$$

From (4.87) we see that the pmd $\{p_n, n = 1, 2, \cdots\}$ strictly improves the cost function. ■

From the proof we see that the larger the value of Δ, the greater the improvement, so that the value of Δ should be chosen to be $\min\left\{\frac{q_i}{r}, \frac{q_k}{1-r}\right\}$ which makes one of p_i or p_k zero.

The following theorem gives the solution to (P).

Theorem 4.9 *If $l \leq c < l+1$ for some integer l, then the optimal solution to (P) is given by*

$$\begin{cases} p_l = 1 - (c - l) \\ p_{l+1} = c - l \\ p_n = 0, \qquad n \neq l, l+1. \end{cases} \tag{4.88}$$

Proof. We first show that any solution to (P) cannot have positive components with indices smaller than l or greater than $l+1$. Let $\{q_n, n = 1, 2, \cdots\}$ be any feasible solution to (P). We observe that if $q_i > 0$ for some $i < l$, then there must be some $k > l$ such that $q_k > 0$, for otherwise (4.81) cannot be satisfied. Thus we can pick any j such that $i < j < k$ and apply Lemma 4.8 to conclude that $\{q_n, n = 1, 2, \cdots\}$ is not optimal. Similarly, we can show that the solution to (P) cannot have positive components with indices greater than $l + 1$. Therefore, the solution to (P) must be of the form

$$p_l \geq 0, \; p_{l+1} \geq 0, \quad \text{and} \quad p_n = 0, \; n \neq l, l+1. \tag{4.89}$$

Together with the feasibility conditions, the optimality of (4.88) is automatic. ■

It is easy to extend Theorem 4.9 to conclude that the pmd given by (4.88) not only minimizes (4.80) under the constraints (4.81) – (4.83) but also minimizes the following quantity under the same constraints

$$\sum_{n=1}^{\infty} \phi(n) p_n, \tag{4.90}$$

with $\phi : \mathbb{R}_+ \to \mathbb{R}$ being any convex function. Thus the minimum variance thinning is actually the minimum variability thinning under much broader sense.

Compare (4.88) with the asymptotic version of the inter-arrival times of the $\boldsymbol{w}^*$-thinning, it is not surprising to find that the $\boldsymbol{w}^*$-thinning is just the minimum variance thinning.

CHAPTER 5

THE PEAKEDNESS FUNCTIONALS

5.1 The Peakedness Functionals

The squared coefficient of variation is a simple and useful measure of burstiness as we have seen from Chapters 2 and 4. It has, however, a drawback – it cannot reflect the correlations between successive inter-arrival times. To overcome this drawback, the squared coefficient of variation can be generalized to involve more than one inter-arrival time in the definition as in the case of the index of dispersion for intervals (IDI) [16, 32]. However, the analysis involving IDI becomes more complicated.

Eckberg [19, 20] proposed another measure for burstiness by generalizing the notion of peakedness factor [59] to the peakedness functional. The peakedness functional of a point process is defined as follows: Consider an infinite server queue with arrivals from a point process. All servers in the infinite server group are independent with the same service time distribution G; let $\bar{G}$ denote the complementary distribution of G, i.e., $\bar{G} = 1 - G$. Denote by $K(t)$ the number of busy servers in the infinite server group at time $t \geq 0$.

Definition 5.1 *The peakedness functional of a point process is given by*

$$z(\bar{G}) = \lim_{t \to \infty} \frac{\mathrm{Var}\,(K(t))}{\mathbb{E}\,[K(t)]} \tag{5.1}$$

provided the limit exists.

It is intuitive that the more bursty the arrival process is, the more variable the number of busy servers in the infinite server group will be, and the more busy servers there will be. This is the basic idea behind the definition of the peakedness functionals which can be viewed as the limit of the product of the squared coefficient of variation of $K(t)$ and the mean of $K(t)$. We see from this definition that the

peakedness functional implicitly reflects the correlation between arrivals through the effect on the infinite server group. From Definition 5.1, The so-called peakedness factor is now obtained from (5.1) by taking G to be an exponential service time distribution, i.e.,

$$G(t) = 1 - e^{-\mu t}, \qquad t \geq 0. \tag{5.2}$$

As a result, we can refer to peakedness factor as the exponential peakedness, and write $z(\mu)$ to denote $z(\bar{G})$ with G being given by (5.2), to emphasize that it is a function of μ. Exponential peakedness arises in the equivalent random method which is well known in teletraffic engineering [59]. The peakedness functionals of stationary processes were derived in [19]. In the sequel, we shall only study the peakedness functionals of renewal processes. In appendix E, we give another derivation of the peakedness functional for renewal processes using a renewal argument. For easy reference, we provide the formula in the following lemma.

Lemma 5.1 *The peakedness functional of a renewal process with arrival epochs* $\{a_n, n = 1, 2, \cdots\}$ *is given by*

$$z(\bar{G}) = 1 - \frac{\lambda}{\mu} + \sum_{n=1}^{\infty} \mathbb{E}\left[H(a_n)\right] \tag{5.3}$$

$$= 1 - \frac{\lambda}{\mu} + \int_0^{\infty} H(x) dM(x) \tag{5.4}$$

where λ *and* μ *are the arrival and service rates, respectively,* M *is the renewal function of the renewal process, and* $H : \mathbb{R}_+ \to \mathbb{R}_+$ *is given by*

$$H(t) = 2\mu \int_0^{\infty} \bar{G}(t+y)\bar{G}(y) dy, \qquad t \geq 0. \tag{5.5}$$

If G is given by (5.2), then

$$H(t) = e^{-\mu t}, \qquad t \geq 0, \tag{5.6}$$

and (5.4) simplifies to yield an explicit formula for the exponential peakedness

Corollary 5.1 *The exponential peakedness for renewal process is given by*

$$z(\mu) = \frac{1}{1 - A^*(\mu)} - \frac{\lambda}{\mu} \tag{5.7}$$

where $A^*(\cdot)$ is the Laplace-Stieltjes transform of the inter-arrival time of the renewal process.

We now state a property of the peakedness functional. Let X and X' be the generic rvs of the inter-arrival times of two renewal processes, and let z and z' be their peakedness functionals.

Theorem 5.1 *If H is a convex function, then*

$$X \leq_{cx} X' \quad \textit{implies} \quad z(\bar{G}) \leq z'(\bar{G}). \tag{5.8}$$

Proof. If $X \leq_{cx} X'$, then by Theorem 2.2.4 or Proposition 2.2.5 in [55] we see that $a_m \leq_{cx} a'_m$ for all $m = 1, 2, \cdots$, and therefore (5.1) follows from (5.3) and from the definition of convex ordering. ■

Since $\mathbb{E}[X] \leq_{cx} X$, Theorem 5.1 is readily translated to the following corollary which validates the intuition that the deterministic arrival process is the least bursty.

Corollary 5.2 *If H is a convex function, then the deterministic arrival process has the smallest $z(\bar{G})$ value among all renewal processes with the same mean.*

There are many examples of cdfs for which H is a convex function, e.g., the exponential distribution and the degenerated deterministic distribution. However, there are cdfs for which the convexity of H fails. For instance, if $\bar{G}$ has the form

$$\bar{G}(x) = (1 - x^2)^+, \qquad x \geq 0, \tag{5.9}$$

then H is neither convex nor concave. In fact, it can be seen that H can never be a concave function.

5.2 Superposition of Point Processes

In this section, we shall derive the peakedness functional for the superposition of point processes. Let $\{\boldsymbol{A}^{(i)},\ i = 1, \cdots, n\}$ be n mutually independent point processes. The superposition of the n point processes is a point process $\boldsymbol{A}$ which counts the arrivals from all the n point processes. It is well known that $\boldsymbol{A}$ is in general not a renewal process [42].

For any distribution function G, let $z(\bar{G})$ denote the peakedness functional of $\boldsymbol{A}$, and let $z_i(\bar{G})$ denote the peakedness functional of $\boldsymbol{A}^{(i)}$, $i = 1, \cdots, n$, respectively. Consider an infinite server group which serves $\boldsymbol{A}$. For $i = 1, \cdots, n$, let $K^{(i)}(t)$ be the number of busy servers in the infinite server group at time $t \geq 0$ serving arrivals from $\boldsymbol{A}^{(i)}$. The total number of busy servers in the infinite server group is then given by

$$K(t) = \sum_{i=1}^{n} K^{(i)}(t) \qquad t \geq 0. \tag{5.10}$$

Because all servers in the infinite server group are independent and identical, and the n point processes are mutually independent, the processes $\{K^{(i)}(t),\ t \geq 0\}$, $i = 1, \cdots, n$ are mutually independent. Therefore, we can view the whole infinite server group as n independent infinite subgroups, each serving arrivals from one of the n point processes.

Let

$$\alpha_i(t) = \frac{\mathbb{E}\left[K^{(i)}(t)\right]}{\sum_{j=1}^{n} \mathbb{E}\left[K^{(j)}(t)\right]}, \qquad t \geq 0,\ i = 1, \cdots, n, \tag{5.11}$$

and denote

$$\alpha_i = \lim_{t \to \infty} \alpha_i(t), \qquad i = 1, \cdots, n \tag{5.12}$$

whenever it exists. In particular, if the n processes are renewal processes, then by Little's law we get

$$\lim_{t \to \infty} \mathbb{E}\left[K^{(i)}(t)\right] = \lambda_i m(G), \qquad i = 1, \cdots, n \tag{5.13}$$

where $m(G)$ is the mean of a rv with distribution G. Therefore, we have

$$\alpha_i = \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j}, \qquad i = 1, \cdots, n \tag{5.14}$$

where λ_i is the arrival rate of $\boldsymbol{A}^{(i)}$, $i = 1, \cdots, n$. Clearly we have $\alpha_i \geq 0, i = 1, \cdots, n$, and $\sum_{i=1}^{n} \alpha_i = 1$.

Using the independence of the rvs $\{K^{(i)}(t), i = 1, \cdots, n\}$, we conclude from (5.10) that

$$\mathbb{E}\left[K(t)\right] = \sum_{i=1}^{n} \mathbb{E}\left[K^{(i)}(t)\right], \qquad t \geq 0 \tag{5.15}$$

and

$$Var\left(K(t)\right) = \sum_{i=1}^{n} Var\left(K^{(i)}(t)\right), \qquad t \geq 0, \tag{5.16}$$

so that

$$\frac{Var\left(K(t)\right)}{\mathbb{E}\left[K(t)\right]} = \sum_{i=1}^{n} \alpha_i(t) \frac{Var\left(K^{(i)}(t)\right)}{\mathbb{E}\left[K^{(i)}(t)\right]}. \tag{5.17}$$

Taking limit in (5.17) as $t \to \infty$, we conclude to the following fact.

Lemma 5.2 *The peakedness functional of the superposition of n mutually independent point processes is given by*

$$z(\bar{G}) = \sum_{i=1}^{n} \alpha_i z_i(\bar{G}). \tag{5.18}$$

So, the peakedness functional of superposition of point processes is the weighted average of the peakedness functional of each of the individual point processes. It worths noting that the weight for each individual contribution to the overall peakedness functional is proportional to the mean number of busy servers serving each point process. In the renewal case, the weight is simply the ratio of the arrival rate independent of the infinite server group.

The following corollaries are immediate consequences of Lemma 5.2.

