BITS OF INTERNET TRAFFIC CONTROL

THÈSE N^{O} 2827 (2003)

PRÉSENTÉE À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS

Institut de systèmes de communication

SECTION DES SYSTÈMES DE COMMUNICATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES TECHNIQUES

PAR

Milan VOJNOVIĆ

Master of Technical Science, University of Split, Croatie
et de nationalité croate

acceptée sur proposition du jury:

Prof. J.-Y. Le Boudec, directeur de thèse
Dr M. Andrews, rapporteur
Prof. M. Grossglauser, rapporteur
Prof. P. Thiran, rapporteur
Prof. D. Towsley, rapporteur

Lausanne, EPFL
2003

Bits of Internet Traffic Control

Equation-Based Rate Control, Increase-Decrease Controls,
Expedited Forwarding, Input-Queued Switch

Milan Vojnović

October 6th, 2003

To my wife Sandra, and our children Srđan and Mirta
— for bearing this work with me

To my mum and sister
— for bearing the distance

In memory of my beloved father

Acknowledgments

First and foremost, I am grateful to Prof. Jean-Yves Le Boudec, my teacher and adviser. I admire his clarity and rigor. I appreciate his caring. He impressed me with his open-mind, organizational excellence, a professional approach to each bit of his work and with his passion for cycling. He gave me a great deal of freedom to work. His advice was always valuable and well-dosed.

I thank Prof. Jean-Pierre Hubaux for his advice and support in the early stages of my Ph.D. work. He encouraged me to go after the problems that intrigued me.

I ended up at EPFL both incidentally and at the last minute. I joined the graduate school in communication systems in October 1998. I applied to the program rather late, after the deadline, but still was lucky enough to be enrolled. I consider invaluable the knowledge gained while attending the school. Especially, I am in dept to a lecture "*Stochastic Models in Communication Systems*" by Prof. Pierre Brémaud who initiated me to and taught me how to admire the subject. I am grateful to EPFL for awarding me the EPFL Ph.D. fellowship, 1999–2002. I could hardly imagine a better place for Ph.D. work, than the one I had the privilege to enjoy.

I carried out most of my work with the Laboratory for Computer Communications and Applications (LCA). I thank Prof. Matthias Grossglauser and Prof. Patrick Thiran, in particular, for their constructive suggestions on how to make some of my bad talks less bad. Over the years, many people left the lab and new people joined. I am thankful for having had such great colleagues as Dr. Chadi Barakat, Sonja Buchegger, Dr. Ljubica Blažević, Dr. Catherine Boutremans, Prof. Levénte Buttyan, Srđan Čapkun, Mario Čagalj, Olivier Dousse, Mathilde Durvy, Felix Farkas, Dr. Jean-Philippe Martin-Flatin, Mirko Franceschinis, Dr. Silvia Giordano, Dr. Maher Hamdi, Dr. Paul Hurley, Dr. Lukas Kencel, Daniel Lungu, Jun Luo, Hung Nguyen, Božidar Radunović, Naouel Ben Salem. Special thanks go to my former office-mate Jean-Philippe, who gave me plentiful advice, by unselfishly taking from his experience.

Many thanks go to the LCA administration staff, Danielle Alvarez, Holly Cogliati, and Angela Devenoge. They made my life a lot easier. Danielle and Angela helped me in many instances to overcome my incompetence in French. I thank equally well the LCA system administrators, Jean-Pierre Dupertuis and Marc-Andre Lüthi. Special thanks go to Holly for correcting part of the English of this thesis, and Olivier for writing for me the "Version abrégée."

I was privileged to undertake an internship with the Mathematics Research Center, Bell Laboratories, Lucent Technologies, Murray Hill, NJ, in August-October 2001. I worked with Dr. Matthew Andrews. Working with Matthew was both fruitful and fun. Some work from that period appears in this thesis.

I would not miss thanking Prof. Anurag Kumar for his support in writing a reference letter for my application for the aforementioned internship. While at Bell Labs, and at some other moments, I found time to visit and enjoy hospitality of many people with various labs and universities in the USA. I hope that those, whose names for brevity I omit to display, would accept my apology and gratitude.

I thank Prof. Emre Telatar for discussions at the beginning of my Ph.D. work, at a time when I was not sure which way to go. I also benefitted from discussions with Prof. Don Towsley during his visits to EPFL, his warm hospitality for a week at the University of Massachusetts, Amherst, and upon some other occasions.

I thank the Ministry of Science and Technology of Republic of the Croatia for a support in an initial phase of my work at EPFL. The support was through the project no. 023–023, led by Prof. Nikola Rožić, FESB, Split, Croatia.

I am grateful to those who created user accounts at their sites for the purpose of the measurements, a part of which appear in the empirical results of this thesis. Those are: Steven Low, X. Wei (CalTech), Vishal Misra (Columbia University), Mounir Hamdi (Hong Kong University of Science and Technology), Chadi Barakat, Thierry Turletti, Thierry Parmentelat (INRIA, Sophia-Antipolis), Gunnar Karlsson, Ignacio M. Ivars (KTH Royal Institute of Technology, Sweden), Daniel R. Figueiredo (University of Massachusetts, Amherst), Darryl Veitch (University of Melbourne). Thanks go to Marc-Andre and Božidar who taught me numerous Linux configuration tweaks. I thank Chadi Barakat, Jörg Widmer, and Vern Paxson for discussions and some software they generously allowed me to use. Last but not least, I appreciate the work of students, Christian Latesch and Thorsten Müller, who worked with me on a daunting task of the Internet measurements.

I thank the committee members of my thesis defense: Dr. Matthew Andrews, Prof. Matthias Grossglauser, Prof. Patrick Thiran, and Prof. Don Towsley, for their time and effort in reviewing a thesis draft, and their valuable suggestions.

— Lausanne, April 28, 2003
(with continued refinements since then)

Version abrégée

Le présent travail traite de quatre problèmes relatifs au contrôle de trafic sur Internet.

Le premier problème est de comprendre quand et pourquoi une source utilisant un contrôle de flux régi par équation est-elle "TCP-friendly" (une source est dite TCP-friendly si dans les mêmes conditions, son débit à long terme ne dépasse pas celui d'une source utilisant TCP). Il est par ailleurs établi que pour éviter la congestion, certaines sources doivent être TCP-friendly. Le contrôle de flux régi par équation fonctionne de la manière suivante: la source estime le taux de perte de paquets ainsi que leur temps de parcours. Elle calcule à l'aide des ces paramètres le débit qu'une source TCP aurait dans ces conditions, et ajuste son propre débit selon le résultat. Ce travail propose une analyse exacte de ce mécanisme et prédit quand la source sera effectivement TCP-friendly. Des preuves expérimentales montrent que dans des cas réalistes, ce type de sources peut grossièrement différer d'une source TCP.

Le deuxième problème concerne une autre famille de contrôles de flux, dite "increase-decrease", dont un cas particulier est l'incrément additif/décrément multiplicatif. Il s'agit ici de calculer le débit moyen à long terme de plusieurs sources concurrentes utilisant ce dernier type de mécanisme, dans un réseau arbitraire où les chemins sont fixés et les temps de parcours connus. Nous montrons quelle répartition des ressources en résulte. Ce résultat est un progrès certain dans la compréhension de la répartition équitable de ressources dans un réseau où les délais sont arbitraires. Un autre problème consiste à concevoir un mécanisme "increase-decrease" qui réalise une certaine fonction débit/taux de perte. Il est montré dans ce travail que si la conception est faite en utilisant un processus de pertes usuel (une séquence d'intervalles constantes entre les pertes), le mécanisme sort de son objectif pour un processus de perte plus général. La motivation pour étudier le problème de conception est de créer un mécanisme "increase-decrease" qui soit TCP-friendly.

Le troisième problème consiste à obtenir des bornes probabilistes à la performance de noeuds qui se conforment au comportement spécifié dans "Expedited Forwarding", un service différencié d'Internet. Sous l'hypothèse que le processus d'arrivée à un noeud est formé de flots régulés individuellement –c'est un lieu commun dans Expedited Forwarding– et que les flux sont stochastiquement indépendants, des bornes à la taille et à la durée de la file d'attente à un noeud, ainsi qu'au taux de perte, sont calculées. Ces bornes permettent un dimen-

sionnement plus efficace que des bornes déterministes calculées pour le pire des cas.

Le quatrième et dernier problème est de calculer le temps de réponse d'un commutateur (switch) avec attente en entrée et utilisant un planificateur à décomposition (decomposition-based scheduler). Partant d'une certaine matrice de demande (contenant le débit à atteindre entre chaque paire de ports), le planificateur décompose cette matrice en un ensemble de matrices de permutations qui décrivent les connections entre les ports. La difficulté est de construire une suite de permutations telle que le temps de réponse soit faible pour chaque paire entrée-sortie. De nouvelles bornes sur ce temps de réponse sont proposées, qui sont, dans bien des cas, meilleures que les bornes connues. Il est utile de concevoir des commutateurs à temps de réaction borné pour obtenir des garanties sur les délais et la gigue.

Abstract

In this work, we consider four problems in the context of Internet traffic control.

The first problem is to understand when and why a sender that implements an equation-based rate control would be TCP-friendly, or not—a sender is said to be TCP-friendly if, under the same operating conditions, its long-term average send rate does not exceed that of a TCP sender. It is an established axiom that some senders in the Internet would need to be TCP-friendly. An equation-based rate control sender plugs-in some on-line estimates of the loss-event rate and an expected round-trip time in a TCP throughput formula, and then at some points in time sets its send rate to such computed values. Conventional wisdom held that if a sender adjusts its send rate as just described, then it would be TCP-friendly. We show exact analysis that tells us when we should expect an equation-based rate control to be TCP-friendly, and in some cases excessively so. We show experimental evidence and identify the causes that, in a realistic scenario, make an equation-based rate control grossly non-TCP-friendly.

Our second problem is to understand the throughput achieved by another family of send rate controls—we termed these "increase-decrease controls," with additive-increase/multiplicative-decrease as a special case. One issue that we consider is the allocation of long-term average send rates among senders that adjust their send rates by an additive-increase/multiplicative-decrease control, in a network of links with arbitrary fixed routes, and arbitrary round-trip times. We show what the resulting send rate allocation is. This result advances the state-of-the-art in understanding the fairness of the rate allocation in presence of *arbitrary* round-trip times. We also consider the design of an increase-decrease control to achieve a given target loss-throughput function. We show that if we design some increase-decrease controls under a commonly used reference loss process—a sequence of constant inter-loss event times—then we know that these controls would overshoot their target loss-throughput function, for some more general loss processes. A reason to study the design problem is to construct an increase-decrease control that would be friendly to some other control, TCP, for instance.

The third problem that we consider is how to obtain probabilistic bounds on performance for nodes that conform to the per-hop-behavior of Expedited Forwarding, a service of differentiated services Internet. Under the assumption that the arrival process to a node consists of flows that are individually regulated (as it is commonplace with Expedited Forwarding) and the flows are

stochastically independent, we obtained probabilistic bounds on backlog, delay, and loss. We apply our single-node performance bounds to a network of nodes. Having good probabilistic bounds on the performance of nodes that conform to the per-hop-behavior of Expedited Forwarding, would enable a dimensioning of those networks more effectively, than by using some deterministic worst-case performance bounds.

Our last problem is on the latency of an input-queued switch that implements a decomposition-based scheduler. With decomposition-based schedulers, we are given a rate demand matrix to be offered by a switch in the long-term between the switch input/output port pairs. A given rate demand matrix is, by some standard techniques, decomposed into a set of permutation matrices that define the connectivity of the input/output port pairs. The problem is how to construct a schedule of the permutation matrices such that the schedule offers a small latency for each input/output port pair of the switch. We obtain bounds on the latency for some schedulers that are in many situations smaller than a best-known bound. It is important to be able to design switches with bounds on their latencies in order to provide guarantees on delay-jitter.

Contents

Chapter 1

Introduction

> This research might seem like a strange mixture. What field is it in? Is it education? Computer science? Psychology? Epistemology? Biology? In my view, it is all of these—and necessarily so. It would be counterproductive to separate one from the others. Only by drawing on all of these domains is it possible to do justice to any of them.
>
> —Mitchel Resnik, "Turtles, Termites, and Traffic Jams," The MIT Press

1.1 Goals and Achievements

The primary goals of this dissertation are:

- Equation-Based Rate Control—to understand when and why an equation-based rate control would be either TCP-friendly or non-TCP-friendly.
- Increase-Decrease Controls—to show what is the long-term send rate allocation in a network with senders that adjust their send rates by an additive-increase/multiplicative-decrease (AIMD) control, when the round-trip time between a sender and a receiver is arbitrary. To understand the design of an increase-decrease send rate control to target a given loss-throughput function.
- Expedited-Forwarding—to obtain bounds on backlog, delay, and loss that hold in probability, for a node that conforms to the per-hop-behavior of Expedited Forwarding, a service of differentiated services Internet. To apply bounds to a network of nodes.
- Input-queued switch—to obtain bounds on the latency for a decomposition-based input-queued switch scheduler better than a best known bound.

The primary achievements are:

- Equation-Based Rate Control.

  - We have identified biases that may cause an equation-based rate control to be non-TCP-friendly, or TCP-friendly but overly conservative;
  - We have obtained sufficient conditions under which we know that some biases drive the control in either a TCP-friendly or a non-TCP-friendly direction;
  - We have provided analysis that points to the biases that result in overly smaller throughput than expected—this explains the *throughput-drop* phenomena that others have observed empirically;
  - We have provided experimental evidence and identified the causes of excessive non-TCP-friendliness in a realistic network scenario.

- Increase-Decrease Controls.
  - We have found how the network bandwidth is shared by adaptive senders that exercise AIMD send rate control, whereas a sender has an arbitrary round-trip time;
  - We have identified classes of increase-decrease controls that overshoot their target loss-throughput function, although they were designed to achieve the given target loss-throughput function for a reference process of loss-events;
  - We have obtained a bound on the throughput of an AIMD sender that depends only on the coefficient of the variation of the square-root of the number of packets sent between two successive loss-events, besides the loss-event rate.

- Expedited Forwarding.
  - We have derived bounds on backlog, delay, and loss for a superposition of independent and individually regulated arrival processes to a node that conforms to the per-hop-behavior of Expedited Forwarding;
  - We have discovered an inaccuracy in a known probabilistic bound on delay for a guaranteed rate node and provided a fix to it.

- Input-Queued Switch.
  - We have obtained bounds on the latency of some schedulers for a decomposition-based input-queued switch that are in many situations better than a previously best known bound.

1.2 Dissertation Overview

This thesis studies several traffic control problems in the Internet. Some of the problems arise in the Internet of the present, and some are studied for the Internet of the future.

1.2.1 Equation-Based Rate Control

Motivation

It is widely believed that the stability of the Internet is largely attributed to TCP, a window-based transmission control exercised end-to-end, between a sender and receiver. TCP was defined in the late 1980s (Jacobson and Karels [63]), and has been continually refined since then. Empirical evidence indicates that TCP traffic is a predominant fraction of traffic in the Internet to date: in 1998, 95% of the bytes, 90% of the packets, and 80% of the flows on a link [65]. More recent empirical evidence, from 2001, suggests that the situation remains largely unchanged [79].

TCP is purely an end-to-end protocol, relying on no particular assumptions on the control elements within a network. A TCP sender infers the state of the network from packet losses, hence feedback is in a sense implicit. With additional control elements implemented in a network, the feedback can be made explicit by conveying binary congestion notifications to a sender. This is called marking. We call "loss-events" either packet drops or marks upon which TCP is expected to react. In steady-state, the basic control law of TCP is AIMD. Roughly speaking, in the absence of loss-events, the window effectively increases by a fixed number of packets per round-trip time (typically, half a packet), otherwise, the window is halved. This control law is exercised in TCP congestion avoidance, a state of TCP finite-state machine. AIMD is not specific only to TCP; it was used in other protocols, earlier [96] and later [105].

Over more than a decade, TCP has proved to serve well applications such as bulk transfers, the web, and more recently, peer-to-peer. These are data applications, largely insusceptible to delay-jitter. On the other side of the spectrum, there are applications such as audio and video streaming that require delay-jitter to be sufficiently small to operate with an acceptable quality-of-service. TCP may not be a good choice for this set of applications. One reason is the large dynamic range of its send rate. This was a motivation to study other transmission controls. In the mid-1990s, another class of transmission controls emerged: Equation-Based Rate Control.

The essential control law of equation-based rate control, to be introduced shortly, was probably the result of a vast work on characterizing the relation of TCP loss-event rate and throughput. We call such a relation, a TCP loss-throughput formula. There have been several TCP formulas derived based on either empirical results or modeling, or both—see, for instance, [92, 85, 93, 4].

It is an established *axiom* by a part of the Internet research community that controls other than TCP should be "TCP-friendly." In this thesis, we adhere to the following definition of TCP-friendliness:

> A flow that is not "TCP-friendly" is one whose long-term arrival rate exceeds that of any conformant TCP in the same circumstances.
>
> — Floyd and Fall, 1999 [43]

In essence, the equation-based rate control law is as follows: The control

on-line estimates the loss-event rate, it uses a given TCP loss-throughput formula, $x \to f(x)$, to set its send rate at some special instants to the function $f(\cdot)$ evaluated at the current estimate of the loss-event rate. The special points in time are mainly at loss-event instants as seen by the sender. A TCP throughput formula would also depend on an expected round-trip time, which is also estimated on-line—locally, this is deliberately ignored for simplicity of discussion. The control law that we described is embodied in some particular protocols, often with some additional, refining control laws. An example is TCP-Friendly Rate Control (TFRC), which first appeared in Floyd *et al.* [45] and, at the time of this writing, advanced to a proposed standard, RFC 3448 [54].

Is Equation-Based Rate Control TCP-Friendly?

The short answer is: *No, not always.* We identified biases that can systematically drive the control to either TCP-friendliness or non-TCP-friendliness. An element of the bias is the non-linearity of the function $x \to f(x)$ and the fact that the loss is random. Another element of the bias is because the control sets the send rate at some *special* points in time; this results in the *sampling bias*. The loss-event rates as seen by different senders may differ. The same may be the case for the average round-trip times. TCP may not conform to the used TCP loss-throughput function $f(\cdot)$.

Our analysis points to some elements of the bias that in some situations can result in *over*-conservativeness of the control. This explains TFRC throughput-drop for large loss-event rates, which others observed empirically.[1]

We show experimental evidence that in a realistic scenario TFRC can be grossly non-TCP-friendly. This can happen when a few TFRC and TCP connections compete in a bottleneck. The observed non-TCP-friendliness is largely due to TFRC, which sees a smaller loss-event rate, and TCP that achieves a smaller throughput than predicted by one of its throughput formulae—used by TFRC as a control element.

TCP-Friendly—To Be or Not to Be?

The concept of TCP-friendliness implies a particular notion of fairness. It imposes one special transmission control, TCP, as a reference. Hence, it is *TCP-centric.* It is legitimate to challenge: Why a notion of fairness would need to be tied to a particular reference control?

Our experimental work indicates that TCP-friendliness is difficult to verify in practice. However, we note that the value of most of our analysis results, if not all, remains intact as for the validity of the concept of TCP-friendliness. Whenever one aims to design a control with the goal to attain "time-average send rate = f(loss-event rate)", for some function $x \to f(x)$, one will be confronted with exactly the same fundamental problems found in this thesis. An

[1] For instance, Bansal *et al.* [10] note: *"..., we also find that in return for smoother transmission rates, slowly-responsive algorithms lose throughput to faster ones (like TCP) under dynamic network conditions."*

evocative example is solving a dual problem in the utility-maximization framework, perhaps introduced first by Kelly, Maullo, and Tan [68] to study Internet congestion controls; a framework built on a notion of fairness, less restrictive than TCP-friendliness requirement.

Engineering Guideline

Our engineering guideline is to separate the elements of the bias (we often refer to as the *factors*) and study them separately. Failure to do so would blur a cause of an observed discrepancy of the throughputs. It may lead a designer to make inappropriate adjustments.

1.2.2 Increase-Decrease Controls

Fairness in Network Bandwidth Sharing Among AIMD Senders with Arbitrary Round-Trip Times

Consider a network of links. Assume that senders in the network adjust their send rates according to an AIMD control. Assume the routing path between a sender and a receiver is arbitrarily fixed in the network. We allow the round-trip time for a sender/receiver pair to be arbitrary. The question is: *"What is the long-term average rate allocation to the senders in the network?"*

We show that rates are allocated according to a specific notion of fairness. The fairness is neither max-min (Chiu and Jain, [31]) nor proportional fairness (Kelly, Maullo, and Tan [68]). It is a generalization of the result by Hurley, Le Boudec, and Thiran [60], obtained under the assumption that the round-trip times are the same for all senders in the network. For a TCP-like setting of AIMD parameters, our result sheds some light on a known bias against long round-trip time TCP connections, Floyd [41].

Analysis Problem: Throughput of an AIMD Sender

We obtain an *upper-bound* on the throughput of an AIMD sender. One merit of the bound is that it depends only on two statistical parameters of the loss process: (1) the loss-event rate, and (2) the coefficient of the variation of the square-root of the number of the packets sent in a loss-event interval. An *exact* throughput expression for an AIMD sender has been previously obtained by Altman, Avrachenkov, and Barakat [4]. Our expression requires knowledge of a fewer number of statistical parameters of the loss process.

Synthesis Problem

Consider a sender that exercises a window control. We say the control is "increase-decrease" if in the absence of loss-events, the window $W(t)$ is increased over time t with the rate of the increase $a(W(t))$, for a positive-valued function $x \to a(x)$. At an instant of a loss-event, the window is reduced to $b(W(t-))$, for a positive-valued function $x \to b(x)$.

The synthesis problem is as follows. We are given a target function, say $x \rightarrow f(x)$, which would relate the loss-event rate p to the time-average window $\bar{w}$. In particular, the function f may be of a TCP sender. The goal is to design an increase-decrease control, or in other words, to determine the functions $a(\cdot)$ and $b(\cdot)$, such that the control verifies $\bar{w} \leq f(p)$, with strict equality, in an ideal case.

We were motivated to study the design problem by diverse, often non-consistent, approaches in some work of others that aims to address the problem for some special instances of the increase-decrease controls; for instance, for AIMD by Rejaie, Handley, and Estrin [97] and Floyd, Handley, and Padhye [44], Highspeed TCP by Floyd [42], and binomial controls by Bansal and Balakrishnan [10]. Our study aims to give a coherent view of the problem.

A Design Method: Design for a Reference Loss Process

One design method is to chose functions $a(\cdot)$ and $b(\cdot)$ such that the control achieves $\bar{w}' = f(p')$, for a *reference* loss process. As a reference, we assume a sequence of *constant* inter-loss event times. This reference loss process has been frequently assumed in the related work, perhaps largely due to tractability and simplicity of the computations.

Our results are as follows. We first show a result for an AIMD sender. Assume for an AIMD sender we know $\bar{w}' = f(p')$. Then we know that for any sequence of the inter loss-event times with the same mean as in the reference process of the loss-events, it holds $\bar{w} \geq f(p)$. To paraphrase, the time-average window for a general set of the loss-events is lower-bounded by $f(p)$. In other words, an AIMD control overshoots its design goal. This result may appear intuitive. It can be derived from [4], as pointed to us by Barakat [13]. We independently arrived at this result by a different, an adversarial argument.

Our next result is new. We show that for a subset of the increase-decrease controls, which are designed such that $\bar{w}' = f(p')$, for any *renewal* point process of the loss-events, it holds $\bar{w} \geq f(p)$. Note that this is the same type of the result as before, but for a more general set of the increase-decrease controls, however, for a less general set of the loss-event processes. We show that, roughly speaking, the last property holds for Highspeed TCP [42].

1.2.3 Expedited Forwarding

Motivation

A proposal to enhance the quality-of-service of the Internet, beyond today's best-effort, is differentiated services. Expedited-Forwarding is a per-hop-behavior of the differentiated services, see RFC 3246 [35]. Expedited-Forwarding was defined, primarily, to offer a service for applications such as audio streaming, which require a low loss and delay-jitter.

For the arrival processes of bits that use the Expedited-Forwarding service, it is standard to assume that the arrival processes are individually regulated

(shaped, constrained) at the network ingress. In practice, a common regulator is a conjunction of two leaky-bucket regulators. A node is said to conform to the per-hop-behavior of Expedited Forwarding if it offers Packet Scale Rate Guarantee (PSRG) with a rate r and latency e, as defined in [35].

An engineering problem is the dimensioning of Expedited-Forwarding networks with a goal to offer a bound on the end-to-end delay-jitter, and virtually no loss in the network. To this end, some worst-case deterministic bounds on backlog, delay, and loss were obtained for a PSRG node, and a network of PSRG nodes, see Charny and Le Boudec [30] and Bennett *et al.* [15]. One problem with the deterministic worst-case approach is that it often results in a non-effective dimensioning. An alternative is to obtain bounds on the performance that hold in probability, which often yields a less pessimistic dimensioning.

Probabilistic Bounds on Backlog, Delay, and Loss

We consider a node that conforms to the per-hop-behavior of Expedited Forwarding. Under the assumptions that the arrival flows to the node are individually regulated and stochastically independent, we obtained probabilistic bounds on backlog, delay and loss. We apply our bounds to a network of nodes, under assumptions that the flows that use the Expedited-Forwarding service are individually regulated and stochastically independent at the network ingress, however, no such assumptions are made for a node in the network.

1.2.4 Input-Queued Switch

Motivation

In recent years, there has been a great deal of work on scheduling algorithms for input-queued switches. The key feature of an input-queued switch is that at each time step, each input can be connected to at most one output and each output can be connected to at most one input. The aim of the scheduler is to determine how to configure the switch at each time step so as to provide high throughput and low delays for the arriving packets.

Consider an input-queued switch. Assume that the long-term arrival rates between input/output port pairs of the switch are known. Then, we can decompose the rate matrix into a convex combination of permutation matrices. If the scheduler configures the switch according to this decomposition then we have stability. We refer to these schedulers as "decomposition-based" schedulers, see [27, 28, 81]. Specific algorithms for performing the decomposition can be derived from results of Birkhoff [16] and von Neumann [112]. Other, different approaches exist for scheduling configuration of input-queued switches; see a discussion in [8] and the references therein.

In [27] Chang, Chen, and Huang derived a worst-case deterministic bound on latency for a Birkhoff-von Neumann decomposition with an on-line scheduler of the permutation matrices. The matrices are scheduled according to Packetized Generalized Processor Sharing (PGPS) of Parekh and Gallager [94] such that

at an instant a matrix is served, it is placed as a new arrival into the PGPS system.

Having a rate-latency characterization of a node, such as a switch, is important for quality-of-service provisioning.

Latency Bounds

We show that by using probabilistic techniques we are able to tighten the bounds of Chang, Chen, and Huang [27] for the worst-case input-output pairs in many scenarios.

1.3 Dissertation Outline

The thesis is organized as follows. The four problems, Equation-Based Rate Control, Increase-Decrease Controls, Expedited-Forwarding, and Input-Queued Switch, are studied, respectively, in Chapter 1 to Chapter 4. Appendix I gives some details about our Internet and lab experiments.

At the beginning of each chapter we give a detailed outline. Whenever we felt appropriate to defer a proof to the end of an ongoing chapter, we did so.

Chapter 2

Equation-Based Rate Control

In this chapter, we study: *Is equation-based rate control TCP-friendly?* We present:

- our analysis;
- our claims suggested by the analysis;
- validation of our claims through numerical examples, ns-2 simulation, Internet and laboratory experiments.

2.1 Introduction and Outline

Suppose $\bar{x}$ is the time-average send rate of a sender that implements equation-based rate control and suppose $\bar{x}'$ is the time-average send rate of a competing TCP sender. We say the equation-based rate control is TCP-friendly iff it holds

(F) $\bar{x} \leq \bar{x}'$.

Suppose the time-average send rate of a TCP sender is characterized by $f(p', r')$, where p' and r' are the loss-event rate and the average round-trip time, respectively, as observed by the TCP sender. Assume $(x, y) \rightarrow f(x, y)$ is a non-increasing function in both x and y.

We *breakdown* the TCP-friendliness problem as follows. First, we check whether the control is *conservative*. That is

(C) $\bar{x} \leq f(p, r)$,

where p and r are the loss-event rate and average round-trip time as seen by *this* protocol. Second, we check whether the control observes the loss-event rate that is not smaller than that of TCP. That is

(P) $p \geq p'$.

Third, we pose the following ordering of the average round-trip times, r and r' as observed, respectively, by the equation-based rate control and the TCP sender,

(R) $r \geq r'$.

Forth, and last, we verify if TCP conforms to its throughput formula. In fact, we relax this condition to the inequality

(T) $\bar{x}' \geq f(p', r')$.

Note that conservativeness is *not* the same as TCP-friendliness, although there often exists a misconception in some of the protocol proposals. The conjunction of the conditions (C), (P), (R), and (T), imply TCP-friendliness. If the control is observed to be non-TCP-friendly, then *it must be* that at least one of the conditions (C), (P), (R), and (T) does not hold. The fact that conjunction of the conditions (C), (P), (R), and (T) is a sufficient condition for (F), the TCP-friendliness condition, follows from a trivial identity

$$\frac{\bar{x}}{\bar{x}'} = \frac{\bar{x}}{f(p,r)} \frac{f(p,r)}{f(p',r')} \frac{f(p',r')}{\bar{x}'},$$

and the hypothesis that $f(x, y)$ is a non-increasing function in both x and y. The breakdown of the TCP-friendliness condition (F) is more than a sufficient condition. The deviations of the terms on the left- and right-hand sides in (C), (P), (R), and (T) tell us not only about the direction of the respective biases, but also about their absolute magnitude.

Under the hypothesis that the round-trip time is fixed, we were able to carry out a detailed analysis of the conservativeness condition (C). Our analysis tells us when to expect conservative behavior and, when not to. We give analytical arguments that show (P) to hold in a limit case, when a sender has a negligible effect on the state of the network, and it is driven by the network loss process. We give experimental evidence that in a situation when a few connections compete, (P) may not hold. We give experimental evidence that TCP may not conform to its throughput formula (T); specifically, when a few connections compete, TCP may not attain the throughput predicted by the function f, with the observed values of the loss-event rate and the average round-trip time.

2.1.1 Outline of the Chapter

In Section 2.2, we define the basic control that we consider, a more comprehensive variant of the control, and the functions $f(\cdot)$ that we take as examples. In Section 2.3, we show the throughput formulae for both control. In the same section, we give our main analysis results on conservativeness, the condition (C) displayed above. We then consider other conditions: TCP's obedience to its formula in Section 2.4 and the deviation of the loss-events rates of TCP

and the equation-based rate control sender in Section 2.5. Section 2.6 validates our claims through numerical and packet-level simulations. In the second part, we remove the assumption on the round-trip times. Section 2.10 shows a throughput formula in this general case, which points to some other factors. We evaluate the factors empirically through Internet and lab experiments, in Section 2.6.3. Section 2.8 gives our conclusions. Some plots of our empirical results are deferred to the end of the chapter.

2.2 Notation and Assumptions

We consider an adaptive source with the send rate at time t equal to $X(t)$. We assume that $X(t)$ can be described by a stationary ergodic process, and thus equate the long-run average with the expected value:

$$\bar{x} = \lim_{t\to\infty} \frac{1}{t} \int_0^t X(s)ds = \mathbf{E}[X(0)].$$

Index n refers to the nth loss-event. We use the following additional notation. T_n is an instant at which a loss-event labeled with n is detected by the source. $S_n = T_{n+1} - T_n$ is the nth inter loss-event time. $X_n = X(T_n)$ is the rate set at the nth loss-event. θ_n is the number of packets sent in $[T_n, T_{n+1})$. Following TFRC, we call θ_n, the loss-event interval. Think of both S_n and θ_n as of loss-event intervals, however, the former is measured in seconds and the latter in packets.

We study long-run behavior of the control, and hence, it is more convenient to work under the convention $\cdots < T_{-1} < T_0 \leq 0 < T_1 < \cdots$. The instant 0 is an arbitrary point in time. We define $N(s,t]$ as the number of loss-events that fall in an interval $(s,t] \in \mathbb{R}$. We assume the point process of loss-events has finite non-null intensity λ. The quantity λ is a loss-event rate, the expected number of loss events on an arbitrary unit time interval. With $\mathbf{E}_N^0[\cdot]$ we denote the expectation with respect to the Palm probability (see Baccelli and Bremaud [9]). For a stationary random process $\psi(t)$, $t \in \mathbb{R}$,

$$\mathbf{E}_N^0[\psi(0)] = \frac{\mathbf{E}[\int_{(0,1]} \psi(s)N(ds)]}{\mathbf{E}[N(0,1]]}.$$

The loss-event rate as observed by the source is

$$p = \frac{1}{\mathbf{E}_N^0[\theta_0]}. \tag{2.1}$$

The quantity p is also a loss-event rate, it is the fraction of loss-events observed in the number of the packets sent over a long time interval. Let $\hat{\theta}_n$ be an estimator of the expected loss-event interval in packets, computed at T_n. We assume

(E) $\hat{\theta}_n$ is an unbiased estimator of $1/p$.

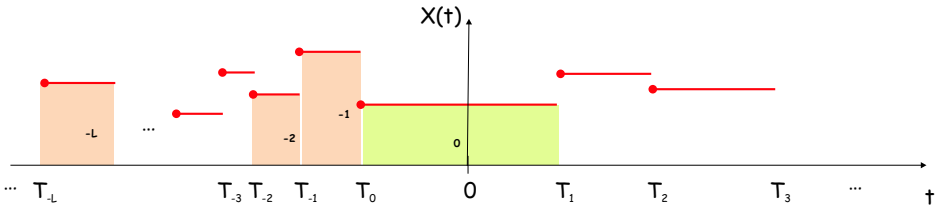


Figure 2.1: A sample-path of the send rate of *the basic control*. The drawing highlights the loss-event intervals $(\theta_{-L}, \ldots, \theta_{-2}, \theta_{-1})$ that are used in computing the estimator $\hat{\theta}_0 = \sum_{l=1}^{L} w_l \theta_{-l}$. θ_0 is the *next* loss-event interval as seen at T_0.

Moreover, we assume that $\hat{\theta}_n$ is a moving-average of the loss-event intervals, for a fixed positive integer L, and some positive-valued weights $(w_1, w_2, \ldots, w_L)$,

$$\hat{\theta}_n = \sum_{l=1}^{L} w_l \theta_{n-l}. \tag{2.2}$$

Note that by (E), the weights sum up to 1. TFRC uses this type of the loss-event interval estimator, for a particular setting of the weights, with w_l all equal for $1 \leq l \leq L/2$, else w_l linearly decreases with l.

As an aside, note that $\mathbf{E}_N^0[1/\hat{\theta}_0] \geq p$, and thus $1/\hat{\theta}_n$ is a *biased* estimator of p. This follows as a direct application of Jensen's inequality and (2.1).

2.2.1 Basic Control

The basic control is defined as follows. For a $t \in [T_n, T_{n+1})$, $n \in \mathbb{Z}$,

$$X(t) = f\left(\frac{1}{\hat{\theta}_n}\right). \tag{2.3}$$

Function f is a loss-throughput formula assumed to be positive-valued and non-increasing. See Figure 2.1 for an example sample-path.

2.2.2 Comprehensive Control

We add an additional control law to the basic control (2.3), and call the resulting system the comprehensive control. The mechanism reflects the send rate increase in the absence of the loss-events, as found in TFRC [54].

Let $\theta(t)$ be the number of the packets sent since the last loss-event observed at t. Then we define the comprehensive control as follows, for $t \in [T_n, T_{n+1})$, $n \in \mathbb{Z}$,

$$X(t) \quad = \quad f\left(\frac{1}{\hat{\theta}(t)}\right), \tag{2.4}$$

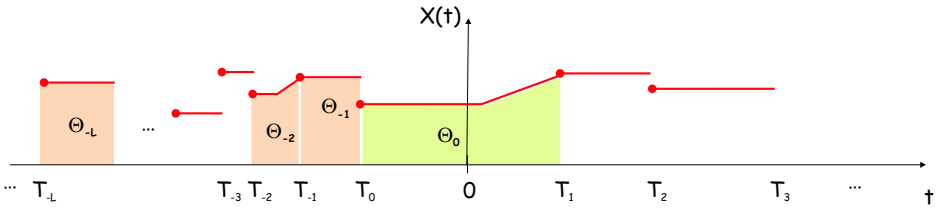


Figure 2.2: The same as in Figure 2.1, but for *the comprehensive control*. Note that the send rate can increase in the absence of loss-events.

$$\hat{\theta}(t) \quad = \quad \left(w_1\theta(t) + \sum_{l=1}^{L-1} w_{l+1}\theta_{n-l} \right) \mathbf{1}_{A_t} + (1 - \mathbf{1}_{A_t})\hat{\theta}_n.$$

Here

$$A_t = \left\{ \theta(t) > \frac{1}{w_1}[\hat{\theta}_n - \sum_{l=1}^{L-1} w_{l+1}\theta_{n-l}] \right\},$$

where $\mathbf{1}_{A_t} = 1$, if A_t is true, else $\mathbf{1}_{A_t} = 0$.

In other words, at an instant t, the loss-event interval estimator $\hat{\theta}(t)$ is updated with $\theta(t)$, if that increases the value of the estimator. If this is not the case, then $\hat{\theta}(t)$ is fixed to $\hat{\theta}_n$. Note that if the condition A_t is true, that is $\theta(t)$ is sufficiently large, the control (2.4) increases its send rate. See Figure 2.2 for an example sample-path.

Note that the send rate dynamics is such that, if $\hat{\theta}_{n+1} \leq \hat{\theta}_n$, then $X(t) = f(1/\hat{\theta}_n)$, all $t \in [T_n, T_{n+1})$. Else, for $\hat{\theta}_{n+1} > \hat{\theta}_n$, the send rate is $X(t) = f(1/\hat{\theta}_n)$, for $t \in [T_n, T_n + U_n]$, and then the rate increases according to (2.4) for $t \in (T_n + U_n, T_{n+1})$. Here, from the definition of A_t,

$$U_n = \frac{1}{w_1 f\left(\frac{1}{\hat{\theta}_n}\right)} \left(\hat{\theta}_n - \sum_{l=1}^{L-1} w_{l+1}\theta_{n-l} \right).$$

2.2.3 Some Functions $x \to f(x)$ used in the Internet

We use the following loss-throughput formulae. We first display the simplest one, "the square-root formula," which we refer to as SQRT [85]

$$f(p) = \frac{1}{c_1 r\sqrt{p}}, \tag{2.5}$$

where c_1 is a positive constant, r is the event-average of the round-trip time; the event-average is by sampling the round-trip times once in a round-trip round.

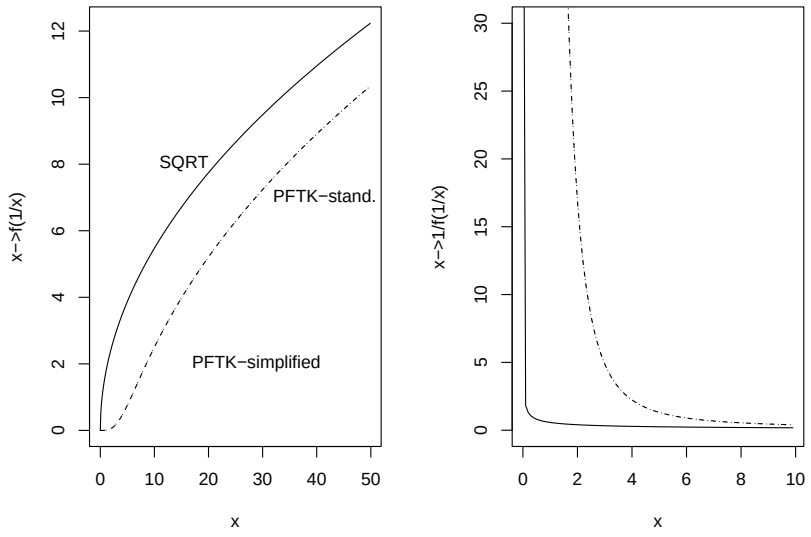


Figure 2.3: (Left) Functions of interest $x \to f(1/x)$ and $x \to 1/f(1/x)$, for SQRT, PFTK-standard, and PFTK-simplified. $r = 1$ s, $q = 4r$. The curves for the two PFTK formulae overlap. Values of x close to 0 correspond to heavy losses. (Right) The plots indicate that the convexity condition (F1) in Theorem 1 would be satisfied in all three cases, but this is strictly true only for SQRT and PFTK-simplified; it also illustrates that convexity is much more pronounced for PFTK-simplified than for SQRT. The left plots illustrate that the concavity condition (F2) of Theorem 2 is true for SQRT; for PFTK-standard and PFTK-simplified it holds only for small loss-event rates; for heavy loss (x small), the curves are convex and thus the opposite condition (F2c) holds.

We next display another well-known function f (Equation (30) in Padhye *et al.* [93]), we refer to as PFTK-standard

$$f(p) = \frac{1}{c_1 r\sqrt{p} + q \min[1, c_2\sqrt{p}](p + 32p^3)}, \tag{2.6}$$

for a positive constant c_2. The parameter q is a value of TCP retransmit timeout. We also consider a simplified version of the last formula, we call, PFTK-simplified

$$f(p) = \frac{1}{c_1 r\sqrt{p} + q c_2(p^{3/2} + 32p^{7/2})}. \tag{2.7}$$

In the formulas, $c_1 = \sqrt{2b/3}$ and $c_2 = 3/2\sqrt{3b/2}$. b is the number of the packets acknowledged by a single acknowledgment; typically $b = 2$.

PFTK formulae are *de-facto* standard. PFTK-simplified is the formula recommended in TFRC standard proposal [54], with $q = 4r$, as a recommendation. Note that, for $p \leq 1/c_2^2$, (2.7) is equal to (2.6), else, it is smaller.

We use the above particular formulae in our examples. Note that most of our findings apply to other functions f as well.

2.3 What Makes the Control Conservative or Not

We first give a throughput formula on which we build our analysis.

2.3.1 Throughput Formulae

Proposition 1 *Throughput of the basic control (2.3) is*

$$\mathbf{E}[X(0)] = \frac{\mathbf{E}_N^0[\theta_0]}{\mathbf{E}_N^0[\frac{\theta_0}{f(\frac{1}{\hat{\theta}_0})}]}. \tag{2.8}$$

For the comprehensive control, in general, we have a bound.

Proposition 2 *Throughput of the comprehensive control (2.4) is such that*

$$\mathbf{E}[X(0)] \geq \frac{\mathbf{E}_N^0[\theta_0]}{\mathbf{E}_N^0[\frac{\theta_0}{f(\frac{1}{\hat{\theta}_0})}]}.$$

The bound implies: If we know the basic control is non-conservative, then we know the comprehensive control is non-conservative as well. The converse is not true.

An exact throughput expression may be obtained for the comprehensive control for particular functions f. We obtain exact throughput expression for SQRT and PFTK-simplified functions f, as given next.

Proposition 3 *Assume f is either SQRT ($c_2 = 0$) or PFTK-simplified. Throughput of the comprehensive control is*

$$\mathbf{E}[X(0)] = \frac{\mathbf{E}_N^0[\theta_0]}{\mathbf{E}_N^0[\frac{\theta_0}{f(\frac{1}{\hat{\theta}_0})}] - \mathbf{E}_N^0[V_0 \mathbf{1}_{\hat{\theta}_1 > \hat{\theta}_0}]}, \tag{2.9}$$

where

$$\begin{aligned} V_n &= \frac{1}{w_1}\Big[-2c_1 r(\hat{\theta}_{n+1}^{\frac{1}{2}} - \hat{\theta}_n^{\frac{1}{2}}) + 2c_2 q(\hat{\theta}_{n+1}^{-\frac{1}{2}} - \hat{\theta}_n^{-\frac{1}{2}}) + \\ &\quad + \frac{64}{5} c_2 q(\hat{\theta}_{n+1}^{-\frac{5}{2}} - \hat{\theta}_n^{-\frac{5}{2}}) + (\hat{\theta}_{n+1} - \hat{\theta}_n)\frac{1}{f(1/\hat{\theta}_n)}\Big]. \end{aligned}$$

Note that, in view of the above propositions and the definition of $\hat{\theta}_n$ (2.2), the throughput of both basic and comprehensive control is expressed in terms of the expected values of some functions of a sequence of $L+1$ loss-event intervals, $\theta_0, \theta_{-1}, \ldots, \theta_{-L}$. Therefore, knowing the joint probability law of $\theta_0, \theta_{-1}, \ldots, \theta_{-L}$ would, at least in theory, enable us to compute the throughput, and explain how the "correlation structure" of the loss process plays a role.

2.3.2 Conditions for the Basic Control to be Conservative

Consider the basic control. We give exact sufficient conditions for conservativeness, or non-conservativeness. The results have interest of their own—they suggest the key factors that can cause conservativeness.

Sufficient Conditions for the Basic Control to be Conservative

Theorem 1 *Assume that*

(F1) $x \to \frac{1}{f(1/x)}$ *is convex,*

(C1) $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0] \leq 0$.

Then the basic control (2.3) is conservative.

Interpretation. The convexity condition (F1) is satisfied by the SQRT loss-throughput formula, and by PFTK-simplified; it is not satisfied by PFTK-standard, but *almost* (we will come back to this in a few lines). This is straightforward to demonstrate and can also be seen in Figure 2.3. The figure also shows that convexity is much more pronounced for PFTK formulae, and thus, we should expect more conservativeness with PFTK than with SQRT formula (this is confirmed numerically in Section 2.3).

Condition (C1) is true in particular when the covariance is 0, which happens when successive loss-event intervals are (stochastically) independent. There are indications in the empirical study by Zhang *et al.* [115] that this may be true, and the theorem says that this would lead to a conservative behavior. We show later in Section 2.6.3 our own experimental evidence that confirms (C1) to be mostly true.

We give the following more explicit statement, which gives a bound on the throughput

$$\mathbf{E}[X(0)] \leq f(p) \frac{1}{1 + \frac{f'(p)p^3}{f(p)} \mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]}. \tag{2.10}$$

This shows that, in most cases, if the covariance is positive but small, there cannot be a significant non-conservativeness of the basic control.

The theorem says more. Remember that $\hat{\theta}_n$ is an incremental estimator of the Palm expectation of the loss-event interval, $1/p$, built on the information available up to the loss-event n, whereas θ_n is the *true* next loss-event interval. Both have the same expectation, as we assumed that $\hat{\theta}_n$ is unbiased. However,

this does not mean that $\hat{\theta}_n$ is a good *predictor* of θ_n. This depends on the joint statistics, in particular the autocovariance of the loss process. The covariance of θ_n and $\hat{\theta}_n$ reflects how good a predictor $\hat{\theta}_n$ is. Condition (C1) means that $\hat{\theta}_n$ is a bad predictor, and, maybe surprisingly, the theorem suggests that this leads to a conservative behavior. Conversely, consider now a hypothetical case where the loss process goes into phases, with slow transitions. Then the loss-event interval becomes highly predictable, that is, $\hat{\theta}_n$ will now be a good predictor of θ_n; the theorem does not say that this alone will make the control non-conservative. However, this may really happen, as we find in Section 2.6. We give another, perhaps more realistic example in Section 2.3.2.

Note that $\hat{\theta}_n$ is the moving-average estimator in (2.2), and thus

$$\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0] = \sum_{l=1}^{L} w_l \mathbf{cov}_N^0[\theta_0, \theta_{-l}]. \tag{2.11}$$

In other words, it depends only on the weighted sum of the autocovariance of the loss-event intervals at the lags 1 to L.

The following corollary was shown in the discussion above.

Corollary 1 *If the convexity condition (F1) holds and the loss-event intervals are independent, then the basic control (2.3) is conservative.*

When Convexity is Almost True. The convexity condition (F1) is not true for PFTK-standard (because of the min term), but almost, as we see now. For a function $x \to g(x)$, we quantify its deviation from convexity by the ratio to its convex closure

$$r = \sup_x \left\{ \frac{g(x)}{g^{**}(x)} \right\}.$$

The convex closure $g^{**}(x)$ is the largest convex function that lower bounds $g(x)$; it is obtained by applying convex conjugation twice [101]. Figure 2.4 shows $g(x) = 1/f(1/x)$ for PFTK-standard and its convex closure; here, we have $r = 1.0026$.

Proposition 4 *Assume that a loss-throughput formula f is such that $1/f(1/x)$ deviates from convexity by a ratio r, and that (C1) holds. Then the basic control (2.3) cannot overshoot by more than a factor equal to r.*

Thus, considering that a fraction of a percent is more than reasonably accurate, we can conclude that for practical purposes, we can proceed as if PFTK-standard satisfies the convexity condition (F1).

When the Sufficient Conditions do not Hold

We give a different set of conditions, which provides additional insights. This new set of conditions applies to some cases where Theorem 1 does not apply.

Theorem 2 *Assume that*

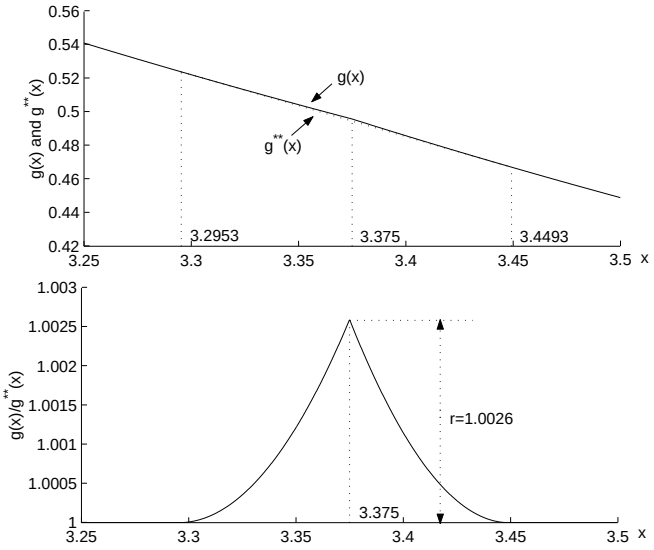


Figure 2.4: The top figure shows $g(x) := 1/f(1/x)$ when $f()$ is PFTK-standard and its convex closure (dotted line). On the interval shown in the top figure, g^{**} is equal to the tangent common to both ends of the graph. Outside the interval it is equal to g. $g(\cdot)$ is not strictly speaking convex, but almost. The bottom figure shows the ratio g/g^{**}, which is bounded by $r = 1.0026$.

(F2) *$x \to f(x)$ is concave,*

(C2) $\mathbf{cov}_N^0[X_0, S_0] \leq 0$.

Then the basic control (2.3) is conservative.
Conversely, if

(F2c) *$x \to f(x)$ is strictly convex,*

(C2c) $\mathbf{cov}_N^0[X_0, S_0] \geq 0$,

(V) *The loss-event estimator $\hat{\theta}_n$ has non-zero variance.*

Then the basic control (2.3) is non-conservative.

Interpretation. The concavity condition (F2) is true for the SQRT formula. In contrast, PFTK-standard and PFTK-simplified are such that concavity (F2) is true for rare losses, but convexity (F2c) is true for frequent losses; see Figure 2.3, the left graph. The covariance condition (C2) is between X_n, the send rate set at the occurrence of the nth loss-event, and S_n, the time until the next loss-event. If the loss process is memory-less and independent of the activity of our source, then the duration S_n of the loss interval is negatively

correlated with the send rate X_n in the given interval (since S_n is counted in real time, not per packet); in such cases, condition (C2) is true, and the basic control is conservative, as long as losses are rare to moderate (or if the SQRT formula is used). This part of Theorem 2 complements Theorem 1.

Consider now the second part of Theorem 2. Assume that $\{S_n\}_n$, a sequence of the inter loss-event times, is independent of the send rate. This may happen, for example, for an audio source that modulates its send rate by varying the packet lengths rather than the packet send rate, and if the packet dropping probability in routers is independent of packet length; for instance, with RED[1] operating in the packet mode. Then (C2c) holds, with equality. Now assume also that PFTK-standard is used, and the network setting happens to be such that the loss-event interval θ_n is mostly in the region where PFTK-standard is convex (that is, heavy losses). The theorem says that the basic control is non-conservative, except in a degenerate case where there is no randomness in the system, i.e. the loss-event interval estimator has converged to a fixed value. We show simulations that illustrate this case in Section 2.6.2.

Another example is for a more traditional source such as TFRC, but when the loss process goes through phases (for example, the network paths used by the flow oscillate between congestion and no congestion), and the send rate roughly follows the phases, that is, it is responsive at the time scale of the loss process. Then, when the network is in a congestion phase, X_n is most often small, and because of congestion, S_n is small. In such a case, condition (C2c) may be true and the basic control may *not* be conservative. In Section 2.6 we show such cases.

Viewpoint Matters. The first part of Theorem 2 illustrates well the importance of Feller's paradox-type of arguments used in this paper; also known as "the bus-stop paradox." The send rate $X(t)$ is updated only at loss-event instances. Consider an observer who picks an arbitrary point in time. This observer is more likely to fall in a large inter loss-event interval S_n. Given that S_n is negatively correlated with X_n, it is thus more likely that on average our observer will see a smaller rate than another observer that would sample the send rate at the loss-event instants. From this we conclude $\mathbf{E}[X(0)] \leq \mathbf{E}_N^0[X(0)]$. Now, the concavity assumption (F2), by Jensen's inequality, shows in turn that $\mathbf{E}_N^0[X(0)] \leq f(p)$. Finally, it follows $\mathbf{E}[X(0)] \leq f(p)$, that is, the control is conservative.

Note that the correlation condition (C2) in Theorem 1 is implied by the condition that the conditional expected duration S_n, given the send rate X_n, decreases with X_n. That is

(C3) $\mathbf{E}_N^0[S_0|X_0 = x]$ is non-increasing with x.

This is a direct consequence of Harris' inequality[2] that (C3) implies the negative correlation condition (C2).

[1] Random Early Discard [46], a popular active queue management scheme.

[2] Harris' inequality says that if $f(x)$ and $g(x)$ are non-decreasing functions, and X is a random variable, then the covariance of $f(X)$ and $g(X)$ is non-negative. See, for example [9], p. 225

Of course, we should expect that the combination of (C2c) and (V) implies that (C1) does not hold. This indeed holds and is shown in the appendix.

It is legitimate to wonder whether Theorem 1 is derived from Theorem 2 or *vice versa*. This does not seem to be the case; we discuss this in the appendix. Note, however, if the concavity condition (F2) holds, then the convexity condition (F1) necessarily also holds. The converse is not true.

2.3.3 What This Tells Us

The analytical results in the previous section are for the basic control. We expect the comprehensive control to give a slightly larger throughput, since it differs by an additional increase during a long loss-event interval. This motivates us to pose as assumptions the following analysis. We confirm our claims later by experiments.

Claim 1 *Assume that the loss-event interval θ_n and the loss-event interval estimator $\hat{\theta}_n$ are slightly positive or negatively correlated. Consider a region where the loss-event interval estimator $\hat{\theta}_n$ takes its values.*

- *The more convex $1/f(1/x)$ is in this region, the more conservative the control is.*
- *The more variable $\hat{\theta}_n$ is, the more conservative the control is.*

Application. For protocols like TFRC, we expect the condition to hold in many practical cases; again, we refer to the empirical findings by Zhang *et al.* [115], and our own experimental evidence shown later. For the three functions we consider in this paper, $x \to 1/f(1/x)$ is more convex for small x, that is, for large loss-event rates p. Thus, the control should be more conservative with high loss than low loss. This effect is more pronounced for PFTK-standard (2.6) and PFTK-simplified (2.7), which are convex and very steep for large p, than for SQRT. (Recall the right graph in Figure 2.3.) This explains the observed *throughput-drop* for the control, with PFTK and heavy losses.

The "variability" of $\hat{\theta}_n$ depends on the variability of the loss-event intervals, and can be controlled by L, the window length of the moving-average estimator. With an appropriate setting of the weights $w_1, w_2, \ldots, w_L$, the larger the window of the estimator L, the smaller the variability of the estimator $\hat{\theta}_n$. We should find that for larger L the control becomes less conservative.

Our second claim concerns the case where the conditions in Claim 1 do not hold.

Claim 2 *Assume that the inter loss-event time S_n and the send rate X_n are negatively correlated or non-correlated.*

- *If $f(1/x)$ is concave in a region where the loss-event interval estimator $\hat{\theta}_n$ takes its values, the control tends to be conservative.*

Conversely, assume S_n and X_n are positively correlated or non-correlated.

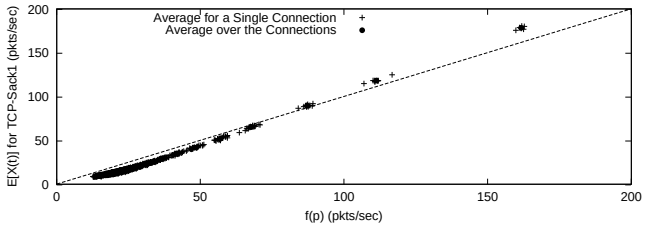


Figure 2.5: TCP Sack 1 versus PFTK-standard formula. Throughput is below the throughput predicted by the formula, except for large throughputs.

- *If $f(1/x)$ is strictly convex in a region where the loss-event interval estimator $\hat{\theta}_n$ takes its values, and $\{\theta_n\}_n$ is not fixed to a constant, the control is* non-conservative.

In both cases, the more variable $\hat{\theta}_n$ is, the more pronounced the effect is.

Application. We expect to have close to no correlation for adaptive audio applications such as [22] when packet losses in RED routers are independent of packet length. Thus, depending on which convexity condition holds, we will find one or the other outcome. For SQRT, the control should always be conservative. The same holds for PFTK with light to moderate losses. The *opposite* holds for either PFTK formulae with heavy losses.

2.4 Does TCP Conform to a TCP Throughput Formula?

The conformance of a TCP implementation to a TCP loss-throughput formula depends on: details of a TCP implementation, the hypotheses under which a TCP loss-throughput formula is derived. The formulas like PFTK are derived under a number of simplifying assumptions and some fixed point approximations. We give a more in-depth analysis of the throughput deviation from a value predicted by a throughput formula in Chapter 3. In the present section, we show a result of ns-2 simulation in Figure 2.5. We observe that PFTK-standard may not be an accurate predictor of the throughput.

We come back to the issue of TCP conformance to PFTK formulae, shortly in this chapter, when we show our empirical results obtained by Internet and laboratory experiments.

2.5 How do the Loss-Event Rates Compare?

We now focus on how the loss-event rates seen by an equation-based rate control and TCP would compare. Unlike the problem of conservativeness, this problem

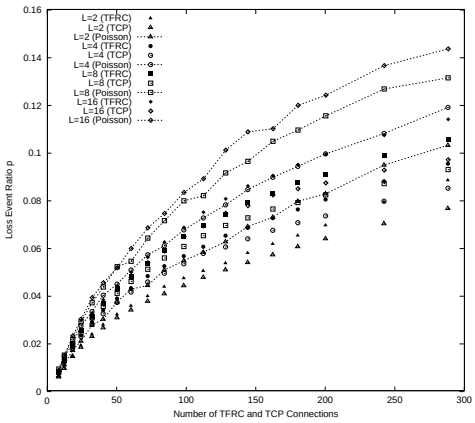


Figure 2.6: Loss-event rate as experienced by TFRC, TCP, and Poisson connections versus N (number of TFRC and number TCP connections in one bottleneck). We have $p' \leq p \leq p''$ as expected. Also, the smoother the TFRC flows (larger L), the larger the loss-event rate.

does not allow us to conclude about the ordering of the loss-event rates, unless some further assumptions are made. Note that the loss-event rates depend on the interaction of the connections that share a network. Yet, a claim can be made in a limiting case.

2.5.1 Many-Sources Regime

Assume that senders in the network are driven by a congestion process $Z(t)$ that evolves in real time, $t \in \mathbb{R}$. This is an approximation that fits with the case of a source with negligible influence on a global network. Assume the congestion process takes values on E, a countable state space. The state transitions are clocked by a point process $\cdots < T'_{-1} < T'_0 \leq 0 < T'_1 < \cdots$. We assume this point process is stationary and has finite non-null intensity λ'. Let N' be the associated counting process.

Let $\pi_i := \mathbf{P}[Z(0) = i]$ be the steady-state probability that the congestion process is in the state $i \in E$. Define

$$p_i = \frac{1}{\mathbf{E}_N^0[\theta_0 | Z(0) = i]}.$$

This is the loss-event rate, given that the congestion process is in the state $i \in E$. Let, also, $\bar{x}_i = \mathbf{E}[X(0)|Z(0) = i]$ be the time-average send rate, conditioning on that the congestion process is in the state i. We show in the appendix

$$p_i = \frac{\sum_{i \in E} b_i p_i \bar{x}_i \pi_i}{\sum_{i \in E} b_i \bar{x}_i \pi_i}, \tag{2.12}$$

where

$$b_i = \frac{\mathbf{E}^0_{N'}[\sum_{n\in\mathbb{Z}} \theta_n \mathbf{1}_{[0,S'_0)}(T_n)|Z(0)=i]}{\mathbf{E}^0_{N'}[\int_0^{S'_0} X(s)ds|Z(0)=i]}.$$

In the limit case, as $\frac{\lambda'}{\lambda_i} \to 0$, $i \in E$, $b_i \to 1$. Here, by definition, $\lambda_i = 1/\mathbf{E}^0_N[S_0|Z(0)=i]$ is the intensity of the loss-events in real time, given the congestion process is in the state $i \in E$. The limit corresponds to a separation of timescales; we assume the congestion process evolves over a larger timescale than is the timescale of the control; remember that the control is clocked by the loss-events. We base our further discussion on the loss-event rate, in this limit case,

$$p \to \frac{\sum_{i\in E} p_i \bar{x}_i \pi_i}{\sum_{i\in E} \bar{x}_i \pi_i}. \tag{2.13}$$

If our source is non-adaptive, say a homogeneous Poisson, then $\bar{x}_i = \bar{x}$ is independent of i. The resulting loss-event rate $p'' = \sum_{i\in E} \pi_i p_i$ can be thought of as the time-average of the *network* loss-event rate. It should be close to what a constant bit rate (CBR) source would experience. Now if, like TCP, our source is very responsive, that is, it follows the congestion process closely, then $\bar{x}_i$ depends on i in the following way: $\bar{x}_i$ is large for "good" states (p_i small) and small for bad states (p_i large). Thus, we should have a smaller p. For TCP, this is confirmed by the measurements of Paxson [95]. The more responsive the sender is, the more pronounced this should be. TCP is expected be more responsive than our adaptive sender, whose responsiveness depends on the averaging window L. We summarize this as follows (see Figure 2.6 for an illustration).

Claim 3 *In the many-sources regime, the loss-event rates of TCP (p'), an equation based-rate controlled sender (p), and a non-adaptive sender (Poisson) (p'') should satisfy the relation*

$$p' \le p \le p''.$$

The more responsive an equation-based rate controlled sender is, the closer p should be to p'.

2.6 Experimental Validation

In this section, we validate our claims by:

- numerical experiments for the basic and comprehensive control;
- ns-2 experiments for TFRC.

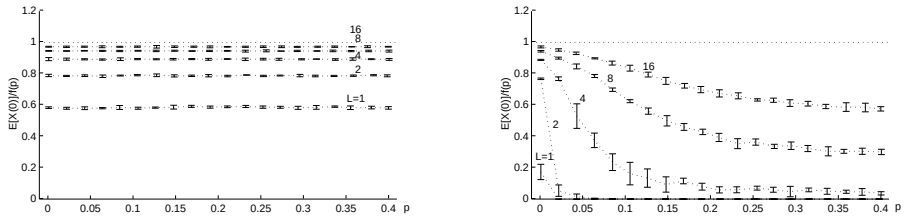


Figure 2.7: Normalized throughput $\mathbf{E}[X(0)]/f(p)$ versus p, for the basic control with $\mathbf{cv}_N^0[\theta_0]$ fixed to 9/1000. (Left) SQRT, (Right) PFTK-simplified with $q = 4r$. The estimator weights are as of TFRC, with length L.

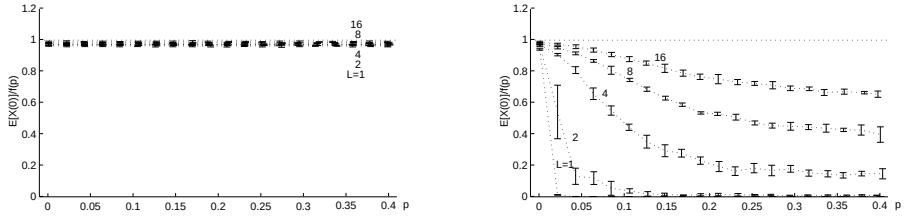


Figure 2.8: The same as in Figure 2.7, but for the comprehensive control.

2.6.1 Numerical Experiments

We validate the claims, we made based on our analysis, by numerical examples. Such a numerical study enables us to isolate individual factors that we expect to contribute to either conservative or non-conservative behavior.

All the results in this section are based on numerical investigations of the basic control and the comprehensive control, with functions SQRT or PFTK-simplified. For PFTK-standard, we rely on ns-2 simulations shown in Section 2.6.2; in view of the claims, the results do not differ significantly.

Validation of Claim 1

We assume the loss-event intervals form a sequence of independent, identically distributed random variables, with θ_0 having the density function $\mu(x) = \lambda \exp(-\lambda(x-x_0))$, for $x \geq x_0$, and $\lambda, x_0 \geq 0$. In other words, θ_0 is a sum of a positive constant x_0 and a random variable with exponential density (λ). We chose μ as defined above because it has some desirable properties: $\mathbf{E}_N^0[\theta_0] = x_0 + 1/\lambda$, the coefficient of the variation $\mathbf{cv}_N^0[\theta_0] = \frac{1/\lambda}{x_0+1/\lambda}$, the skewness and kurtosis[3] equal to 2 and 6, respectively. Note that μ has two degrees of freedom. It allows

[3]Skewness quantifies the skewness of a probability distribution; it is the ratio of the third-order centered moment and the standard deviation to the power three. Kurtosis parameter quantify the sharpness of a probability distribution; it is the ratio of the forth-order centered moment and the standard deviation to the power four.

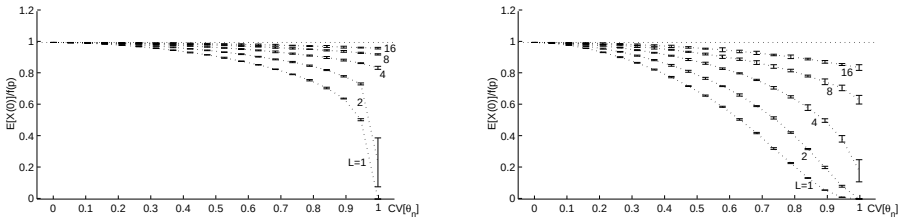


Figure 2.9: Normalized throughput $\mathbf{E}[X(0)]/f(p)$ of the basic control versus the coefficient of the variation of $\{\theta_n\}_n$, with p fixed to (Left) 0.01, (Right) 0.1. Function f is PFTK-simplified with $q = 4r$. The estimator weights set as of TFRC.

us to vary either $\mathbf{E}_N^0[\theta_0]$ or $\mathbf{cv}_N^0[\theta_0]$, whereas the other of these two parameters is kept fixed. At the same time, skewness and kurtosis parameters remain fixed. Thus μ enables us to separate the effects due to convexity of $1/f(1/x)$ and variability of $\hat{\theta}_n$. With some other distributions, geometric (p), for instance, we would have $\mathbf{cv}_N^0[\theta_0] = \sqrt{1-p}$. In this case, the variability of θ_0 would decrease as we increase p.

We compute the throughput $\mathbf{E}[X(0)]$ numerically for the basic and comprehensive control from Equation (2.8) and (2.9), respectively. The results are obtained by Monte Carlo simulations. We run 5 independent replications, each of 10000 samples. The confidence intervals are for 0.95 confidence.

Our first goal is to evaluate the impact of convexity of the function $1/f(1/x)$. To that end, we fix $\mathbf{cv}_N^0[\theta_0]$. In Figure 2.7, we show the normalized throughput $y := \mathbf{E}[X(0)]/f(p)$ versus p for the basic control with SQRT and PFTK-simplified functions f. The values of y are not larger than 1, which corresponds to the conservative behavior. For SQRT function, we observe that for each fixed value of the averaging window L, the normalized throughput y seems not to depend on p. This is indeed true, as the next simple analysis shows. By Taylor development of $1/f(1/x) = K/\sqrt{x}$, for a positive constant K, around $1/p$, we obtain

$$y = \frac{1}{\sum_{n=0}^{\infty}(-1)^n \frac{1\cdot 3\cdot 5\cdots(2n-1)}{2^n n!}\mathbf{E}[p\theta_0(p\hat{\theta}_0 - 1)^n]},$$

where we assume we can interchange the infinite sum and the expectation. For $\{\theta_n\}_n$, an i.i.d. sequence of random variables with mean $1/p$, we have

$$\mathbf{E}[p\theta_0(p\hat{\theta}_0 - 1)^n] = \mathbf{E}[(p\hat{\theta}_0 - 1)^n].$$

Clearly, if the distribution of $p\theta_0$ does not depend on p, then for SQRT, y does not depend on p. The last property indeed holds for the distribution in our example. Recall that in the example, θ_0 is equal in distribution to $x_0 + Y$, where x_0 is a positive constant, and Y a random variable with $\mathbf{P}[Y > y] = \exp(-\lambda y)$, $y \geq 0$. Indeed, $\mathbf{P}[p\theta_0 > t] = \mathbf{P}[Y > (t - px_0)/p] = \exp(-\lambda(t - px_0)/p)$. Now, recall, or observe directly that $1/p = x_0 + 1/\lambda$, which can be re-written as

$1/\lambda = (1 - px_0)/p$. Recall, also, that we fix the coefficient of the variation to a positive constant c, which gives $1 - px_0 = c$. Putting the pieces together, we obtain $\mathbf{P}[p\theta_0 > t] = \exp(-(t - 1 + c)/c)$, a function that does not depend on p. One may compute y explicitly for a particular loss process. An example is for $\{\theta_n\}_n$ an i.i.d. sequence of random variables, with θ_0 exponentially distributed. Then, for SQRT function and uniform weights of the loss-event estimator, a simple calculation reveals

$$y = \frac{(L-1)!}{\sqrt{L}\Gamma(L - \frac{1}{2})}.$$

As expected, y does not depend on p. Here $\Gamma(\cdot)$ is Euler's gamma function.

From Figure 2.7, we observe that for PFTK-simplified, y decreases toward 0 as p becomes larger. This explains the throughput-drop for heavy losses.

In Figure 2.8, we show the corresponding results for the comprehensive control. The results are qualitatively the same as the respective results for the basic control shown in Figure 2.7. For SQRT, the normalized throughputs are less, but fairly near to the ideal value 1. For PFTK-simplified function, the results are somewhat less conservative than for the basic control.

Next we investigate the effect of the variability of $\hat{\theta}_n$. To that end, we fix $p = 0.01$ and 0.1. See Figure 2.9, for numerical results for the basic control with PFTK-simplified. We observe, the larger the variability of $\hat{\theta}_n$, the more conservative the control is. This is indeed more pronounced for larger p due to the larger convexity and steepness of $1/f(1/x)$ for small x (large p) with the PFTK-simplified function.

Lastly, note from Figure 2.7, Figure 2.8, and Figure 2.9 that the normalized throughput depends on the averaging window L of the loss-event interval estimator in the following manner: The larger the L, the smaller the variability of $\hat{\theta}_n$, and, consequently, the larger the normalized throughput y. This is in accordance with Claim 1.

Validation of Claim 2

We do additional experiments to verify Claim 2, which, incidentally, also provide some examples of non-conservative behavior. Assume there exists a hidden Markov chain (HMC) that governs loss-events. We define the HMC $\{Z_n\}_n$ to be clocked at the loss-event instances. Assume $\{Z_n\}_n$ takes values on a countable state space E. Let $\mathbf{P} = [p_{ij}]$ be the matrix of transition probabilities, assumed to be irreducible and positive recurrent. Let $\boldsymbol{\pi}_i = \mathbf{P}[Z_0 = i]$, $i \in E$, denote the stationary distribution. Assume

$$\mathbf{P}[\theta_n = m | Z_n = i, Z_{n-1}, \ldots, \theta_{n-1}, \theta_{n-2}, \ldots] = \mathbf{P}[\theta_n = m | Z_n = i].$$

In other words, conditional on that at the nth loss-event the HMC is in the state i, θ_n is independent of all the past. Let $g_i(m) := \mathbf{P}[\theta_n = m | Z_n = i]$. Note that $\{\theta_n, Z_n\}_n$ is a Markov renewal process with $\mathbf{P}[Z_{n+1} = j, \theta_n = m | Z_n = i] = p_{ij} g_i(m)$.