Corollary 5.3 *The peakedness functional of the superposition of n mutually independent point processes satisfies the bounds*

$$\min\{z_i(\bar{G}),\ i = 1, \cdots, n\} \leq z(\bar{G}) \leq \max\{z_i(\bar{G}),\ i = 1, \cdots, n\}. \tag{5.19}$$

In particular, if the n point processes are identically distributed, then we have

Corollary 5.4 *Superposition of independent and identically distributed point processes preserves the peakedness functional, i.e.,*

$$z(\bar{G}) = z_i(\bar{G}), \qquad i = 1, \cdots, n. \tag{5.20}$$

If the n i.i.d. processes are Poisson processes, then the superposition is still a Poisson process with different rate, and Corollary 5.4 is clear. If the n i.i.d. processes are non-Poisson processes, the resulting superposition process will usually have different distribution. In this case, Corollary 5.4 implies that although superposition might have changed the "variability" of the process, this effect cannot be seen by the servers in the infinite server group. In this sense, the peakedness functional is insensitive to superposition.

Now we consider the effect of correlation between the n point processes on the peakedness functional of the superposition process. To this end, we assume that the n point processes have the same marginal distributions as in the case discussed before but that they may be correlated. For simplicity, we take $n = 2$. This time, the two infinite server subgroups are no longer independent, with the rvs $K^{(1)}(t)$ and $K^{(2)}(t)$ possibly correlated for all $t \geq 0$. Calculations yield

$$\frac{Var\,(K(t))}{\mathbb{E}\,[K(t)]} = \sum_{i=1}^{2} \alpha_i(t) \frac{Var\left(K^{(i)}(t)\right)}{\mathbb{E}\,[K^{(i)}(t)]} + \frac{Cov(K^{(1)}(t), K^{(2)}(t))}{\mathbb{E}\,[K^{(1)}(t)] + \mathbb{E}\,[K^{(2)}(t)]}, \qquad t \geq 0. \tag{5.21}$$

Assume the existence of appropriate limits, with

$$\gamma_{12} = \lim_{t \to \infty} \frac{Cov(K^{(1)}(t), K^{(2)}(t))}{\mathbb{E}\,[K^{(1)}(t)] + \mathbb{E}\,[K^{(2)}(t)]}. \tag{5.22}$$

By taking limit in (5.21) we get

$$z(\bar{G}) = \alpha_1 z_1(\bar{G}) + \alpha_2 z_2(\bar{G}) + \gamma_{12}, \tag{5.23}$$

or equivalently

$$z(\bar{G}) = \tilde{z}(\bar{G}) + \gamma_{12}, \tag{5.24}$$

where $\tilde{z}(\bar{G})$ is the peakedness functional of the superposition process assuming the two processes to be independent.

Another measure which have very similar behavior as the peakedness functional in characterizing the superposition process is the index of dispersion for counts (IDC) [16, 32]. We shall not discuss IDC further in this thesis.

5.3 p-Thinning

Let $z(\bar{G})$ be the peakedness functional of the original point process $\boldsymbol{A}$, and for $0 < p \leq 1$, let $z_p(\bar{G})$ be the peakedness functional of the p-thinning $\boldsymbol{A}_p$ of $\boldsymbol{A}$. Assume that the original point process is a renewal process. To find the peakedness functional $z_p(\bar{G})$ of the p-thinning, we observe the following. The effect of p-thinning on the number of busy servers in the infinite server group is equivalent to forcing the arrivals which are thinned out to have zero service time in the original system. Thus the following relationship holds

$$z_p(\bar{G}) = z(\bar{G}_p), \qquad 0 < p \leq 1 \tag{5.25}$$

with G_p denoting the distribution function of the modified service time σ_p given by

$$\sigma_p = U_p \sigma \tag{5.26}$$

where σ is the original service time with distribution function G and U_p is a Bernoulli rv with parameter p independent of σ. Simple calculations yield

$$\bar{G}_p(x) = p\bar{G}(x), \qquad x \geq 0. \tag{5.27}$$

For this modified system, we find that

$$\frac{1}{\mu_p} = \int_0^\infty \bar{G}_p(x)dx = \frac{p}{\mu} \tag{5.28}$$

and

$$\begin{aligned} H_p(t) &= 2\mu_p \int_0^\infty \bar{G}_p(x+t)\bar{G}_p(x)dx \\ &= \frac{2\mu}{p} \int_0^\infty p^2 \bar{G}(x+t)\bar{G}(x)dx \\ &= pH(t), \qquad t \geq 0. \end{aligned} \tag{5.29}$$

Therefore, substituting (5.28) and (5.29) into (5.3) we get

$$\begin{aligned} z(\bar{G}_p) &= 1 - \frac{\lambda}{\mu_p} + \sum_{m=1}^{\infty} \mathbb{E}\left[H_p(a_m)\right] \\ &= (1-p) + p - p\frac{\lambda}{\mu} + p\sum_{m=1}^{\infty} \mathbb{E}\left[H(a_m)\right]. \end{aligned} \tag{5.30}$$

By defining $z_0(\bar{G}) = 1$, and after simplifying, we have

Lemma 5.3 *The peakedness functional of p-thinning is given by*

$$z_p(\bar{G}) = 1 - p + pz(\bar{G}) \qquad 0 \leq p \leq 1. \tag{5.31}$$

It is then readily seen that

Lemma 5.4 *The function $p \to z_p(\bar{G})$ is monotone and continuous on* $[0, 1]$.

For $p = 1$, we find

$$z_1(\bar{G}) = z(\bar{G}) \tag{5.32}$$

as expected. For $p = 0$, the resulting process is an empty process with no arrivals. However, if we view $z_0(\bar{G})$ as the limit of the peakedness functionals resulting from repeatedly random thinning, this limit is one. This fact suggests that after repeatedly random thinning, the resulting process behaves more and more like a Poisson process.

From (5.31), it is easy to conclude to the following theorem.

Theorem 5.2 *For $0 < p \leq 1$, we have*

$$z_p(\bar{G}) \leq z(\bar{G}) \quad \textit{iff} \quad z(\bar{G}) \geq 1, \tag{5.33}$$

with equality holding if and only if $z(\bar{G}) = 1$.

Note also from (5.31) that $z(\bar{G}) \geq 1$ implies $z_p(\bar{G}) \geq 1$, and $z(\bar{G}) \leq 1$ implies $z_p(\bar{G}) \leq 1$. This fact will be used later. Theorem 5.2 can be interpreted in the following way. Random thinning of point process can reduce peakedness if and only if the original point process is burstier than the Poisson process, otherwise, it can only increase peakedness. If the original process is a Poisson process, then $z(\bar{G}) = 1$ and the p-thinning is still a Poisson process, whence $z_p(\bar{G}) = 1$ for all $0 < p \leq 1$. This result parallels our previous result (Theorem 4.1) in Chapter 4.

5.4 w-Thinning

We have not found a way to characterize the peakedness for general $\boldsymbol{w}$-thinnings as defined in Chapter 4. In this section, we only consider the case where $p = \dfrac{v}{u+v}$ with $v = 1$ and $u = m$ for some $m = 1, 2, \cdots$. In this case, $\boldsymbol{w}$ is a scalar given by

$$\boldsymbol{w} = m + 1, \tag{5.34}$$

and we write z_m for the peakedness functional of the $\boldsymbol{w}$-thinning instead of $z_{\boldsymbol{w}}$. We already know from (4.17) that the resulting $\boldsymbol{w}$-thinning is a renewal process with arrival epochs $\{a'_n = a_{(m+1)n},\ n = 1, 2, \cdots\}$. For this renewal process, we find

$$\rho_m = \frac{\rho}{m+1}, \tag{5.35}$$

and

$$H_m(a'_n) = H(a_{(m+1)n}), \qquad n = 1, 2, \cdots. \tag{5.36}$$

Substituting (5.35) and (5.36) into (5.3), we find

Lemma 5.5 *The peakedness functional $z_m(\bar{G})$ of the $\boldsymbol{w}$-thinning with $\boldsymbol{w}$ defined by (5.34) is given by*

$$z_m(\bar{G}) = 1 - \frac{\rho}{m+1} + \sum_{n=1}^{\infty} \mathbb{E}\left[H(a_{(m+1)n})\right]. \tag{5.37}$$

The exponential peakedness is then easily evaluated.

Corollary 5.5 *The exponential peakedness $z_m(\mu)$ of the $\boldsymbol{w}$-thinning with $\boldsymbol{w}$ defined by (5.34) is given by*

$$z_m(\mu) = \frac{1}{1 - (A^*(\mu))^{m+1}} - \frac{\rho}{m+1}. \tag{5.38}$$

Unlike $z_p(\bar{G})$, it is not true in general that $z_m(\bar{G})$ is monotonic in m. In fact, for $A^*(\mu)$ not too small, $z_m(\mu)$ is first decreasing in m, and then increasing in m for m large enough. And

$$\lim_{m \to \infty} z_m(\mu) = 1. \tag{5.39}$$

Now we look at how the deterministic realization (5.34) of random thinning affects the peakedness functional. We consider only the exponential peakedness. From (5.7) and (5.38), we observe that

$$\begin{aligned} z(\mu) - z_m(\mu) &= \left(\frac{1}{1-A^*(\mu)} - \rho\right) - \left(\frac{1}{1-(A^*(\mu))^{m+1}} - \frac{\rho}{m+1}\right) \\ &= \left(\frac{A^*(\mu)}{1-A^*(\mu)} - \frac{(A^*(\mu))^{m+1}}{1-(A^*(\mu))^{m+1}}\right) - \frac{m\rho}{m+1} \\ &= \frac{A^*(\mu)(1-(A^*(\mu))^m)}{(1-A^*(\mu))(1-(A^*(\mu))^{m+1})} - \frac{m\rho}{m+1} \end{aligned} \tag{5.40}$$

in preparation to the following theorem.