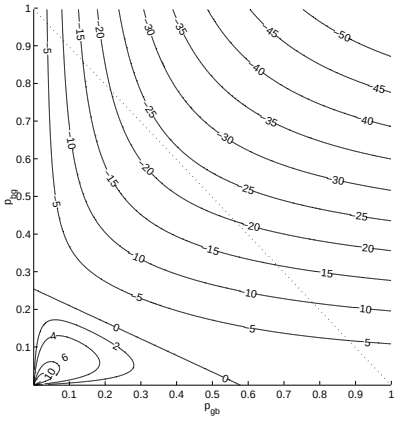


Figure 2.10: The graph shows contour plot of $\mathbf{cov}_N^0[X_0, S_0]/(\mathbf{E}_N^0[X_0]\mathbf{E}_N^0[S_0])$ versus p_{gb} and p_{bg}; $n_g = 200$ and $n_b = 50$. Function f is PFTK-simplified with $r = 100$ ms and $q = 4r$.

For the basic control, from (2.8),

$$\mathbf{E}[X(0)] = \frac{\sum_{i \in E} e(i_0)\pi_i}{\sum_{\underline{i} \in E^{L+1}} e(i_0)g(i_1, \ldots, i_L)p_{i_0 i_1} \cdots p_{i_{L-1} i_L}\pi_{i_L}},$$

where

$$g(i_1, \ldots, i_L) = \mathbf{E}_N^0\left[\frac{1}{f(1/\hat{\theta}_0)} | Z_{-1} = i_1, \ldots, Z_{-L} = i_L\right],$$

and $e(i) = \mathbf{E}_N^0[\theta_0 | Z_0 = i]$. Likewise, one obtains the throughput expression for the comprehensive control.

We consider a special, but instructive case: a 2-state HMC and $L = 1$. Without loss of generality, we call one state the *good* state, and the other, the *bad* state. We respectively label the state space as $E = \{g, b\}$. In addition, we assume two fixed integers $n_g \geq n_b$ such that $g_a(n_a) = 1$, $g_b(n_b) = 1$. In other words, when the HMC is in the good state (resp. bad state), the loss-event interval is fixed to n_g (resp. fixed to n_b).

Under the above assumptions, we have

$$\mathbf{E}[X(0)] = \frac{p_{bg}n_g + p_{gb}n_b}{p_{bg}\frac{n_g}{f(1/n_g)} + p_{gb}\frac{n_b}{f(1/n_b)} + p_{gb}p_{bg}h(n_g, n_b)},$$

where, for the basic control,

$$h(n_g, n_b) = \left(\frac{1}{f(1/n_b)} - \frac{1}{f(1/n_g)}\right)(n_g - n_b),$$

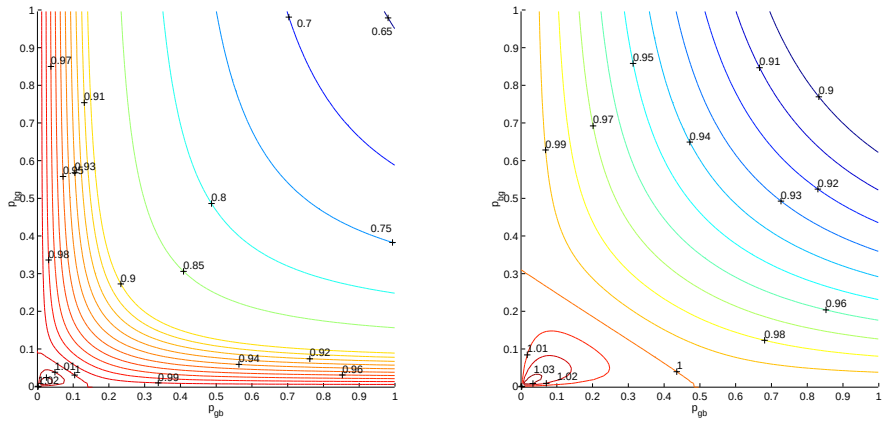


Figure 2.11: (Left) basic control and (Right) comprehensive control. The graphs show the contour plots of the normalized throughput $\mathbf{E}[X(0)]/f(p)$ versus p_{gb} and p_{bg}. $n_g = 200$ and $n_b = 50$. The function f is PFTK-simplified with $r = 100$ ms and $q = 4r$.

and, for the comprehensive control,

$$h(n_g, n_b) = 2c_1 r(n_g^{\frac{1}{2}} - n_b^{\frac{1}{2}}) - 2c_2 q(n_g^{-\frac{1}{2}} - n_b^{-\frac{1}{2}}) - \frac{64}{5} c_2 q(n_g^{-\frac{5}{2}} - n_b^{-\frac{5}{2}}) - \frac{n_g - n_b}{f(1/n_g)}.$$

We next examine the covariance $\mathbf{cov}_N^0[X_0, S_0]$ for our 2-state HMC. Indeed, X_n and S_n being negatively correlated or non-correlated is equivalent to $\mathbf{cov}_N^0[X_0, S_0] \leq 0$. In Figure 2.10, we show a plot of $\mathbf{cov}_N^0[X_0, S_0]$ versus the transition probabilities p_{gb} and p_{bg} . Observe that the covariance is positive for small values of p_{gb} and p_{bg}, which corresponds to a slow dynamics of our 2-state HMC. Note that for the slow HMC limit, and f the SQRT function, we have $\mathbf{cov}_N^0[X_0, S_0] \to \mathbf{var}_N^0[\sqrt{\theta_0}]$; thus a positive value, increasing in the variability of θ_0. In view of our Claim 2, we expect to find non-conservative behavior when the dynamics of the HMC is slow, which we confirm next.

We first consider the basic control with PFTK-simplified formula. In Figure 2.11, we show the normalized throughput $\mathbf{E}[X(0)]/f(p)$ versus the transition probabilities p_{gb} and p_{bg} of the HMC. n_g and n_b are set to 200 and 50, which correspond to the loss-event rates 5/1000 and 2/100, while in the good and the bad state, respectively. Note that we do find some slight overshoot in the lower left corner of the graphs (normalized throughput greater than one).

Note that for the given values of n_g and n_b the function $f(1/x)$ is concave with x in the region where x takes its values. Further, observe from Figure 2.10 and Figure 2.11 that whenever $\mathbf{cov}_N^0[X_0, S_0]$ is not positive, the control is conservative. The last two observations together confirm the first statement of Claim 2. The second statement of Claim 2 we do not verify here, but by ns-2 simulation in Section 2.6.2. Further numerical examples, with another model

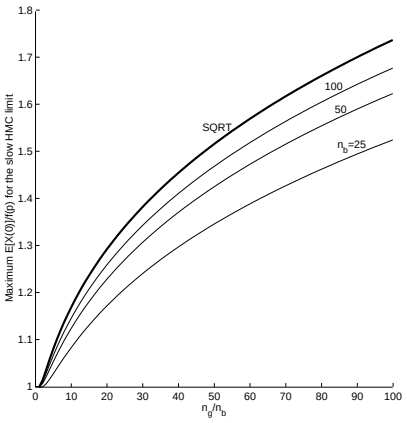


Figure 2.12: The maximum normalized throughput $\mathbf{E}[X(0)]/f(p)$ attained in the slow HMC limit versus n_g/n_b; thick line is for SQRT; thin lines are for PFTK-simplified ($r = 100$ ms, $q = 4r$); n_b is set as indicated in the graph.

that verify the hypotheses of Claim 2 can be found in [108].

We give some further observations. By Corollary 1 we should find the conservative behavior for $p_{gb} + p_{bg} = 1$ (note that this is a degenerate case, $\{\theta_n\}_n$ i.i.d.), which we confirm to be the case. We note that very conservative behavior occurs for $p_{gb} + p_{bg} > 1$, where $\mathbf{cov}_N^0[X_0, S_0]$ is negative, but also $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]$ may be negative; this is to be expected from the bound on the throughput in Equation (2.10). Another observation is that non-conservative behavior happens for positively correlated loss-event intervals ($p_{gb} + p_{bg} \leq 1$), in particular, for small values of p_{gb} and p_{bg} (the slow HMC). In the remainder of this section, we discuss this limit case in some more detail.

We show that for a slow dynamics of the HMC, the control may have a substantial overshoot, as opposed to a modest overshoot observed in Figure 2.11. We define the slow HMC limit as $p_{gb}, p_{bg} \to 0$, $p_{gb} = up_{bg}$, for a fixed $u > 0$. Then, for both the basic and comprehensive control we obtain

$$\mathbf{E}[X(0)] \to \frac{p_{bg}n_g + p_{gb}n_b}{p_{bg}\frac{n_g}{f(1/n_g)} + p_{gb}\frac{n_b}{f(1/n_b)}} = \frac{\mathbf{E}_N^0[\theta_0]}{\mathbf{E}_N^0[\frac{\theta_0}{f(1/\theta_0)}]}.$$

The normalized throughput $\mathbf{E}[X(0)]/f(p)$ of the slow HMC limit is

$$\bar{x}^0(u) = \frac{1}{f(\frac{u+1}{un_g+n_b})} \frac{un_g + n_b}{u\frac{n_g}{f(1/n_g)} + \frac{n_b}{f(1/n_b)}}. \tag{2.14}$$

For a given function f, one may compute u^* at which the global maximum of $\bar{x}^0(u)$ is attained. For SQRT function, we obtain $u^* = \sqrt{n_b/n_g}$, that is

$$\frac{p_{bg}}{p_{gb}} = \sqrt{\frac{n_b}{n_g}}.$$

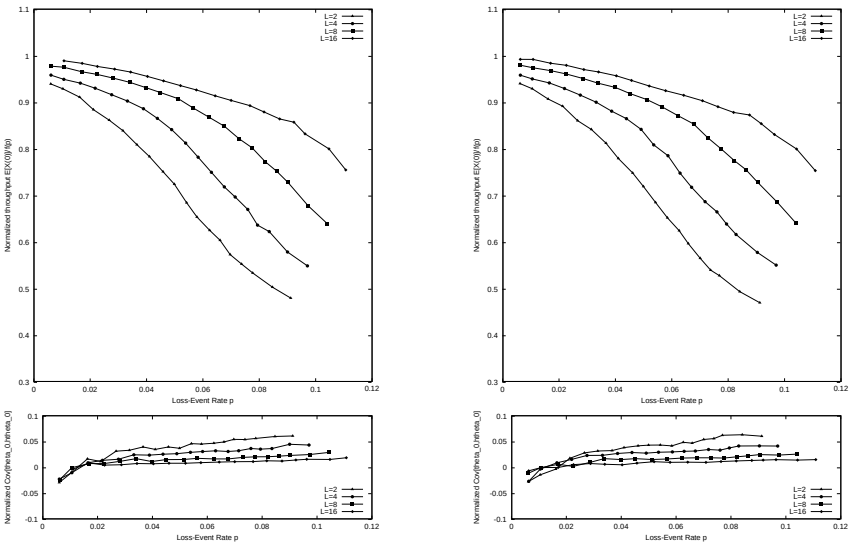


Figure 2.13: (Left) f is PFTK-standard and (Right) PFTK-simplified. The upper graphs show the normalized throughput $\mathbf{E}[X(0)]/f(p)$ attained by TFRC versus the loss-event rate p. The lower graph shows $\mathbf{cov}_N^0[\hat{\theta}_0, \theta_0]p^2$.

Notice that, the larger the deviation of the good and the bad state, the smaller the relative number of transitions from the bad to good state. The last implies that the HMC resides, most of the time, in the bad state, with occasional short excursions to the good state. This dynamics of the HMC gives rise to significant non-conservative behavior (overshoot).

For SQRT function f, the maximum value of $\bar{x}^0$ is

$$\bar{x}^* = \frac{1}{2}\sqrt{2 + \sqrt{\frac{n_g}{n_b}} + \frac{1}{\sqrt{\frac{n_g}{n_b}}}}. \tag{2.15}$$

Note that the right-hand side is increasing with n_g/n_b.

We show in Figure 2.12 numerical values of $\bar{x}^*$ (2.15) versus the ratio n_g/n_b, which we recall is for SQRT function. We also show the results for PFTK-simplified function obtained by numerical computation of the maximum of (2.14). We observe that for sufficiently large values of n_g/n_b we can have a substantial non-conservative behavior.

2.6.2 ns-2 Experiments

We conduct ns-2 simulation experiments to validate the claims made in Section 2.3.3. Unless otherwise indicated, we consider a link shared by TFRC and TCP Sack1 connections. The link implements RED queue management and has a

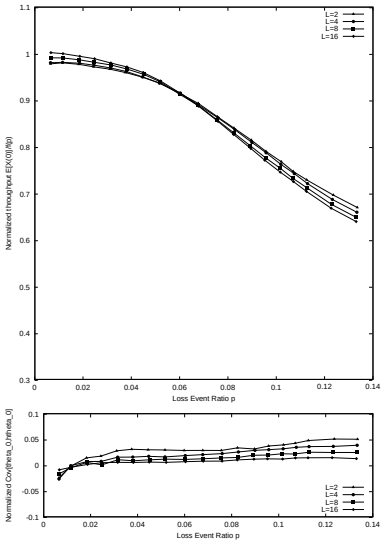


Figure 2.14: Same setting as in Figure 2.13, but function f is SQRT.

rate of 15 Mb/s; we set the buffer length, min_thresh, and max_thresh to 2.5, 0.25 and 1.25 times the bandwidth delay product, respectively. The round-trip time is about 50 ms. We mimic this setting from Bansal *et al.* [11].

Validation of Claim 1

In Figure 2.13, left, we show the normalized throughput for the PFTK-standard function. We verify, the larger the loss-event rate is, the more conservative the control is. We also note that the larger the averaging window L of the loss-event interval estimator is, the less conservative the control is. Next, for PFTK-simplified (Figure 2.13, right) we observe that the results are very close to those with PFTK-standard. We verify in Figure 2.6.1, the conservativeness with SQRT formula is less pronounced and less dependent on L. In all the cases, the covariance of the loss-event estimator and the next sample of the loss-event interval is small.

Validation of Claim 2

We consider a source that sends packets at regular time intervals (20 ms), but controls packet lengths. The source has a connection established through a loss module that drops packets with a fixed probability; this allows us to tune the packet drop rate. For such a source, we have that the covariance of the send rate and the interval between two loss-events is equal to zero. Thus, by Claim 2 we expect our source to be conservative for $f(1/x)$ concave with x; conversely, non-conservative for $f(1/x)$ convex with x.

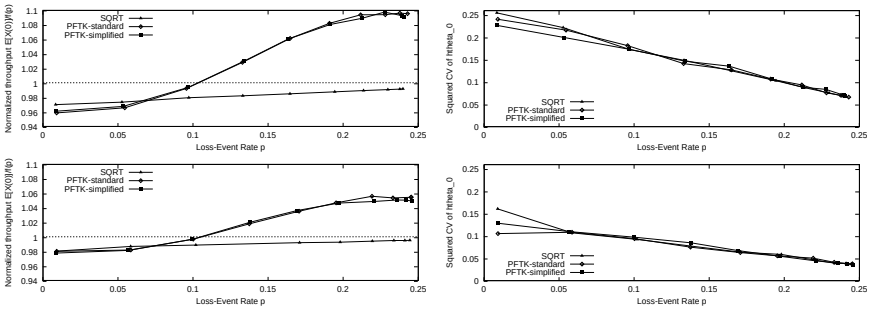


Figure 2.15: (Top-Left) normalized throughput $E[X(0)]/f(p)$ versus the loss-event rate of a source with a constant packet send rate, but controlled packet lengths. The connection goes through a loss module, with a fixed packet drop probability–Bernoulli dropper. $L = 4$. (Top-Right) squared coefficient of the variation of $\hat{\theta}_0$. (Bottom) The same as at the top, but $L = 8$

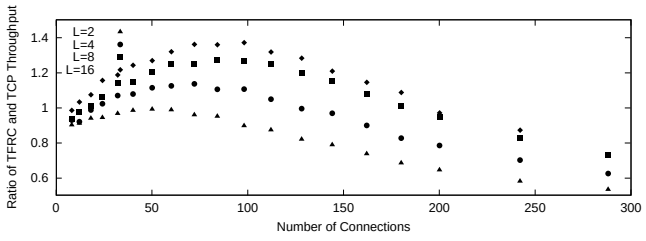


Figure 2.16: The graph shows ratio of TFRC and TCP Sack1 throughputs versus the number of connections.

We show the results for two values of the averaging-window $L = 4$ and 8, see Figure 2.6.2. We verify that the control with SQRT is always conservative. For PFTK-standard and PFTK-simplified the same holds for a low loss rate, however, for a high loss rate the functions are convex, and thus the control exhibits a non-conservative behavior in this region. Observe from Figure 2.6.2, as the loss-event rate increases, the coefficient of the variation of $\hat{\theta}_0$ becomes smaller. Smaller variability of the loss-event estimator makes the control either less conservative or less non-conservative, depending on which behavior is in action. On the other hand, a larger variability of $\hat{\theta}_0$ exaggerates either conservative or non-conservative behavior.

Putting Things Together

Claim 3 tells us that our adaptive source sees a larger loss-event rate than TCP, which drives it in the TCP-friendly direction, on top and above the factors mentioned earlier. Assuming (as is most common) that the conditions for con-

Receiver	c (Mb/s)	Hops	RTT (ms)	SACK	OS
INRIA	100	13	30	no	FreeBSD 4.1
UMASS	100	15	97	yes	Linux 2.4.19
KTH	10	20	46	yes	Linux 2.4.19
UMELB	10	24	350	yes	Linux 2.2.14-tsc

Table 2.1: Some basic facts about our receiver hosts and connections to them from EPFL. The round-trip time estimates are rounded versions of the original values obtained by `traceroute` [62]. c is the access rate of a host.

servativeness in Section 2.3 apply, we would have $\bar{x} \leq f(p) \leq f(p')$, (the latter is because f is decreasing). This makes our adaptive source TCP-friendly under the assumption that TCP does satisfy its equation. Unfortunately, this is only approximately true. Figure 2.5 shows an experiment where TCP throughput is below the formula PFTK-standard for light load and above for high loads. Figure 2.16 shows that, as a result, TFRC flows have larger throughput for medium load than TCP. This is despite TFRC being conservative (Figure 2.13) and experiencing larger loss than TCP (Figure 2.6), as predicted by our theory. This illustrates the importance of separating the factors that effect TCP-friendliness.

2.6.3 Internet and Lab Experiments

In this section, we:

- validate the claims made in Section 2.3.3 through Internet and lab experiments;
- check if TFRC is TCP-friendly, or not, and check the effect and significance of the individual factors (identified at the beginning of this chapter) on TCP-friendliness.

Toward these goals, we run a series of designed experiments over the Internet and in a lab environment. The experiments consist of (1) LAN to LAN Internet measurements; (2) LAN to cable modem Internet measurements; (3) laboratory experiments. Some details of our experiments are deferred to Appendix 2.6.3.

2.6.4 Internet Experiments: LAN to LAN

The Internet measurements were taken from a sender host at EPFL to receivers located at four other locations: (INRIA) Sophia-Antipolis, France, (UMASS) University of Massachusetts, Amherst, US, (KTH) Stockholm, Sweden, and University of Melbourne, Melbourne, Australia (UMELB). Both INRIA and UMASS have access rates equal to 100 Mb/s. The access rates of KTH and UMELB are 10 Mb/s. Our sender at EPFL also has an access rate of 100 Mb/s. We display some further details in Table 2.1.

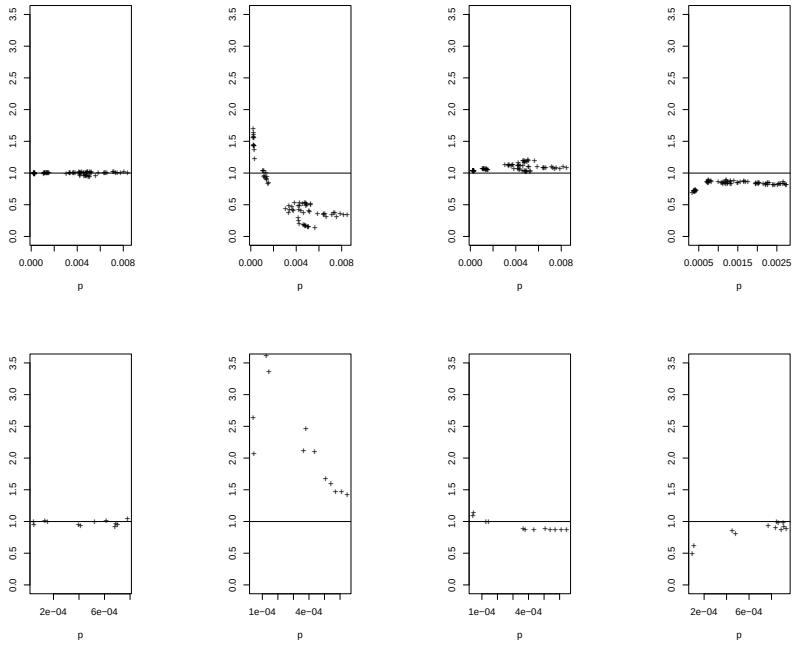


Figure 2.17: (Top) INRIA and (Bottom) KTH. Breakdown the TCP-friendliness condition into: (1st column) the ratio of $\bar{x}$ and $f(p, r)$; (2nd column) the ratio of p' and p; (3rd column) the ratio of r' and r; (4th column) the ratio of $\bar{x}'$ and $f(p', r')$.

We considered six configurations: For the experiments $1, 2, 3, 4, 5, 6$, we fixed the number of TFRC and TCP connections to $1, 2, 4, 6, 8, 10$, respectively. One goal for this design was to evaluate performance over a wider range of loss-event rates than one would typically observe by running a single connection at a time. The function f is PFTK-standard, the averaging-window of the loss-event interval is set as $L = 8$.

We first discuss some statistics of the loss process as observed by TCP and TFRC in our experiments. See Figure 2.33. We observe: The loss-event rates are small; in all the cases, they are less than 1%. We next consider the covariance of θ_0 and $\hat{\theta}_0$ for TFRC, and the coefficient of the variation of θ_0 for both TFRC and TCP; see Figure 2.33. We observe: (1) $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]$ is in most cases almost zero; it slightly inclines to negative values; for UMELB, it is noticably less than zero[4], (2) $\mathbf{cv}_N^0[\theta_0]$ is mostly slightly less than one, for all the hosts; except for UMELB,

[4]By inspection of the loss-event intervals, we observed that negative covariance is a result of the loss-events that alternate between a large loss-event interval and a series of small-valued loss-event intervals. In other words, the loss-events arrive in batches.

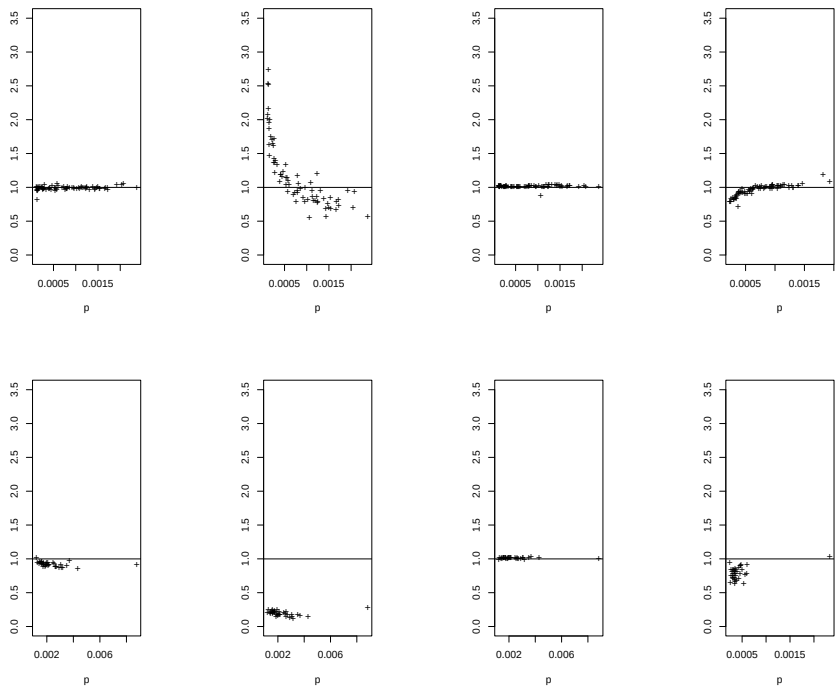


Figure 2.18: (Top) UMASS and (Bottom) UMELB. Same as in Figure 2.17.

in which case it is noticably larger. In the view that the covariance $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]$ is *not* significantly positive, a hypothesis of Claim 1 is verified. From Claim 1 we anticipate observing that TFRC is conservative, in particular, for UMELB we expect it to be more conservative. Another metric of interest, in view of our Claim 2, is the covariance of the send rate X_n and the inter loss-event time S_n; see Figure A.4 in Appendix A. We observe: $\mathbf{cov}_N^0[X_0, S_0]$ is typically small; in most cases, it increases with the loss-event rate. The observation verifies a hypothesis of our Claim 2. Therefore, we expect TFRC to be conservative. Note that this conclusion is in line with the one made earlier from our Claim 1.

We show the empirical estimates of the four factors of the TCP-friendliness breakdown, as indicated in the caption of Figure 2.17, see, also, Figure 2.18.

Validation of Claim 1 and Claim 2. See the leftmost plots in Figure 2.17 and 2.18. We observe that the deviation in either a conservative or non-conservative direction is mostly negligible. Recall that the loss-event rates are small, and hence by Claim 1 we do expect the conservativeness to be moderate. Indeed, Claim 1 says that a stronger conservativeness is to be expected for a larger convexity of the function $x \to 1/f(1/x)$. Now, for small loss-event rates (as it is exactly the cases in our experiments), the function $1/f(1/x)$ is

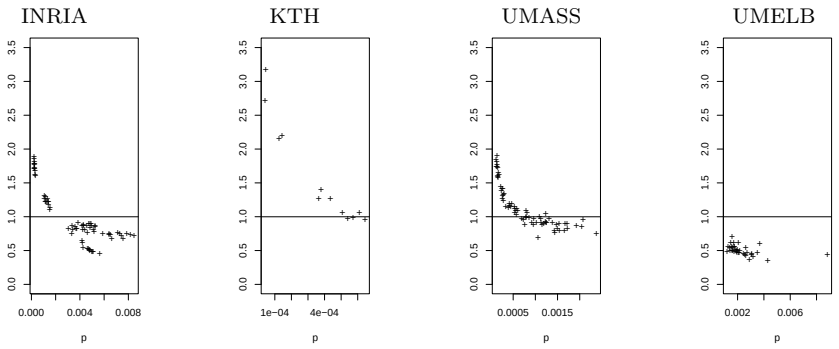


Figure 2.19: Internet experiments: check is TFRC TCP-friendly. The graphs show the ratio of $\bar{x}$ and $\bar{x}'$, respectively, the throughputs of TFRC and TCP, versus p. The values not larger than one indicate TCP-friendliness, else, non-TCP-friendliness.

effectively $1/\sqrt{x}$, which is mildly steep and convex for large x, that is, for small loss-event rate.

Breakdown of TCP-Friendliness into Sub-Conditions. Check if TFRC is TCP-friendly in Figure 2.19. For INRIA, KTH, and UMASS, the answer is negative. More specifically, for small loss-event rates (a few competing TCP and TFRC connections), TFRC can be significantly non-TCP-friendly. We now proceed by checking the factors of the TCP-friendliness breakdown to reveal a cause of the observed non-TCP-friendliness. Check the factors in Figure 2.17 and 2.18. We observe that the loss-event rate p' as observed by TCP can be significantly larger than p, the loss-event rate observed by TFRC. In particular, this seems recurrent when a few connections compete for a bottleneck. The observed discrepancy of the loss-event rates drives TFRC to a non-TCP-friendly direction. Another cause that drives TFRC into the same direction is that TCP attains smaller throughput than predicted by the formula; see the rightmost plots in Figure 2.17 and 2.18. We stress that in our Internet experiments the deviation of the loss-event rates is the *dominant factor* that drives TFRC to non-TCP-friendliness.

2.6.5 Internet Experiments: LAN to Cable Modem

We also designed experiments with a receiver connected to the Internet via a cable modem. The receiver was located at EPFL. We set up two sender hosts at EPFL and UMASS, both connected to the Internet via 100 Mb/s. Note that unlike our LAN to LAN experiments, the experiments with the modem are unpaired, we ran exactly one, either a TCP or TFRC connection in an experiment. The plots of our empirical results are deferred to Appendix 2.6.3, for the benefit of the space in this chapter.

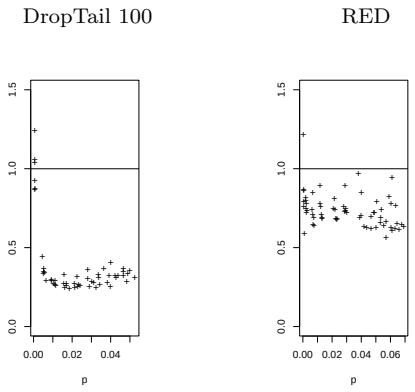


Figure 2.20: Lab experiments: check is TFRC TCP-friendly. The graphs show the ratio of TFRC and TCP throughputs versus p.

We first go through the statistics of the loss and delay process. From Figure A.6, we observe: The loss-event rate is small, not larger than 0.4%; for UMASS, the average-round trip time is large, about 1.5 s. From Figure A.6, we observe: for EPFL, $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]$ is negative or slightly positive; for UMASS, it is always negative, and this is non negligible.

Validation of Claim 1. The observed loss-event statistics confirms a hypothesis made in Claim 1, and hence, we expect the control to be conservative. We found this to be true as seen next. From Figure A.7, we note: TFRC is conservative, in all the cases.

Breakdown of TCP-Friendliness into Sub-Conditions. We recall that our cable-modem experiments are un-paired, and thus we cannot directly compare TCP and TFRC throughputs. In Figure A.9 we compare TCP and TFRC by plotting their throughputs, the loss-event rates, and the average round-trip times against the time-of-the-day. The observations that we make here should be taken with some caution. From Figure A.8, we observe: TCP overshoots the value predicted by its formula, but slightly. From Figure A.9, we observe: The loss-event rate seen by TFRC is larger than TCP's, in most cases; TFRC sees smaller average round-trip time; TFRC is TCP-friendly, in most of the cases.

2.6.6 Lab Experiments

We designed a series of lab experiments that to some extent mimic our Internet setup, see Appendix 2.6.3 for a detailed description. Our motivation to perform lab experiments was to compare TCP and TFRC over a larger range of the loss-event rates, than it would be practically possible in the Internet. Note that, in contrast to our other experiments, the lab experiments are for TFRC with the basic control, defined in Section 2.2.1.

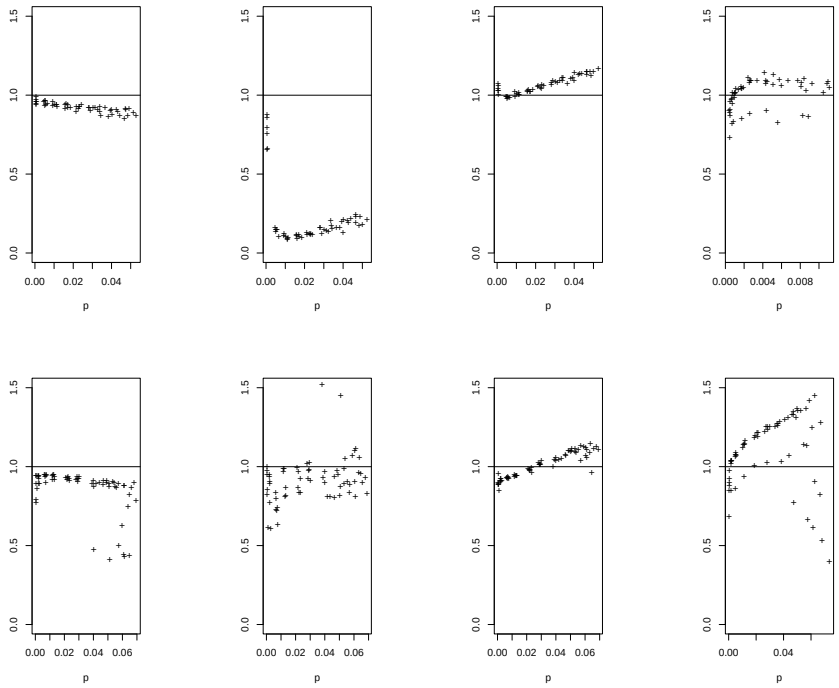


Figure 2.21: Lab experiments: (Top) DropTail and (Bottom) RED. Breakdown the TCP-friendliness condition into: (1st column) the ratio of $\bar{x}$ and $f(p, r)$; (2nd column) the ratio of p' and p; (3rd column) the ratio of r' and r; (4th column) the ratio of $\bar{x}'$ and $f(p', r')$.

We first check statistics of the loss process. From Figure 2.34, we observe: the loss-event rates roughly cover the interval $0 - 9\%$ for RED, and $0 - 6\%$ for DropTail[5].

Validation of Claim 1. See Figure 2.21, the leftmost graphs. The empirical results indicate: The larger the loss-event rate, the stronger the conservativeness. Note that the covariance condition in Claim 1 seems to hold, as indicated in Figure 2.34, the leftmost graph.

Breakdown TCP-Friendliness into Sub-Conditions. Check if TFRC is TCP-friendly in Figure 2.20. The answer is positive. However, we observe that it is mostly overly TCP-friendly. This is more emphasized for DropTail, than for RED. We now separately examine the factors to reveal a cause of the observed overly TCP-friendliness. First, we consider whether TFRC is conservativeness, see the leftmost plots in Figure 2.21. We observe that the answer is positive;

[5]A popular name for a finite buffer FIFO queue.

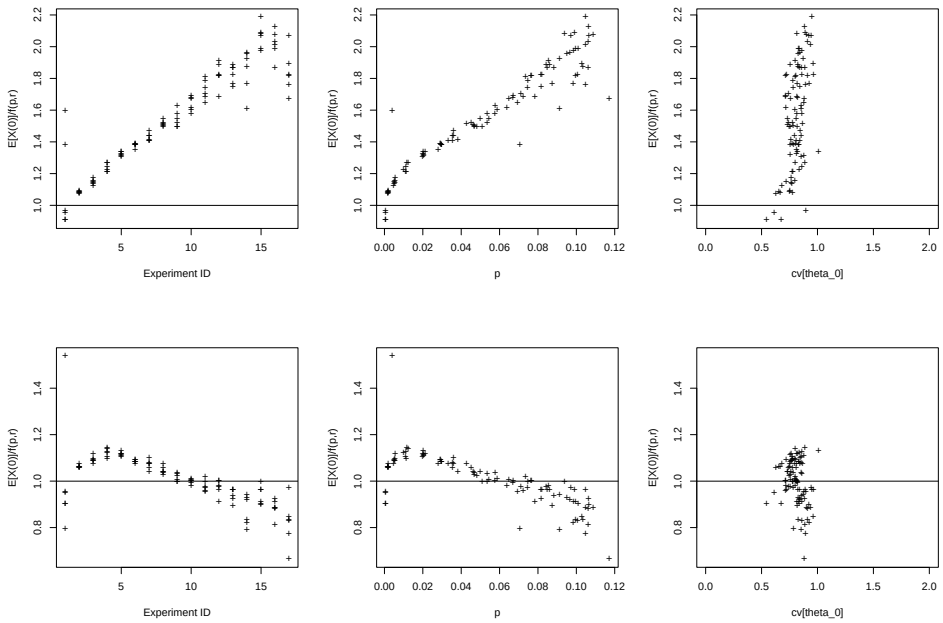


Figure 2.22: Lab experiments for RED: Does TCP conform to its formula? (Top) PFTK-full, (Bottom) SQRT.

the larger the loss-event rate, the more it is conservative. Second, we check how the loss-event rates compare, see the second column of the plots in Figure 2.21. We observe that the loss-event rate seen by TFRC is mostly larger than as seen by TCP; for DropTail, this deviation is excessive. Third, we look at the average round-trip times, see the plots in the thrid colum of Figure 2.21. We observe that in some cases the average round-trip time seen by TCP is larger than that of TFRC, sometimes it is opposite. Last, we check how TCP conforms to PFTK-standard, see the rightmost plots in Figure 2.21. We observe that for small loss-event rates (a few competing connections), TCP seems to attain smaller throughput than given by the formula. In contrast, for a non-small loss-event rate, TCP attains a larger throughput than given by the formula. The last deviation is large for PFTK-standard; the larger the loss-event rate, the larger the deviation. In summary, for DropTail, the dominant factor of the excessive TCP-friendliness is a grossly larger loss-event rate of TFRC than of TCP; for RED, the dominant factor seems to be the undershoot of TCP of PFTK-standard formula.[6] (See also Figure 2.22.)