Theorem 5.3 *For $m = 1, 2, \cdots$, and $\rho > 0$, there is a unique α_m in $(0,1)$ such that $A^*(\mu) = \alpha_m$ implies $z(\mu) = z_m(\mu)$. Furthermore,*

$$z(\mu) < z_m(\mu) \quad \textit{iff} \quad 0 \le A^*(\mu) < \alpha_m \tag{5.41}$$

and

$$z(\mu) > z_m(\mu) \quad \textit{iff} \quad \alpha_m < A^*(\mu) < 1. \tag{5.42}$$

Proof. Fix $m = 1, 2, \cdots$ and define

$$h_m(\alpha) = \frac{\alpha(1-\alpha^m)}{(1-\alpha)(1-\alpha^{m+1})}, \qquad 0 \le \alpha < 1. \tag{5.43}$$

Lemma 5.6 below shows that $\alpha \to h_m(\alpha)$ is a strictly monotone increasing mapping $[0,1) \to [0,\infty)$. Invoking the strict monotonicity of h_m, we see that for any fixed $\rho > 0$, the equation $h_m(\alpha) = \dfrac{m\rho}{m+1}$ has a unique solution α_m so that

$$h_m(\alpha) < \frac{m\rho}{m+1} \quad \text{iff} \quad 0 \le \alpha < \alpha_m \tag{5.44}$$

and

$$h_m(\alpha) > \frac{m\rho}{m+1} \quad \text{iff} \quad \alpha_m < \alpha < 1. \tag{5.45}$$

Theorem 5.3 then follows by setting $\alpha = A^*(\mu)$. ■

Lemma 5.6 *The mapping $\alpha \to h_m(\alpha)$ defined in (5.43) is a strictly monotone increasing mapping $[0,1) \to [0,\infty)$.*

Proof. Set

$$g_1(\alpha) = \frac{1-\alpha^m}{1-\alpha} \quad \text{and} \quad g_2(\alpha) = \frac{\alpha}{1-\alpha^{m+1}}, \qquad 0 \le \alpha < 1. \tag{5.46}$$

On the range $(0,1)$, $g_1(\alpha) > 0$, $g_2(\alpha) > 0$ and simple calculations show that

$$g_1'(\alpha) = \left(\frac{1-\alpha^m}{1-\alpha}\right)' = \left(\sum_{i=1}^{m}\alpha^i\right)' = \sum_{i=1}^{m} i\alpha^{i-1} > 0 \tag{5.47}$$

and

$$g_2'(\alpha) = \left(\frac{\alpha}{1-\alpha^{m+1}}\right)' = \frac{1-\alpha^{m+1}+\alpha(m+1)\alpha^m}{(1-\alpha^{m+1})^2} = \frac{1+m\alpha^{m+1}}{(1-\alpha^{m+1})^2} > 0. \tag{5.48}$$

In conclusion, we get

$$h_m'(\alpha) = g_1'(\alpha)g_2(\alpha) + g_1(\alpha)g_2'(\alpha) > 0, \qquad 0 < \alpha < 1. \tag{5.49}$$

So, $\alpha \to h_m(\alpha)$ is strictly increasing with $h_m(0) = 0$ and $\lim_{\alpha \nearrow 1} h_m(\alpha) = +\infty$. ■

In terms of $z(\mu)$, (5.41) and (5.42) are equivalent to

$$z(\mu) < z_m(\mu) \ \text{ iff } \ z(\mu) < 1 - \rho + \frac{\alpha_m}{1-\alpha_m}, \tag{5.50}$$

and

$$z(\mu) > z_m(\mu) \ \text{ iff } \ z(\mu) > 1 - \rho + \frac{\alpha_m}{1-\alpha_m}, \tag{5.51}$$

respectively

Recall Theorem 5.2 and the remark immediately after it. Random thinning can reduce peakedness only if the original process has peakedness greater than one. Furthermore, the resulting peakedness can never be less than one. However, as we now show, the deterministic realization of random thinning can reduce peakedness even if the original peakedness is already less than one. This is seen from (5.51) and the following lemma.

Lemma 5.7 *With the notation of Theorem 5.3, we always have*

$$\frac{\alpha_m}{1-\alpha_m} < \rho. \tag{5.52}$$

Proof. Since $\rho = \dfrac{m+1}{m} h_m(\alpha_m)$ by the definition of α_m, we find

$$\begin{aligned} 1-\left(1-\rho+\frac{\alpha_m}{1-\alpha_m}\right) &= \rho - \frac{\alpha_m}{1-\alpha_m} \\ &= \frac{m+1}{m}\frac{\alpha_m(1-\alpha_m^m)}{(1-\alpha_m)(1-\alpha_m^{m+1})} - \frac{\alpha_m}{1-\alpha_m} \\ &= \frac{(m+1)\alpha_m(1-\alpha_m^m) - m\alpha_m(1-\alpha_m^{m+1})}{m(1-\alpha_m)(1-\alpha_m^{m+1})}. \end{aligned} \tag{5.53}$$

We observe that

$$\begin{aligned} (m+1)(1-\alpha_m^m) - m(1-\alpha_m^{m+1}) &= 1-(m+1)\alpha_m^m + m\alpha_m^{m+1} \\ &= (1-\alpha_m^m) - m\alpha_m^m(1-\alpha_m) \\ &= (1-\alpha_m)\left[\sum_{i=0}^{m-1}\alpha_m^i - m\alpha_m^m\right] \\ &= (1-\alpha_m)\sum_{i=0}^{m-1}(\alpha_m^i - \alpha_m^m) \\ &> 0, \end{aligned} \tag{5.54}$$

and the lemma follows from (5.53). ∎

We have seen from Theorem 5.3 that if $A^*(\mu) > \alpha_m$, then the peakedness can be reduced through $\boldsymbol{w}$-thinning. Combining with the comment made after Corollary 5.5, we see that it is possible to reduce the peakedness to less than one. We shall soon see that if the original process is already smooth enough (i.e., $z(\mu) < 1-\rho+\dfrac{\alpha_m}{1-\alpha_m} < 1$), then the peakedness after $\boldsymbol{w}$-thinning will remain less than one.

Lemma 5.8 *If $A^*(\mu) < \alpha_m$, then we always have $z_m(\mu) < 1$.*

Proof. From (5.38), we want to show that

$$\frac{(A^*(\mu))^{m+1}}{1-(A^*(\mu))^{m+1}} - \frac{\rho}{m+1} < 0. \tag{5.55}$$

Since $A^*(\mu) < \alpha_m$ by Theorem 5.3 and $\rho = \frac{m+1}{m} h_m(\alpha_m)$, we have

$$\begin{aligned}
\frac{(A^*(\mu))^{m+1}}{1-(A^*(\mu))^{m+1}} - \frac{\rho}{m+1} &< \frac{\alpha_m^{m+1}}{1-\alpha_m^{m+1}} - \frac{\rho}{m+1} \\
&= \frac{\alpha_m^{m+1}}{1-\alpha_m^{m+1}} - \frac{1}{m} h_m(\alpha_m) \\
&= \frac{\alpha_m^{m+1}}{1-\alpha_m^{m+1}} - \frac{1}{m} \frac{\alpha_m(1-\alpha_m^{m+1})}{(1-\alpha_m)(1-\alpha_m^{m+1})} \\
&= \frac{\alpha_m}{m(1-\alpha_m^{m+1})} [m\alpha_m^m - \sum_{i=0}^{m} \alpha_m^i] \\
&< 0. \qquad (5.56)
\end{aligned}$$

■

5.5 Comparison of p-Thinning and w-Thinning

Now we compare random thinning with its deterministic realization. There is no general result for $z_p(\bar{G})$ and $z_m(\bar{G})$ for all G. Comparison, however, is available in some cases when H is a convex function. We have already shown that the inter-arrival times of the processes resulting from random thinning and from its deterministic realization are distributed according to the rvs Y_p and Y_m defined by (4.2) and (4.17), respectively. Let $\{Y_p^{(i)},\ i = 1, 2, \cdots\}$ and $\{Y_m^{(i)},\ i = 1, 2, \cdots\}$ be two sequences of i.i.d. rvs distributed according to Y_p and Y_m, respectively. Then the n^{th} arrival epoch after random thinning (resp. its deterministic realization) are given according to the rv

$$\sum_{i=1}^{n} Y_p^{(i)} \ (\text{resp. } \sum_{i=1}^{n} Y_m^{(i)}), \qquad n = 1, 2, \cdots. \tag{5.57}$$

A similar argument as in the proof of Lemma 8.6.7 in [52] shows that

Lemma 5.9 *We have*

$$\sum_{i=1}^{n} Y_m^{(i)} \leq_{cx} \sum_{i=1}^{n} Y_p^{(i)}, \qquad n = 1, 2, \cdots. \tag{5.58}$$

Lemma 5.9 establishes relations between the arrival epochs in the two processes and leads to the following theorem.

Theorem 5.4 *If H is convex, then we have*

$$z_p(\bar{G}) \geq z_m(\bar{G}). \tag{5.59}$$

Proof. By the convexity of H and Lemma 5.9, with $p = \dfrac{1}{m+1}$, we find that

$$\begin{aligned} z_p(\bar{G}) &= 1 - \frac{\rho}{m+1} + \sum_{n=1}^{\infty} \mathbb{E}\left[H(\sum_{i=1}^{n} Y_p^{(i)})\right] \\ &\geq 1 - \frac{\rho}{m+1} + \sum_{n=1}^{\infty} \mathbb{E}\left[H(\sum_{i=1}^{n} Y_m^{(i)})\right] \\ &= z_m(\bar{G}). \end{aligned} \tag{5.60}$$

■

Thus, in the case where H is a convex function, the deterministic realization of random thinning with $p = \dfrac{1}{m+1}$ is always better than the random thinning.

As an example, we consider the exponential peakedness. From (5.7), (5.31), and (5.38), we compute the difference

$$z_p(\mu) - z_m(\mu) = \left(\frac{1}{m+1}\right)\left(\frac{A^*(\mu)}{1-A^*(\mu)}\right) - \frac{(A^*(\mu))^{m+1}}{1-(A^*(\mu))^{m+1}} \tag{5.61}$$

with the consequence that

$$z_p(\mu) \geq z_m(\mu) \quad \text{iff} \quad \left(\frac{1}{m+1}\right)\left(\frac{A^*(\mu)}{1-A^*(\mu)}\right) \geq \frac{(A^*(\mu))^{m+1}}{1-(A^*(\mu))^{m+1}}. \tag{5.62}$$

Since $0 \leq A^*(\mu) \leq 1$, this condition is always true as we now show. Define

$$f_n(\alpha) = \frac{1}{n}\frac{\alpha}{1-\alpha} - \frac{\alpha^n}{1-\alpha^n}, \quad 0 \leq \alpha \leq 1, \ n = 1, 2, \cdots. \tag{5.63}$$

For $n = 1$, we have $f_1(\alpha) = 0$, while for $n = 2, 3, \cdots$, we can rewrite $f_n(\alpha)$ as

$$\begin{aligned} f_n(\alpha) &= \frac{\alpha}{1-\alpha}\left[\frac{1}{n} - \frac{\alpha^{n-1}}{1+\alpha+\cdots+\alpha^{n-1}}\right] \\ &= \left(\frac{\alpha}{1-\alpha}\right)\frac{1+\alpha+\cdots+\alpha^{n-1}-n\alpha^{n-1}}{n(1+\alpha+\cdots+\alpha^{n-1})} \geq 0. \end{aligned} \tag{5.64}$$

CHAPTER 6

APPLICATIONS OF THE PEAKEDNESS FUNCTIONALS

6.1 When to Perform Rate Control? — The Random Case

With the tools developed in Chapter 5, we now study a problem which arises in network design. It is widely accepted that in high-speed networks rate control should be employed to avoid congestion from occurring. The question is whether we should perform rate control to each individual source, or to perform rate control at the trunk level [30]. This problem was discussed earlier in Chapter 3 in the context of leaky bucket. In this chapter, we consider the same systems, system I and system II, with rate control being performed by thinning (see Fig. 3.5). In system I, we multiplex individual sources on a trunk and then perform rate control on the resulting traffic. In system II, we perform rate control on each individual source before multiplex them.

Multiplexing can be modeled by superposition. In this section, we model the rate control by random thinning. There are many considerations about this problem, e.g., implementation costs, fairness, and the burstiness of the resulting traffic. We shall consider only the burstiness of the resulting traffic. We assume that the n sources being considered are mutually independent, each modelied by a renewal process. Throughout this chapter, we use the notation of Chapter 5 without further explanation.