[6]The reasons of the observed discrepancy may be at least twofold. First, PFTK-standard is used with the retransmit timeout parameter set to $\max[4r, 0.2]$; the lower bound 0.2 s is fixed to match the lower bound of Linux TCP, the value $4r$ is according to TFRC standard [45]. Sec-

2.7 Discussion

In this section we discuss some refinements of our analysis.

Variable Round-Trip Times. In our analysis of the conservativeness we have assumed that the round-trip times are fixed to a constant. In Section 2.10, we derive a throughput expression that accounts for the variability of the round-trip times and dependency with the loss process. The throughput expression in Section 2.10 accounts, also, for the bias due to estimation of the loss-event rate at the receiver. The throughput expression reveals us exact expressions of the bias terms. We showed some empirical estimates of the bias terms in Section 2.10.

Impact of Rate Limitation. TCP protocol imposes a limit on TCP congestion window size through the receiver advertised window. On the other hand, some equation-based rate control protocols use TCP throughput formulae that do not account for the receiver-window limitation; an example is TFRC. We observed in our pilot experiments that in many cases TCP receiver window is effective in limiting TCP throughput. Apparently, an empirical verification of TCP-friendliness of an equation-based rate control by comparing with a receiver-window limited TCP will be *unfair*. For this reason, we deliberately increased the receiver buffer sizes in our experiments so that the TCP receiver-window was ineffective; see Appendix 2.6.3. We next discuss that some hypotheses of our analysis results of conservativeness remain to hold for some TCP loss-throughput formulae that account for TCP receiver-window limitation. Consider an equation-based rate control that uses

$$p \to \min\{f(p), x_0\},$$

as the loss-throughput formula, where $p \to f(p)$ is a non-increasing function, for instance, one of those displayed in Section 2.2.3, x_0 is a positive-valued number. The above displayed correction has been proposed as a correction of PFTK-standard formula considered in this chapter; see Equation (33), Padhye *et al.* [93]. In the context of [93], x_0 is defined as the ratio of TCP receiver-window and the event-average of the round-trip time. In our context, think of x_0 as of a fixed rate limitation. We now note that some hypotheses of our conservativeness analysis remain to hold by modifying a loss-throughput formula by the map $x \to \min\{x, x_0\}$. First, note that if in Theorem 1, (F1) holds for $1/f(1/x)$, then it holds also for $1/\min\{f(1/x), x_0\}$. Second, if in Theorem 2, (F2) holds for $f(x)$, then it holds for $\min\{f(x), x_0\}$ as well. However, the same implication does not hold for (F2c).

ond, it is possible that the given TCP behaves almost as an additive-increase/multiplicative-decrease and the effects due to TCP retransmit timeouts are negligible. Note that, for the range of the loss-event rates in the experiments of the present discussion, the values of both PFTK functions are significantly smaller than the values of SQRT; recall Figure 2.3.

2.8 Conclusions

- Our study should help designers of TCP-friendly equation-based rate controls better understand the trade-offs that have to be taken.

- It is important to separately verify different factors, which we recall from the beginning of this chapter:

  (C) conservativeness;

  (P) TCP's loss-event rate versus this protocol's loss-event rate;

  (R) TCP's average round-trip time versus this protocol's average round-trip time;

  (T) TCP's obedience to its own formula.

- We should be aware of the strong dependency on the nature of function f; SQRT behaves differently than PFTK. If PFTK is used, and under the conditions on the loss process in Claim 1, very pronounced conservativeness should be expected for heavy loss. Under the conditions of Claim 2, the opposite may hold. In any case, the more variable the estimator is, the more pronounced the effect is.

- Our experiments indicate that in today's Internet, even with additional background traffic to increase the load of an Internet channel, the loss process seen by TFRC is such that the estimator of the expected loss-event interval is almost independent, or negatively correlated to the subsequent sample of the loss-event intervals. Under these properties and the fact that all the function $f(\cdot)$ considered in our work are effectively such that $1/f(1/x)$ is convex with x, our Claim 1 tells us to expect conservativeness.

- We showed experimental evidence that in some cases, for TFRC, the factors (P) and (T) can be predominant in determining the final outcome—whether or not the control is TCP-friendly. If a few TCP and TFRC connections compete for a bottleneck, TCP's loss-event rate may be larger. Under the same circumstances, TCP achieves smaller throughput than predicted by PFTK formula. Given that in today's Internet, typical loss-event rates are small, the effects of conservativeness are moderately pronounced. The overall effect is that in the present situation TFRC can be grossly *non-TCP-friendly*.

- Conversely, to the last described limit case, when a substantial number of TFRC and TCP connections compete for a well-provisioned resource, then we observed tendency of TCP to achieve larger throughput than predicted by PFTK formula. At the same time, we observed tendency of TFRC's loss-event rate to be larger than that of a TCP sender. Both of these factors lead to TCP-friendliness. In the on-going scenario with a high-multiplex of connections, we confirm our Claim 3.

- Our lab experiments with RED as a queue discipline show us that for non-small loss-event rates, PFTK-standard, as suggested by the standard [54], may overly under-predict TCP throughput. In the same situation, TCP better conforms to SQRT. We note that TCP throughput may in some cases be grossly off a prediction by PFTK; revealing the true reason of this discrepancy is beyond the scope of our study.

- Our engineering guideline is to check the factors individually. Failing to do so may hide a true cause of an observed non-TCP-friendliness or excessive TCP-friendliness. This might lead a protocol designer to change some parameters of her protocol, in order to correct either effect. Understanding why and when the effects occur is essential to avoid undesired corrections.

- We favour conservativeness as a design objective, instead of TCP-friendliness. One argument is an experienced difficulty to verify TCP-friendliness in practice, while conservativeness is amenable to a formal verification. Another argument is that the concept of conservativeness moves us away from TCP-centric design, where the design of a control is restricted to take TCP as a reference control.

2.8.1 Possible Directions of Future Work

- An important problem is to understand better interaction of a few competing TCP and TFRC-like connections. A particular problem is to understand better the observed phenomena that in some situations described above, TCP's loss-event rate can be larger than TFRC's.

- With regard to our experiments (detailed in Appendix), it would be worthwile to carry on the same type of experiments we did, but for a kernel implementation of TFRC, in order to allow a more fair comparison of the two protocols.

2.9 Proofs

2.9.1 Proof of Proposition 1

The starting point is the Palm inversion formula [9]. We have

$$\mathbf{E}[X(0)] = \lambda \mathbf{E}_N^0[\int_0^{T_1} X(s)ds]. \tag{2.16}$$

We can think of (2.16) as the ratio of the expected number of packets sent in-between two successive loss-events and the expected inter loss-event time. However, it is important to remember the expected values are with respect to the Palm probability, that is, as seen at the loss-event instants.

For the basic control this gives

$$\mathbf{E}[X(0)] = \frac{\mathbf{E}_N^0[X_0 S_0]}{\mathbf{E}_N^0[S_0]}. \tag{2.17}$$

From (2.3), $\theta_n = X_n S_n$ and $X_n = f(1/\hat{\theta}_n)$. Hence, $S_n = \frac{\theta_n}{f(1/\hat{\theta}_n)}$. Combining the last three identities into (2.17) we obtain (2.8).

2.9.2 Proof of Proposition 2

Define $Y_n := \theta_n / S_n$, all $n \in \mathbb{Z}$. The physical meaning of Y_n is the average send rate over the interval $[T_n, T_{n+1})$, for $n \in \mathbb{Z}$. Again from Palm inversion formula

$$\mathbf{E}[X(0)] = \frac{\mathbf{E}_N^0[\theta_0]}{\mathbf{E}_N^0[\frac{\theta_0}{Y_0}]}$$

Now, by the definition of comprehensive control, $X(t) \geq X_n$, for all $t \in [T_n, T_{n+1})$, and hence, $Y_n \geq X_n$. Replacing Y_0 in the above display with its lower bound X_0, and recalling that by definition $X_n = f(1/\hat{\theta}_n)$, $n \in \mathbb{Z}$, recovers the asserted lower bound.

2.9.3 Proof of Proposition 3

Case 1 ($\hat{\theta}_{n+1} \leq \hat{\theta}_n$). In this case $\theta_n = X_n S_n$, and hence, $S_n = \frac{\theta_n}{f(1/\hat{\theta}_n)}$.

Case 2 ($\hat{\theta}_{n+1} > \hat{\theta}_n$) In this case, for $T_n \leq t \leq T_n + U_n$, $\theta(t) = t f(1/\hat{\theta}_n)$. Else, for $T_n + U_n < t < T_n + S_n$,

$$\theta(t) = \theta(T_n + U_n) + \int_{T_n + U_n}^{t} X(s)ds.$$

By the definition of comprehensive control (2.4), we obtain the ordinary differential equation, $T_n + U_n \leq t < T_n + S_n$, $\theta(T_n + U_n) = (\hat{\theta}_n - W_n)/w_1$,

$$\frac{d\theta(t)}{dt} = f\left(\frac{1}{w_1 \theta(t) + W_n}\right), \tag{2.18}$$

where $W_n = \sum_{l=1}^{L-1} w_{l+1}\theta_{n-l}$.

Now, we solve $\theta(T_n + S_n-) = \theta_n$ for S_n. To that end, we solve (2.18) for PFTK-simplified formula (2.7). Plugging PFTK-simplified function f into (2.18), and a simple re-arragement, we obtain

$$c_1 r \int_{T_n+U_n}^{T_n+S_n} \frac{d\theta(t)}{\sqrt{w_1\theta(t)+W_n}} + c_2 q \int_{T_n+U_n}^{T_n+S_n} \frac{d\theta(t)}{\sqrt{(w_1\theta(t)+W_n)^3}} + \\ +32c_2 q \int_{T_n+U_n}^{T_n+S_n} \frac{d\theta(t)}{\sqrt{(w_1\theta(t)+W_n)^7}} \quad = \quad S_n - U_n.$$

Use the substitution $y = w_1\theta(t) + W_n$. Note that $d\theta(t) = dy/w_1$ and that with this substitution the boundaries of the integrals, $T_n + U_n$ and $T_n + S_n$, respectively, are equal to $\hat{\theta}_n$ and $\hat{\theta}_{n+1}$. We re-write the last display as

$$\frac{c_1 r}{w_1} \int_{\hat{\theta}_n}^{\hat{\theta}_{n+1}} \frac{dy}{\sqrt{y}} + \frac{c_2 q}{w_1} \int_{\hat{\theta}_n}^{\hat{\theta}_{n+1}} \frac{dy}{\sqrt{y^3}} + \frac{32c_2 q}{w_1} \int_{\hat{\theta}_n}^{\hat{\theta}_{n+1}} \frac{dy}{\sqrt{y^7}} = S_n - U_n.$$

Solving the elementary integrals, we obtain

$$S_n = U_n + \frac{2c_1 r}{w_1}(\hat{\theta}_{n+1}^{1/2} - \hat{\theta}_n^{1/2}) - 2\frac{c_2 q}{w_1}(\hat{\theta}_{n+1}^{-1/2} - \hat{\theta}_n^{-1/2}) - \frac{64}{5}\frac{c_2 q}{w_1}(\hat{\theta}_{n+1}^{-5/2} - \hat{\theta}_n^{-5/2})$$

For convinience of notation, let $B_n := S_n - U_n$. Recall,

$$U_n = \frac{\hat{\theta}_n - W_n}{w_1 f(\frac{1}{\hat{\theta}_n})}.$$

Finally, we have

$$\begin{aligned} S_n &= \frac{\theta_n}{f(\frac{1}{\hat{\theta}_n})}\mathbf{1}_{\hat{\theta}_{n+1}\le\hat{\theta}_n} + \left(B_n + \frac{\hat{\theta}_n - W_n}{w_1 f(\frac{1}{\hat{\theta}_n})}\right)\mathbf{1}_{\hat{\theta}_{n+1}>\hat{\theta}_n} \\ &= \frac{\theta_n}{f(\frac{1}{\hat{\theta}_n})} + \left(B_n - \frac{\hat{\theta}_{n+1} - \hat{\theta}_n}{w_1 f(\frac{1}{\hat{\theta}_n})}\right)\mathbf{1}_{\hat{\theta}_{n+1}>\hat{\theta}_n} \\ &= \frac{\theta_n}{f(\frac{1}{\hat{\theta}_n})} - V_n\mathbf{1}_{\hat{\theta}_{n+1}>\hat{\theta}_n}. \end{aligned}$$

The last identity follows directly by definitions of V_n and B_n. It remains only to use Palm inversion formula $\mathbf{E}[X(0)] = \mathbf{E}_N^0[\theta_0]/\mathbf{E}_N^0[S_0]$, and plug-in the expression of S_n displayed above to show (2.9).

2.9.4 Proof of Theorem 1

Define $g(x) := \frac{1}{f(\frac{1}{x})}$. Also call $m = \frac{1}{p}$, thus $\mathbf{E}_N^0[\theta_0] = \mathbf{E}_N^0[\hat{\theta}_0] = m$. From Equation (2.8), conservativeness is equivalent to

$$\mathbf{E}_N^0[\theta_0 g(\hat{\theta}_0)] \ge m g(m). \tag{2.19}$$

Function g is convex, thus it is above its tangents:

$$g(x) \geq (x - m)g'(m) + g(m).$$

Apply the above to $x = \hat{\theta}_0$, multiply by θ_0 and take the expectation. After some calculus, this shows Equation (2.10).

Now f is decreasing. Since $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0] \leq 0$, it follows from Equation (2.10) that the control is conservative.

2.9.5 Proof of Proposition 4

Use the same notation as in the proof of Theorem 1. By Equation (2.8) the ratio of the throughput to $f(p)$ is equal to

$$\rho := \frac{mg(m)}{\mathbf{E}_N^0\left[\theta_0 g(\hat{\theta}_0)\right]}. \tag{2.20}$$

Now we have

$$g^{**}(x) \leq g(x) \leq rg^{**}(x).$$

and thus $\rho \leq r$.

2.9.6 Proof of Theorem 2

Use the same notation as in the proof of Theorem 1.
Part 1. By (C2)

$$\mathbf{E}_N^0[\theta_0 g(\hat{\theta}_0)] \geq \frac{m}{\mathbf{E}_N^0\left[\frac{1}{g(\hat{\theta}_0)}\right]}, \tag{2.21}$$

now (F2) means that $\frac{1}{g}$ is concave, thus by Jensen's inequality:

$$\mathbf{E}_N^0\left[\frac{1}{g(\hat{\theta}_0)}\right] \leq \frac{1}{g(\mathbf{E}_N^0[\hat{\theta}_0])}, \tag{2.22}$$

which combined with the previous equation shows that the control is conservative.
Part 2. By (C2c) and (F2c) we have the reverse inequalities in Equation (2.21) and Equation (2.22), but the inequality is strict in Equation (2.22) because convexity is strict and $\hat{\theta}_n$ is not a degenerate random variable.

2.9.7 Derivation of Equation (2.12)

We start from Equation (2.1). By Neveu's exchange formula ([9], Sec. 3.3.4) and a simple conditioning

$$p = \frac{1}{\mathbf{E}_N^0[\theta_0]}$$

$$= \frac{\mathbf{E}^0_{N'}[\sum_{n\in\mathbb{Z}} \mathbf{1}_{[0,S'_0)}(T_n)]}{\mathbf{E}^0_{N'}[\sum_{n\in\mathbb{Z}} \theta_n \mathbf{1}_{[0,S'_0)}(T_n)]}$$
$$= \frac{\sum_{i\in E} \mathbf{E}^0_{N'}[\sum_{n\in\mathbb{Z}} \mathbf{1}_{[0,S'_0)}(T_n)|Z(0)=i]}{\sum_{i\in E} \mathbf{E}^0_{N'}[\sum_{n\in\mathbb{Z}} \theta_n \mathbf{1}_{[0,S'_0)}(T_n)|Z(0)=i]}. \quad (2.23)$$

We show that the above is equivalent to Equation (2.12).

By applying Palm inversion formula to $X(0)\mathbf{1}_{Z(0)=i}$, we obtain

$$\bar{x}_i = \mathbf{E}[X(0)|Z(0)=i] = \frac{\mathbf{E}^0_{N'}[\int_0^{S'_0} X(s)ds|Z(0)=i]}{\mathbf{E}^0_{N'}[S'_0|Z(0)=i]},$$

where we also use Palm inversion on $\mathbf{1}_{Z(0)=i}$ to obtain

$$\pi_i = \mathbf{P}[Z(0)=i] = \frac{\mathbf{E}^0_{N'}[S'_0|Z(0)=i]}{\mathbf{E}^0_{N'}[S'_0]} \mathbf{P}^0_{N'}[Z(0)=i].$$

From Neveu's exchange formula applied to $\theta_0 \mathbf{1}_{Z(0)=i}$, we have

$$\frac{1}{p_i} = \mathbf{E}^0_N[\theta_0|Z(0)=i] = \frac{\mathbf{E}^0_{N'}[\sum_{n\in\mathbb{Z}} \theta_n \mathbf{1}_{[0,S'_0)}(T_n)|Z(0)=i]}{\mathbf{E}^0_{N'}[\sum_{n\in\mathbb{Z}} \mathbf{1}_{[0,S'_0)}(T_n)|Z(0)=i]},$$

where we use the identity obtained by Neveu's exchange formula applied to $\mathbf{1}_{Z(0)=i}$,

$$\mathbf{P}^0_N[Z(0)=i] = \frac{\mathbf{E}^0_{N'}[\sum_{n\in\mathbb{Z}} \mathbf{1}_{[0,S'_0)}(T_n)|Z(0)=i]}{\mathbf{E}^0_{N'}[\sum_{n\in\mathbb{Z}} \mathbf{1}_{[0,S'_0)}(T_n)]} \mathbf{P}^0_{N'}[Z(0)=i].$$

Finally, by plugging the above expressions for $\bar{x}_i$, π_i, and p_i into Equation (2.12) we recover Equation (2.23).

2.9.8 Comparison of Conditions in Theorem 1 and Theorem 2

We use the same notation as in the proof of Theorem 1. Notice that by the assumption that f is non-increasing (Section 2.2), g is non-increasing as well. For technical convenience, suppose g is strictly decreasing at m, that is $g'(m) < 0$.

Proposition 5 *Assume (F2c), (C2c), and (V) hold, i.e., the second part of Theorem 2 applies. Then, in Theorem 1, if (F2) is true, it must be that (C1) does not hold.*

Proof 1 *Note the equivalence*

$$\mathbf{cov}^0_N[X_0, S_0] > 0 \Leftrightarrow \mathbf{E}^0_N[\theta_0 g(\hat{\theta}_0)] < \frac{m}{\mathbf{E}^0_N[f(1/\hat{\theta}_0)]}.$$

Under (F2), by the same argument as in Theorem 1,

$$\mathbf{E}_N^0[\theta_0 g(\hat{\theta}_0)] \geq g'(m)\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0] + mg(m).$$

Suppose (C2c) and (F2) are true, then from the last two inequalities, we conclude that the following is implied:

$$\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0] > \frac{m}{g'(m)} \left(\frac{1}{\mathbf{E}_N^0[f(1/\hat{\theta}_0)]} - \frac{1}{f(1/m)} \right).$$

Finally, if $f(1/x)$ is strictly convex with x, that is (F2c) holds, and (V) holds, then the right-hand side in the above inequality is strictly positive, and thus (C1) does not hold.

2.9.9 An Intermediate between Theorem 1 and Theorem 2

The following theorem is intermediate between Theorem 1 and Theorem 2.

Theorem 3 *If (F1) and*

(C3) $\mathbf{cov}_N^0[X_0 S_0, \frac{1}{X_0}] \geq 0$,

the basic control is conservative.

The proof is similar to that of Theorem 2 and is not given here. If the convexity condition (F1) is almost true, then the same as in Proposition 4 holds.

This theorem is intermediate between Theorem 1 and Theorem 2. Indeed $(F2) \Rightarrow (F1)$ and $(C3) \Rightarrow (C2)$. The former is straightforward; a proof of the latter implication uses the convexity of $1/x$. Thus Theorem 3 is with a weaker condition on the function f than Theorem 2, but this comes at the expense of having a stronger condition on the statistics of $\{\theta_n\}_n$. A natural question is whether both Theorem 3 and the first part of Theorem 2 derive from a more general theorem, which would state that under the combination of the less restrictive conditions (F1) and (C2), the control would be conservative. But this is not true; a counter-example is the case presented in the second paragraph of the interpretation of Theorem 2.

2.9.10 When is $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]$ Slightly Positive?

In Claim 1, we posed as assumption that $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]$ is slightly positive or negative. The former property is indeed qualitiative, and we would need to know how to qualify the given $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]$ as either slightly positive or non-slightly positive. This can be done with respect to the effect the normalized covariance has on the throughput bound in (2.10). We make this precise by the following simple analysis. Fix a loss-event rate p. The bound on the throughput in (2.10) can be re-written as

$$\frac{\mathbf{E}[X(0)]}{f(p)} \leq h(\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2),$$

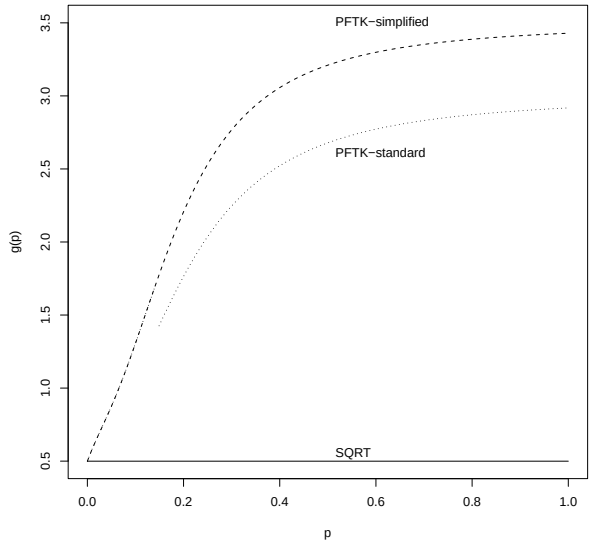


Figure 2.23: Function $p \to g(p)$. For SQRT, $g(p) = 1/2$, for PFTK formulae $g(p) \geq 1/2$.

where

$$h(x) = \frac{1}{1 - g(p)x},$$

and $g(p) := -f'(p)p/f(p)$. Hence, for a given loss-event rate p, the amount of the overshoot is bounded by the function $h(\cdot)$ of the normalized covariance $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$. For SQRT, the function $h(\cdot)$ does not depend on p; in this case, $g(p) = 1/2$. This means that for SQRT the throughput bound (2.10) is *only* in terms of $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$. In contrast, for PFTK formulae, the function $h(\cdot)$ does depend on p. For PFTK-simplified, the function $h(\cdot)$ is with

$$g(p) = \frac{1}{2}\frac{4r + 9qp(3 + 224p^2)}{4r + 9qp(1 + 32p^2)}$$

Lastly, for PFTK-standard, the function $g(p)$ is equal to the last display on the interval $[0, 1/c_2^2)$, else

$$g(p) = \frac{1}{2}\frac{c_1 r + 2q\sqrt{p}(1 + 96p^2)}{c_1 r + q\sqrt{p}(1 + 32p^2)}$$

We now look at the ways when we can easily conclude whether $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$ is slightly positive, or not. The next observation may allow us to easily conclude that a given $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$ is non-slightly positive.

Observation 1 *If we know that* $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$ *is equal to* x*, then we know that the overshoot function* $h(\cdot)$ *is at least* $1 + g(p)x$*.*

If $g(p)x$ is non-small, then x may be qualified to be non-slightly positive. The $g(p)x$ would be qualified non-small depending on the significance of the overshoot $1 + g(p)x$ from an engineering perspective. (For SQRT, $g(p) = 1/2$.) For PFTK-formulae, $g(p) \geq 1/2$. Hence, in the last above statement, we can replace $1 + g(p)x$ with $1 + x/2$, that is we can use $x/2$ as *a rule of thumb*. The rule of thumb is that an observed normalized covariance x is considered non-slightly positive, if the overshoot $1 + x/2$ is not considered negligible.

The last rule of thumb may allow us to easily conclude that a given value of the covariance $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$ is non-slightly positive, but not conversely. In the remainder of the forgeoing discussion, we show a rule to qualify the normalized covariance as slightly-positive or not. Fix a positive $\epsilon \geq 0$. Let c_ϵ be the largest x such that $h(x) \leq 1 + \epsilon$ holds. The term c_ϵ is given by

$$c_\epsilon = \frac{1}{g(p)} \frac{\epsilon}{1+\epsilon}.$$

We qualify $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$ to be slightly positive according to the following definition.

- We say that $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$ is slightly positive, in the sense that the overshoot is at most $1 + \epsilon$, iff $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$ is not larger than c_ϵ.

For SQRT, $c_\epsilon = 2\epsilon/(1 + \epsilon)$, for PFTK formulae, $c_\epsilon \leq 2\epsilon/(1 + \epsilon)$. For PFTK formulae, it is thus correct to say that: If $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$ is larger than $2\epsilon/(1+\epsilon)$, then $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$ is *not* slightly positive. The last is a sufficient condition, and thus a smaller value of the normalized covariance, than $2\epsilon/(1 + \epsilon)$, may be qualified as non-slightly positive. For a small loss-event rate, for PFTK formulae, $c_\epsilon \approx 2\epsilon/(1 + \epsilon)$.

2.10 Other Factors that Effect TCP-Friendliness

2.10.1 Estimation of TCP Retransmit Timeout

The TCP formulae that we consider as examples depend on the loss-event rate p and the average round-trip time r. In addition, the PFTK formulae depend on the average TCP retransmit timeout (q in the formulae of Section 2.2.3). We call $\hat{R}_n$ the estimator of the round-trip time at time T'_n. Some equation-based rate control protocols estimate TCP retransmit timeout at time T'_n as $\beta\hat{R}_n$, for a positive constant β. In particular, for TFRC $\beta = 4$; in [54], this value is claimed to work reasonably well, however, no reference is given to any empirical evidence. Consider q as the *true* value of the expected TCP retransmit timeout and βr as the expected value of its estimator. A question of interest is: How do q and βr compare? Our ultimate interest is to compare $f(p, r, q)$ and

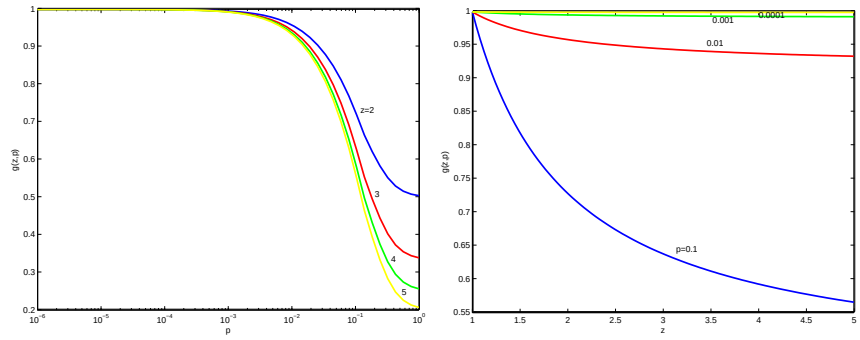


Figure 2.24: $g(z,p)$ is the deviation of the PFTK-standard formula with the TCP retransmit timeout parameter equal to q and with its estimator βr. $z = q/(\beta r)$, r is the average round-trip time, and p is the loss-event rate.

$f(p,r,\beta r)$. (Here we redefine $f : [0,1) \times \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$.) To that end, consider PFTK-standard, we have that the ratio of $f(p,r,\beta r)$ and $f(p,r,q)$ is equal to

$$g(z,p) = \frac{zh(p)+1}{z(h(p)+1)},$$

where

$$h(p) = \frac{c_1}{\beta \min[\frac{1}{\sqrt{p}}, c_2](p+32p^3)},$$

and $z := \beta r/q$. The function $g(z,p)$ is the deviation of the PFTK-standard formula for the given loss-event rate p, and the given bias of the q-estimator z. If $z \geq 1$, then $g(z,p) \leq 1$. In other words, the bias of the q-estimator is safe in the sense that $f(p,r,q) \geq f(p,r,\beta r)$. Conversely, for $z < 1$, $g(z,p) > 1$. In the view of the TCP-friendliness requirement, we would like the q-estimator to satisfy $z \geq 1$. For a fixed $z \geq 1$, the function $p \to g(z,p)$ is decreasing with p, that is, the larger the p, the more conservative $f(p,r,\beta r)$ is with respect to $f(p,r,q)$. (See Figure 2.24.) This deviation is of engineering significance only when the loss-event rate is sufficiently large; in these situations the deviation can be significant.

We consider now the bias parameter z. Consider the definition of TCP retransmit timeout as found in [99] (see Section 21.3 in the book). From this definition, we have

$$q = \left(1 + 4\frac{\mathbf{E}[|R_0 - \hat{R}_0|]}{r}\right) r. \tag{2.24}$$

Note that, for TCP retransmit timeout as defined in [99], $z \geq 1$ is equivalent to saying

$$\frac{\mathbf{E}[|R_0 - \hat{R}_0|]}{r} \leq \frac{\beta - 1}{4}.$$

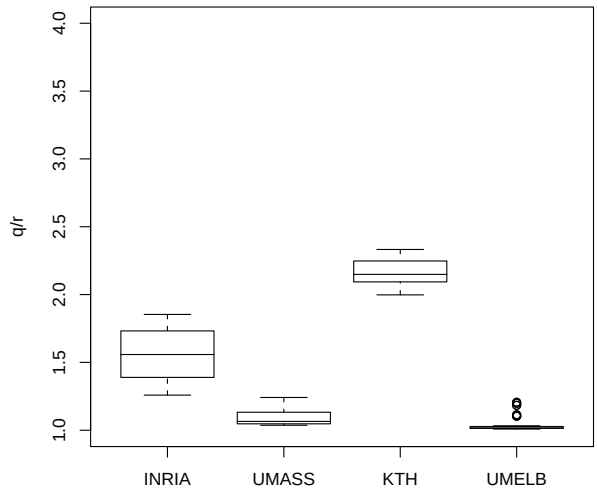


Figure 2.25: Boxplots of the ratio of the empirical estimates of q (2.24) and r inferred from TCP of our Internet experiments.

The left-hand side is the average absolute prediction error of the round-trip time estimator that is normalized by the average round-trip time. In particular, for TFRC, the right-hand side is equal to 3/4. Whether or not the last inequality would hold depend on statistics of the round-trip times. We give an empirical example in Figure 2.25 that we take from our Internet experiments; the results indicate that setting β to 4 is a conservative choice.[7]

2.10.2 Variable Round-Trip Times and Receiver-Estimated Loss-Event Rate

In our analysis of conservativeness we have assumed that the round-trip times are fixed to a constant, and under this assumption, we derived the throughput formulae in Section 2.3.1, and carried out our analysis of the conservativeness in Section 2.3.2. In this section we derive a throughput representation that accounts for variability of the round-trip times. We take one step further to also

[7]The empirical results in Figure 2.25 were obtained by sampling the round-trip time once in a round-trip round in order to obtain a time series of the round-trip times. It is noteworthy that in all the experiments, the sender TCP was of a Linux host. We should bear in mind that Linux TCP samples the round-trip times per-each packet transmitted and that the round-trip time estimator is Linux-specific.

account for the fact that the loss-event rate may be estimated at the receiver; for example, as found in TFRC.

Variable Round-Trip Times. In reality, the round-trip times are variable. TCP throughput formulae depend on a particular *event-average* of the round-trip times; obtained by sampling the round-trip times once in a round-trip round. A sender that implements equation-based rate control, on-line estimates this event-average. The variability of the round-trip time and its stochastic dependency with the loss process may have an effect on the conservativeness of the control.

Receiver-Estimated Loss-Event Rate. With some protocols found in practice (e.g. TFRC) the loss-event rate is estimated at the receiver, and *not* at the sender as assumed in our analysis thus far. Let D_n denote the loss-event interval between the nth and $n+1$th loss-events as observed at the receiver. We assume the loss-event rate is now estimated as before, but replacing θ_n with D_n in the definition of the moving-average estimator $\hat{\theta}_n$ (2.2). We denote with p^* the loss-event rate estimated at the receiver. If the system is stationary, then $p^* = p$; see Figure 2.26 for an illustration and the caption of the figure for a discussion.

In the sequel, we use the notation $Z_n := \theta_n - D_n$, $n \in \mathbb{Z}$. With this new notation, we can write

$$\frac{1}{p} = \frac{1}{p^*} + \mathbf{E}_N^0[Z_0].$$

We use the definition of Z_n, $n \in \mathbb{Z}$, later, in our empirical evaluations.

A Refined Throughput Formula

We redefine $f : (0,1] \times (0,\infty) \to \mathbb{R}_+$. The function f maps the loss-event rate p and the average round-trip time r to the throughput $f(p,r)$.

Consider a sender that updates its control state at some instants T'_n, $n \in \mathbb{Z}$. We assume the standard convention $\cdots < T'_{-1} < T'_0 \leq 0 < T'_1 < \cdots$, and assume that the instants are a realization of a stationary point process on $\mathbb{R}$ that has a finite non-null intensity. With N' we denote the counting measure of the point process. In practice, for instance with TFRC, T'_n is an instant when the nth *report* is received by the sender from the receiver. Following this, we call T'_n the nth report instant. At the report instants, the sender updates the loss-event rate and the round-trip time state variables. Note that the instants of the loss-events as seen by the sender is a *subsequence* of the report instants. Let $\hat{R}_n$ be the value of the round-trip time estimator at T'_n. If $t \in \mathbb{R}$ falls in the interval $[T'_n, T'_{n+1})$, $n \in \mathbb{Z}$, then we define $\hat{R}(t) = \hat{R}_n$.

The next proposition gives us a throughput formula that generalizes the throughput formulae in Section 2.3.1.

Proposition 6 *For the comprehensive control, we have*

$$\mathbf{E}[X(0)] = \frac{\mathbf{E}_N^0[D_0]}{\mathbf{E}_N^0[\frac{D_0}{f(1/\hat{\theta}_0, r)}]} z_1 z_2 z_3 z_4 z_5, \tag{2.25}$$

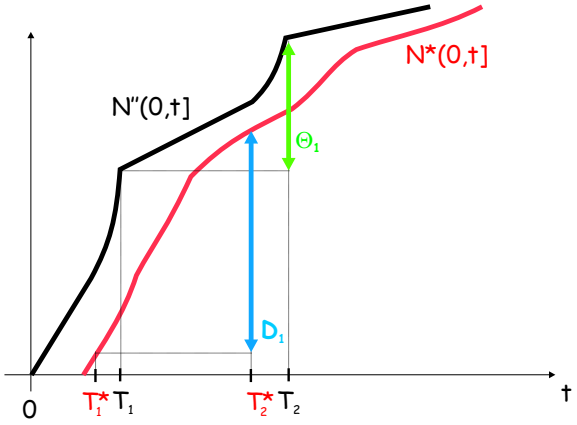


Figure 2.26: A sketch for the receiver-estimated loss-event rate. $N''(s,t]$ is the number of packets sent by the sender in an interval $(s,t]$. $N^*(s,t]$ is the difference between the highest packet sequence number observed in $(s,t]$ and the highest packet sequence number observed before s *by the receiver*. If there were no packet losses, then $N^*(s,t]$ would be the number of packets received by the receiver on $(s,t]$. T_n is an instance when the sender is notified about the nth loss-event, which was detected by the receiver at the instant T_n^*. We have $\theta_n = N''(T_n, T_{n+1}]$ and $D_n = N^*(T_n^*, T_{n+1}^*]$, $n \in \mathbb{Z}$. θ_n is not in general equal to D_n, see the drawing. We have $\theta_n = D_n + N''(T_{n+1} - R_{n+1}, T_{n+1}] - N''(T_n - R_n, T_n]$. R_n is the sum of the packet delay from the sender to the receiver of the packet that arrives at T_n^* at the receiver, and $T_{n+1} - T_{n+1}^*$. Evidently, if the system is stationary, then $\mathbf{E}_N^0[D_0] = \mathbf{E}_N^0[\theta_0]$, that is $p^* = p$.

where $r := \mathbf{E}_{N'}^0[\hat{R}(0)]$ *and*

$$z_1 = 1 + \frac{\mathbf{E}_N^0[Z_0]}{\mathbf{E}_N^0[D_0]}, \tag{2.26}$$

$$z_2 = \frac{1}{1 + \frac{\mathbf{E}_N^0[\frac{Z_0}{X_0}]}{\mathbf{E}_N^0[\frac{D_0}{X_0}]}}, \tag{2.27}$$

$$z_3 = \frac{\mathbf{E}[\frac{\theta_0}{X_0}]}{\mathbf{E}_N^0[S_0]}, \tag{2.28}$$

$$z_4 = \frac{\mathbf{E}_N^0[\frac{D_0}{f(1/\hat{\theta}_0, r)}]}{\mathbf{E}_N^0[\frac{D_0}{f(1/\hat{\theta}_0, \hat{R}(0))}]}. \tag{2.29}$$

$$\tag{2.30}$$

The rational in (2.25) is *exactly* the expression we have analyzed in the earlier sections, whereas we assumed the round-trip times to be fixed. The z-terms are additional bias terms. The proposition follows as a direct application of the Palm inversion formula, and some appropriate factorizations of the resulting expression. The proof is simple, and hence omitted.