We now fix $0 < p \leq 1$. Denote the peakedness functional for each individual source by z_i, $i = 1, \cdots, n$. For system I, the peakedness functional for the traffic on the trunk, denoted by z_T, is given by Lemma 5.2 in the form

$$z_T(\bar{G}) = \sum_{i=1}^{n} \alpha_i z_i(\bar{G}) \tag{6.1}$$

where α_i, $i = 1, \cdots, n$, are given by (5.14). After rate control, the peakedness

functional of system I, denoted by z_I, is then given via Lemma 5.3 by

$$\begin{aligned} z_I(\bar{G}) &= 1-p+pz_T(\bar{G}) \\ &= 1-p+p\sum_{i=1}^{n}\alpha_i z_i(\bar{G}) \\ &= \sum_{i=1}^{n}\alpha_i[1-p+pz_i(\bar{G})] \end{aligned} \tag{6.2}$$

On the other hand, in system II, we perform rate control before multiplexing. By Lemma 5.3, the peakedness functionals, $\tilde{z}_i$, $i=1,\cdots,n$, after applying rate control on each of the n sources are given by

$$\tilde{z}_i(\bar{G}) = 1-p+pz_i(\bar{G}), \qquad i=1,\cdots,n. \tag{6.3}$$

After multiplexing, we use Lemma 5.2 to conclude that the peakedness functional of system II, denoted by z_{II}, is

$$\begin{aligned} z_{II}(\bar{G}) &= \sum_{i=1}^{n}\tilde{\alpha}_i\tilde{z}_i(\bar{G}) \\ &= \sum_{i=1}^{n}\tilde{\alpha}_i[1-p+pz_i(\bar{G})], \end{aligned} \tag{6.4}$$

where

$$\tilde{\alpha}_i = \lim_{t\to\infty}\frac{\mathbb{E}\left[\tilde{K}^{(i)}(t)\right]}{\sum_{i=1}^{n}\mathbb{E}\left[\tilde{K}^{(i)}(t)\right]}, \qquad i=1,\cdots,n, \tag{6.5}$$

with $\tilde{K}^{(i)}(t)$ denoting the number of busy servers at time t serving the i^{th} source after rate control. Fix $i=1,\cdots,n$, we are going to find $\tilde{\alpha}_i$. Let $\{u_m, m=1,2,\cdots\}$ be a sequence of i.i.d. Bernoulli rvs with parameter p, which are independent of the i^{th} point process and of the service times of the infinite server group. Therefore, we can write

$$\tilde{K}^{(i)}(t) = \sum_{m=1}^{\infty}\mathbf{1}\{\sigma_m > t-a_m\}u_m, \qquad t\geq 0. \tag{6.6}$$

Thus

$$\begin{aligned} \mathbb{E}\left[\tilde{K}^{(i)}(t)\right] &= \mathbb{E}\left[\sum_{m=1}^{\infty}\mathbb{E}\left[\mathbf{1}\{\sigma_m > t-a_m\}|a_1,a_2,\cdots\right]\mathbb{E}\left[u_m|a_1,a_2,\cdots\right]\right] \\ &= \mathbb{E}\left[\sum_{m=1}^{\infty}\mathbb{E}\left[\mathbf{1}\{\sigma_m > t-a_m\}|a_1,a_2,\cdots\right]p\right] \end{aligned}$$

$$
\begin{aligned}
&= p\,\mathbb{E}\left[\sum_{i=1}^{\infty}\mathbf{1}\{\sigma > t - a_i\}\right] \\
&= p\,\mathbb{E}\left[K^{(i)}(t)\right], \qquad t \geq 0. \qquad (6.7)
\end{aligned}
$$

Substituting in the definition of $\tilde{\alpha}_i$ we conclude that

$$
\begin{aligned}
\tilde{\alpha}_i &= \lim_{t\to\infty}\frac{\mathbb{E}\left[\tilde{K}^{(i)}(t)\right]}{\mathbb{E}\left[\tilde{K}(t)\right]} \\
&= \lim_{t\to\infty}\frac{p\,\mathbb{E}\left[K^{(i)}(t)\right]}{p\,\mathbb{E}\left[K(t)\right]} \\
&= \alpha_i, \qquad i = 1,\cdots,n. \qquad (6.8)
\end{aligned}
$$

Finally, combining (6.2), (6.4) and (6.8), we have the following result.

Proposition 6.1 *If the rate control is performed through random thinning, then systems I and II are equivalent in that we have*

$$z_I(\bar{G}) = z_{II}(\bar{G}). \qquad (6.9)$$

This result is not surprising, since everything here being totally random, the order of doing things does not matter. However, things will be different if we want to discriminate some users.

To that end, we consider the following situation. The n sources are identical with rate $\frac{\lambda}{n}$ and peakedness functional z. Suppose we want to thin each source in system II with a different probability $\{p_i,\ i = 1,\cdots,n\}$. To get the same amount of thinning in system I, we must choose $\{p_i,\ i = 1,\cdots,n\}$ so that

$$p = \frac{1}{n}\sum_{i=1}^{n} p_i. \qquad (6.10)$$

By applying the results in Chapter 5, we find that

$$z_I(\bar{G}) = 1 - p + pz(\bar{G}) \qquad (6.11)$$

and

$$z_{II}(\bar{G}) = 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{p_i^2}{p} + \left(\frac{1}{n}\sum_{i=1}^{n}\frac{p_i^2}{p}\right) z(\bar{G}). \qquad (6.12)$$

Therefore, we get from (6.11) and (6.12) that

$$p(z_I(\bar{G}) - z_{II}(\bar{G})) = (1 - z(\bar{G}))\left(\left(\frac{1}{n}\sum_{i=1}^{n} p_i^2\right) - p^2\right). \tag{6.13}$$

By Jensen's inequality, we have

$$\left(\frac{1}{n}\sum_{i=1}^{n} p_i^2\right) - p^2 > 0. \tag{6.14}$$

Therefore, if the source is smooth enough (no more bursty than a Poisson source), then $z_I(\bar{G}) \geq z_{II}(\bar{G})$, and it is advantageous to discriminate some of the sources to benefit the others. However, if the sources are very bursty, there is no reason to discriminate any sources. We conclude to the following proposition.

Proposition 6.2 *If the rate controls for systems I and II are performed through random thinning with parameters* $\{p_i,\ i = 1, \cdots, n\}$ *for system I and* p *(defined by (6.10)) for system II, respectively, then we have*

$$z_I(\bar{G}) \geq z_{II}(\bar{G}) \quad \textit{iff} \quad z(\bar{G}) \leq 1. \tag{6.15}$$

6.2 When to Perform Rate Control? — The Deterministic Case

In this section, we use the setup of Section 6.1 except that the rate control is now assumed to be the deterministic realization of random thinning. We assume that each source produces a Poisson process with rate λ_i, $i = 1, \cdots, n$, and we set $\lambda = \sum_{i=1}^{n} \lambda_i$. Also we assume that $p = \dfrac{1}{m+1}$ for some $m = 1, 2, \cdots$. Under these assumptions, we have $\boldsymbol{w} = m+1$, so to allow us to use results obtained in Chapter 5. In system I, the trunk is seen to be a Poisson process with rate λ, and by Corollary 5.5, we have

$$z_I(\mu) = \frac{(1+\rho)^{m+1}}{(\rho+1)^{m+1} - \rho^{m+1}} - \frac{\rho}{m+1}. \tag{6.16}$$

In system II, we further assume $\lambda_i = \dfrac{\lambda}{n}$, $i = 1, \cdots, n$, the so-called balanced case. By Corollaries 5.4 and 5.5, we have

$$z_{II}(\mu) = \frac{(n+\rho)^{m+1}}{(n+\rho)^{m+1} - \rho^{m+1}} - \frac{\rho}{(m+1)n}. \tag{6.17}$$

From (6.16) and (6.17), we have the following result.

Proposition 6.3 *For n independent and identical Poisson sources, system I outperforms system II if the rate control is performed through the deterministic realization of p-thinning with $p = \dfrac{1}{m+1}$, $m = 1, 2, \cdots$, i.e., we have*

$$z_I(\mu) \leq z_{II}(\mu). \tag{6.18}$$

Proof. From (6.16) and (6.17), we have

$$z_I(\mu) - z_{II}(\mu) \tag{6.19}$$

$$\begin{aligned} &= -\frac{(n-1)\rho}{n(m+1)} + \frac{\rho^{m+1}\left[(\rho+n)^{m+1} - (\rho+1)^{m+1}\right]}{\left[(\rho+1)^{m+1} - \rho^{m+1}\right]\left[(\rho+n)^{m+1} - \rho^{m+1}\right]} \\ &= \frac{(n-1)\rho}{n(m+1)} \left\{ \frac{\rho^m n(m+1)\left[(\rho+n)^{m+1} - (\rho+1)^{m+1}\right]}{(n-1)\left[(\rho+n)^{m+1} - \rho^{m+1}\right]\left[(\rho+1)^{m+1} - \rho^{m+1}\right]} - 1 \right\}. \end{aligned} \tag{6.20}$$

Therefore to show $z_I(\mu) \leq z_{II}(\mu)$, it suffices to show that $B_{m,n}(\rho) \leq 0$ where

$$\begin{aligned} B_{m,n}(\rho) \triangleq\ & \rho^m n(m+1)\left[(\rho+n)^{m+1} - (\rho+1)^{m+1}\right] - \\ & (n-1)\left[(\rho+n)^{m+1} - \rho^{m+1}\right]\left[(\rho+1)^{m+1} - \rho^{m+1}\right]. \end{aligned} \tag{6.21}$$

Upon using the binomial expansion, as the terms in ρ^{m+1} cancelled out, we get

$$\begin{aligned}
B_{m,n}(\rho) &= n(m+1)\rho^m \left[\sum_{i=1}^{m+1} C_i^{m+1} n^i \rho^{m+1-i} - \sum_{i=1}^{m+1} C_i^{m+1} \rho^{m+1-i}\right] \\
&\quad -(n-1)\left[\sum_{i=1}^{m+1} C_i^{m+1} n^i \rho^{m+1-i}\right]\left[\sum_{i=1}^{m+1} C_i^{m+1} \rho^{m+1-i}\right] \\
&= n(m+1)\rho^m \left[\sum_{i=1}^{m+1} C_i^{m+1} n^i \rho^{m+1-i} - \sum_{i=1}^{m+1} C_i^{m+1} \rho^{m+1-i}\right] \\
&\quad -(n-1)\sum_{i=1}^{m+1}\sum_{j=1}^{m+1} C_i^{m+1} C_j^{m+1} n^i \rho^{2m+2-(i+j)} \\
&= \sum_{i=1}^{m+1} C_i^{m+1}(m+1) n^{i+1} \rho^{2m+2-(i+1)} - \sum_{j=1}^{m+1} C_j^{m+1}(m+1) n \rho^{2m+2-(j+1)} \\
&\quad -\sum_{i=1}^{m+1}\sum_{j=1}^{m+1} C_i^{m+1} C_j^{m+1} n^{i+1} \rho^{2m+2-(i+j)} + \sum_{i=1}^{m+1}\sum_{j=1}^{m+1} C_i^{m+1} C_j^{m+1} n^i \rho^{2m+2-(i+j)} \\
&= -\sum_{i=1}^{m+1}\sum_{j=2}^{m+1} C_i^{m+1} C_j^{m+1} n^{i+1} \rho^{2m+2-(i+j)} + \sum_{i=2}^{m+1}\sum_{j=1}^{m+1} C_i^{m+1} C_j^{m+1} n^i \rho^{2m+2-(i+j)} \\
&= \sum_{k=3}^{2m+2} \rho^{2m+2-k} \left[\sum_{i+j=k, 1\le i\le m+1, 2\le j\le m+1} C_i^{m+1} C_j^{m+1}(n^j - n^{i+1})\right] \\
&= \sum_{k=3}^{2m+2} \rho^{2m+2-k} \left[\sum_{i=\max\{1,k-m-1\}}^{\min\{k-2,m+1\}} C_i^{m+1} C_{k-i}^{m+1}(n^{k-i} - n^{i+1})\right]. \qquad (6.22)
\end{aligned}$$

Now we show that

$$C_k \triangleq \sum_{i=\max\{1,k-m-1\}}^{\min\{k-2,m+1\}} C_i^{m+1} C_{k-1}^{m+1}(n^{k-1} - n^{i+1}) \le 0, \qquad k = 3, \cdots, 2m+2. \quad (6.23)$$

For $k = 3$,

$$C_3 = C_1^{m+1} C_2^{m+1}(n^2 - n^2) = 0. \qquad (6.24)$$

For $4 \le k < m+3$,

$$\begin{aligned}
C_k &= \sum_{i=1}^{k-2} C_i^{m+1} C_{k-i}^{m+1}(n^{k-i} - n^{i+1}) \\
&= \sum_{l=1}^{u(k)} (n^l - n^{k+1-l}) \left[C_{k-l}^{m+1} C_l^{m+1} - C_{k-l+1}^{m+1} C_{l-1}^{m+1}\right], \qquad (6.25)
\end{aligned}$$

where

$$u(k) = \begin{cases} \dfrac{k-1}{2}, & \text{if } k \text{ is odd,} \\ \dfrac{k}{2}, & \text{if } k \text{ is even.} \end{cases} \qquad (6.26)$$

Since $k-l \geq l, l=2,\cdots,u(k)$, we find

$$n^l - n^{k+1-l} < 0, \tag{6.27}$$

and therefore $C_k < 0$ if we can show that

$$C_{k-l}^{m+1} C_l^{m+1} \geq C_{k-l+1}^{m+1} C_{l-1}^{m+1}. \tag{6.28}$$

This follows from the following general inequality

$$C_i^n C_j^n > C_{i+1}^n C_{j-1}^n, \qquad 1 \leq j \leq i < n, \tag{6.29}$$

which can be easily verified by evaluating

$$\begin{aligned} \frac{C_i^n C_j^n}{C_{i+1}^n C_{j-1}^n} &= \frac{(i+1)!(n-i-1)!(j-1)!(n-j+1)!}{i!(n-i)!j!(n-j)!} \\ &= \frac{(i+1)(n-j+1)}{(n-i)j} \\ &= 1 + \frac{(n+1)(i-j+1)}{j(n-i)} \\ &> 1. \end{aligned} \tag{6.30}$$

For $m+3 \leq k \leq 2m+2$,

$$C_k = \sum_{i=k-m-1}^{m+1} C_i^{m+1} C_{k-i}^{m+1} (n^{k-i} - n^{i+1}). \tag{6.31}$$

There are $2m-k+3$ terms. For $l = 0, 1, \cdots, \left\lfloor m - \frac{k-3}{2} \right\rfloor$, the sum of the l^{th} term and the $(2m-k+3-l)^{th}$ term is

$$\begin{aligned} & C_{k-m-1+l}^{m+1} C_{m+1-l}^{m+1} (n^{m+1-l} - n^{k-m+l}) + C_{m+1-l}^{m+1} C_{k-m-1+l}^{m+1} (n^{k-m-1+l} - n^{m+1-l+1}) \\ = \; & C_{k-m-1+l}^{m+1} C_{m+1-l}^{m+1} (n^{m+1-l} + n^{k-m-1+l})(1-n) \\ < \; & 0. \end{aligned} \tag{6.32}$$

If $2m-k+3$ is odd (whence k is even), the middle term is

$$C_{\frac{k}{2}}^{m+1} C_{\frac{k}{2}}^{m+1} (n^{\frac{k}{2}} - n^{\frac{k}{2}+1}) = (C_{\frac{k}{2}}^{m+1})^2 n^{\frac{k}{2}} (1-n) < 0, \tag{6.33}$$

and we get $C_k < 0$. ■

6.3 Comparison of Scheduling Policies

Consider n sources which are independent of each other. Each source produces jobs with the time between two successive jobs forming an i.i.d. sequence of $\mathbb{R}_+$-valued rvs. All the jobs are identical. There are n identical servers to provide services to the jobs. The problem is how to schedule the jobs produced by these sources to the servers. Our goal is to make the job flow to each server as smooth as possible, where we use the peakedness functional of the job flow to each server as the measure of smoothness. We consider the following policies.

Policy I: Dedicated server (**DS**).

Under this policy, each source sends its jobs to a dedicated server.

Policy II: Distributed Bernoulli dispatching (**DBD**).

Under this policy, each source sends its jobs to each of the servers with equal probability.

Policy III: Centralized Bernoulli dispatching (**CBD**).

Under this policy, each source sends its jobs to a centralized dispatcher which then sends the job to each of the servers with equal probability.

Policy IV: Distributed Round Robin (**DRR**).

Under this policy, each source sends its jobs to each of the servers cyclically.

Policy V: Centralizes Round Robin (**CRR**).

Under this policy, each source sends its jobs to a centralized dispatcher which then sends the job to each of the servers cyclically.

It is easily seen that all the policies except **DS** have the property of balancing the load between the servers. If in addition we assume the n sources to be identical, policy **DS** has this property, too. Suppose each source has peakedness functional $z_i(\bar{G})$, $i = 1, \cdots, n$. for some distribution function G. We say that policy **A** is better

than (resp. equivalent to) policy **B** with respect to G, and we write $\mathbf{A} \prec_G \mathbf{B}$ (resp. $\mathbf{A} =_G \mathbf{B}$), if $z_A(\bar{G}) \leq z_B(\bar{G})$ (resp. $z_A(\bar{G}) = z_B(\bar{G})$). We consider the following conditions.

(A) All sources are identical;

(B) All sources are Poisson;

(C) G is the distribution function of an exponential rv; and

(D) G is such that H (defined by (5.5)) is a convex function.

Then we have the following results.

Proposition 6.4 *We have*

(1) $\mathbf{DBD} =_G \mathbf{CBD}$.

(2) Under condition (A), we have $\mathbf{DS} \prec_G \mathbf{DBD}$ *and* $\mathbf{DS} \prec_G \mathbf{CBD}$ *if and only if* $z(\bar{G}) \leq 1$.

(3) Under conditions (A) and (C), we have $\mathbf{DS} \prec_G \mathbf{DRR}$ *if and only if* $0 \leq A^*(\mu) \leq \alpha_{n-1}$, *where* A^* *is the Laplace-Stieltjes transform of the inter-arrival time of the job flow from the source, and* α_{n-1} *is defined in Theorem 5.3.*

(4) Under condition (D), we have $\mathbf{DRR} \prec_G \mathbf{DBD}$ *and* $\mathbf{DRR} \prec_G \mathbf{CBD}$.

(5) Under conditions (A) and (B), we have $\mathbf{CRR} \prec_G \mathbf{DS}$.

(6) Under conditions (A), (B) and (D), we have $\mathbf{CRR} \prec_G \mathbf{DBD}$ *and* $\mathbf{CRR} \prec_G \mathbf{CBD}$.

(7) Under conditions (A), (B) and (C), we have $\mathbf{CRR} \prec_G \mathbf{DRR}$.

Proof. To prove the proposition we observe the following. The job flow to each server is under policy **DS** the job flow from each source. Under other policies, job

flows from the sources undergo both superposition and thinning before they get to the server. For the centralized policies, the superposition happens at the dispatcher, while for the distributed policies, the superposition occurs at each server. To each server, Bernoulli dispatching is equivalent to random thinning, and Round Robin is equivalent to deterministic thinning. Thus under policy **DBD**, the job flows undergo first random thinning with probability $\frac{1}{n}$, and then superposition; under policy **CBD**, the job flows undergo first superposition and then random thinning with probability $\frac{1}{n}$; under policy **DRR**, the job flows undergo first deterministic thinning with probability $\frac{1}{n}$, and then superposition; and under policy **CRR**, the job flows undergo first superposition and then deterministic thinning with probability $\frac{1}{n}$. Therefore, (1) follows from Proposition 6.1; (2) follows from Theorem 5.2 and Corollary 5.4; (3) follows from Theorem 5.3 and Corollary 5.4; (4) follows from Theorem 5.4 and Lemma 5.2; (5) follows from (5.50) and Lemma 5.7; (6) follows from Theorem 5.4; and (7) follows from Proposition 6.3. ■

APPENDIX A

MAJORIZATION

In this appendix, we list some of the definitions and properties of majorization used in the thesis. For details on majorization and its applications, the reader is referred to the monograph by Marshall and Olkin [46].

For any vector $\boldsymbol{x} = (x_1, \ldots, x_n)$ in $\mathbb{R}^n$ we denote by $x_{[i]}$ the i^{th} largest element of $\boldsymbol{x}$, $i = 1, \cdots, n$, i.e., we have

$$x_{[1]} \geq x_{[2]} \geq \cdots \geq x_{[n]}. \tag{A.1}$$

Definition A.1 *For vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$ in $\mathbb{R}^n$ such that*

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \tag{A.2}$$

we say that $\boldsymbol{y}$ is majorized by $\boldsymbol{x}$, and write $\boldsymbol{y} \prec \boldsymbol{x}$, if

$$\sum_{i=k}^{n} y_{[i]} \geq \sum_{i=k}^{n} x_{[i]}, \qquad k = 1, \cdots, n. \tag{A.3}$$

Moreover, we say that $\boldsymbol{y}$ is weakly supermajorized by $\boldsymbol{x}$, and write $\boldsymbol{y} \prec^w \boldsymbol{x}$, if only (A.3) holds.

From this definition we see that majorization is a property of the order statistics of two vectors, and is therefore invariant under permutations.

Majorization and weak majorization are related by the following lemma [46, p. 11].

Lemma A.1 *For $\boldsymbol{x}$ and $\boldsymbol{y}$ in $\mathbb{R}^n$, $\boldsymbol{y} \prec^w \boldsymbol{x}$ iff there exists a vector $\boldsymbol{z}$ in $\mathbb{R}^n$ such that $\boldsymbol{y} \prec \boldsymbol{z}$ and $\boldsymbol{z} \geq \boldsymbol{x}$, where $\boldsymbol{z} \geq \boldsymbol{x}$ means $z_i \geq x_i$, $\quad i = 1, \cdots, n$.*

The following lemma presents a closure property of majorization under concatenation [46, Proposition 5.A.7, p. 121].

Lemma A.2 *For vectors* $\boldsymbol{x}$, $\boldsymbol{y}$ *in* $\mathbb{R}^n$, *and* $\boldsymbol{b}$, $\boldsymbol{c}$ *in* $\mathbb{R}^m$, *define vectors* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *in* $\mathbb{R}^{n+m}$ *by*

$$\boldsymbol{u} = (x_1, \ldots, x_n, b_1, \cdots, b_m) \quad \textit{and} \quad \boldsymbol{v} = (y_1, \ldots, y_n, c_1, \cdots, c_m), \tag{A.4}$$

then

$$\boldsymbol{y} \prec \boldsymbol{x} \quad \textit{and} \quad \boldsymbol{c} \prec \boldsymbol{b} \quad \textit{imply} \quad \boldsymbol{v} \prec \boldsymbol{u}. \tag{A.5}$$

The result remains true if $\prec$ *is replaced by* $\prec^w$.