Our goal is to understand which factors would lead $\mathbf{E}[X(0)]$ to deviate from $f(p,r)$, and in which direction. It is natural to consider $y = \mathbf{E}[X(0)]/f(p,r)$. We use the additional notation

$$z_0 = \frac{\mathbf{E}_N^0[D_0]}{\mathbf{E}_N^0[\frac{D_0}{f(1/\hat{\theta}_0,r)}]}, \tag{2.31}$$

$$z_6 = \frac{f(p^*,r)}{f(p,r)}. \tag{2.32}$$

Indeed, $y = z_0 \prod_{i=1}^{6} z_i$.

We discuss and give intuitive interpretations of the above terms. A part of the discussion is limited to

(R) $f(p,r) = g(p)/r$, for a non-decreasing function $g : (0,1] \to \mathbb{R}_+$.

This is a non-restrictive assumption for the functions f used in practice, e.g. SQRT, and PFTK- formulae with the retransmit timeout equal to a linear function of the round-trip time, a setting suggested in TFRC specification [54].

In order to better understand the product $z_3 z_4$, we will also factor out it further. To that end, we introduce an additional assumption. Let T_n'' be the transmission time of the nth packet, with the standard convention $\cdots < T_{-1}'' < T_0'' \leq 0 < T_1'' < T_2'' < \cdots$.

(S) If the transmission of a packet labeled $n+1$ is scheduled at T_{n+1}'', then this packet transmission is *not* re-scheduled, even though the sender control state (the estimators of the loss-event rate and the average round-trip time) may have been updated in the interval (T_n'', T_{n+1}'').

The control state can be updated in an interval (T_n'', T_{n+1}''), if a report is received in that interval. This is illustrated in Figure 2.27.

We now proceed with discussion of the z terms.

(z_1) This term comprises the deviation of the expected loss-event intervals as observed by the sender and receiver.

(z_2) The term reflects the covariance of the difference of the loss-event intervals as seen at the sender and the receiver, and the send rate at the loss-event instant.

(z_3) This factor comprises more then one effect. If the round-trip time were fixed, then $z_3 \geq 1$. Equality occurs when the control is the basic control. Indeed,

$$\mathbf{E}_N^0\left[\frac{D_0}{X_0}\right] = \mathbf{E}_N^0\left[\frac{1}{X(0)}\sum_{n\in\mathbb{Z}} X(T_n')S_n^*\mathbf{1}_{[0,T_1)}(T_n')\right].$$

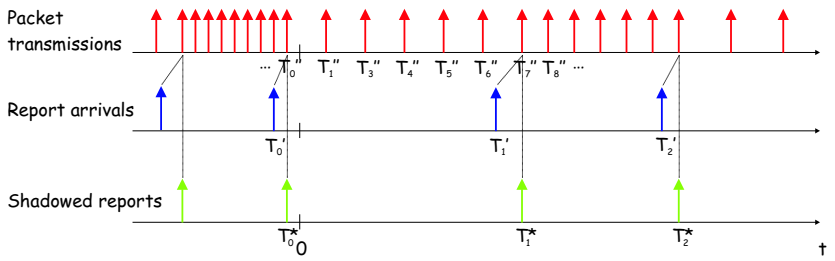


Figure 2.27: A report arrives at the sender at an instant T'_n. This instant is "shadowed" to T^*_n, the first packet transmission instant after T'_n. The instants of the shadowed report arrivals is a subsequence of the packet transmission instants.

By definition, $X(t) = f(1/\hat{\theta}(t), \hat{R}(t))$, and hence

$$\mathbf{E}^0_N\left[\frac{D_0}{X_0}\right] = \mathbf{E}^0_N\left[\frac{1}{f(1/\hat{\theta}_0, \hat{R}_0)}\sum_{n\in\mathbb{Z}} S^*_n f(1/\hat{\theta}(T'_n), \hat{R}(T'_n))\mathbf{1}_{[0,T_1)}(T'_n)\right].$$

Now, it is useful to gain insight for the special case of the basic control, but with variable round-trip times. Then, $\hat{\theta}(t) = \hat{\theta}_n$, for all $t \in [T_n, T_{n+1})$, $n \in \mathbb{Z}$. Assume, in addition, that the function f is of the form (R). Then

$$\begin{aligned}\mathbf{E}^0_N\left[\frac{D_0}{X_0}\right] &= \mathbf{E}^0_N\left[\hat{R}_0\sum_{n\in\mathbb{Z}}\frac{1}{\hat{R}(T'_n)}S^*_n\mathbf{1}_{[0,T_1)}(T'_n)\right] \\ &= \mathbf{E}^0_N[\hat{R}_0 N'(0,T_1]] + \mathbf{E}^0_N\left[\hat{R}_0\sum_{n\in\mathbb{Z}}\left(\frac{S^*_n}{\hat{R}(T'_n)} - 1\right)\mathbf{1}_{[0,T_1)}(T'_n)\right].\end{aligned}$$

The first summation element on the right-hand side comprises the covariance of the reciprocal of the round-trip time estimator and the number of the round-trip time rounds in a loss-event interval with respect to $\mathbf{P}^0_N$. The second element is an additional bias term; for the controls like TFRC, whereas the reports are generated per-each estimated round-trip time, we expect this bias term to be near zero. We proceed with the general case. Note that $\hat{\theta}(t) \geq \hat{\theta}_n$, for all $t \in [T_n, T_{n+1})$, $n \in \mathbb{Z}$. Hence,

$$\mathbf{E}^0_N\left[\frac{D_0}{X_0}\right] \geq \mathbf{E}^0_N\left[\frac{1}{f(1/\hat{\theta}_0, \hat{R}_0)}\sum_{n\in\mathbb{Z}} S^*_n f(1/\hat{\theta}_0, \hat{R}(T'_n))\mathbf{1}_{[0,T_1)}(T'_n)\right].$$

Under (R), it holds

$$\mathbf{E}^0_N\left[\frac{D_0}{X_0}\right] \geq \mathbf{E}^0_N\left[\hat{R}_0\sum_{n\in\mathbb{Z}}\frac{S^*_n}{\hat{R}(T'_n)}\mathbf{1}_{[0,T_1)}(T'_n)\right]. \tag{2.33}$$

Note from the above derivations that the inequality holds with strict equality for the basic control. We use the last lower bound shortly to factor out the bias term z_3.

(z_4) The factor reflects both the variability of the round-trip times and their correlation to the loss-process. We re-write

$$z_4 = \frac{r}{\mathbf{E}_N^0[\hat{R}(0)]} \times \frac{\mathbf{E}_N^0[\hat{R}(0)]}{r} \frac{\mathbf{E}_N^0\left[\frac{\theta_0}{f(1/\hat{\theta}_0,r)}\right]}{\mathbf{E}_N^0\left[\frac{\theta_0}{f(1/\hat{\theta}_0,\hat{R}(0))}\right]}. \tag{2.34}$$

Our motivation for the last factorization is evident under (R). Then the second factor in the right-hand side of (2.34) is equal to

$$\frac{1}{1 + \frac{\mathbf{cov}_N^0\left[\hat{R}(0), \frac{\theta_0}{g(1/\hat{\theta}_0)}\right]}{\mathbf{E}_N^0[\hat{R}(0)]\mathbf{E}_N^0[\frac{\theta_0}{g(1/\hat{\theta}_0)}]}}.$$

The last factor reflects a correlation of the smoothed round-trip time and the loss process. Another motivation for (2.34) comes in the factorization of z_3 and z_4 which we do next.

We use a simplifying notation $Y_n := \sum_{k\in\mathbb{Z}} \frac{S_k^*}{\hat{R}_k} \mathbf{1}_{[T_n,T_{n+1})}(T_k')$, $n \in \mathbb{Z}$. We factor out the lower bound in (2.33) as

$$\frac{\mathbf{E}_N^0[Y_0]\mathbf{E}_N^0[\hat{R}(0)]}{\mathbf{E}_N^0[S_0]} \times \frac{\mathbf{E}_N^0[\hat{R}(0)Y_0]}{\mathbf{E}_N^0[Y_0]\mathbf{E}_N^0[\hat{R}(0)]}. \tag{2.35}$$

Now, the factorization in the last expression and (2.34) suggest the following

$$z_3 z_4 = y_1 y_2 y_3 y_4,$$

where

$$y_1 = \frac{\mathbf{E}_N^0[Y_0]r}{\mathbf{E}_N^0[S_0]}, \tag{2.36}$$

$$y_2 = \frac{\mathbf{E}_N^0[\hat{R}(0)Y_0]}{\mathbf{E}_N^0[\hat{R}(0)]\mathbf{E}_N^0[Y_0]}, \tag{2.37}$$

$$y_3 = \frac{\mathbf{E}_N^0[\frac{D_0}{X_0}]}{\mathbf{E}_N^0[\hat{R}(0)Y_0]}, \tag{2.38}$$

$$y_4 = \frac{\mathbf{E}_N^0[\hat{R}(0)]}{r} \frac{\mathbf{E}_N^0[\frac{\theta_0}{f(1/\hat{\theta}_0,r)}]}{\mathbf{E}_N^0[\frac{\theta_0}{f(1/\hat{\theta}_0,\hat{R}(0))}]}. \tag{2.39}$$

$$\tag{2.40}$$

We discuss the factors y_1, y_2, y_3, and y_4, as follows.

(y_1) Note that this factor can be written as

$$y_1 = \mathbf{E}^0_{N'} \left[\frac{S^*_0}{\hat{R}(0)} \right] / \left(\frac{\mathbf{E}^0_{N'}[S^*_0]}{\mathbf{E}^0_{N'}[\hat{R}(0)]} \right).$$

For some protocols (TFRC, for instance), roughly speaking, the reports are sent by the receiver per-each estimated round-trip time. In this case, the bias term y_1 can be interpreted as an expectation involding the delay.

(y_2) This factor can be re-written as a covariance of $\hat{R}(0)$ and $N'(0, T_1]$ with respect to $\mathbf{P}^0_N$ plus some additional terms.

(y_3) This factor comprises more than one effect. One effect is due to the send rate increase in the absence of loss-events, as allowed with comprehensive control. Note that if the control were basic control, then $y_3 = 1$.

(y_4) This factor is similar to z_4, but different. Note that z_4 encompasses the term r; the expectation of the round-trip time with respect to $\mathbf{P}^0_{N'}$. In constrast, y_4 depends only on the expectations with respect to $\mathbf{P}^0_N$.

Empirical Evaluation

In this section, we show empirical estimates of the bias terms identified in the preceding section. The bias terms reveal us what makes the control conservative, or not.

Internet Experiments: LAN to LAN

We first consider the z-factors, defined in (2.31), (2.26)–(2.29), see Figure 2.28. We observe: The dominant factors are z_0, z_3, and z_4; other factors are non-negligible only for KTH. The factor z_0 explains roughly about 5% of the undershoot, for all the receivers, except for UMELB, where it explains a larger amount, roughly about 15%.The absolute values of the factors z_3 and z_4 are mostly in the order of the factor z_0. A peculiar case is KTH, for this set of experiments only, we observe that z_1 is significantly positive. Recall that a positive z_1 means that the event-average of the loss-event interval as seen at the sender is larger than as seen at the receiver. The cause of this remains unknown to us.

We now consider the y-factors; recall the y-factors are a factorization of the product $z_3 z_4$. See Figure 2.29. We observe: For both KTH and UMELB, the factor y_4, which recall is a correlation of the loss process and the round-trip time, tends to be *positive*. Recall that the receivers at KTH and UMELB are connected via a low access rate of 10 Mb/s, while the sender is connected with a rate of 100 Mb/s. A similar observation will be made shortly for a cable modem.

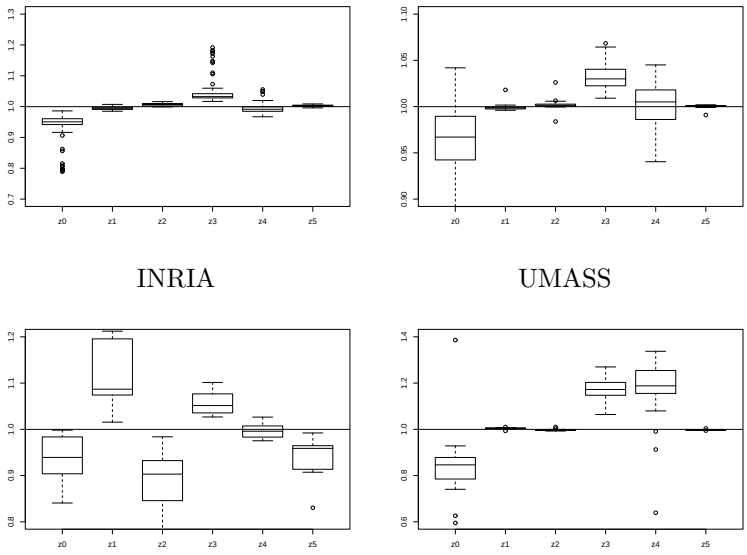


Figure 2.28: Internet experiments: what makes the control conservative. In all the cases, the factor z_0 contributes to conservativeness. Other non-negligible factors are z_3 and z_4.

Internet Experiments: LAN to Cable modem

Consider the z-factors and y-factors displayed in Figure 2.30. We observe: The non-negligible factors are z_0, z_3, and z_4. The factorization of $z_3 z_4$ into y-factors indicates: Both y_3 and y_4 are siginificant and contribute to overshoot. Their absolute values are of the same order as the factor z_0. Recall that the factor y_4 is a correlation of the delay and loss process, see (2.39) for a precise definition. We indeed observed that with cable modem experiments the round-trip delay and the loss-event intervals are strongly correlated; the evidence is shown in Figure A.10.

Laboratory Experiments

We consider the z-factors in Figure 2.31. We observe: The non-negligible factors are z_0, z_3 and z_4. Now, consider the factorization of $z_3 z_4$ into y-factors, also displayed in Figure 2.31. We note: The non-negligible factors are y_3 and y_4; in most cases, y_3 contributes to overshoot; in contrast, y_4 contributes to undershoot.

We now discuss the factors for RED in Figure 2.31, left, in more quantitative terms. The factor z_0 contributes to an undershoot of about 10%. The factors

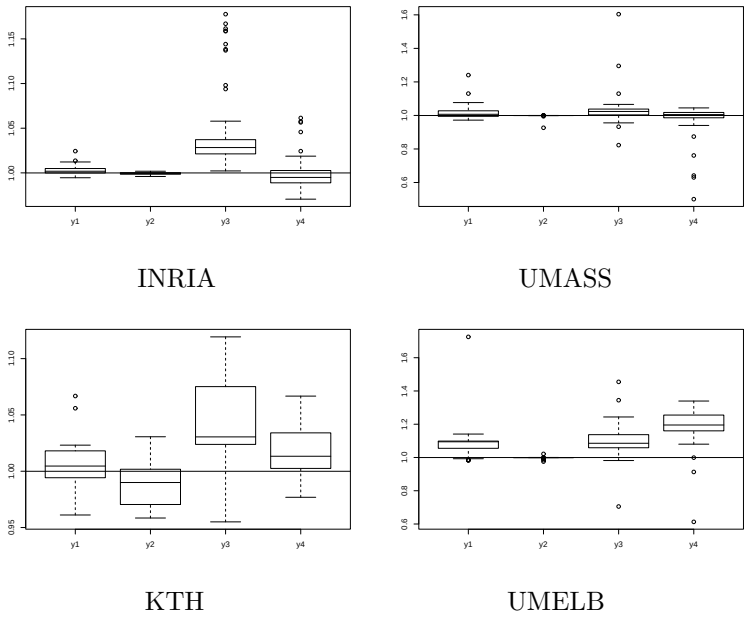
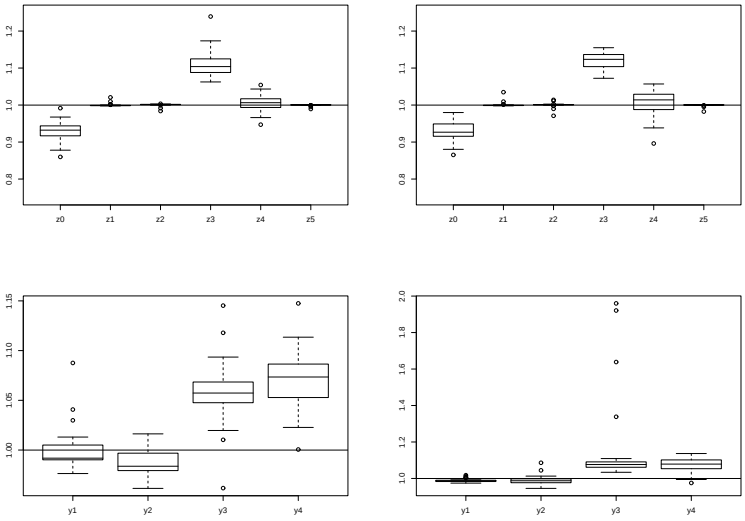


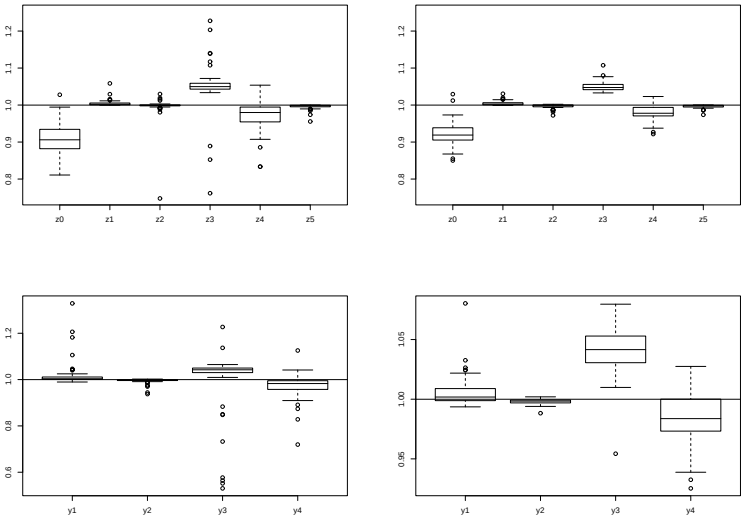
Figure 2.29: Internet experiments: what makes the control conservative revisited. The y-factors indicate that y_3 and y_4 are non-negligible.

y_3 and y_4 are in absolute value mostly less than 5%. The product $y_3 y_4$ seems to be near to 1. In summary, we observe that the conservativeness seems to be mostly due the term z_0, that is due to the reasons found in Section 2.3. The same qualitative observations remain unchanged for DropTail in Figure 2.31, right.



EPFL UMASS

Figure 2.30: Cable modem: what makes the control conservative.



RED DropTail

Figure 2.31: Lab experiments: What makes the control conservative.

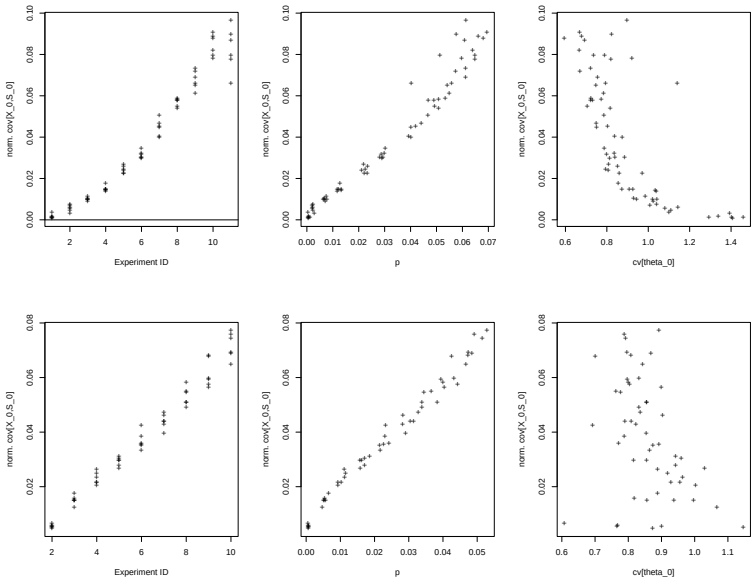


Figure 2.32: Lab experiments: $\mathbf{cov}_N^0[X_0, S_0]/(\mathbf{E}_N^0[X_0]\mathbf{E}_N^0[S_0])$ for (Top) RED and (Bottom) DropTail.

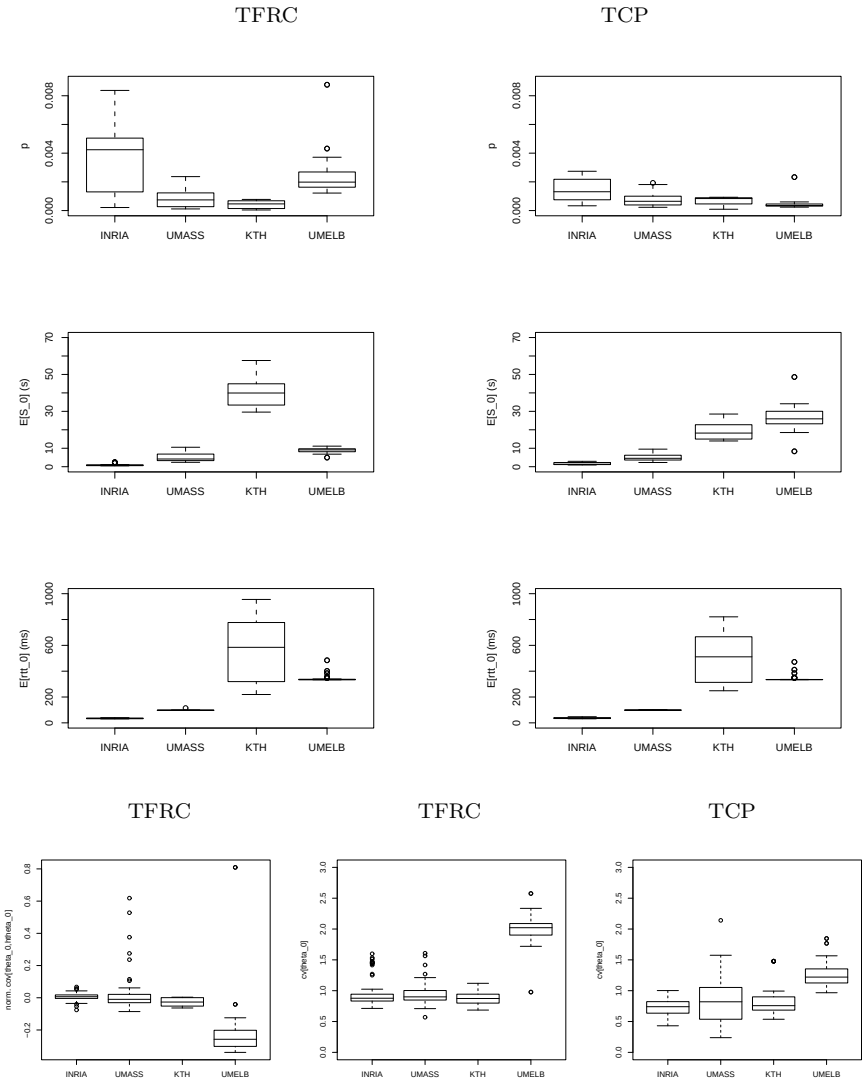


Figure 2.33: Internet experiments: boxplots of the empirical estimates of (First Row) the loss-event rate, (Second Row) the average inter loss-event time, (Third Row) the average round-trip time, (Forth Row) (Left) covariance $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$ for TFRC, (Middle) coefficient of the variation $\mathbf{cv}_N^0[\theta_0]$ for TFRC, (Right) coefficient of the variation $\mathbf{cv}_N^0[\theta_0]$ for TCP.

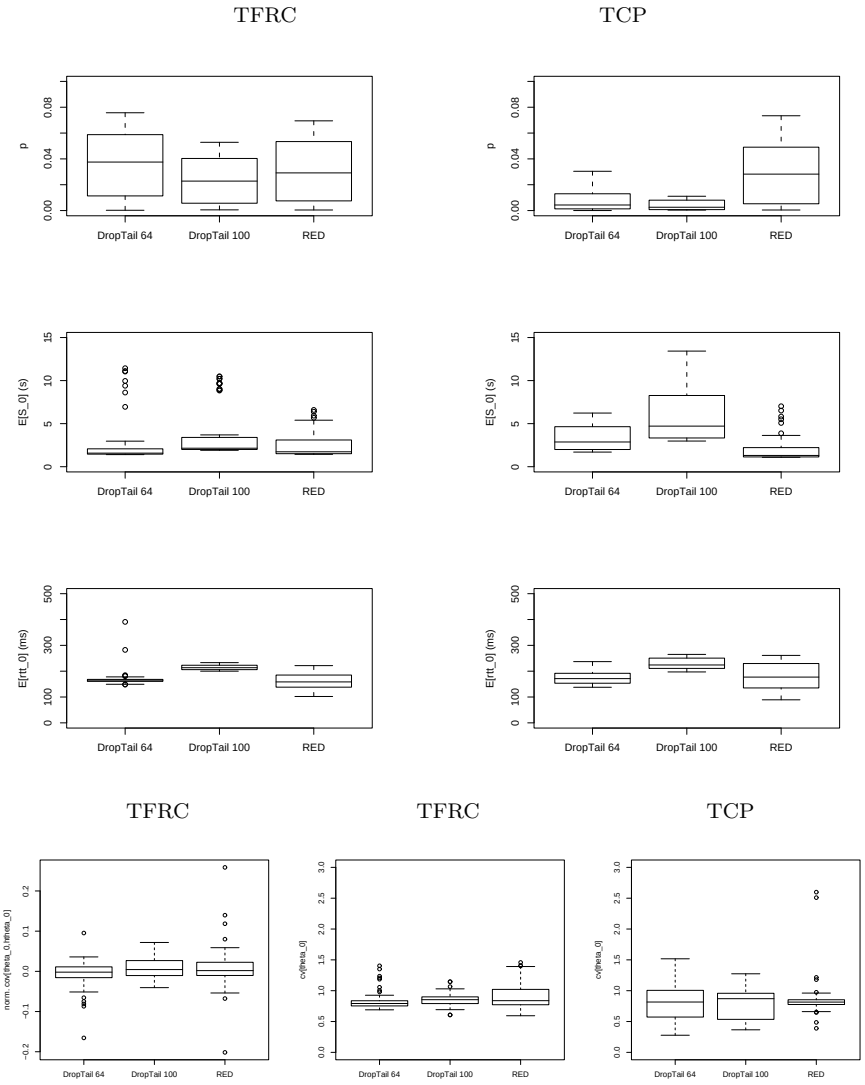


Figure 2.34: Lab experiments: boxplots of the empirical estimates of (First Row) the loss-event rate, (Second Row) average inter loss-event time, (Third Row) average round-trip time, (Forth Row) (Left) covariance $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$ for TFRC, (Middle) coefficient of the variation $\mathbf{cv}_N^0[\theta_0]$ for TFRC, (Right) coefficient of the variation $\mathbf{cv}_N^0[\theta_0]$ for TCP.

Chapter 3

Increase-Decrease Controls

In this chapter, we:

- identify the notion of fairness according to which the send rates are allocated to AIMD senders, in a network, with arbitrarily fixed routes and arbitrary round-trip times between a sender/receiver pair;
- obtain a new bound on the time-average window of an AIMD sender, which depends on only *two* statistical parameters of the loss process;
- study design rules for the synthesis of an increase-decrease control; identified a set of increase-decrease controls, which designed for a *reference* loss processes, overshoot their design loss-throughput formula for a more general loss process. (Highspeed TCP [42] belongs to this set.)

3.1 Introduction and Outline

The fairness of bandwidth sharing in a network of links, where senders in the network adjust their send rates in a classical additive-increase/multiplicative-decrease (AIMD) manner (Chiu and Jain [31]), is only *partially* understood, under some special assumptions. Hurley, Le Boudec, and Thiran [61] showed that a rate allocation achieved by AIMD senders, in a network, is obtained by maximizing a system-wide objective function F_A subject to the capacity constraints of the network. The authors termed this notion of fairness F_A-fairness. The F_A-fairness is neither max-min (Chiu and Jain [31]) nor proportional fairness (Kelly, Maullo, and Tan [68]). However, the result of [61] was obtained under a restrictive assumption that all sender/receiver pairs in the network have *equal* round-trip times. We generalize F_A-fairness to hold for *arbitrary* round-trip times. The result helps us to better understand a known bias against long round-trip time TCP connections.

Most of the results in the above context provide predictions of the throughput attained by an adaptive sender in a network. The method in the above

context does not, in general, give exact value of the throughput, but an approximation. There has been a lot of work on obtaining precise loss-throughput formulas for senders such as TCP, see Padhye *et al.* [93], Altman, Avrachenkov, and Barakat [4], and the references therein. We show a result on the time-average window of an AIMD sender. The result is an *upper-bound*, and it is *new*. It requires only *two* statistical parameters of the loss process. Knowing a throughput formula for senders such as AIMD, which would well-approximate a TCP sender, is important for various applications; an example is the design of equation-based rate controls, studied in Chapter 2.

An *inverse* problem, in a sense defined later in this chapter, to the problem of deriving a relation of the throughput and the loss-event rate for a given send rate or window control (we call "an analysis problem") is the "synthesis problem." In the synthesis problem, one is given a target loss-throughput formula, and the goal is to construct a send rate or window control such that, loosely speaking, the attained throughput compares well with the given target function evaluated at the loss-event rate as seen by *this* control. To our knowledge, solving the synthesis problem appeared first as an attempt to design TCP-friendly controls. Some examples are Rejaie, Handley, and Estrin [97], Bansal and Balakrishnan [10], Floyd, Handley, and Padhye [44], more recently Jin *et al.* [64]. One application of these controls is send rate control of audio and video streaming sources in the Internet. Another application is for control of high-speed bulk data transfers; for example, Highspeed TCP [42]. Again, a goal is to design a TCP-friendly control.

We pose the synthesis problem such that a design requirement is that the throughput of a control not be larger than the value of a given target function evaluated at the loss-event rate of this protocol. We show two design rules and provide sufficient conditions under which a control designed with these rules, solves the synthesis problem. Our final results in this chapter are on identifying a set of controls, which designed by another design rule—design for a reference loss processes—do not, in general, solve the synthesis problem. This means that such controls would *a priori* overshoot their design target, by following the given design method. Given that, typically, a deterministic process is taken as a reference, these type of results are intimately connected with a known phenomena that in many circumstances a determinism is an extremal.

3.1.1 Outline of the Chapter

We present our results on the fairness of network bandwidth sharing in Section 3.2. In Section 3.3, we introduce our notation and assumptions that are used in subsequent sections. Section 3.4 gives an upper-bound on time-average window of an AIMD sender. Our analysis of the synthesis problem is given in Section 3.5. The chapter ends with our conclusions in Section 3.6.

3.2 Fairness of Network Bandwidth Sharing

Consider a network of nodes that serve packets. Assume the senders in the network are adaptive, that they adjust their send rates according to an AIMD control. An AIMD sender adjusts the send rate at some instants of time, such that if in the last rate update interval no loss-event is detected, the rate is increased by a fixed increment, else, the send rate is reduced to a fixed fraction of the current send rate. This is a traditional control found in networks [31], in particular, it is a control law of TCP [63].

Assume each connection between a sender and a receiver in the network traverses an arbitrary fixed sequence of nodes. Assume the average round-trip time for a sender/receiver pair is arbitrary. The problem is the following: *How are the long-term average send rates allocated to the senders?*

We find that the rates are allocated according to F_A-fairness that depends on the values of the round-trip times. This result generalizes [61], obtained under hypothesis that the round-trip times are the same for all sender/receiver pairs. We display our result in the next section.

3.2.1 Assumptions and Notations

Consider a sender i that belongs to a collection of senders $\mathcal{I}$. Fix a positive real $\epsilon > 0$. Let $X^\epsilon_{i,n}$ be the send rate of the sender i at a real-time instant $T_{i,n}$, $n \in \mathbb{N}$. We use the convention $T_{i,0} = 0 < T_{i,1} < T_{i,2} < \cdots$, and the definition $S_{i,n} := T_{i,n+1} - T_{i,n}$, $n \in \mathbb{N}$. We consider the sender i that adjusts its send rate as, for some $X^\epsilon_{i,0} \geq 0$,

$$X^\epsilon_{i,n+1} = X^\epsilon_{i,n} + \epsilon[a_i(X^\epsilon_{i,n}) - (X^\epsilon_{i,n} - b_i(X^\epsilon_{i,n}))Z^\epsilon_{i,n}],\ n = 0, 1, 2, \ldots \qquad (3.1)$$

Here $Z^\epsilon_{i,n} = 1$, if there is a loss-event on the rate update interval $[T_{i,n}, T_{i,n+1})$, else, $Z^\epsilon_{i,n} = 0$. The functions $a_i(\cdot)$ and $b_i(\cdot)$ are given, assumed to be positive-valued and $b_i(x) < x$. The recurrence (3.1) is a particular instance of a stochastic approximation algorithm, see Kushner and Yin [77].

We base our main result on Theorem 3.1 (Chapter 12, [77]), which is in our adaptation under the hypotheses given shortly. First for an instant t' in real-time, we associate an instant $t = \epsilon t'$ in *virtual-time*. In particular, we denote $T^\epsilon_{i,n} = \epsilon T_{i,n}$, $n \in \mathbb{N}$. We define

$$X^\epsilon_i(t) = X^\epsilon_{i,n},\ T^\epsilon_{i,n} \leq t < T^\epsilon_{i,n+1}.$$

See Figure 3.1 for an illustration. We proceed to display the assumptions under which weak convergence to an ODE holds.

(A1) $\{a_i(X^\epsilon_{i,n}) - (X^\epsilon_{i,n} - b_i(X^\epsilon_{i,n}))Z^\epsilon_{i,n}, T_{i,n+1} - T_{i,n}\}$ is uniformly integrable[1].

[1]The uniform integrability of a sequence of vector-valued random-variables $\{Y_n\}$ is equivalent to saying $\lim_{m\to\infty} \sup_n \mathbf{E}[|Y_n|\mathbf{1}_{|Y_n|\geq m}]=0$ [77].

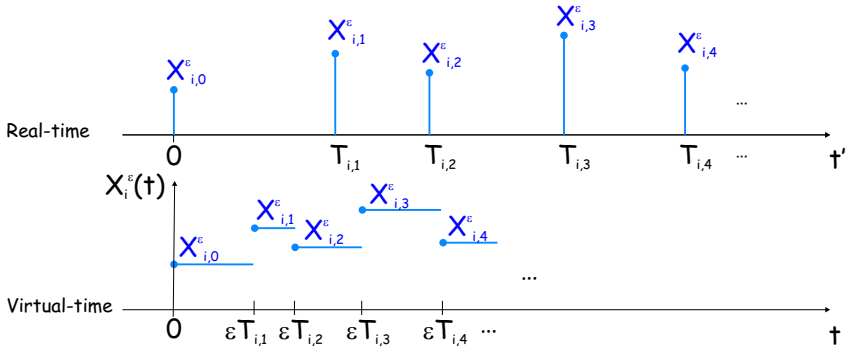


Figure 3.1: The interpolation on virtual-time.

(A2) There are real-valued functions $z^\epsilon_{i,n}(\cdot)$, continuous in n and ϵ, and random variables $\beta^\epsilon_{i,n}$ and non-negative random variables $D^\epsilon_{i,j,n}$ such that

$$\mathbf{P}[Z^\epsilon_{i,n} = 1 | \mathcal{F}^\epsilon_{i,n}] = z^\epsilon_{i,n} \left(\begin{bmatrix} X_{1,n}(T^\epsilon_{1,n} - D^\epsilon_{i,1,n}) \\ \vdots \\ X_{|\mathcal{I}|,n}(T^\epsilon_{|\mathcal{I}|,n} - D^\epsilon_{i,|\mathcal{I}|,n}) \end{bmatrix} \right) + \beta^\epsilon_{i,n},$$

where $\{\beta^\epsilon_{i,n},\ n, i, \epsilon\}$ is uniformly integrable, and for each $t > 0$, in probability as $\epsilon \to 0$,

$$\sup_{n \le t/\epsilon} D^\epsilon_{i,j,n} \to 0.$$

Here $D^\epsilon_{i,j,n} = \epsilon D_{i,j,n}$, where $D_{i,j,n}$ is a random delay in real-time.

(A3) There are real-valued strictly positive functions $s^\epsilon_{i,n}(\cdot)$, and continuous uniformly in n and ϵ, such that

$$\mathbf{E}[T^\epsilon_{i,n+1} - T^\epsilon_{i,n} | \mathcal{F}^\epsilon_{i,n}] = s^\epsilon_{i,n} \left(\begin{bmatrix} X_{1,n}(T^\epsilon_{1,n} - D^\epsilon_{i,1,n}) \\ \vdots \\ X_{|\mathcal{I}|,n}(T^\epsilon_{|\mathcal{I}|,n} - D^\epsilon_{i,|\mathcal{I}|,n}) \end{bmatrix} \right).$$

(A4) There are continuous real-valued functions $\bar{z}_i(\cdot)$ such that for each $x \ge 0$,

$$\lim_{m,n,\epsilon} \frac{1}{m} \sum_{k=n}^{n+m-1} \left(z^\epsilon_{i,k}(x) - \bar{z}_i(x) \right) = 0.$$

(A5) There are continuous real-valued functions $\bar{s}_i(\cdot)$ such that for each $x \ge 0$,

$$\lim_{m,n,\epsilon} \frac{1}{m} \sum_{k=n}^{n+m-1} \left(s^\epsilon_{i,k}(x) - \bar{s}_i(x) \right) = 0.$$

(A6)

$$\lim_{m,n,\epsilon} \frac{1}{m} \sum_{k=n}^{n+m-1} \mathbf{E}[\beta_{i,k}^{\epsilon} | \mathcal{F}_{i,n}^{\epsilon}] = 0.$$

3.2.2 The Limit-Mean ODE

Having displayed the assumptions (A1)-(A6), it follows from Theorem 3.1 [77].

Theorem 4 (Kushner and Yin [77]) *Assume (A1)-(A6). Suppose that* $\{X_{i,n}\}$ *is bounded with probability one, then for any* $\mu > 0$*, there exists a* $t_\mu > 0$ *such that for* $t \geq t_\mu$*,* $||X^{\epsilon}(t) - x(t)|| < \mu$*. Also, for any* $t > t_\mu$*,*

$$\limsup_{\epsilon} \mathbf{P} \left[\sup_{t_\mu \leq s \leq t} ||X^{\epsilon}(t) - x(t)|| \geq \mu \right] = 0,$$

where, $x(t) = [x_1(t), x_2(t), \ldots, x_{|\mathcal{I}|}(t)]^T$*,*

$$\frac{dx_i(t)}{dt} = \frac{a_i(x_i(t)) - [a_i(x_i(t)) - b_i(x_i(t))]\bar{z}_i(x(t))}{\bar{s}_i(x(t))}, \ i = 1, 2, \ldots, |\mathcal{I}|, \qquad (3.2)$$

is assumed to be globally asymptotically stable[2]*.*

What This Means. The convergence of the recursive sequence (3.1) to the solution of the ODE (3.2) on the appropriately scaled time (virtual-time) is in a sample-path sense. Informally speaking, for any sufficiently large time instant, as $\epsilon \to 0$, the continuous-time interpolation of the sequence (3.1) is in the μ-tube around the solution of the ODE (3.2). In order to solve the ODE, it remains to know $\bar{z}_i(\cdot)$ and $\bar{s}_i(\cdot)$, $i \in \mathcal{I}$. Note that $\bar{z}_i(x)$ is the probability of receiving a negative feedback in a rate adaptation interval, given that the send rates of the senders are fixed to x. $\bar{s}_i(x)$ is the average rate adaptation interval, given that the sender send rates are fixed to x.

Without loss of generality, we use[3]

$$p_i(x) = \frac{\bar{z}_i(x)}{x_i \bar{s}_i(x)}, \ i = 1, 2, \ldots, |\mathcal{I}|.$$

Then, we can re-write the ODE (3.2) as

$$\frac{dx_i}{dt} = x_i[a_i(x_i) - b_i(x_i)] \left(\frac{a_i(x_i)}{x_i[a_i(x_i) - b_i(x_i)]\bar{s}_i(x)} - p_i(x) \right). \qquad (3.3)$$

[2]For completeness, we recall some definitions regarding stability of an ODE [55]. A solution of an ODE $x : \mathbb{R}_+ \to \mathbb{R}^n$, is said to be *asymptotically stable*, if for any $\epsilon > 0$ and any $t_0 \geq 0$ there exists a $\delta > 0$ such that for any solution $y : \mathbb{R}_+ \to \mathbb{R}^n$ with $||y(t_0) - x(t_0)|| < \delta$, it holds $||y(t) - x(t)|| < \epsilon$, all $t \geq t_0$, and $\lim_{t\to\infty} ||y(t) - x(t)|| = 0$. The *basin of abstraction* of an asymptotically stable solution $x : \mathbb{R}_+ \to \mathbb{R}^n$ is the set of $y_0 \in \mathbb{R}$ such that the solution $y : \mathbb{R}_+ \to \mathbb{R}^n$ with $y(0) = y_0$ satisfies $\lim_{t\to\infty} ||y(t) - x(t)|| = 0$. Finally, a solution $x : \mathbb{R}_+ \to \mathbb{R}^n$ is said to be *globally asymptotically stable* if its is asymptotically stable and its basin of attraction is the whole space $\mathbb{R}^n$.

[3]Think of this as of the loss-event rate.

We consider a network of links in $\mathcal{L}$. Assume that the routing paths in the network are fixed, defined by a $|\mathcal{I}| \times |\mathcal{L}|$ matrix A, such that $A_{i,l} = 1$, if the sender i sends through the node l, else $A_{i,l} = 0$. Suppose a link l in the network is associated with a function $g_l(\cdot)$, assumed to be positive-valued, non-decreasing convex. It is standard to assume

$$p_i(x) = \sum_{l \in \mathcal{L}} A_{i,l} g_l \left(\sum_{j \in \mathcal{I}} A_{j,l} x_j \right). \tag{3.4}$$

The imposed function $p_i(\cdot)$ corresponds to "the rate proportional feedback," as introduced by Hurley, Le Boudec, and Thiran [61].

3.2.3 The Asymptotic Solution of the ODE

We next discuss the asymptotic solution of the ODE, $\lim_{t \to \infty} x(t)$. Let $J(\cdot)$ be such that

$$\frac{\partial}{\partial x_i} J(x) = \frac{a_i(x_i)}{x_i[a(x_i) - b_i(x_i)]\bar{s}_i(x)} - p_i(x). \tag{3.5}$$

Let

$$F(x) = \sum_{i \in \mathcal{I}} \int \frac{a_i(x_i)}{x_i[a(x_i) - b_i(x_i)]\bar{s}_i(x)} dx_i.$$

Then, we can write

$$J(x) = F(x) - \sum_{l \in \mathcal{L}} \int_0^{\sum_{j \in \mathcal{I}} A_{j,l} x_j} g_l(s) ds. \tag{3.6}$$

Under the hypothesis that $J(\cdot)$ is a strictly concave function, it is a Lyapunov function of the ODE (3.3). The limit solution of (3.3) $\lim_{t \to \infty} x(t) = x^*$ is the solution of the following optimization problem

$$\text{maximize } J(x)$$

subject to $x \geq 0$.

The summation term in (3.6) acts a role of a penalty function that keeps a solution within the capacity limits of the system. If the function $g_l(x)$, all $l \in \mathcal{L}$, can be assumed to be arbitrary close to δ_{c_l}, $\delta_{c_l}(x) = 0$, $x < c_l$, and else $\delta_{c_l}(x) = 1$. Here c_l is the service rate of a link l. Then, the optimization problem can be stated as follows

$$\text{maximize } F(x)$$

subject to $x \geq 0$ and the capacity constraints $\sum_{j \in \mathcal{I}} A_{j,l} x_j \leq c_l$, $l \in \mathcal{L}$.

3.2.4 Fairness with AIMD Senders

Assume the senders in the network adjust their send rates according to an AIMD control, that is $a_i(x) = \alpha_i$ and $b_i(x) = \alpha_i + \beta_i x$, for some $\alpha_i > 0$, $0 < \beta_i < 1$. We assume the AIMD recurrence is "clocked" per-each round-trip time interval, that is to say $T_{i,n+1} - T_{i,n}$ is the nth round-trip time of the sender i. Assume (3.4) and for an arbitrarily fixed strictly positive real number r_i,

$$\bar{s}_i(x) = r_i. \tag{3.7}$$

Note that this assumption *does not* mean that for the sender i, the round-trip times are fixed to r_i. The assumption implies that the *average* round-trip time of the sender i is r_i. Recall that $\bar{s}_i(x)$ can be interpreted as the conditional expected round-trip time, given that the send rates are fixed to x. The assumption means that $\bar{s}_i(x)$ is independent of x. In particular, in practice, this would correspond to networks where the propagation delay is a dominant component of the round-trip delay and the queueing delays at the links are negligible. The assumption applies more generally. Suppose that we choose the constants r_i's such that $\bar{s}_i(x^*) = r_i$, for all i, where $x^* = \lim_{t\to\infty} x(t)$. Note that now $\bar{s}_i(x)$ is allowed to depend on x. The problem is that "the equilibrium" r_i's are unknown. The next result holds in general for r_i, the equilibrium round-trip time of the sender i.

Proposition 7 *The function $F(\cdot)$ in (3.6) is*

$$F_A(x) = \sum_{i\in\mathcal{I}} \frac{1}{r_i} \ln \frac{x_i}{\alpha_i + \beta_i x_i}. \tag{3.8}$$

This result generalizes [61] to hold for arbitrary round-trip times.

F_A-limits

We can write (3.8) in the following form

$$F_A(x) = \sum_{i\in\mathcal{I}} \frac{1}{r_i} \ln \frac{1}{\beta_i} - \sum_{i\in\mathcal{I}} \frac{1}{r_i} \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n} \left(\frac{\alpha_i}{\beta_i x_i}\right)^n.$$

By the limited development into Taylor series, for $x_i \gg \alpha_i/\beta_i$, we obtain that the rate allocation x is such that x maximizes

$$F_A^+(x) = \sum_{i\in\mathcal{I}} -\frac{\alpha_i}{r_i \beta_i x_i}, \tag{3.9}$$

subject to the capacity constraints. In the opposite case, $x_i \ll \alpha_i/\beta_i$, all $i \in \mathcal{I}$, by simple manipulation we obtain another objective function

$$F_A^-(x) = \sum_{i\in\mathcal{I}} \frac{1}{r_i} \ln x_i. \tag{3.10}$$

We use (3.9) and (3.10) in Section 3.2.6 to understand the bias of TCP against long round-trip time connections.

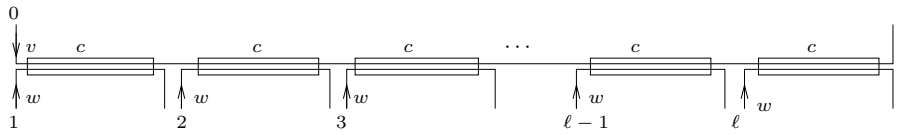


Figure 3.2: A multiple-bottleneck network.

Example: A Multiple-Bottleneck Network

In this section we consider a network consisting of a sequence of ℓ links as depicted in Fig. 3.2. In this scenario we distinguish the connections that traverse all the links (class-0), and the connections that traverse a single link (class-i, for connections that traverse the link i). Let n_i be the number of the connections that belong to the class i. The ith link offers capacity c_i, $i \in \mathcal{L}$.

Note that the capacity constraints are $n_0 x_0 + n_i x_i = c_i$, for $i = 1, 2, \ldots, \ell$. We re-define (3.8) as a function of x_0, as

$$F_A(x_0) = \frac{n_0}{\tau_0} \ln \frac{x_0}{\alpha_0 + \beta_0 x_0} + \sum_{i=1}^{\ell} \frac{n_i}{r_i} \ln \frac{c_i - n_0 x_0}{\alpha_i n_i + \beta_i (c_i - n_0 x_0)}. \tag{3.11}$$

We restrict our example to the following setup of the parameters

$$c_i = c,\ \alpha_i = \alpha_1,\ \beta_i = \beta_1,\ n_i = w,\ i = 1, 2, \ldots, \ell,\ n_0 = v. \tag{3.12}$$

Then, we can write

$$F_A(x_0) = \frac{v}{r_0} \ln \frac{x_0}{\alpha_0 + \beta_0 x_0} + \frac{\ell}{\bar{r}} w \ln \frac{c - v x_0}{\alpha_1 n_1 + \beta_1 (c - v x_0)}, \tag{3.13}$$

where $1/\bar{r} = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{1}{r_i}$.

Lemma 1 *The F_A–fair rate allocation for the multiple-bottleneck scenario is*

$$x_0 = \frac{-B - \sqrt{B^2 - 4AC}}{2A}, \tag{3.14}$$

where

$$\begin{aligned} A &= v^2 \alpha_0 \beta_1 - w^2 \alpha_1 \beta_0 \frac{\ell r_0}{\bar{r}} \\ B &= -\alpha_0 (v(w\alpha_1 + 2c\beta_1) + w^2 \alpha_1 \frac{\ell r_0}{\bar{r}}) \\ C &= \alpha_0 c (w\alpha_1 + c\beta_1), \end{aligned}$$

for $v^2 \alpha_0 \beta_1 - w^2 \alpha_1 \beta_0 \ell r_0 / \bar{r} \neq 0$, else

$$x_0 = \frac{c(w\alpha_1 + c\beta_1)}{v(w\alpha_1 + 2c\beta_1) + w^2 \alpha_1 \frac{\ell r_0}{\bar{r}}}. \tag{3.15}$$

Then,

$$x_i = \frac{c - vx_0}{w},\ i = 1, 2, \ldots, \ell.$$

Proof 2 *Proof is along the same lines as in [61].*

We gain more insight by considering the first–order approximation (3.9), which is for $x_0 \gg \alpha_0/\beta_0$ and $x_i \gg \alpha_1/\beta_1$, all $i = 1, 2, \ldots, \ell$,[4]

$$\frac{x_0}{c} = \frac{1}{v + w\sqrt{\frac{\alpha_1\beta_0\ell r_0}{\alpha_0\beta_1\bar{r}}}}, \tag{3.16}$$

Even a simpler expression is obtained for a particular case, $r_i = r_1$, all $i = 1, 2, \ldots, \ell$, so that $\bar{r} = r_1$,

$$\frac{x_0}{c} = \frac{1}{v + w\sqrt{\ell\frac{\alpha_1\beta_0 r_0}{\alpha_0\beta_1 r_1}}}. \tag{3.17}$$

We make a connection with the early work of Floyd [41]. We note that for the particular case of the multiple-bottleneck scenario, the send rate expression (3.17) of the class-0 senders coincides with an expression in [41] [5]. Let $h(\cdot)$ be a strictly-positive real-valued function, then for TCP we define $\alpha_i = h(r_i)/r_i$. Then (3.17) becomes

$$\frac{x_0}{c} = \frac{1}{v + w\sqrt{\ell}\frac{r_0}{r_1}\sqrt{\frac{h(r_1)}{h(r_0)}}}.$$

This is exactly the expression found in [41].

In Table 3.1 we show that for our multiple-bottleneck example network F_A-fair rate allocation may coincide with some other definitions of a fair rate allocation, depending on the values of the round-trip times.

3.2.5 Simulation Results

We show results of numerical simulations for the multiple-bottleneck scenario shown in Figure 3.2. The delays of the links that connect the senders and receivers to nodes are varied as defined for specific cases later. Note that the multiple-bottleneck scenario encompasses as special cases the single bottleneck case with arbitrary round-trip times, and a multiple bottleneck case; the two being considered separately in [41].

A sender i adjusts its send rate at instants $T_{i,n} = r_i n$, where r_i is a fixed round-trip time, as

$$X_{i,n+1} = X_{i,n} + \alpha_i - (\alpha_i + \beta_i X_{i,n})Z_{i,n}.$$

[4] In [61] it is referred to this case as $\lim_{c\to\infty} x_0/c$, which is encompassed in $x_i \gg \alpha_i/\beta_i$, for $\alpha_i < \infty$ and $\beta_i > 0$.

[5] The expression in [41] was obtained by an argument based on throughput formula of an AIMD sender

Table 3.1: Fraction of capacity c given to class–0 flows.

Fairness	x_0/c	parameters setup
F_A	$\frac{1}{v+w\sqrt{\ell}}$	$\alpha_0=\alpha_1$, $\alpha_0=\alpha_1$; $\alpha_0=Kr_0$, $\alpha_1=Kr_1$
Proportional	$\frac{1}{v+w\ell}$	$\alpha_0=\alpha_1$, $r_0=\ell r_1$
Max-min	$\frac{1}{v+w}$	$\alpha_0=K\ell r_0$, $\alpha_1=Kr_1$
TCP-Reno	$\frac{1}{v+w\ell\sqrt{\ell}}$	$\alpha_0=1/r_0$

(First row) F_A–like result [61] is achieved for equal round-trip times, and additive-increase proportional to a round-trip time. (Second row) The rates are distributed according to the proportional fairness for the HETRTT setting and equal values of α_i and β_i for all sources. (Third row) Max-min fairness rate distribution is obtained for indicated parameters. (Forth row) TCP Reno rate distribution for the HETRTT setting (def. Section 3.2.5), which complies to the other results mentioned in Section 3.2.5.

This is a special case of our original AIMD control that we analyzed earlier, for $\epsilon = 1$. Given that now ϵ if fixed, we omit the superscripts. Note that by setting $\epsilon = 1$ we consider the system with the original values of the additive-increase and multiplicative-decrease parameters; we do not consider the system in the limit when the parameters become asymptotically small.

We assume

$$\mathbf{P}[Z_{i,n}=1|\mathcal{F}_{i,n}] = 1-\left(1-\sum_{l\in\mathcal{L}} A_{i,l}g_l\left(\sum_{j\in\mathcal{I}} A_{j,l}X_j(T_{i,n}-D_{i,j})\right)\right)^{X_{i,n-1}r_i}. \tag{3.18}$$

The delay $D_{i,j}$ is the fixed delay from the sender i along the forward route to the receiver i, then back to the node that serves both i and j and which appears first on the forward route of the flow i, and then, finally, from that node on the backward route to the sender j. This is illustrated in Figure 3.3.

The route matrix A is for the multiple-bottleneck scenario set as, for a sender i that belongs to the class-0, $A_{i,l}=1$, all $l=1,2,\ldots,\ell$, for a sender i that belongs to the class-m, $A_{i,l}=\mathbf{1}_{l=m}$, $m=1,2,\ldots,\ell$.

Note that for small values of the summation in (3.18) (the rare-negative feedback) it holds

$$\mathbf{P}[Z_{i,n}=1|\mathcal{F}_{i,n}] \approx X_{i,n-1}r_i\sum_{l\in\mathcal{L}} A_{i,l}g_l\left(\sum_{j\in\mathcal{I}} A_{j,l}X_j(T_{i,n}-D_{i,j})\right).$$

Note that in the link cost functions in (3.18), the send rate of the sender i is precisely $X_{i,n-1}$. In [61], the authors consider a special case when all the round-trip times are equal, then in the link cost functions in (3.18), the send rate of the sender j would be precisely $X_{j,n-1}$; [61] takes as a hypothesis that

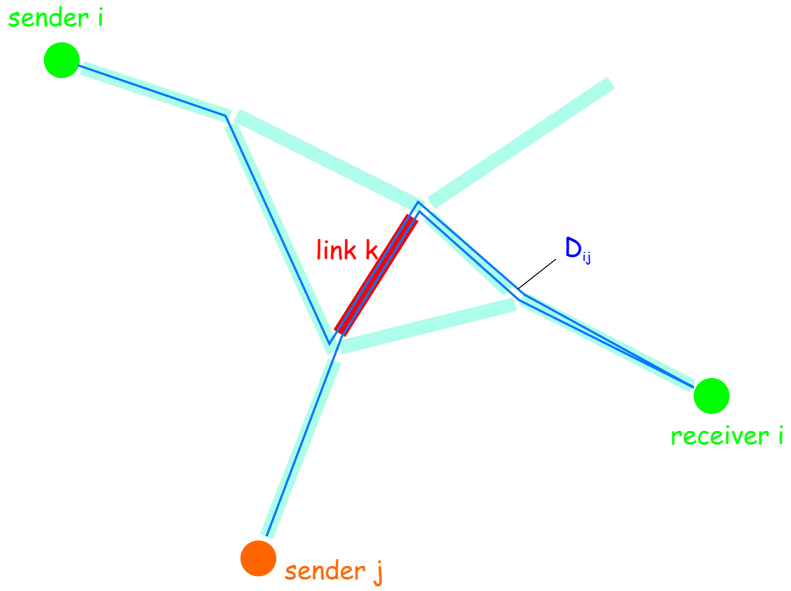


Figure 3.3: The feedback delay D_{ij}. If at time t the sender i receives a feedback that comes from the link k, the feedback was generated at the link k with the send rate of the sender j equal to $X_j(t - D_{ij})$, as perceived at the link.

those rates are equal to $X_{j,n}$. We call this "the stolen lag." We show simulation results for both systems with the stolen and non-stolen lag. The limit mean ODE for both systems, with the stolen lag or not, are the same. For non-small additive-increase and multiplicative-decrease parameters, these two systems are different.

For convenience of our computations, all the link delays, and hence the round-trip times, are set to integer multiples of a $\varepsilon > 0$, which is set to the smallest delay of a link in the network. This allows us to simulate our system over discrete points in time $\varepsilon, 2\varepsilon, 3\varepsilon, \ldots$. We assume the same link cost function as in [61]. For the link l, we define

$$g_l(x) = \begin{cases} 0, & x < 0 \\ \left(\frac{x/c-d}{1-d}\right)^p, & d \le x \le 1 \\ 1, & x > 1 \end{cases},$$

where c is the link capacity, $d \in [0, 1]$, $p > 0$. We consider all combinations of the sets of the parameters $\ell \in \{2, 5\}$, $v, w \in \{1, 2, 6, 12\}$, $c \in \{250, 625\}$, $d \in \{0, 0.5, 1\}$, and $p \in \{1, 2, 5, 10\}$. This amounts to 768 settings, in total. Total simulation time is set equal to 500 times the largest round-trip time. Each average value is computed over four simulation runs, excluding the initial 20% of a trace with the aim to eliminate the initial transient. All confidence

intervals are computed as 95% of confidence. For all our simulation results, the results obtained for $d = 1$ noticeably deviate from results of other settings; this was also observed in [61]. For this reason, we plot the results obtained for $d = 1$ in a different style.

In our simulation examples we mostly consider two specific settings of the sender and receiver access delays (resp. delay from a sender to/from its access node and delay from a receiver to/from its access node).

HOMRTT All the receiver access delays are set to zero; the sender access delays are set such that the round-trip times for all sender/receiver pairs are equal.

HETRTT All sender and receiver access delays are set to zero.

In the sequel, we show scatter plots of the send rate of class-0 flows predicted by the F_A-fairness against the respective simulation results.

In Figure 3.4 we show the empirical estimates of the time-average send rate against the F_A-fair rate. The network is HOMRTT. See the caption of Figure 3.4 for the setup of some parameters. We show the results for both the stolen lag, Figure 3.4, left part, and non-stolen lag, Figure 3.4, the right part. We observe:

- the empirical time-average send rate conforms well with the F_A-fair rate;
- there is no noticeable difference in the conformance with respect to whether or not the lag is stolen;
- in many cases, for $d = 1$, there seem to exist a bias of the F_A-fair rate to over-predict.

The results in Figure 3.5, the left part, validate conformance of the F_A-fair rates with empirical companions for the HETRTT setting. Figure 3.5, the right part, illustrates that the fair rate allocation of [61] does not compare well with empirical companions; this is to be expected given that the hypothesis that the round-trip times are equal for all senders [61] is not true in this example.

In Figure 3.6 we show results with the additive–increase of a sender set proportional to the round-trip time of that sender. This was proposed as a correction for the bias against long round-trip time connections [41]. The empirical time-average send rates for this particular setting of the additive-increase compare well with the respective F_A-fair rates.

Our next set of experiments illustrates that for some values of the round-trip times, F_A-fair rate allocation may coincide with some other notions of fairness. We fixed number of the links to $\ell = 2$. All the access delays are set to zero, except the egress link of the class-1 flows and the ingress link of the class-2 flows. See Figure 3.7 for the results. The non-zero access delays are set to be equal and varied such that the network setting gradually evolves from the HETRTT case (the leftmost point on abscissa in Figure 3.7) to a HETRTT case, where r_1 and r_2 are twice as large as r_0 (the rightmost point on abscissa in Figure 3.7). The results validate predictions of F_A-fairness.

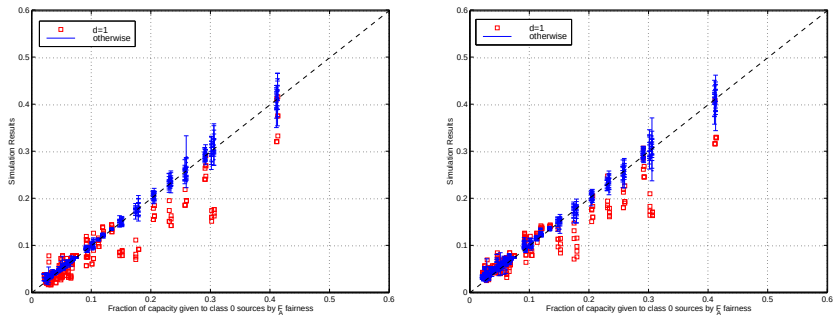


Figure 3.4: HOMRTT. Empirical time-average send rate against the F_A-fair rate (Left) with the stolen lag, (Right) with non-stolen lag. $r_i = 0.2$ s, $\alpha_i = 5$, $\beta_i = 1/2$, all flows.

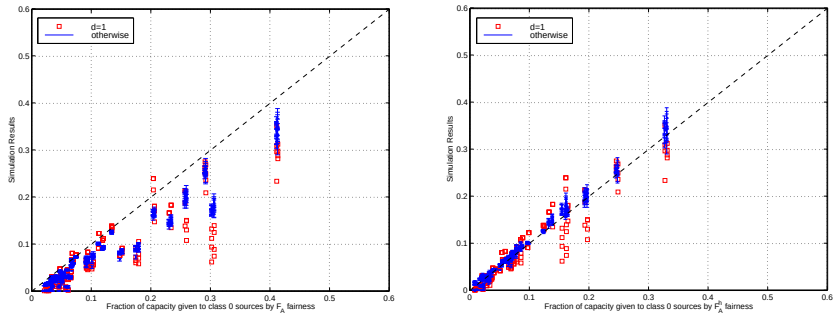


Figure 3.5: HETRTT. Empirical time-average send rate against (Left) the F_A-fair rate of [61] obtained under assumption that round-trip times are the same for all senders, (Right) F_A-fair rate. $r_0 = 0.2$ s, $r_i = r_0/\ell$, all $i = 1, 2, \ldots, \ell$, and $\alpha_i = 5$, $\beta_i = 1/2$, all flows.

Lastly, we compare the empirical estimates of the throughput of a sender that runs an AIMD recurrence as defined above, against the ODE prediction. See Figure 3.8. The results indicate that the ODE method over-predicts. Remember that the ODE method is an "equilibrium analysis," it is based on fixed-point approximations. The discrepancy of the observed values should be attributed to the stochastic bias.

3.2.6 Bias Against Long Round-Trip Time Connections

Our finding enables us to understand better the bias of TCP against connections with long round-trip times.

First, for any congestion control mechanism, if the distribution of rates tends to maximize a concave utility function, then flows with many hops are likely

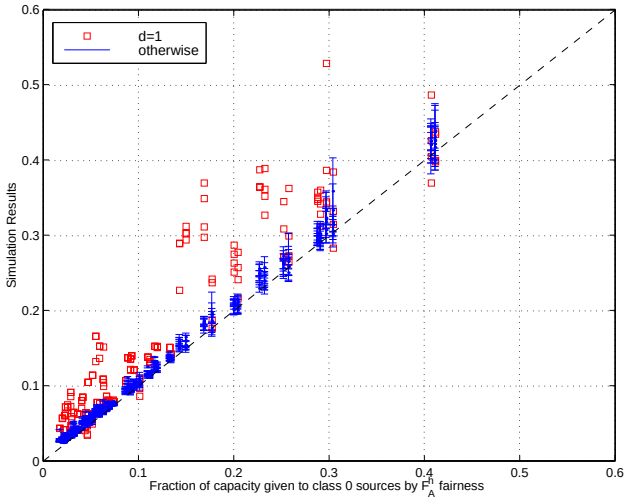


Figure 3.6: HETRTT. A correction for the bias against long-round trip time connections—the additive-increase parameter of a sender is set proportional to this sender's round-trip time—$\alpha_i = Kr_i$, $K = 25$. Other parameters are set as $\beta_i = 1/2$, $r_0 = 0.2$ s, $r_i = r_0/\ell$, $i = 1, 2, \ldots, \ell$.

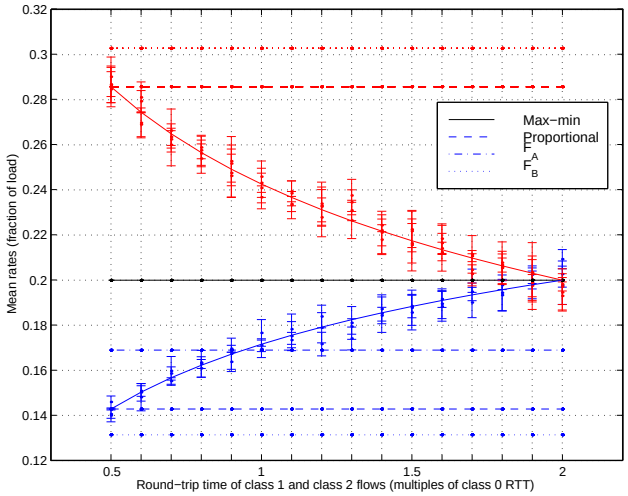


Figure 3.7: F_A-rate allocation obtained for some values of round-trip times may coincide with the rate allocation of some other notions of fairness. $c = 250$, $\ell = 2$, $v = 3$, $w = 2$, $r_0 = 0.2$ s, $r_i = 0.1 - 0.4$ s.

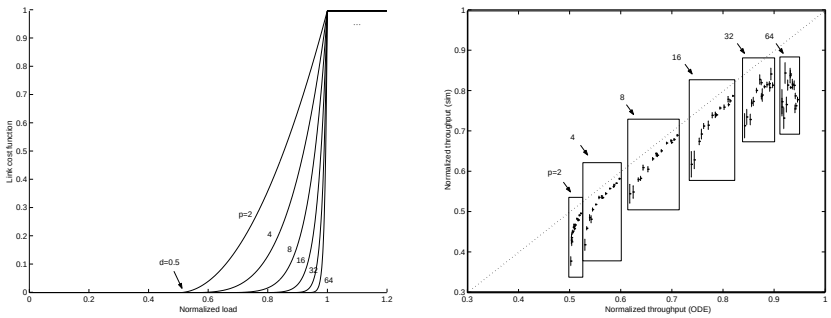


Figure 3.8: (Left) The link cost functions against the load of a link. (Right) Scatter plot of the throughput normalized by the link rate obtained as the solution of the ODE against the discrete-time recurrence; different boxes correspond to different link cost functions; within a box the results are given for varying number of competing connections (mostly, the more right a point is, the larger the number of connections). The example is for the multiple-bottleneck network with $\ell = 2$, $v, w = \{1, 2, 6, 12\}$, $\alpha_0 = \epsilon/r_0$, $\alpha_1 = 1/r_1$, $r_0 = 0.2$ s, $r_1 = 0.1$ s, $\beta_0 = \beta_1 = 1/2$.

to receive a small rate [68]. Since F_A is concave with x_i, this is true with our system, whatever the rate adaptation parameters α_i and β_i are. This is probably a desired bias, since flows with many hops use more network resources. In practice, many hops often mean larger round-trip times, but not always.

Second, both the specific values of the rate adaptation parameters, α_i and β_i, and the update frequency $1/r_i$ also play a role; recall, r_i is the round-trip time for sender i. With TCP–NewReno, with no delayed acknowledgments, we have:

- $\alpha_i = m_i/r_i$, where m_i is packet length of the ith sender; in the congestion avoidance phase, with no delayed acknowledgments, the window is effectively increased by one packet per round-trip time (resp. 1/2 packets for delayed acknowledgments);
- $\beta_i = 1/2$, when a loss is detected, the target window size is divided by a factor 2.

This results in an obvious bias against TCP connections with long round-trip times; the increase element α_i is smaller, and less frequent for long round-trip times.

3.2.7 Related Work

The mechanism to identify the limit-mean ODE as we described in Section 3.2.2 can be applied to identify the limit mean ODEs for some other variants of AIMD. One variant is defined by $a_i(x) = \alpha_i$ and $b_i(x) = \beta_i x$. This means that for all the

rate updates, the send rate is increased, and the decrease is only at a loss-event. Again, assume (3.7).

Proposition 8 *The function $F(\cdot)$ in (3.6) is*

$$F_B(x) = \sum_{i \in \mathcal{I}} -\frac{\alpha_i}{r_i \beta_i} \frac{1}{x_i}.$$

This result was obtained by Kunniyur and Srinkant [75]. Another AIMD variant differs from the original AIMD algorithm that we considered only in the definition of the instants at which the send rate is updated. Assume

$$\bar{s}_i(x) = \frac{1}{x_i}, \ i = 1, 2, \ldots, |\mathcal{I}|.$$

This would model a sender that receives acknowledgments per each packet transmitted and adjust its send rate according to AIMD at those instants.

Proposition 9 *The function $F(\cdot)$ in (3.6) is*

$$F_C(x) = \sum_{i \in \mathcal{I}} \frac{1}{\sqrt{\beta_i} r_i} \arctan(\sqrt{\beta_i} r_i x_i).$$

To our knowledge, this result was first obtained by Kelly [67].

The system of limit-mean ODEs can be identified for an extended system where queues at the links are also taken as the state variables (see [107] for a demonstration). It turns out that another approach by Misra, Gong, and Towsley [88] that derives *stochastic* differential equations for the evolution of the *expected* send rate over time amounts to solving almost the same system as obtained by the ODE method; this is an *artifact* of some fixed-point approximations in [88].

3.3 Increase-Decrease Controls

In this section we proceed with the *analysis* and *synthesis* problem for the increase-decrease controls, the two problems introduced at the beginning of this chapter. In contrast to the last section, we now concentrate on *one* sender and are concerned with the throughput of this sender and its relation to the loss-event rate as observed by the sender. The limit-mean ODE method is *not* appropriate for solving the problems of this section; we need more refined tools. Remember that the limit-mean ODE is an "equilibrium-analysis." The send rate obtained by the ODE method approximately attains the zero conditional expected drift of the send rate. The ODE method is exact only asymptotically, when the increments and decrements of the rate adjustment become asymptotically small. We may think of the ODE method as of a *small-gain limit* or a *fixed-point* analysis. In reality, the increments and decrements of the increase-decrease controls are *not* small.

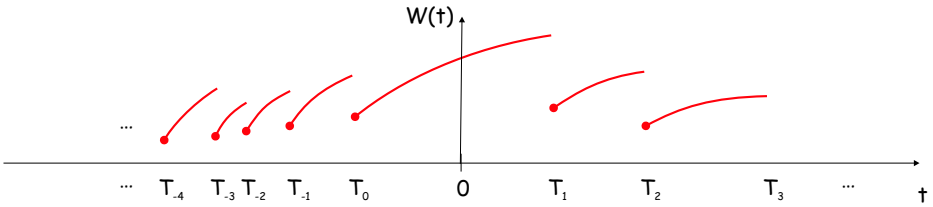


Figure 3.9: A sample-path of the window evolution of an increase-decrease control.