An important characterization of majorization is contained in the following lemma [46, Proposition 4.B.1, p. 108].

Lemma A.3 *For vectors* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *in* $\mathbb{R}^n$, *we have*

$$\boldsymbol{x} \prec \boldsymbol{y} \quad \textit{iff} \quad \sum_{i=1}^{n} \phi(x_i) \leq \sum_{i=1}^{n} \phi(y_i), \tag{A.6}$$

for all convex functions $\phi : \mathbb{R} \to \mathbb{R}$.

For weak majorization, we have a similar characterization which is an immediate consequence of [46, Theorem 3.A.8, p. 59].

Lemma A.4 *For vectors* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *in* $\mathbb{R}^n$, *we have*

$$\boldsymbol{x} \prec^w \boldsymbol{y} \quad \textit{iff} \quad \sum_{i=1}^{n} \phi(x_i) \leq \sum_{i=1}^{n} \phi(y_i), \tag{A.7}$$

for all decreasing convex functions $\phi : \mathbb{R} \to \mathbb{R}$.

APPENDIX B

STOCHASTIC ORDERING

In this appendix, we review the definitions and properties of the stochastic orderings we use in the thesis. For more information, the reader is referred to [52, 55]. Throughout, X and Y denote two $\mathbb{R}$-valued rvs.

Definition B.1 *The rvs X and Y are equal stochastically or in distribution, and we write $X \stackrel{d}{=} Y$, if*

$$\mathbb{P}\{X \leq a\} = \mathbb{P}\{Y \leq a\}, \qquad a \in \mathbb{R}. \tag{B.1}$$

Definition B.2 *We say that X is smaller than Y in the convex ordering, and we write $X \leq_{cx} Y$, if*

$$\mathbb{E}[\phi(X)] \leq \mathbb{E}[\phi(Y)] \tag{B.2}$$

for all convex functions $\phi : \mathbb{R} \to \mathbb{R}$ whenever the expectations exist in (B.2).

Note that for any $\mathbb{R}_+$-valued rvs, $X \leq_{cx} Y$ implies

$$\mathbb{E}[X] = \mathbb{E}[Y] \tag{B.3}$$

and

$$\mathbb{E}\left[X^k\right] \leq \mathbb{E}\left[Y^k\right], \qquad k = 1, 2, \cdots. \tag{B.4}$$

As a result

$$c^2(X) \leq c^2(Y) \tag{B.5}$$

where for any $\mathbb{R}_+$-valued rv Z, $c^2(Z)$ is the squared coefficient of variation of Z, defined by

$$c^2(Z) \triangleq \frac{Var(Z)}{(\mathbb{E}[Z])^2}. \tag{B.6}$$

Similarly, we define the increasing (resp. decreasing) convex ordering as follows.

Definition B.3 *We say that X is smaller than Y in the increasing (resp. decreasing) convex ordering, and we write $X \leq_{icx} Y$ (resp. $X \leq_{dcx} Y$), if*

$$\mathbb{E}[\phi(X)] \leq \mathbb{E}[\phi(Y)] \tag{B.7}$$

for all increasing (resp. decreasing) convex functions $\phi : \mathbb{R} \to \mathbb{R}$ whenever the expectations exist in (B.2).

APPENDIX C

WEAK CONVERGENCE

This appendix is borrowed from Appendix A of [56]. In this Appendix, we collect the weak convergence theorems used in the thesis. In the sequel, $D[0,1]$ will denote the space of all right continuous functions on $[0,1]$ having left-hand limits. We denote by m the Skorohod metric on $D[0,1]$. For a brief description of the space $D[0,1]$ and of the Skorohod metric, the reader is referred to [7]. Let $\mathcal{W}$ be a Wiener process defined on $[0,1]$. We also denote $\Longrightarrow$ for weak convergence of probability measures induced by random functions in $D[0,1]$ or by rvs in $\mathbb{R}$, and $\xrightarrow{P}$ for convergence in probability.

Theorem C.1 *(Converging together) Let* $\{X_n, n = 0,1,\cdots\}$ *and* $\{Y_n, n = 0,1,\cdots\}$ *be two sequences of rvs taking values in a separable metric space* S *with metric* ρ. *Suppose that for each* $n = 1,2,\cdots$, *the rvs* X_n *and* Y_n *are defined on the probability space so that* $\rho(X_n, Y_n)$ *is a real-valued rv. If* $X_n \Longrightarrow X$ *and* $\rho(X_n, Y_n) \xrightarrow{P} 0$, *then* $Y_n \Longrightarrow X$.

APPENDIX D

HEAVY TRAFFIC FOR GI/GI/1 QUEUES

Most part of the appendix is from Section 2.2 of [56]. With the notations introduced in Appendix C, we shall discuss waiting time, queue size, and virtual waiting time processes in that order.

D.1 The Probabilistic Setting

Let $(\Omega, \mathcal{F}, P)$ be a probability space rich enough to support a sequence $\{G^r, r = 1, 2, \cdots\}$ of GI/GI/1 queues with FCFS service discipline. For each $r = 1, 2, \cdots$, G^r is represented by two independent sequences of i.i.d. nonnegative random variables, $\{A_n^r, n = 0, 1, \cdots\}$ and $\{B_n^r, n = 0, 1, \cdots\}$. For $n = 0, 1, \cdots$, we interpret B_n^r as the n^{th} service time and A_{n+1}^r as the time between the n^{th} and the $(n+1)^{st}$ arrivals in G^r with the understanding that A_0^r is the arrival time of the first customer. We assume that for each $r = 1, 2, \cdots$, the first customer in G^r arrivals at time $t = 0$, i.e., $A_0^r = 0$, to an empty queue.

We now introduce assumption (A1) on any double sequence $\{Y_n^r, n, r = 1, 2, \cdots\}$ of $\mathbb{R}$- valued rvs defined on the probability space $(\Omega, \mathcal{F}, P)$.

Assumption (A1):

1. For each $r = 1, 2, \cdots$, $\{Y_n^r, n = 1, 2, \cdots\}$ is an i.i.d. sequence of rvs with mean $\underline{a}rY^r$ and standard deviation σ_Y^r;

2. As $r \to \infty$, $\bar{Y}^r \to \bar{Y} < \infty$ and $0 < \sigma_Y^r \to \sigma_Y$ with $0 < \sigma_Y < \infty$; and

3. For some $\epsilon > 0$, $\sup_r \mathbb{E}\left[|Y_n^r|^{2+\epsilon}\right] < \infty$.

The sequences $\{A_n^r, n = 0, 1, \cdots\}$ and $\{B_n^r, n = 0, 1, \cdots\}$ are both assumed to satisfy Assumption (A1) defined in Appendix C. If we define

$$X_n^r = B_{n-1}^r - A_n^r, \qquad n, r = 1, 2, \cdots, \tag{D.1}$$

then clearly $\{X_n^r, n, r = 1, 2, \cdots\}$ also satisfies Assumption (A1).

We impose the following conditions

$$\bar{X}^r\sqrt{r} \to C \quad \text{with} \quad -\infty < C < 0 \quad \text{as} \quad r \to \infty; \tag{D.2}$$

$$0 < \bar{A}^r \to \bar{A} \quad \text{and} \quad 0 < \bar{B}^r \to \bar{B} \quad \text{with} \quad 0 < \bar{A}, \bar{B} < \infty. \tag{D.3}$$

We define the continuous reflection mapping $f : D[0,1] \to D[0,1]$ by

$$f(y)(t) = \sup_{0 \le s \le t}[y(t) - y(s)], \qquad 0 \le t \le 1, y \in D[0,1]. \tag{D.4}$$

D.2 The Actual Waiting Time Process

For each $r = 1, 2, \cdots$, let $\{W_n^r, n = 0, 1, \cdots\}$ be the waiting time of the n^{th} customer $(n = 0, 1, \cdots)$ in G^r, and let

$$\omega^r(t) = \frac{1}{\sqrt{r}} W^r_{\lfloor rt \rfloor}, \qquad 0 \le t \le 1. \tag{D.5}$$

We have

$$\omega^r \Longrightarrow f(\sigma_X \mathcal{W} + d), \tag{D.6}$$

where

$$d(t) = Ct, \qquad 0 \le t \le 1. \tag{D.7}$$

D.3 The Queue Size Process

For each $r = 1, 2, \cdots$, let $\{Q^r(t),\ t \ge 0\}$ be the queue size process of G^r, and define

$$\chi^r(t) = \frac{1}{\sqrt{r}} Q^r(rt), \qquad 0 \le t \le 1. \tag{D.8}$$

Then it can be shown that

$$\chi^r \Longrightarrow f(\gamma \mathcal{W} + e), \tag{D.9}$$

where

$$\gamma^2 = \sigma_A^2/\bar{A}^3 + \sigma_B^2/\bar{B}^3 \quad \text{and} \quad e(t) = Ct, \quad 0 \le t \le 1. \tag{D.10}$$

APPENDIX E

PEAKEDNESS FUNCTIONALS OF RENEWAL PROCESSES

In this appendix, we use notations in Sections 2 and 5. Denote by $N(s,t)$ the number of arrivals in the time interval $[s,t)$. It is easily verified that the number of busy servers in the $GI/G/\infty$ queue can be expressed by

$$K(t) = \sum_{i=1}^{N(t)} \mathbf{1}\{\sigma_i > t - a_i\},$$

where $\{\sigma_n, n = 1, 2, \cdots\}$ are i.i.d. rvs with common distribution function G.

Denote $k_1(t) = \mathbb{E}[K(t)]$. Following the basic renewal argument [38, p. 187], we have

$$\begin{aligned} &\mathbb{E}[K(t)|X_1 = x] \\ &= \begin{cases} 0, & \text{if } x > t; \\ \mathbb{E}[\mathbf{1}\{\sigma_1 > t - x\}] + \mathbb{E}\left[\sum_{i=2}^{N(t)} \mathbf{1}\{\sigma_i > t - a_i\}|X_1 = x\right], & \text{if } x \le t. \end{cases} \end{aligned} \tag{E.1}$$

Since

$$\mathbb{E}[\mathbf{1}\{\sigma_1 > t - x\}] = \mathbb{P}\{\sigma_1 > t - x\} = \bar{G}(t - x) \tag{E.2}$$

and

$$\begin{aligned} &\mathbb{E}\left[\sum_{i=2}^{N(t)} \mathbf{1}\{\sigma_i > t - a_i\}|X_1 = x\right] \\ &\quad = \mathbb{E}\left[\sum_{i=2}^{1+N(x,t)} \mathbf{1}\{\sigma_i > t - x - \sum_{j=2}^{i} X_j\}|X_1 = x\right] \\ &\quad = \mathbb{E}\left[\sum_{l=1}^{N(t-x)} \mathbf{1}\{\sigma_l > t - x - a_l\}\right] = k_1(t - x), \end{aligned} \tag{E.3}$$

we have

$$\mathbb{E}[K(t)|X_1 = x] = \begin{cases} 0, & \text{if } x > t; \\ \bar{G}(t - x) + k_1(t - x), & \text{if } x \le t. \end{cases} \tag{E.4}$$

By the law of total probability, we get

$$\begin{aligned} k_1(t) &= \int_0^t [\bar{G}(t-x) + k_1(t-x)]dA(x) \\ &= \int_0^t \bar{G}(t-x)dA(x) + \int_0^t k_1(t-x)dA(x). \end{aligned} \tag{E.5}$$

So, $k_1(t)$ satisfies the renewal equation

$$k_1(t) = a(t) + \int_0^t k_1(t-x)dA(x) \tag{E.6}$$

with

$$a(t) = \int_0^t \bar{G}(t-x)dA(x). \tag{E.7}$$

Since $a(t)$ is bounded, the solution to (E.5) is then by [38, Theorem 4.1] given by

$$k_1(t) = a(t) + \int_0^t a(t-x)dM(x) \tag{E.8}$$

where $M(x)$ is the renewal function of the renewal process, i.e., $M(x) = \mathbb{E}[N(x)]$. It is well known that

$$M(x) = \sum_{i=1}^{\infty} A_i(x) \tag{E.9}$$

where $A_i(x)$ is the i^{th} convolution of the inter-arrival time distribution by itself.