3.3.1 Notation and Assumptions

Consider a sender that exercises a window control. Let $W(t)$ be the window value at an instant $t \in \mathbb{R}$. Assume the window takes values on $\mathbb{R}_+$. Let T_n be an instant when a loss-event labeled with n is detected by the sender. Assume a standard convention, $\cdots < T_{-1} < T_0 \leq 0 < T_1 < T_2 < \cdots$. Let S_n be the time interval between the nth and $n+1$th loss-event, that is $S_n = T_{n+1} - T_n$, all $n \in \mathbb{Z}$. We assume $W(t)$ is a right-continuous function of time t, with left-hand limits. This implies $W(T_n)$ is the window just after the nth loss-event. Let $N(s,t]$ be the number of loss-events that fall in an interval $(s,t]$.

Let $a : \mathbb{R}_+ \to \mathbb{R}_+$ and $b : \mathbb{R}_+ \to \mathbb{R}$ be some positive-valued functions. Assume $b(x) < x$, all $x \in \mathbb{R}_+$.

Definition 1 *We say a window control is* increase-decrease, *if for* $W(0) \geq 0$,

$$W(t) = W(0) + \int_0^t a(W(s))ds - \int_0^t [W(s-) - b(W(s-))]N(ds),\ t \geq 0. \quad (3.19)$$

In other words, for a t in an interval of time that contains no loss-events, the window is increased with the rate of the increase $a(W(t))$. At an instant T_n, the window is decreased to $b(W(T_n-))$. See Figure 3.9 for an example sample-path of the window process. We call $a(\cdot)$ an "increase" function, and $b(\cdot)$ a "decrease" function.

An AIMD window control is a special case, for $a(x) = \alpha$, $b(x) = \beta x$, $x \in \mathbb{R}_+$, $\alpha > 0$, $0 < \beta < 1$. We use the notation AIMD(α, β), for an AIMD window control with parameters α and β. We use additional notation

$$\theta_n = \int_{T_n}^{T_{n+1}} W(s)ds,\ n \in \mathbb{Z}.$$

We call θ_n, a loss-event interval. The loss-event rate is $p = 1/\mathbf{E}_N^0[\theta_0]$.

We assume the system (3.19) is stable. We assume 0 is an arbitrary point in time and $W(t)$, $t \in \mathbb{R}$, is a stationary ergodic process. This allows us to write

$$\bar{w} = \lim_{t\to\infty} \frac{1}{t} \int_0^t W(s)ds = \mathbf{E}[W(0)].$$

That is, we can equate the time-average window with the expected steady-state window.

3.3.2 Why do we Study Time-Average Window and Not Throughput?

Our analysis is concerned with the time-average window $\bar{w}$ and its relation to the loss-event rate p. One may legitimately argue that it is the throughput $\bar{x}$, i.e. the time-average send rate, that is in practice of importance. We consider time-average window for technical convenience as explained next. Consider an increase-decrease window control with the window $\tilde{W}(t)$ at the instant $t \in \mathbb{R}$. Suppose $\cdots < T'_{-1} < T'_0 \leq 0 < T'_1 < \cdots$ is a stationary ergodic point process on $\mathbb{R}$, such that $\tilde{W}(s) \leq \tilde{W}(t)$, for all $T_n \leq s \leq t < T_{n+1}$, $n \in \mathbb{Z}$. In other words, $\tilde{W}(t)$, $t \in \mathbb{R}$, is non-decreasing on the epochs of the point process. Let $W(t')$ denote transformation of $\tilde{W}(t)$ in *virtual-time* $t' \in \mathbb{R}$, defined as

$$W(t') = \tilde{W}\left(T_{\lfloor t' \rfloor} + (T_{\lceil t' \rceil} - T_{\lfloor t' \rfloor})(t' - \lfloor t' \rfloor)\right), \ t' \in \mathbb{R}.$$

Here $\lfloor x \rfloor$, $x \in \mathbb{R}$ is the largest integer not larger than x, and $\lceil x \rceil$ is the smallest integer not smaller than x. Note that the transformation, in particular, maps $\tilde{W}(T_n)$ to $W(n)$. Now, by Palm inversion, and the monotonicity assumption on the window $\tilde{W}(t)$, $t \in \mathbb{R}$, we have, for s', an arbitrary point in $\mathbb{R}$,

$$\mathbf{E}[W(s')] = \mathbf{E}[\int_0^1 W(s)ds] \geq \mathbf{E}[W(0)]. \tag{3.20}$$

The strict equality holds if $\tilde{W}(t)$, $t \in \mathbb{R}$, is piece-wise constant over the epochs of the point process.

Now, let the points T'_n and T'_{n+1} mark the n-th round-trip time round; $R_n := T_{n+1} - T_n$, $n \in \mathbb{Z}$, is the nth round-trip time. It is standard to assume that the number of the data units sent on the interval $[T'_n, T'_{n+1})$ is equal to $\tilde{W}(T'_n)$. Under this assumption, by Palm inversion, we have

$$\bar{x} = \frac{\mathbf{E}[W(0)]}{\mathbf{E}[R_0]}.$$

Hence, in view of the last equality, in order to evaluate the throughput, we need to consider the expected value of the discrete-time sequence $W(n)$, $n \in \mathbb{Z}$. In our study, we evaluate $\mathbf{E}[W(s')]$, where s' is an arbitrary point in $\mathbb{R}$, for the sake of technical convenience to work in continuous-time under the continuity assumption introduced in the earlier section. From (3.20), we know precisely that $\bar{x} \leq \mathbf{E}[W(s')]/\mathbf{E}[R_0]$. Note that the window process defined in the preceding section, and the results in the remainder of this chapter are for the window process in the *virtual time*.

3.3.3 Analysis and Synthesis

We define the analysis problem as: Given some increase and decrease functions, for some process of loss-events, find $x \to f(x)$ such that $\bar{w} = f(p)$.

The analysis problem has been an object of study by many (see [93, 4] and the references therein). Their particular interest was on the loss-throughput relation of an AIMD control, which has been largely motivated to derive TCP loss-throughput formula.

The synthesis problem we define as: The objective is $\bar{w} \leq f(p)$, for a given non-increasing function $x \to f(x)$. Find an increase and a decrease function such that the objective is achieved for a set of processes of the loss-events. Note that ideally we would like to have $\bar{w} = f(p)$, but this would not hold for all loss processes that induce the loss-event rate p. As a relaxation, we aim to design *conservative* controls, such that $\bar{w}$ is not larger than $f(p)$.

The synthesis problem may be seen as an *inverse problem*[6] to the analysis problem. (It is exactly an inverse problem for strict equality in the synthesis problem, and a fixed process of the loss-events for both problems.)

In the next section we show a result related to the analysis problem. In the remainder of the chapter we consider the synthesis problem.

3.4 Analysis: Two-Sided Bounds for AIMD

This is the main result of this section.

Theorem 5 *Consider an AIMD(α, β) window control, $0 < a < \infty$, $0 < \beta < 1$. Assume $\{\theta_n\}_n$ is a stationary ergodic sequence of loss-event intervals with a mean $0 < 1/p < \infty$ and $\mathbf{P}_N^0[\theta_0 > 0] = 1$. We have*

$$\sqrt{\frac{\alpha}{2}\frac{1+\beta}{1-\beta}}\frac{1}{\sqrt{p}} \leq \bar{w} \leq \sqrt{\frac{\alpha}{2}\frac{1+\beta}{1-\beta}}\frac{1-\beta}{\frac{1}{\sqrt{1+\mathbf{cv}_N^0[\sqrt{\theta_0}]^2}}-\beta}\frac{1}{\sqrt{p}}. \tag{3.21}$$

The upper bound is finite iff $\mathbf{cv}_N^0[\sqrt{\theta_0}]^2 < \frac{1}{\beta^2} - 1$.

The upper-bound is obtained from an exact expression showed in the proof of the theorem,

$$\bar{w} = \sqrt{\frac{\alpha}{2}}\frac{1}{\mathbf{E}_N^0[\sqrt{\sum_{k=0}^{\infty}\beta^{2k}\theta_{-k}}] - \mathbf{E}_N^0[\sqrt{\sum_{k=1}^{\infty}\beta^{2k}\theta_{-k}}]}\frac{1}{p}. \tag{3.22}$$

Comments. The lower-bound is exactly the well-known "square-root" formula that would be obtained under the assumption that the inter-loss event times are fixed to a *constant*. We prove the lower-bound based on a suggestion by Barakat [13]. We had independently arrived to the same lower-bound

[6] Two problems are said to be inverses of one another if the formulation of each involves all or part of the solution of the other [66].

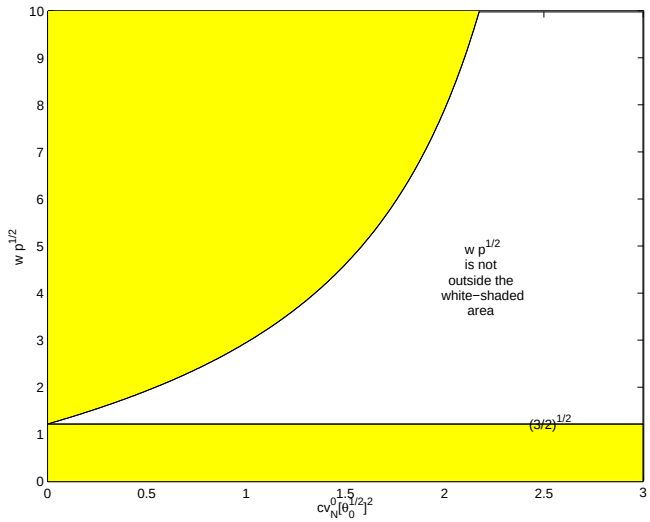


Figure 3.10: The constant of the square-root formula is not outside the blank area. The abscissa is the squared coefficient of variation of $\sqrt{\theta_0}$. The plot is for AIMD with $\alpha = 1$ and $\beta = 1/2$.

by using an *adversarial* argument; see Section 3.7.3. The upper-bound is a *new* result. The upper bound holds for any stationary ergodic loss-event intervals, broadly speaking, with bounded "variability." Strictly speaking, it holds for any stationary ergodic processes of loss-event intervals for which it holds $\mathbf{cv}_N^0[\sqrt{\theta_0}]^2 < \frac{1}{\beta^2} - 1$.

Note that our bound depends only on two statistical parameters of the loss process: The loss-event rate $p = 1/\mathbf{E}_N^0[\theta_0]$ and the coefficient of the variation $\mathbf{cv}_N^0[\sqrt{\theta_0}]$. Both parameters are of the *marginal* distribution of the loss-event intervals. The bound is increasing in "variability" of the loss-event interval, $\mathbf{cv}_N^0[\sqrt{\theta_0}]$. It is somewhat surprising that the bound on the throughput is *solely* in terms of the parameters of the *marginal* distribution of the loss-event intervals.

The result adds to our knowledge about the value of the constant in the square-root formula. It tells us that $\sqrt{\frac{\alpha}{2}\frac{1+\beta}{1-\beta}}$, which would be the value of the constant in the square-root formula, for a sequence of constant inter loss-event times, is never exceeded by more than the factor

$$\frac{1-\beta}{\frac{1}{\sqrt{1+\mathbf{cv}_N^0[\sqrt{\theta_0}]^2}} - \beta}.$$

See Figure 3.10 for a plot of the upper-bound (3.21) for TCP-like AIMD$(1, 1/2)$.

Contrasting with Exact Formula of [4]. Our result is an upper bound. An exact expression was obtained by Altman, Avrachenkov, and Barakat [4],

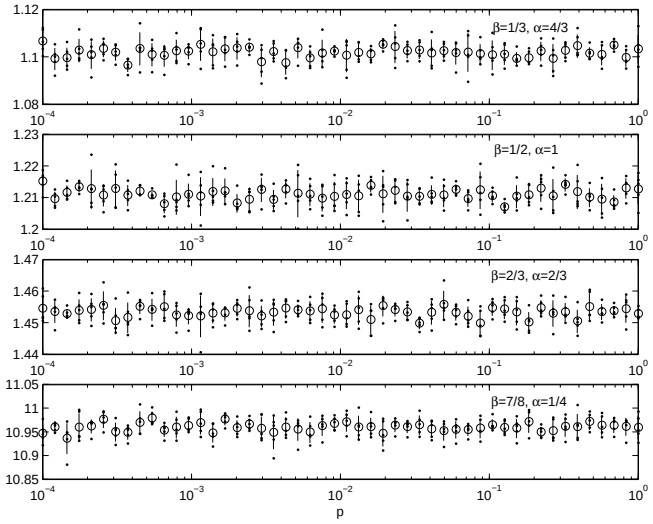


Figure 3.11: The ratio of the upper-bound on time-average window (3.21), and the time-average window $\bar{w}$. The less responsive the control is, the more conservative the bound is.

which we re-write as follows

$$\bar{w} = \left(\frac{1}{2} \frac{1+\beta}{1-\beta} + \frac{1}{2} \mathbf{var}_N^0[S_0]\lambda^2 + \sum_{k=1}^{\infty} \beta^k \mathbf{cov}_N^0[S_0, S_k]\lambda^2 \right)^{\frac{1}{2}} \frac{\sqrt{\alpha}}{\sqrt{p}}. \tag{3.23}$$

This result gives us the exact value of the constant in the square-root formula that depends on the variance and autocovariance of the inter loss-event times. We now compare the hypotheses under which (3.23) and (3.21) are obtained. The hypothesis in obtaining (3.23) is that $\{S_n\}_n$ is a stationary point process with $0 < \mathbf{E}_N^0[S_0] < \infty$. We assume $\{\theta_n\}_n$ is a stationary ergodic process with $0 < \mathbf{E}_N^0[\theta_0] < \infty$. Note that $\mathbf{E}_N^0[\theta_0] < \infty$ implies $\mathbf{E}_N^0[S_0] < \infty$. The converse is not true. The direct implication follows from (3.39),

$$\mathbf{E}_N^0[S_0] \leq \frac{1}{\alpha} \mathbf{E}_N^0 \sqrt{\mathbf{E}_N^0[W_0^2] + 2\alpha \mathbf{E}_N^0[\theta_0]} = \sqrt{\frac{2\beta^2}{\alpha(1-\beta^2)}} \sqrt{\mathbf{E}_N^0[\theta_0]} < \infty.$$

The last equality follows from (3.40). Note from (3.39) that $\{\theta_0 > 0\} = \{S_0 > 0\}$. Hence, the condition $\mathbf{P}[\theta_0 > 0] = 1$ means that $\{T_n\}_n$ is a *simple* point process. From (3.39), we note that $\mathbf{P}_N^0[\theta_0 > 0] = 1$ implies $\mathbf{E}_N^0[S_0] > 0$. The alert reader would note that the conditions of (3.21) are stronger than that of (3.23). However, in view that our interest is the time-average window equal to $\mathbf{E}_N^0[\theta_0]/\mathbf{E}_N^0[S_0]$ this is not restrictive.

Finally, we give two examples in order to demonstrate how the upper-bound compares with the exact value.

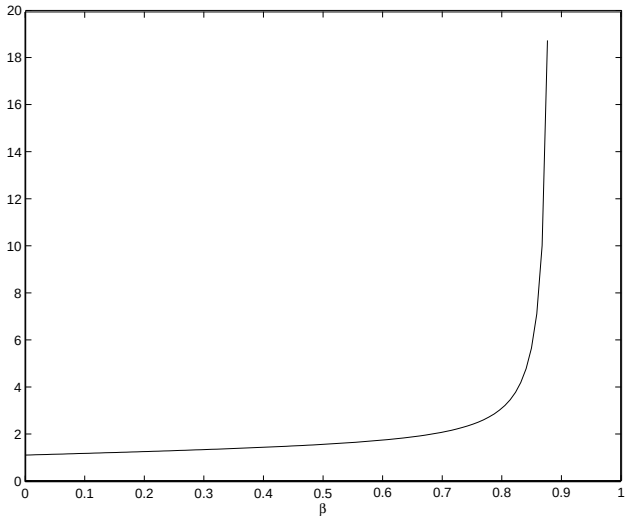


Figure 3.12: The upper bound of Theorem 5 versus β with $\alpha = 2(1-\beta)$ and $\mathbf{cv}_N^0[\sqrt{\theta_0}] = 4/\pi - 1$.

Example 1 *Consider a sequence of the loss-event intervals, a stationary renewal process with exponential $(1/p)$ marginal distribution. The example is a continuous analog of discrete memoryless (Bernoulli) losses. In this case we have*

$$\mathbf{cv}_N^0[\sqrt{\theta_0}]^2 = \frac{4}{\pi} - 1 \simeq 0.2732.$$

We ran our experiments for different values of β and α. In Figure 3.11, we show the ratios of our upper bound and the estimate of $\bar{w}$. We observe:

- *the less responsive (larger β, smaller α) an AIMD control is, the more conservative the bound is.*

We further support the last observation with a plot of our bound with respect to the parameter β, for $\alpha = 2(1-\beta)$; see Figure 3.12.

We should bear in mind that the last example is only for independent loss-event intervals with an exponential marginal distribution. If the loss-event intervals would be non-independent, with the same marginal distribution, then our bound may be less pessimistic.

The next example is more realistic; we re-use some measurement data that we obtained by lab experiments with RED queue discipline in Chapter 2, see also Appendix 2.6.3.

Example 2 (TCP lab experiments) *In the limit case of rare losses, TCP congestion window control would be mostly by TCP congestion avoidance, which uses an AIMD control with $\alpha = 1/2$ and $\beta = 1/2$.*

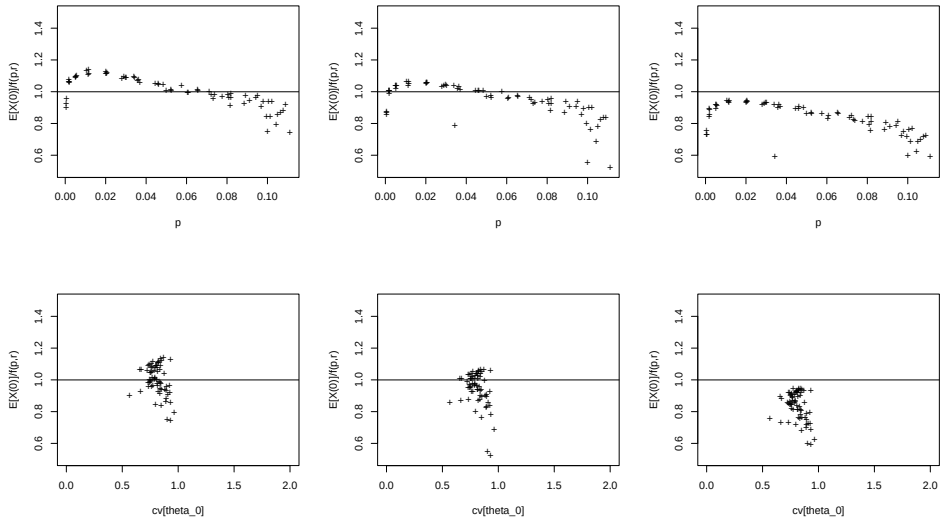


Figure 3.13: Lab experiments with RED. The plots show TCP throughput divided with $f(\hat{p}, \hat{r})$. $\hat{p}$ and $\hat{r}$ are the estimates of p and r, respectively. Function f is: (Left) The lower-bound of Theorem 5, (Middle) (3.23) [4], (Right) The upper-bound of Theorem 5.

In this example, we want to see how TCP throughput compares with a prediction obtained by our upper–bound in Theorem 5. We also compare with the lower-bound in the same theorem, as well as with the throughput formula in [4]. We are interested in two specific questions: (1) "Is our upper-bound an upper-bound in practice?", (2) "How close is it to TCP throughput?". From Figure 3.13, we observe:

- *TCP throughput is smaller than our upper-bound in all the cases;*
- *in some cases, our upper-bound is fairly close to TCP throughput. The upper-bound follows qualitatively the predictions of the other two formulas.*

3.5 Synthesis

We recall the definition of the synthesis problem introduced in Section 3.3. We define the synthesis problem as: The objective is $\bar{w} \leq f(p)$, for a given non-increasing function $x \rightarrow f(x)$; find an increase and a decrease function such that the objective is achieved for a set of loss-event processes.

First, we introduce some new notation. Let W_n be the value of the window just after the nth loss-event, that is $W_n = W(T_n)$, $n \in \mathbb{Z}$. From (3.19),

$$W(t) = A^{-1}(A(W_n) + t),\ t \in [T_n, T_{n+1}),\ n \in \mathbb{Z}.$$

Here A is a primitive of $1/a$. That is, for a constant $C \in \mathbb{R}$,

$$\int \frac{dx}{a(x)} = A(x) + C,$$

With A^{-1} we denote the inverse of A. The window embedded just after a loss-event evolves as follows

$$W_{n+1} = b(A^{-1}(A(W_n) + S_n)),\ n \in \mathbb{Z}. \tag{3.24}$$

This is a stochastic recursive sequence, in particular, an instance of generalized autoregression [18].

We next show two design rules to choose an increase and a decrease function, and give conditions under which such chosen increase and decrease functions solve the synthesis problem.

3.5.1 First Design Rule

Our first design rule is suggested by the following lemma.

Lemma 2 *$\bar{w} \le f(p)$ is equivalent to saying*

$$b(\bar{w}) \le A^{-1}(A(\bar{w}) - \frac{\chi}{\bar{w} f^{-1}(\bar{w})}),$$

where

$$\chi = \frac{\varphi(\bar{w})}{\mathbf{E}_N^0[\varphi(W(0-))]}. \tag{3.25}$$

By definition, $\varphi(x) = A(x) - A(b(x))$.

Note that χ is dimensionless; we discuss it shortly. First, we give an immediate implication from Lemma 2.

Lemma 3 *If there exists a $\Delta > 0$ such that $\chi \le \Delta$, and $a(\cdot)$ and $b(\cdot)$ are such that*

$$b(x) \le A^{-1}(A(x) - \frac{\Delta}{x f^{-1}(x)}),\ \text{all } x \ge 0, \tag{3.26}$$

then $\bar{w} \le f(p)$.

The last lemma gives us a "design rule." The non-trivial part of the design rule is to find a bound on χ, which we discuss next.

By definition of the stochastic intensity $\lambda(\cdot)$ (see [9]) for the loss-events, we can write

$$\chi = \frac{\lambda \varphi(\bar{w})}{\mathbf{E}[\varphi(W(0-))\lambda(0)]}.$$

Think of $\lambda(t)$ as of density that there would happen a loss-event at t, given the history of the system before t. By the definition of covariance, we can further write

$$\frac{1}{\chi} = \frac{\mathbf{E}[\varphi(W(0))]}{\varphi(\bar{w})} + \frac{\mathbf{cov}[\varphi(W(0-)), \lambda(0)]}{\mathbf{E}[\varphi(W(0-))]\lambda}. \tag{3.27}$$

The first element on the right-hand side would deviate from 1 based purely on the convexity nature of the function $x \to \varphi(x)$. Before providing some sufficient conditions for $\chi \leq 1$ to hold, we go through a few examples.

Constant inter loss-event times Assume inter loss-event times are fixed to a constant. Then,

$$\chi = \frac{\varphi(\mathbf{E}[W(0)])}{\varphi(\mathbf{E}_N^0[W(0-)])}.$$

In this case, we know $\mathbf{E}_N^0[W(0-)] \geq \mathbf{E}[W(0)]$. Hence, $\chi \leq 1$, iff $x \to \varphi(x)$ is non-decreasing, a *monotonicity* property.

Poisson loss-events in real-time Assume the point process of the loss-events is a homogeneous Poisson process. Then, $\mathbf{cov}[\varphi(W(0-)), \lambda(0)] = 0$. In this case,

$$\chi = \frac{\varphi(\mathbf{E}[W(0)])}{\mathbf{E}[\varphi(W(0))]}.$$

This is a consequence of PASTA.[7] In this case, it is the *convexity* nature of $x \to \varphi(x)$ that entirely determines whether $\chi \leq 1$, or not.

Poisson loss-events in transmission-time Assume the point process of the loss-events is a homogeneous Poisson process in transmission-time[8] with intensity p. Now, the stochastic intensity is $\lambda(t) = W(t-)p$. Hence, it holds

$$\frac{1}{\chi} = \frac{\mathbf{E}[\varphi(W(0))]}{\varphi(\bar{w})} + \frac{\mathbf{cov}[\varphi(W(0-)), W(0-)]}{\mathbf{E}[\varphi(W(0-))]\bar{w}}.$$

In this case, both *convexity and monotonicity* of $x \to \varphi(x)$ play a role. For $x \to \varphi(x)$, non-decreasing convex, it indeed holds $\chi \leq 1$. A special case, for which $x \to \varphi(x)$ is non-decreasing convex, is AIMD. Then,

$$\frac{1}{\chi} = 1 + \mathbf{cv}[W(0)]^2.$$

Indeed, $\chi \leq 1$. The less variable the steady-state window is, the closer χ is to 1.

Observation 2 *It is the convex nature of the function $x \to \varphi(x)$ that would in part determine the value of χ.*

[7] Poisson Arrivals See Time Averages, see [9].

[8] We call "transmission-time," a virtual-time that is clocked by the units of data sent by the sender.

We continue with giving sufficient conditions for $\chi \leq 1$ to hold. From (3.27), we directly conclude

Theorem 6 *If (F) $x \to \varphi(x)$ is convex and (C)* $\mathbf{cov}[\varphi(W(0-)), \lambda(0)] \geq 0$*, then* $\chi \leq 1$.

The theorem gives us a conjunction of two conditions that are sufficient for $\chi \leq 1$. Note that if the sufficient conditions hold, then using the design rule (3.26) with $\Delta = 1$, we know $\bar{w} \leq f(p)$.

When (C) Holds. We consider a few particular processes of loss-events. Our focus is on the sign of

$$\mathbf{cov}[\varphi(W(0-)), \lambda(0)].$$

Let $\ell(x) := \mathbf{E}[\lambda(0)|W(0-) = x]$. In other words, $\ell(x)$ is the intensity that a loss-event would happen at an instant 0 given that we know the window at $0-$ is x. Now, we have

$$\mathbf{cov}[\varphi(W(0-)), \lambda(0)] = \mathbf{cov}[\varphi(W(0)), \ell(W(0))].$$

Proposition 10 *If both $\varphi(\cdot)$ and $\ell(\cdot)$ are either non-decreasing or non-increasing, then* $\mathbf{cov}[\varphi(W(0-)), \lambda(0)] \geq 0$*. Then,*

$$\chi \leq \frac{\varphi(\mathbf{E}[W(0)])}{\mathbf{E}[\varphi(W(0))]}.$$

What This Tells Us. We expect $\ell(\cdot)$ to be non-decreasing in many cases, but not always. An example is a network where loss-events are driven by a hidden process that evolves slowly over time, in other words, when the intensity of the loss-events is modulated by a hidden process that evolves on a larger timescale than the increase-decrease control. Then, knowing that at an arbitrary point in time the window is large, it would be likely that the intensity of loss-events is small, and *vice versa*. In situations, where $\ell(\cdot)$ is non-decreasing, the proposition suggests choosing the functions $a(\cdot)$ and $b(\cdot)$ such that $\varphi(\cdot)$ is a non-decreasing convex function. Then, using the design rule (3.26) with $\Delta = 1$, we have $\bar{w} \leq f(p)$.

3.5.2 Second Design Rule

Lemma 4 *$\bar{w} \leq f(p)$ is equivalent to saying*

$$a(\bar{w}) \leq \bar{w} f^{-1}(\bar{w})(\bar{w} - b(\bar{w}))\chi',$$

where

$$\chi' = \frac{a(\bar{w})}{\mathbf{E}[a(W(0))]} \frac{\mathbf{E}_N^0[W(0-) - b(W(0-))]}{\bar{w} - b(\bar{w})}. \tag{3.28}$$

The lemma suggests the following design rule.

Lemma 5 *Assume there exists a* $\Delta > 0$ *such that* $\chi' \geq \Delta$. *If* $a(\cdot)$ *and* $b(\cdot)$ *are such that*

$$a(x) \leq xf^{-1}(x)(x - b(x))\Delta, \ x \geq 0, \tag{3.29}$$

then $\bar{w} \leq f(p)$.

Comments. This design rule has been used in some past work. For instance, binomial controls found in [10] were derived for strict equality in (3.29) and $\Delta = 1$. However, there seems to be no justification that $\chi' \geq 1$ would always hold. This design rule is intimately related to the "small-gain limit," a limiting regime used in Section 3.2 of this chapter. Then $\chi' \approx 1$. This limit case entirely leaves aside the stochastic bias due to difference of the stationary and Palm distributions of the window, and non-linearity of the increase and decrease functions.

Interpretation of χ'. The value of the first rational in (3.28) depends on the convexity of the increase function, and the distribution of the steady-state window. The second rational in (3.28) is more intricate. It can be interpreted as the ratio of the expected value of the window decrease at a loss-event and a virtual window decrease computed at the expected steady-state window. In a particular case, if $x \to x - b(x)$ is non-decreasing, that is $b'(x) \leq 1$, then the second rational in (3.28) is greater than or equal to 1, iff $\mathbf{E}_N^0[W(0-)] \geq \bar{w}$. The last is indeed true for constant inter loss-event times. It is also true with strict equality if loss-events are a homogeneous Poisson process in real-time with intensity λ (PASTA). In general, $\mathbf{E}_N^0[W(0-)] \geq \bar{w}$ may not hold.

When χ' **is Not Larger than One**. For the design rule of the present section, a problem is to find Δ, a lower bound on χ'. The next result gives us sufficient conditions under which we can use $\Delta = 1$.

Proposition 11 *If*

(A) $x \to a(x)$ *is concave,*

(B) $x \to b(x)$ *is concave,*

(C) $\mathbf{cov}[W(0-) - b(W(0-)), \lambda(0)] \geq 0$,

then, $\chi' \geq 1$.

What This Tells Us. The conditions (A) and (B) are on the convexity of the increase and decrease functions. The condition (C) is on a covariance of the window and the stochastic intensity of loss-events. If $x \to x - b(x)$ is non-decreasing, then (C) holds as long as $\ell(x) := \mathbf{E}[\lambda(0)|W(0-) = x]$ is a non-decreasing function. This is true in some cases. For instance, it is true for loss-events, a homogeneous Poisson process in either real- or transmission-time; see the examples in Section 3.5.1. It may not be true for the intensity of the loss-events modulated over larger timescale by a hidden modulator; the same type of arguments as in Section 3.5.1.

3.5.3 Synthesis for a Reference Loss Process

In Section 3.5.1 and Section 3.5.2 we showed two design rules for choosing the increase and decrease functions, such that under some conditions, a design rule solves the synthesis problem.

Another design rule can be deduced from some past work, for instance, Floyd *et al.* [42, 44]. It consists in choosing the increase and decrease functions such that, for a given target function $f(\cdot)$, the resulting increase-decrease control satisfies $\bar{w}' = f(p')$, where $\bar{w}'$ and p' are the time-average window and loss-event rate for a reference process of loss-events. Typically, the reference is taken to be a sequence of *constant* inter-loss event times

$$\mathcal{S}'_\lambda = \left(\ldots, \frac{1}{\lambda}, \frac{1}{\lambda}, \frac{1}{\lambda}, \ldots \right). \tag{3.30}$$

To our knowledge, this particular reference had been chosen primarily for tractability and simplicity of computations.

The question is: *If an increase-decrease control solves the synthesis problem for the reference loss process, does it solve the synthesis problem for some other loss process?* In general, the answer is: *no.*

A check whether an increase-decrease control, designed to solve the synthesis problem for the reference loss process $\mathcal{S}'_\lambda$, would solve the synthesis problem for some other loss process $\mathcal{S}_\lambda$, is to compute $\bar{w}$ and check whether $\bar{w} \leq f(p)$ holds. The problem is that, in general, it may not be possible to obtain an exact expression for $\bar{w}$ in terms of some statistics of the loss process $\mathcal{S}_\lambda$. Fortunately, in some situations it is possible to check whether $\bar{w} \geq f(p)$, without explicitly computing $\bar{w}$.

In this section we give *converse* results to the condition of the synthesis problem. We assume an increase-decrease control is designed such that the synthesis problem is solved with $\bar{w}' = f(p')$, for the reference loss process $\mathcal{S}'_\lambda$. We then show conditions on the increase and decrease functions for which the resulting increase-decrease control verifies $\bar{w} > f(p)$, for a set of loss processes $\mathcal{S}_\lambda$. In other words, we give conditions under which an increase-decrease control designed to solve the synthesis problem for the reference loss process, does not solve the synthesis problem for a set of loss processes.

We assume the reference loss process $\mathcal{S}'_\lambda$ is defined by (3.30), a sequence of constant inter loss-event times.

Analysis Problem. Assume $\mathcal{S}'_\lambda$. Given some functions $a(\cdot)$ and $b(\cdot)$, find the function $f(\cdot)$, such that $\bar{w} = f(p)$. The problem is to solve

$$\begin{aligned} w_0 &= b(A^{-1}(A(w_0) + 1/\lambda)) \\ \frac{1}{p} &= \int_0^{1/\lambda} A^{-1}(A(w_0) + s)ds \\ \bar{w} &= \frac{\lambda}{p}. \end{aligned} \tag{3.31}$$

We have three equations and four unknowns $\bar{w}$, w_0, λ, and p, a well-posed problem.

Synthesis Problem. Assume $\mathcal{S}'_\lambda$. Given some function $f(\cdot)$, find functions $a(\cdot)$ and $b(\cdot)$ such that $\bar{w} = f(p)$. The problem is to solve

$$\begin{aligned} w_0 &= b(A^{-1}(A(w_0) + 1/\lambda)) \\ \frac{1}{p} &= \int_0^{1/\lambda} A^{-1}(A(w_0) + s)ds \\ \frac{\lambda}{p} &= f(p). \end{aligned} \tag{3.32}$$

We have three equations and five unknowns λ, p, w_0, $a(\cdot)$, and $b(\cdot)$. In general, solving the design problem may be non-trivial.

The design rule found in Highspeed TCP [42] can be seen as solving the design problem above, but approximately. There, the authors fixed a function $b(\cdot)$. An exact computation seems to be non-trivial; it requires solving the functional with respect to $A^{-1}(\cdot)$ in the synthesis problem.

Synthesis of an AIMD under the Reference Loss Process

Consider we are given the target function $f(p) = K/\sqrt{p}$, for a fixed constant $K > 0$. We apply the design rule of the present section to "synthesize" an AIMD(α,β) control for the given $f(\cdot)$. The design rule requires α and β to be chosen such that

$$K = \sqrt{\frac{\alpha}{2}\frac{1+\beta}{1-\beta}}.$$

We know from the lower bound in Theorem 3.7.1, that for a process of loss-events in $\mathcal{S}_\lambda$, defined as a set of all stationary ergodic sequences of inter loss-event times that have an arbitrary non-null finite intensity λ, $\bar{w} \geq f(p)$ holds. In other words, an AIMD control, designed under the reference loss-events, overshoots its target for any stationary ergodic sequence of inter loss-event times with a finite non-null mean, and non-null variance. Hence, in general, it *does not* solve the synthesis problem. This tells us that the reference system of constant inter loss-event times is *extremal* for AIMD. It is precisely the *worst-case*. This implies that if we apply the design rule of the present section, then we do a worst-case design, such that $\bar{w}$ is always at least $f(p)$, for any stationary ergodic sequence of the inter-loss event times. We defer showing this extremal property with another approach to Section 3.7.3.

We next demonstrate by two examples that $\bar{w}$ can be significantly larger than $f(p)$. We first give a numerical example, and then a more realistic example from our lab experiments with TCP.