Next denote $k_2(t) = \mathbb{E}[K^2(t)]$. Simple calculations yield

$$\begin{aligned} K^2(t) &= \left(\sum_{i=1}^{N(t)} \mathbf{1}\{\sigma_i > t - a_i\}\right)^2 \\ &= K(t) + 2\sum_{i=1}^{N(t)} \sum_{j=i+1}^{N(t)} \mathbf{1}\{\sigma_i > t - a_i\}\mathbf{1}\{\sigma_j > t - a_j\}. \end{aligned} \tag{E.10}$$

Again by the renewal argument, we have

$$\begin{aligned} &\mathbb{E}\left[K^2(t)|X_1 = x\right] \\ &= \begin{cases} \mathbb{E}[K(t)|X_1 = x] \\ \quad +2\,\mathbb{E}\left[\sum_{i=1}^{N(t)} \sum_{j=i+1}^{N(t)} \mathbf{1}\{\sigma_i > t - a_i\}\mathbf{1}\{\sigma_j > t - a_j\}|X_1 = x\right], & x \le t, \\ 0, & x > t. \end{cases} \end{aligned} \tag{E.11}$$

By the independence of σ_i and σ_j for $i \neq j$, we get

$$
\begin{aligned}
& 2\,\mathbb{E}\left[\sum_{i=1}^{N(t)}\sum_{j=i+1}^{N(t)} \mathbf{1}\{\sigma_i > t-a_i\}\mathbf{1}\{\sigma_j > t-a_j\}|X_1 = x\right] \\
&= 2\,\mathbb{E}\left[\sum_{j=2}^{N(t)} \mathbf{1}\{\sigma_1 > t-x\}\mathbf{1}\{\sigma_j > t-a_j\}|X_1 = x\right] \\
&\quad +2\,\mathbb{E}\left[\sum_{i=2}^{N(t)}\sum_{j=i+1}^{N(t)} \mathbf{1}\{\sigma_i > t-a_i\}\mathbf{1}\{\sigma_j > t-a_j\}|X_1 = x\right] \\
&= 2\,\mathbb{E}\left[\sum_{j=2}^{1+N(t-x)} \mathbf{1}\{\sigma > t-x\}\mathbf{1}\{\sigma_j > t-x-a_{j-1}\}\right] \\
&\quad +2\,\mathbb{E}\left[\sum_{i=2}^{1+N(t-x)1+N(t-x)}\sum_{j=i+1} \mathbf{1}\{\sigma_i > t-x-a_{i-1}\}\mathbf{1}\{\sigma_j > t-x-a_{j-i}\}\right] \\
&= 2\,\mathbb{E}\left[\sum_{i=1}^{N(t-x)} \mathbf{1}\{\sigma_i > t-x-a_i\}\right] P\{\sigma > t-x\} - \mathbb{E}[K(t-x)] \\
&\quad +2\,\mathbb{E}\left[\sum_{m=1}^{N(t-x)N(t-x)}\sum_{n=m+1} \mathbf{1}\{\sigma_m > (t-x)-a_m\}\mathbf{1}\{\sigma_n > (t-x)-a_n\}\right] + \mathbb{E}[K(t-x)] \\
&= k_2(t-x) + 2k_1(t-x)\bar{G}(t-x) - k_1(t-x). \qquad \text{(E.12)}
\end{aligned}
$$

Together with (E.4), we get

$$\mathbb{E}\left[K^2(t)|X_1 = x\right] = \bar{G}(t-x) + k_2(t-x) + 2k_1(t-x)\bar{G}(t-x). \qquad \text{(E.13)}$$

Therefore, $k_2(t)$ also satisfies a renewal equation

$$k_2(t) = b(t) + \int_0^t k_2(t-x)dA(x) \qquad \text{(E.14)}$$

with

$$b(t) = \int_0^t \bar{G}(t-x)[1 + 2k_1(t-x)]dA(x). \qquad \text{(E.15)}$$

It is not difficult to verify that $b(t)$ is also bounded, the solution to (E.15) is then given by

$$k_2(t) = b(t) + \int_0^t b(t-x)dM(x). \qquad \text{(E.16)}$$

Invoking the Basic Renewal Theorem [38, Theorem 5.1], we get

$$k_1 \triangleq \lim_{t\nearrow\infty} k_1(t) = \lambda \int_0^\infty a(x)dx \qquad \text{(E.17)}$$

and

$$k_2 \triangleq \lim_{t \nearrow \infty} k_2(t) = \lambda \int_0^\infty b(x)dx \tag{E.18}$$

with λ being the arrival rate. Now we further calculate k_1 and k_2. Substituting (E.7) into (E.17), we get

$$\begin{aligned} \int_0^\infty a(x)dx &= \int_0^\infty \int_0^x \bar{G}(x-y)dA(y)dx \\ &= \int_0^\infty \int_y^\infty \bar{G}(x-y)dxdA(y) \\ &= \int_0^\infty \left[\int_0^\infty (\bar{G}(z))dz\right] dA(y) \\ &= \int_0^\infty \bar{G}(z)dz = \frac{1}{\mu}. \end{aligned} \tag{E.19}$$

As is expected, we have

$$k_1 = \frac{\lambda}{\mu} \triangleq \rho. \tag{E.20}$$

Now substituting (E.15) into (E.18), we get

$$\int_0^\infty b(x)dx = \int_0^\infty \int_0^x \bar{G}(x-y)dA(y)dx + \int_0^\infty \int_0^x 2\bar{G}(x-y)k_1(x-y)dA(y)dx. \tag{E.21}$$

From previous calculation, we know that the first term in (E.21) is $\frac{1}{\mu}$. For the second term, we have

$$\begin{aligned} &\int_0^\infty \int_0^x 2\bar{G}(x-y)k_1(x-y)dA(y)dx \\ &= \int_0^\infty \int_0^\infty 2\bar{G}(z)k_1(z)dzdA(y) \\ &= \int_0^\infty 2\bar{G}(z)k_1(z)dz \\ &= \int_0^\infty 2\bar{G}(z)\left\{a(z) + \int_0^z a(z-x)dM(x)\right\} dz \\ &= \int_0^\infty 2\bar{G}(z)a(z)dz + \int_0^\infty \int_0^z 2\bar{G}(z)a(z-x)dM(x)dz. \end{aligned} \tag{E.22}$$

By defining

$$H(t) = 2\mu \int_0^\infty \bar{G}(t+y)\bar{G}(y)dy, \tag{E.23}$$

we have

$$\int_0^\infty 2\bar{G}(z)a(z)dz$$

$$
\begin{aligned}
&= \int_0^\infty \int_0^z 2\bar{G}(z-x)\bar{G}(z)dA(x)dz \\
&= \int_0^\infty \int_x^\infty 2\bar{G}(z)\bar{G}(z-x)dzdA(x) \\
&= \frac{1}{\mu}\int_0^\infty H(x)dA(x) \\
&= \frac{1}{\mu}\mathbb{E}\left[H(X_1)\right] \;=\; \frac{1}{\mu}\mathbb{E}\left[H(a_1)\right]
\end{aligned} \tag{E.24}
$$

and

$$
\begin{aligned}
&\int_0^\infty \int_0^z 2\bar{G}(z)a(z-x)dM(x)dz \\
&\quad= \int_0^\infty \int_x^\infty 2\bar{G}(z)a(z-x)dzdM(x) \\
&\quad= \int_0^\infty \int_0^\infty 2\bar{G}(x+y)a(y)dydM(x) \\
&\quad= \int_0^\infty \left[\int_0^\infty 2\bar{G}(x+y)\int_0^y \bar{G}(y-z)dA(z)dy\right] dM(x) \\
&\quad= \int_0^\infty \left[\int_0^\infty \int_z^\infty 2\bar{G}(x+y)\bar{G}(y-z)dydA(z)\right] dM(x) \\
&\quad= \int_0^\infty \left[\int_0^\infty \left\{\int_0^\infty 2\bar{G}(x+z+w)\bar{G}(w)dw\right\} dA(z)\right] dM(x) \\
&\quad= \int_0^\infty \int_0^\infty \frac{1}{\mu}H(x+z)dA(z)dM(x) \\
&\quad= \sum_{l=1}^\infty \int_0^\infty \int_0^\infty \frac{1}{\mu}H(x+z)dA(z)dA_l(x) \\
&\quad= \frac{1}{\mu}\sum_{l=1}^\infty \mathbb{E}\left[H(a_{l+1})\right] \;=\; \frac{1}{\mu}\sum_{m=2}^\infty \mathbb{E}\left[H(a_m)\right].
\end{aligned} \tag{E.25}
$$

Combining these results, we finally get

$$
\int_0^\infty b(x)dx = \frac{1}{\mu} + \frac{1}{\mu}\sum_{m=1}^\infty \mathbb{E}\left[H(a_m)\right] = \frac{1}{\mu}\left\{1 + \sum_{m=1}^\infty \mathbb{E}\left[H(a_m)\right],\right\} \tag{E.26}
$$

and therefore,

$$
k_2 = \rho\left\{1 + \sum_{m=1}^\infty \mathbb{E}\left[H(a_m)\right]\right\}. \tag{E.27}
$$

By the definition of peakedness functionals, we conclude

$$
\begin{aligned}
z(\bar{G}) &= 1 - \rho + \sum_{m=1}^\infty \mathbb{E}\left[H(a_m)\right] && \text{(E.28)} \\
&= 1 - \rho + \int_0^\infty H(x)dM(x). && \text{(E.29)}
\end{aligned}
$$

It is easy to verify that $H(t)$ has the following properties which will be used later.

(i) $0 \leq H(t) \leq 2$;

(ii) $H(t)$ is monoton decreasing in t;

(iii) $H(t)$ may be a convex function, but it can never be a concave function.

Eckberg showed [19] that the peakedness function of a stationary point process is given by

$$z(\bar{G}) = 1 + \int_0^\infty H(x)dU(x) - \rho \tag{E.30}$$

where $U(x)$ is the expectation function of the point process, i.e.,

$$U(x) = E[\text{ number of arrivals following, and no later than a time } x \text{ from, an arbitrarily chosen arrival }], \qquad x \geq 0.$$

If the point process is a renewal process, then $U(x) = M(x)$. Thus the result we have just proved is just a special case of (E.30).

REFERENCES

[1] H. Ahmadi, R. Guérin and K. Sohraby, "Analysis of leaky bucket access control mechanism with batch arrival process," in *Proceedings of the GLOBECOM'90*, San Diego (CA), December 1990, paper 400B.1.

[2] V. Anantharam and P. Konstantopoulos, "Burst reduction properties of the leaky bucket flow control scheme in ATM networks," to appear in IEEE Transactions on Communications.

[3] J. J. Bae and T. Suda, "Survey of traffic control schemes and protocols in ATM networks," *Proceedings of the IEEE* **79(2)**, February 1991, pp. 170–189.

[4] K. Bala, I. Cidon and K. Sohraby, "Congestion control for high speed packet switch," in *Proceedings of the INFOCOM'90*, San Francisco (CA), June 1990, pp. 520–526.

[5] A. W. Berger, "Overload control using rate control throttle: selecting token bank capacity for robustness to arrival rates," *IEEE Transactions on Automatic Control* **36(2)**, February 1991, pp. 216–219.

[6] A. W. Berger, "Performance analysis of a rate-control throttle where tokens and jobs queue," *IEEE Journal on Selected Areas in Communications* **9(2)**, February 1991, pp. 165–170.

[7] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York (NY) (1968).

[8] P. Billingsley, *Probability and Measure*, Second Edition, John Wiley & Sons, New York (NY) (1986).

[9] U. Briem, T. H. Theimer and H. Kroner, "A general discrete-time queueing model: analysis and applications," in *Proceedings of the 13th International Teletraffic Congress*, Copenhagen (Denmark) (1991).

[10] K. C. Budka and D. D. Yao, "Monotonicity and convexity properties of rate control throttles," in *Proceedings of the 29th IEEE Conference on Decision and Control*, Honolulu (Hawaii), December 1990, pp. 883–884.

[11] M. Butto, E. Cavallero and A. Tonietti, "Effectiveness of the "leaky bucket" policing mechanism in ATM networks," *IEEE Journal on Selected Areas in Communications* **9(3)**, April 1991, pp. 335–342.

[12] CCITT SG.XVIII, Draft recommendation I.211.

[13] C. S. Chang, "Smoothing point processes as a means to increase throughput," IBM Technical Report **RC-16866**, IBM, Inc. (1991).

[14] I. Cidon and I. S. Gopal, "PARIS: an approach to integrated high-speed private networks," *International Journal of Digital & Analog Cabled Systems* **1(2)**, April-June 1988, pp. 77–86.

[15] E. G. Coffman, Jr. and M. I. Reiman, "Diffusion approximations for computer/communication systems," in G. Iazeolla, P. J. Courtois, and A. Hordijk, editors, *Mathematical Computer Performance and Reliability*, Elsevier Science Publishers B. V. (North-Holland) (1984), pp. 33–53.

[16] D. R. Cox and P. A. W. Lewis, *The Statistical Analysis of Series of Events*, Methuen & CO LTD, London (UK) (1966).

[17] F. Denissen, E. Desmet and G. H. Petit, "The policing function in ATM networks," in *Proceedings of the 1990 International Zurich Seminar on Digital Communications*, Zürich (Switzerland) (1990), pp. 131–144.

[18] L. Dittmann, S. B. Jacobsen and K. Moth, "Flow enforcement algorithms for ATM networks," *IEEE Journal on Selected Areas in Communications* **9(3)**, April 1991, pp. 343–350.

[19] A. E. Eckberg, "Generalized peakedness of teletraffic processes," in *Proceedings of the 10th International Teletraffic Congress*, Montreal (Canada), June 1983, paper 44B3.

[20] A. E. Eckberg, "Approximations for bursty (and smoothed) arrival queueing delays based on generalized peakedness," *Teletraffic Issues in an Advanced Information Society*, 11th International Teletraffic Congress, Kyoto (Japan), September 1985, pp. 331–335.

[21] A. E. Eckberg, D. T. Luan and D. M. Lucantoni, "Bandwidth management: a congestion control strategy for broadband packet networks - characterizing the throughput-burstiness filter," in *Proceedings of the International Teletraffic Congress Specialist Seminar*, Adelaide (Australia) (1989), paper 4.4.

[22] A. E. Eckberg, D. T. Luan and D. M. Lucantoni, "Meeting the challenge: Congestion and flow control strategies for broadband information transport," in *Proceedings of the GLOBECOM'89*, Dallas (TX), November 1989, paper 49.3.

[23] A. E. Eckberg, D. T. Luan and D. M. Lucantoni, "An approach to controlling congestion in ATM networks," *International Journal of Digital & Analog Cabled Systems* **3** (1990).

[24] A. Fukuda, "Analysis of input control based on discrete monitoring in mixed input queueing model," *Electronic Communications Japan* **68(12)** (1985), pp. 57–65.

[25] A. Fukuda, "Input regulation control based on periodical monitoring using call gapping control," *Electronic Communications Japan* **69(11)** (1986), pp. 84–93.

[26] G. Gallassi, G. Rigolio and L. Fratta, "ATM: Bandwidth assignment and bandwidth enforcement policies," in *Proceedings of the GLOBECOM'89*, Dallas (TX), November 1989, paper 49.6.

[27] E. Garetti, R. Melen and A. Tonietti, "Efficiency of leaky-bucket mechanism in the framework of ATM resource allocation," in *3rd RACE 1022 Workshop*, (1989).

[28] R. Guérin, H. Ahmadi and M. Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE Journal on Selected Areas in Communications* **9(7)**, September 1991, pp. 968–981.

[29] R. Guérin and L. Gün, "A unified approach to bandwidth allocation and access control in fast packet-switched networks," in *Proceedings of the INFOCOM'92*, Florence (Italy) (1992).

[30] L. Gün, Private discussions.

[31] L. Gün and R. Guérin, "A framework for bandwidth management and congestion control in high speed networks," in *Proceedings of the TriCom'92 – High Speed Networks*, Research Triangle Park (NC), February 1992, pp. 39–54.

[32] R. Gusella, "Characterizing the variability of arrival processes with index of dispersion," *IEEE Journal on Selected Areas in Communications* **9(2)** (1991), pp. 203–211.

[33] M. Hirano and N. Watanabe, "Traffic characteristics and a congestion control scheme for an ATM network," *International Journal of Digital & Analog Cabled Systems* **3** (1990), pp. 211–217.

[34] D. S. Holtsinger and H. G. Perros, "Performance analysis of leaky bucket policing mechanisms," in *Proceedings of the of TriComm'92 – High Speed Networks*, Research Triangle Park (NC), February 1992, pp. 55–89.

[35] J. Y. Hui, "Resource allocation for broadband networks," *IEEE Journal on Selected Areas in Communications* **6(9)** (1988), pp. 1598–1608.

[36] D. L. Iglehart and W. Whitt, "Multiple channel queues in heavy traffic I," *Advances in Applied Probability* **2** (1970), pp. 150–177.

[37] R. Jain, "Congestion control in computer networks: issues and trends," *IEEE Network Magazine* **4(3)**, May 1990, pp. 24–30.

[38] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, Second Edition, Academic Press, New York (NY) (1975).

[39] K. Kawashima, "Queueing analysis for input regulation method employing periodic monitoring and control," *European Journal of Operational Research* **23** (1986), pp. 100–107.

[40] L. Kleinrock, *Queueing Systems*, volume II: *Computer Applications*, John Wiley & Sons, New York (NY) (1976).

[41] W. Kowalk and R. Lehnert, "The 'policing function' to control user access in ATM networks–definition and implementation," in *Proceedings of the International Symposium on Subscriber Loops and Services*, Boston (Mass) (1988), paper 12.2.

[42] A. J. Lawrance. "Dependency of intervals between events in superposition processes," *J. R. Statist. Soc.* **B 35** (1973), pp. 306–315.

[43] A. J. Lemoine, "Networks of queues – a survey of weak convergence results," *Management Science* **24** (1978), pp. 1175–1193.

[44] S. Low and P. Varaiya, "A simple theory of traffic and resource allocation in ATM," in *Proceedings of the GLOBECOM'91*, Phoenix (AZ), December 1991.

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

[46] A. W. Marshall and I. Olkin, *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York (NY) (1979).

[47] M. Murata, Y. Oie, T. Suda and H. Miyahara, "Analysis of a discrete-time single-server queue with bursty input for traffic control in ATM networks," *IEEE Journal on Selected Areas in Communications* **8(3)**, April 1990, pp. 447–458.

[48] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models – An Algorithmic Approach*, The Johns Hopkins University Press, Baltimore (MD) and London (UK) (1981).

[49] M. F. Neuts, *Structured Stochastic Matrices of M/G/1 Type and Their Applications*, Marcel Dekker, Inc., New York (NY) (1989).

[50] E. P. Rathgeb, "Modeling and performance comparison of policing mechanisms for ATM networks," *IEEE Journal on Selected Areas in Communications* **9(3)**, April 1991, pp. 325–334.

[51] CCITT Recommendations, *Blue Book*, Geneva (Switzerland) (1989).

[52] S. M. Ross, *Stochastic Processes*, John Wiley & Sons, New York (NY) (1983).

[53] M. Sidi, W. Liu, I. Cidon and I. Gopal, "Congestion control through input rate regulation," in *Proceedings of the GLOBECOM'89*, Dallas (TX), November 1989, paper 49.2.

[54] K. Sohraby and M. Sidi, "On the performance of bursty and correlated sources subject to leaky bucket rate-based access control schemes," in *Proceedings of the INFOCOM'91*, Bal Harbour (FL) (1991), paper 4D.3.

[55] D. Stoyan, *Comparison Methods for Queues and Other Stochastic Models*, John Wiley & Sons, New York (NY) (1983).

[56] Tedijanto, *Nonexhaustive Policies in Polling Systems and Vacation Models: Qualitative and Approximate Approach*, PhD thesis, University of Maryland, College Park (MD) (1990).

[57] J. S. Turner, "New directions in communications (or which way to the information age?)," *IEEE Communication Magazine* **24(10)**, October 1986, pp. 8–15.

[58] W. Whitt, "Heavy traffic limit theorems for queues: A survey," in A. B. Clarke, editor, *Mathematical Methods in Queueing Theory*, Springer-Verlag, Berlin (Germany) and New York (NY) (1974), pp. 307–350.

[59] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, Englewood Cliffs (NJ) (1989).

[60] G. M. Woodruff, R. Kositpaiboon, G. Fitzpatrick and P. Richards, "Control of ATM statistical multiplexing performance," in *Proceedings of the International Teletraffic Congress Specialist Seminar*, Adelaide (Australia) (1989).

[61] G. M. Woodruff, R. G. H. Rogers and P. S. Richards, "A congestion control framework for high-speed integrated packetized transport," in *Proceedings of the GLOBECOM'88*, Hollywood (FL), November 1988, paper 7.1.

[62] M. Yamamoto, H. Nakanishi, Y. Tezuka, I. Akiyoshi and H. Sanada, "Approximate analysis for window-controlled packet network with finite input queue and optimal window allocation," *Electronic Communications Japan* **72(2)** (1989), pp. 99–111.