Example 3 (Period-Two Loss-Events) *Consider a sequence of inter loss-event times defined as, for some fixed* $0 \leq \eta \leq 1$*, and an even positive integer* m*,*

$$s_m^{\lambda,\eta} = \left(\eta\frac{2}{\lambda}, (1-\eta)\frac{2}{\lambda}, \eta\frac{2}{\lambda}, (1-\eta)\frac{2}{\lambda}, \ldots, \eta\frac{2}{\lambda}, (1-\eta)\frac{2}{\lambda}\right) \text{ in } \mathbb{R}_+^m.$$

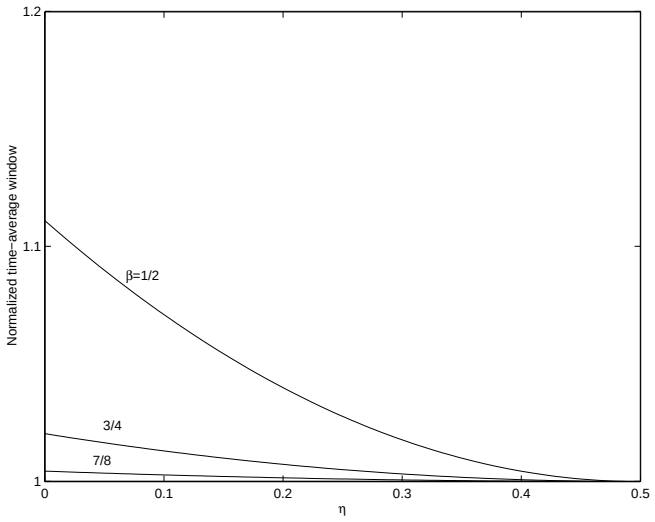


Figure 3.14: The plot shows the ratio of the time-average window attained for period-two loss-events and the time-average window obtained under the reference of constant inter-loss times, versus the parameter η. $\alpha = 1$, $\beta = 1/2$, $\lambda = 1$.

The time-average window can be computed to be equal to

$$h_{\lambda,w_0}(s_\infty^{\lambda,\eta}) = \frac{\alpha}{\lambda(1-\beta^2)}\left[1+\beta^2-2(1-\beta(2-\beta))\eta(1-\eta)\right].$$

Indeed, $h_{\lambda,w_0}(s_\infty^{\lambda,\eta}) \geq h_{\lambda,w_0}(s_\infty^{\lambda,1/2})$*, where the right-hand side is the time-average window attained for the reference sequence of constant inter loss-event times. We show numerical values of the ratio of the left and right-hand side in the last inequality, see Figure 3.14. The result illustrates well that the reference loss process of constant inter loss-event times is the worst-case. In this case the overshoot is moderate, but non-negligible; the maximum possible overshoot is* $2(1+\beta^2)/(1+\beta)^2$*, for* $\beta = 1/2$*, this amounts to* $10/9 = 1.\dot{1}$.

Example 4 *We review the results of our lab experiments obtained for TCP and RED, in Figure 3.13. Look at Figure 3.13, left part. The plots compare TCP throughput against the square-root formula obtained under the assumption that inter loss-event times are fixed to a constant. Note that over a wide range of loss-event rates, TCP throughput overshoots the formula. The overshoot is not larger than* 20%*. For larger loss-event rates, the throughput tends to be smaller than the prediction made by formula; this is a limit case of no interest, given that for non-rare losses, TCP may not be well approximated by an AIMD sender.*

3.5.4 Increase-Decrease Controls that Do Not Solve the Synthesis Problem

In this section, we identify a subset of increase-decrease controls, which if obtained by the synthesis under the reference loss process of constant inter loss-event times, attain $\bar{w} > f(p)$, or *almost*, for some loss processes, defined shortly. Highspeed TCP [42] belongs to this subset.

Let $\mathcal{I}_\epsilon$, a fixed $\epsilon \geq 0$, be a subset of increase-decrease controls that for a given function $x \to f(x)$ obeys to the design rule of the present section, and the following conditions are true;

(A1) $x \to A(b(A^{-1}(x)))$ is increasing convex.

(A2) There exists a convex function $x \to \varphi(x)$ such that

$$\varphi(x) \leq A^{-1}(x) \leq (1+\epsilon)\varphi(x), \text{ all } x \geq 0.$$

(A3) $x \to xf(x)$ is non-decreasing.

The main result in this section is;

Theorem 7 *For all increase-decrease controls in $\mathcal{I}_\epsilon$ and a renewal sequence of loss-events, the following inequality holds*

$$\bar{w} \geq \frac{1}{1+\epsilon} f\left(\frac{1}{1+\epsilon} p\right).$$

Comments. If $\epsilon = 0$, then $\bar{w} \geq f(p)$, the time-average window $\bar{w}$ is lower-bounded by $f(p)$. If $\varepsilon > 0$, but small, then $\bar{w}$ is *almost* lower-bounded by $f(p)$. It is evident from the proof (see Equation (3.50) in Section 3.7.4) that in the theorem *strict* inequality holds for any independent, identically distributed sequence of inter loss-event times with a finite non-null mean and a non-zero variance.

Corollary 2 *A direct implication of the theorem is*

$$\bar{w} \geq \left(\frac{1}{1+\epsilon} \inf_{x \in (0,1]} \frac{f(\frac{1}{1+\epsilon}x)}{f(x)}\right) f(p).$$

When (A3) is True. This hypothesis is verified for some loss-throughput functions f, but not all. It is indeed true for $f(p) = K/p^c$, $K > 0$, $c \leq 1$. However, it can be checked that the hypothesis is not verified for a loss-throughput formula in [93]. We give a particular result for (A2) replaced with

(A2') $x \to a(x)$ is non-decreasing.

Note that (A2) is weaker than (A2').[9] Hence, we have the following corollary from Theorem 7.

Corollary 3 *Replace (A2) with (A2'), then with all other assumptions remaining unchanged, the conclusion of Theorem 7 is* $\bar{w} \geq f(p)$.

Supplemental results

The last theorem is obtained by a conjunction of two lemmas that we display next. Suppose $\mathcal{S}_\lambda$ is the set of all independent, identically distributed inter-loss event times with non-zero variance and finite non-null mean $1/\lambda$.

Lemma 6 *Consider an increase-decrease control that obeys (A1) and (A2). The time-average window* $\bar{w}'$ *attained under* $\mathcal{S}'_\lambda$*, and the time-average window attained under* $\mathcal{S}_\lambda$*, are related as* $\bar{w} \geq \frac{1}{1+\epsilon}\bar{w}'$.

The result implies that, if we know that $\bar{w}' \geq f(p')$,[10] then we know that under the assumptions of Lemma 6, we have $\bar{w} > \frac{1}{1+\varepsilon} f(p')$. It still remains to show when the last implies $\bar{w} > \frac{1}{1+\varepsilon} f(\frac{1}{1+\varepsilon}p)$.

Lemma 7 *Assume* $\bar{w} \geq \frac{1}{1+\varepsilon}\bar{w}'$, $\bar{w}' \geq f(p')$*, and (A3). Then* $\bar{w} \geq \frac{1}{1+\varepsilon} f(\frac{1}{1+\varepsilon}p)$.

As a by-product, from the proof of Theorem 6 we obtain almost for free the following result, which may be of independent interest. Note that the next theorem applies for a more general set of inter loss-event times, than in Theorem 6. Let

$$w_0 = \lim_{m\to\infty} \frac{1}{m} \sum_{k=0}^{m-1} W(T_k).$$

We re-define $\mathcal{S}_\lambda$ to be the set of all stationary ergodic sequences of the inter loss-event times with non-zero variance and with finite non-null mean $1/\lambda$.

Theorem 8 *Consider an increase-decrease control that satisfies (A1) and (A2). Then, the window event-average* w'_0 *attained under* $\mathcal{S}'_\lambda$*, and the window event-average* w_0 *attained under* $\mathcal{S}_\lambda$*, are related as* $w_0 \geq \frac{1}{1+\epsilon} w'_0$.

Corollary 4 *Replace (A2) in Theorem 8 with (A2'), then the conclusion of the theorem is* $w_0 \geq w'_0$.

[9] To see this note that $x \to A^{-1}(x)$ is convex iff $x \to a(x)$ is non-decreasing. Hence, if (A2') holds, then (A2) indeed holds with $\epsilon = 0$.

[10] Ideally, we would design an increase-decrease control such that $\bar{w}' = f(p')$, under the reference loss process. However, in some situations we may not be able to exactly solve the synthesis problem, even under the reference loss process. This may really happen as we found in the design of Highspeed TCP; see Claim 5.

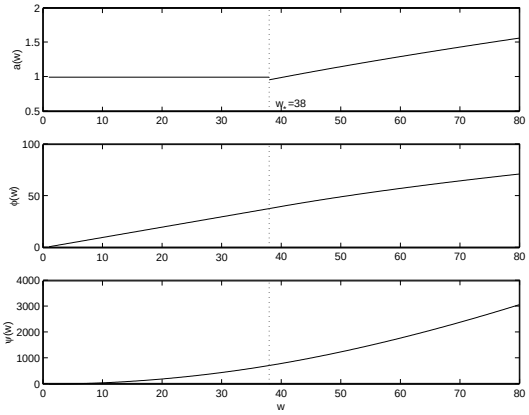


Figure 3.15: Highspeed TCP. Function (Top) $a(\cdot)$, (Middle) $A(\cdot)$ (primitive of $1/a$), (Bottom) $\psi(\cdot)$ (primitive of A).

3.5.5 Application to Highspeed TCP

Highspeed TCP definition as found in [42] may be seen as a design of an increase-decrease control for the given function

$$f(p) = \begin{cases} \sqrt{K}\frac{1}{p^\gamma}, & p \le p_* \\ \sqrt{\frac{3}{2}}\frac{1}{\sqrt{p}}, & p > p_* \end{cases}$$

where $p_* := \frac{3}{2}w_*^2$, $w_* := 38$, $K \simeq 0.12$, and $\gamma \simeq 1/1.2$.

The decrease function is fixed to $b(w) = (1 - \beta(w))w$,

$$\beta(w) = \begin{cases} \frac{1}{2}, & w \le w_* \\ c \ln w + d, & w > w_* \end{cases}$$

where c and d are some constants specified in [42], $c \simeq -0.052$, $d \simeq 0.689$.

Then, by some arguments found in [42], the increase function is set to

$$a(w) = \begin{cases} 1, & w \le w_* \\ \frac{1}{K^\gamma} w^{\frac{2\gamma-1}{\gamma}} \frac{2\beta(w)}{2-\beta(w)}, & w > w_*. \end{cases}$$

See Figure 3.15, top part, for a plot of the function $a(\cdot)$.

Design of Highspeed TCP [42]. The argument used in [42] to obtain $w \to a(w)$ for the fixed $b(\cdot)$ and $f(\cdot)$ can be seen as solving the design problem under $\mathcal{S}'_\lambda$, but approximately, assuming that $a(\cdot)$ and $\beta(\cdot)$ are almost constants—that is, that the control is almost AIMD. The argument in [42] would yield an exact solution if $a(\cdot)$ and $\beta(\cdot)$ would be functions of the time-average window, hence, a constant, for a given steady-state. We make no attempt to solve the design problem posed in Section 3.5.3. Rather, we take $a(\cdot)$ and $b(\cdot)$ as defined

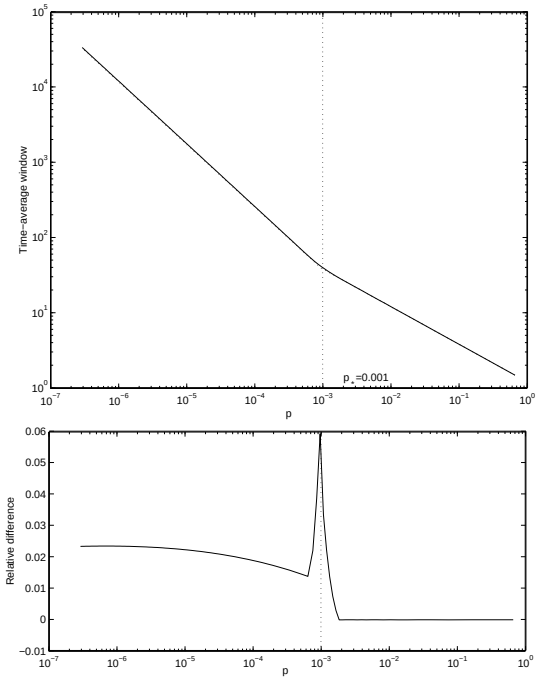


Figure 3.16: (Top) $p' \to \bar{w}'$ for Highspeed TCP. The dotted line is the target function f, (Bottom) $(\bar{w}' - f)/f$. We observe that $\bar{w}'$ is never smaller than f, and never larger than f by the factor 1.06.

in [42] and solve the analysis problem under $\mathcal{S}'_\lambda$, that is, we compute $\bar{w}'$. We then compare $\bar{w}'$ with $f(p')$. If $a(\cdot)$ would be the exact solution of the analysis problem, then, we would find $\bar{w}' = f(p')$. We show later that this is not the case.

In Section 3.7.7, we show that idealized Highspeed TCP verifies the hypotheses of Theorem 7. Given that the verification is done by numerical computations, we pose the result as a claim.

Claim 4 *For idealized Highspeed TCP,*

- *the statement of Theorem 7 is true for some* $\epsilon \in (0, 0.0012)$*,*
- $\bar{w} \geq (1 - \varepsilon'')f(p)$*, for some* $\epsilon'' \in (0, 0.0023)$*, under any random renewal loss-events.*

The above claim is based on the following observation, from a numerical computation in Figure 3.16.

Claim 5 *For idealized Highspeed TCP, the time-average window $\bar{w}'$ and loss-event rate p', under $\mathcal{S}'_\lambda$, are related as, for some $\epsilon' \in (0, 0.06)$,*

$$f(p') \leq \bar{w}' \leq (1+\epsilon')f(p').$$

3.5.6 A Catalog of Internet Control Examples Whereas a Determinism is Extremal

The extremal property of a sequence of *constant* inter loss-event times (a determinism) found in this chapter, may not come as a surprise to many, in particular, to those familiar with a folk theorem of queueing theory. The theorem says that under some conditions, determinism minimizes waiting time in queues, e.g. see Humblet [58], Baccelli and Bremaud [9], Chapter 4.

We showed in Section 3.5.3 that determinism minimizes the time-average window for AIMD. We showed in Section 3.5.4 that the same holds for a broader set of increase-decrease controls, but a smaller set of loss processes. We next give a list of some further examples, found in the context of Internet congestion controls, where determinism is an extremal.

AIMD with Loss-Events Arriving in Batches. Consider an AIMD(α, β) window control with loss-events that arrive in batches. Let $\ldots, Z_{n-1}, Z_n, Z_{n+1}, \ldots$ be a stationary sequence of batch sizes that take values on $\mathbb{N}$. It is a straightforward exercise to extend the result in [4] to hold for batch loss-events

$$\bar{w} = \alpha\lambda \left(\frac{1}{2}\mathbf{E}[S_0^2] + \sum_{k=1}^{\infty} \mathbf{E}[\beta^{Z_{-1}+Z_{-2}+\cdots+Z_{-k}} S_0 S_{-k}] \right).$$

Now, assume that the sequence of batch sizes $\ldots, Z_{n-1}, Z_n, Z_{n+1}, \ldots$ is independent of the point process of batch arrivals. Then, by convexity of $x \to \beta^x$, it follows

$$\bar{w} \geq \alpha\lambda \left(\frac{1}{2}\mathbf{E}[S_0^2] + \sum_{k=1}^{\infty} \beta^{k\mathbf{E}[Z_0]} \mathbf{E}[S_0 S_{-k}] \right).$$

The inequality is strict for batch sizes with non-zero variance. To paraphrase, under the present assumptions, the worst-case is deterministic, with batch sizes fixed to their mean value. A related result was obtained in Guillemin, Robert, and Zwart [52].

AIMD and Variable Round-Trip Times. [11] Altman, Barakat, and Ramos Ramos, in a rephrase of Section V in [6], showed that:

[11] This example was revised on December 16, 2003. The revised version contains a new result (Theorem 9) that supersedes a previous statement, which does not apply to AIMD (discussed at the end of the example in the revised version). The author would like to kindly thank Eitan Altman and Chadi Barakat for pointing him to a problem with the previous statement and discussion.

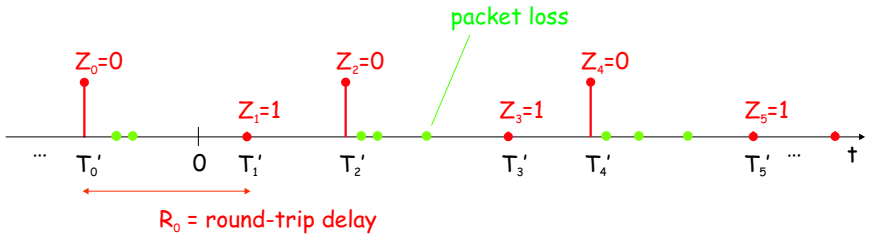


Figure 3.17: The construction of the loss-events from packet loss events. A loss-event happens in the nth round-trip round $(T'_n, T'_{n+1}]$ if at least one packet loss lands in that interval.

> An AIMD window controller that experiences random packet losses in time according to a homogeneous Poisson process (with a finite intensity), and experiences the round-trip delays that are a sequence of independent, identically distributed (i.i.d.) random variables (with finite non-null mean), the throughput is smaller or equal, than if the controller was driven by a sequence of i.i.d. random round-trip times that is larger in the stochastic convex order[12] than the former sequence of the round-trip times.

The result, in particular, implies that fixing the round-trip times to the mean of an i.i.d. random sequence of the round-trip times, the throughput is not larger. This result exhibits yet another example whereas *determinism* is extremal, precisely speaking, the *worst-case*.

See Figure 3.17 for an illustration of the construction of the loss-events from the packet loss events and the round-trip times.

We show a new result that, in particular, extends the statement of [6] under a *weaker* assumption: the round-trip times, *a sequence of stationary random variables* with a finite non-null mean. From a practical perspective, this generalization has a merit, given that in many situations, the hypothesis that a sequence of the round-trip times is an i.i.d. random sequence may not be justified. Our assumption on the relation of the loss process and the process of the round-trip times ((A1) below) may not be mild, however, it is general enough to accommodate the Poisson assumption in [6]. In summary, the result shown below orders the throughputs of AIMD with respect to the variability of a stationary random sequence of the round-trip times.

We consider an AIMD(α, β) window control indexed with $\alpha \in (0, \infty)$ and $\beta \in (0, 1)$. We observe the window size W_n just before time T_n, for some $n \in \mathbb{Z}$. The sequence of the windows evolves as

$$W_{n+1} = B_n W_n + \alpha, \tag{3.33}$$

[12]For two random variables X and Y that take values on $\mathbb{R}^n$, we say X is larger in the stochastic convex order than Y, iff $\mathbf{E}[f(X)] \geq \mathbf{E}[f(Y)]$, for $f : \mathbb{R}^n \to \mathbb{R}$, any convex mapping. We use the notation $X \geq_{cx} Y$, when X is larger than Y in the stochastic convex order.

where $B_n = \beta + (1-\beta)Z_n$. The elements of the sequence $\{Z_n\}_n$ take values on $\{0,1\}$. We interpret $Z_n = 0$, if there is a loss-event in the interval $[T_n, T_{n+1})$, else $Z_n = 1$.

Assume that $\{Z_n\}_n$ is a stationary ergodic sequence and $\mathbf{E}[Z_0] < 1$, then it is known that W_n converges in distribution to (e.g. see [6] and references therein)

$$W_n^* = \alpha \sum_{k=0}^{\infty} \prod_{m=n-k}^{n-1} B_m. \tag{3.34}$$

Let $R_n = T_{n+1} - T_n$, $n \in \mathbb{Z}$, and attach the physical meaning to R_n as the duration of the n-th round-trip time. Under assumption that $\{R_n\}_n$ is a stationary sequence with finite non-null mean r, it is standard to define the throughput as

$$x = \frac{1}{r}\mathbf{E}[W_0^*].$$

For this reason, our object of interest is $\mathbf{E}[W_0^*]$ and its dependence on the probability law of the loss indications. We consider a particular class of the loss indications whose probability law is related to the probability law of the round-trip times as introduced now:

(A1) There exists a family of the functions $g_n : \mathbb{R}_+^n \to [0,1]$, $n = 1, 2, \ldots$ such that for a random sequence of the round-trip times $R_{i_1}, R_{i_2}, \ldots, R_{i_n}$, $(i_1, i_2, \ldots, i_n) \in \mathbb{Z}^n$,

$$\mathbf{E}\left[\prod_{j=1}^{n} Z_{i_j}\right] = \mathbf{E}\left[g_n(R_{i_1}, R_{i_2}, \ldots, R_{i_n})\right].$$

(A2) $g_n : \mathbb{R}_+^n \to [0,1]$ is a convex function, $n = 1, 2, \ldots$.

Theorem 9 *Consider two AIMD(α, β) controllers with the embedded windows denoted as W_n^* and $\tilde{W}_n^*$. The two controllers experience the round-trip delays $\{R_n\}_n$ and $\{\tilde{R}_n\}_n$, respectively, assumed to be stationary random sequences with a finite non-null mean. Assume (A1) and (A2) hold. Then, if $R_0 \geq_{cx} \tilde{R}_0$, then*

$$\mathbf{E}[W_0^*] \geq \mathbf{E}[\tilde{W}_0^*].$$

Corollary 5 *The statement of the theorem holds, in particular, when $\tilde{R}_n = \mathbf{E}[R_0]$, all $n \in \mathbb{Z}$. (Note that $R_0 \geq_{cx} \tilde{R}_0$ implies $\mathbf{E}[R_0] = \mathbf{E}[\tilde{R}_0]$.)*

The last result tells us that for an AIMD source, under the assumptions (A1) and (A2), the round-trip times fixed to a finite non-null constant r is the *worst-case* over the space of all stationary sequences of the round-trip times with the mean r. This an instance whereas *determinism* is *extremal*.

Remark 1 *For Poisson (λ) packet losses in [6],* $\mathbf{P}[Z_{i_1}=1, Z_{i_2}=1,\dots,Z_{i_n}=1|R_{i_1},R_{i_2},\dots,R_{i_n}] = e^{-\lambda R_{i_1}}e^{-\lambda R_{i_2}}\cdots e^{-\lambda R_{i_n}}$, $(i_1,i_2,\dots,i_n)\in\{\mathbb{Z}^n : i_j\neq i_k, \text{ all } j,k\in\{1,2,\dots,n\}\}$, *which we accommodate by defining*

$$g_n(r_0,r_1,\dots,r_{n-1}) = \exp\left(-\lambda(r_0+r_1+\cdots+r_{n-1})\right).$$

This function is indeed convex (shown below), and hence, the case considered in [6], is a corollary of Theorem 9.

Proof 3 (Theorem 9) *Let $f_n:\mathbb{R}_+^n\to[0,1]$, $n\in\mathbb{N}$, be defined as*

$$f_n(x) = \prod_{k=0}^{n-1}(\beta+(1-\beta)x_k),\ x\in\mathbb{R}_+^n.$$

Indeed, $f_1(x)=(1-\beta)(\beta/(1-\beta)+x_0)$, and

$$f_n = f_{n-1}(1-\beta)\left(\frac{\beta}{1-\beta}+x_{n-1}\right),\ n=2,3,\dots.$$

After expanding the last recurrence, we obtain

$$f_n(x) = \beta^n\sum_{j=0}^{n}\left(\frac{1}{\beta}-1\right)^j\sum_{(i_1,i_2,\dots,i_j)\in\mathcal{C}_j^n}x_{i_1}x_{i_2}\cdots x_{i_j},$$

where $\mathcal{C}_j^n$ is the set of all j-combinations of the elements $\{0,1,\dots,n-1\}$.

From (3.34) and the last characterization, we can write (the interchange of the expectation and infinite sum allowed by Lebesgue monotone convergence theorem)[13]

$$\mathbf{E}[W_0^*] = \alpha\sum_{k=0}^{\infty}\beta^k\sum_{j=0}^{k}\left(\frac{1}{\beta}-1\right)^j\sum_{(i_1,i_2,\dots,i_j)\in\mathcal{C}_j^k}\mathbf{E}[Z_{i_1}Z_{i_2}\cdots Z_{i_j}]. \tag{3.35}$$

But by the assumption (A1), we have

$$\mathbf{E}[Z_{i_1}Z_{i_2}\cdots Z_{i_j}] = \mathbf{E}[g_j(R_{i_1},R_{i_2},\dots,R_{i_j})],$$

and by (A2) and $R_0\geq_{cx}\tilde{R}_0$, it holds, for all $j\in\mathbb{N}$,

$$\mathbf{E}[g_j(R_{i_1},R_{i_2},\dots,R_{i_j})]\geq\mathbf{E}[g_j(\tilde{R}_{i_1},\tilde{R}_{i_2},\dots,\tilde{R}_{i_j})].$$

The last three displays yield the statement of the theorem.

[13] In view of (3.35) and the definition of $f_n(\cdot)$, we can replace the identity in (A1) with $\sum_{k=0}^n\beta^k\mathbf{E}[f_k(Z_0,Z_1,\dots,Z_{k-1})] = \mathbf{E}[g_n(R_0,R_1,\dots,R_{n-1})]$, and the theorem remains to hold.

For completeness, for our Remark 1, we need to show that the following function is convex

$$g_n(r_0, r_1, \ldots, r_{n-1}) = \exp(-\lambda n \bar{r})$$

where $\bar{r} = (r_0 + r_1 + \cdots + r_{n-1})/n$. The Hessian of this function is

$$H(r) = \lambda^2 e^{-\lambda n \bar{r}} J_n,$$

where J_n is the $n \times n$ matrix with all the elements equal to 1. Indeed,

$$y^T H(r) y = \lambda^2 e^{-\lambda n \bar{r}} (\mathbf{1}^T y)^2 \geq 0, \ y \in \mathbb{R}^n,$$

where $\mathbf{1}$ is the column-vector with all the elements equal to 1. In other words, $H(r)$ is positive semi-definite for all $r \in \mathbb{R}^n_+$, hence, g_n is convex.

Before finishing the example of AIMD with variable round-trip times, we point out to an alternative approach, which allows us to arrive to a similar conclusion, under a weaker assumption, then the assumptions (A1) and (A2). Unfortunately, a hypothesis of the analysis ((A1') below) does not accommodate AIMD. Consider, again, the window observed just before time T_n, but assume the window evolves as, for some $h : \mathbb{R}_+ \times [0,1] \to \mathbb{R}_+$,

$$W_{n+1} = h(W_n, Z_n), \ n \in \mathbb{Z}. \tag{3.36}$$

We assume that, $h(x,1) > x > h(x,0)$, for for all $x \in \mathbb{R}_+$. By the last assumption, the stochastic recurrence (3.36) well-defines an increase-decrease control, which upon receiving positive feedback, increases the window; otherwise, it decreases the window. Assume that the recurrence (3.36) is stable, that is there exists a stationary process $\{W_n^*\}$, which verifies the recurrence (3.36), and W_n convergences in distribution to W_n^*. Our interest is the steady-state, hence, we assume that W_0 is equal in distribution to W_0^*.

Assume:

(A1') $(w, z) \to h(w, z)$ is increasing convex;

(A2') There exists a convex function $g : \mathbb{R}_+ \to [0,1]$ such that for a stationary random sequence of the round-trip times $\{R_n\}_n$,

$$\mathbf{P}[Z_0 = 1] = \mathbf{E}[g(R_0)].$$

Proposition 12 *Under (A1') and (A2'), we have*

$$\{\tilde{W}_0^*, \tilde{R}_0, \tilde{R}_1, \ldots\} \leq_{icx} \{W_0^*, R_0, R_1, \ldots\} \Rightarrow \{\tilde{W}_0^*, \tilde{W}_1^*, \ldots\} \leq_{icx} \{W_0^*, W_1^*, \ldots\}.$$

The result follows from a known stochastic ordering [9]: under (A1'),

$$(\tilde{W}_0^*, \tilde{Z}_0^*, \tilde{Z}_1, \ldots) \leq_{icx} (W_0^*, Z_0, Z_1, \ldots) \Rightarrow (\tilde{W}_0^*, \tilde{W}_1^*, \ldots) \leq_{icx} (W_0^*, W_1^*, \ldots).$$

For two random variables X and Y, the notation $X \leq_{icx} Y$ means that $\mathbf{E}[f(X)] \leq \mathbf{E}[f(Y)]$, for any increasing, convex mapping $f : \mathbb{R}^n \to \mathbb{R}$. When

$X \leq_{icx} Y$, we say that X is smaller than Y, in the stochastic increasing convex order. (Note that Z_n and Z'_n are two binary random variables, hence $Z'_n \leq_{icx} Z_n$, iff $\mathbf{P}[Z'_n = 1] \leq \mathbf{P}[Z_n = 1]$.)

The hypothesis (A1') does not accommodate AIMD. To check this, consider

$$h(x, z) = (\beta + (1 - \beta)z)(x + \alpha).$$

This function is neither convex nor concave, its Hessian is

$$H = \begin{bmatrix} 0 & 1-\beta \\ 1-\beta & 0 \end{bmatrix}$$

which is indefinite. This was falsely concluded in the earlier version of this example.

Equation-Based Rate Control. In Chapter 2, we found that under some conditions (see Theorem 1, Section 2.2.1), a sequence of constant loss-event intervals maximizes the time-average send rate of equation-based rate control. This is a yet another example, found in the context of Internet congestion controls, whereas determinism is an extremal.

3.6 Conclusions

- We identified the fairness objective of a network bandwidth sharing with senders that adjust their send rates with AIMD, where the round-trip time for a sender/receiver pair is arbitrary. Our results help us to better understand the fairness in a network of TCP-like senders, in particular, a known bias of TCP against long round-trip time connections.

- We obtained a new characterization of the throughput of an AIMD sender, which is an upper bound. The bound is *parsimonious* in the sense that it requires only *two* statistical parameters of the loss process.

- Our analysis of the synthesis problem of an increase-decrease control suggests two design rules. We gave sufficient conditions on analytical properties of the increase and decrease functions and some conditions on the loss process, under which the design rules yield an increase-decrease control that solves the synthesis problem.

- We analyzed a design rule that constructs an increase-decrease control such that the control attains the time-average window that is exactly equal to the given function of the loss-event rate, under a reference loss process. The reference is taken to be commonly used, a sequence of constant inter loss-event times. We identified subsets of the increase-decrease controls, which constructed by following *this* design rule, in general, do not solve the synthesis problem. The examples are AIMD and Highspeed TCP.

- Our analysis of the synthesis problem should help protocol designers to overcome difficulties while trying to "synthesize" an increase-decrease control to a given target loss-throughput formula. An example is the work of Rejaie, Handley, and Estrin [97], where the difficulty was encountered in the synthesis of an AIMD control to conform to TCP loss-throughput relation.

- We showed that for AIMD, over all sequences of inter loss-event times with an arbitrary fixed mean, the sequence of inter loss-event times fixed to this mean is *the worst-case*, in the sense that it *minimizes* the time-average window. This has a consequence on TCP-friendly rate controls that use a loss-throughput formula that is derived under the reference loss process of constant inter loss-event times. It turns out that some loss-throughput formulae used by TCP-friendly protocols are in fact exact under this reference loss process. In particular, a formula termed SQRT in Chapter 2 is exact for an AIMD control under a sequence of constant inter loss-event times. It follows from a result in Chapter 2 (Theorem 1, Section 2.2.1) that, under some conditions, a sequence of constant inter loss-event times is extremal for an equation-based rate control, but it is in contrast, *the best case*. The outcome is a safe bias in the sense that an

AIMD control is non-conservative, whereas an equation-based rate control is conservative.[14]

3.6.1 Possible Directions of Future Work

- We note that the design rules pointed out in this chapter, and their analysis, remain largely at the level of theory. It would be of interest to carry out a more quantitative study of the design rules for some instances of increase-decrease controls of interest. We believe that our analysis would provide a good stepping stone in that direction.
- The same type of the analysis of the synthesis problem can be carried out for the controls other than the increase-decrease controls, e.g. for those defined in Jin *et al.* [64].

[14]The concept of conservativeness is defined at the beginning of Chapter 2.

3.7 Proofs

3.7.1 Proof of Theorem 5

Lower Bound. The recurrence (3.24) defines evolution of the window embedded just after a loss-event, which for an AIMD(α, β) can be re-written as

$$W_{n+1} = \beta(W_n + \alpha S_n),\ n \in \mathbb{Z}. \tag{3.37}$$

Under the hypotheses of the theorem on the parameters α and β, and $\{\theta_n\}_n$, the recurrence (3.40) is stable (see, e.g. Borovkov [18], Corollary 8.1, Chapter 2). We have

$$\mathbf{E}_N^0[W_0] = \alpha\frac{\beta}{1-\beta}\frac{1}{\lambda}. \tag{3.38}$$

Note that $\theta_n = W_n S_n + \frac{\alpha}{2}S_n^2$, $n \in \mathbb{Z}$. Solving the last quadratic function in terms of S_n and taking the non-negative solution, we obtain

$$S_n = \frac{1}{\alpha}\left(\sqrt{W_n^2 + 2\alpha\theta_n} - W_n\right),\ n \in \mathbb{Z}. \tag{3.39}$$

Substituting the last identity in (3.37), we obtain

$$W_{n+1}^2 = \beta^2(W_n^2 + 2\alpha\theta_n),\ n \in \mathbb{Z}. \tag{3.40}$$

Take expectation on both sides to obtain

$$\mathbf{E}_N^0[W_0^2] = \alpha\frac{2\beta^2}{1-\beta^2}\frac{1}{p}.$$

By Palm inversion formula $\bar{w} = \lambda/p$, combining with the last display, we have

$$\bar{w} = \lambda\frac{1-\beta^2}{2\alpha\beta^2}\mathbf{E}_N^0[W_0^2].$$

Now by convexity of $x \to x^2$, we have

$$\begin{aligned}\bar{w} &\geq \lambda\frac{1-\beta^2}{2\alpha\beta^2}\mathbf{E}_N^0[W_0]^2 \\ &= \frac{1}{\lambda}\frac{\alpha}{2}\frac{1+\beta}{1-\beta},\end{aligned}$$

where the last equality is by (3.38). Use the substitution $\lambda = \bar{w}p$ to re-write the last inequality as

$$\bar{w} \geq \sqrt{\frac{\alpha}{2}\frac{1+\beta}{1-\beta}}\frac{1}{\sqrt{p}}.$$

Note that given that $x \to x^2$ is strictly convex, the inequality is *strict* for loss-event intervals with non-zero variance.

Upper Bound. From (3.40), it follows

$$W_n = \sqrt{2\alpha \sum_{k\geq 1} \beta^{2k} \theta_{n-k}}.$$

Substituting the last expression in (3.39), we have

$$S_n = \sqrt{\frac{2}{\alpha}} \left(\sqrt{\sum_{k\geq 0} \beta^{2k} \theta_{n-k}} - \sqrt{\sum_{k\geq 1} \beta^{2k} \theta_{n-k}} \right), \ n \in \mathbb{Z}. \tag{3.41}$$

Conjunction of the last identity and $\bar{w} = \lambda/p$ reveals

$$\bar{w} = \sqrt{\frac{\alpha}{2}} \frac{1}{p} \frac{1}{\mathbf{E}_N^0[\sqrt{\sum_{k=0}^{\infty} \beta^{2k} \theta_{-k}}] - \mathbf{E}_N^0[\sqrt{\sum_{k=1}^{\infty} \beta^{2k} \theta_{-k}}]}. \tag{3.42}$$

The last is an exact expression of the time-average window in terms of the Palm expectations that involve loss-event intervals. (Distinguish this from another exact expression obtained in [4], which is in terms of some expectations of the *inter loss-event times*.)

By concavity of $x \to \sqrt{x}$,

$$\begin{aligned} \mathbf{E}_N^0 \left[\sqrt{\sum_{k=0}^{\infty} \beta^{2k} \theta_{-k}} \right] &\geq \frac{1}{\sqrt{1-\beta^2}} \sum_{k=0}^{\infty} (1-\beta^2) \beta^{2k} \mathbf{E}_N^0[\sqrt{\theta_{-k}}] \\ &= \frac{1}{\sqrt{1-\beta^2}} \mathbf{E}_N^0[\sqrt{\theta_0}], \end{aligned}$$

and

$$\mathbf{E}_N^0 \left[\sqrt{\sum_{k=1}^{\infty} \beta^{2k} \theta_{-k}} \right] \leq \frac{\beta}{\sqrt{1-\beta^2}} \sqrt{\mathbf{E}_N^0[\theta_0]}.$$

(Note that all the summation elements in the above infinite sums are non-negative, which allows us to interchange the expectations and the infinite sums by Lebesgue monotone convergence theorem.)

It is only left to substitute the last two bounds in (3.42), and to use the definition of $\mathbf{cv}_N^0[\sqrt{\theta}_0]$, in order to recover the stated inequality.

3.7.2 Proof of Theorem 2

Recall the evolution of the embedded window (3.24)

$$W(T_{n+1}-) = A^{-1}(A(W(T_n)) + S_n), \ n \in \mathbb{Z}.$$

Hence

$$A(W(T_{n+1}-)) = A(b(W(T_n-))) + S_n, \ n \in \mathbb{Z}.$$

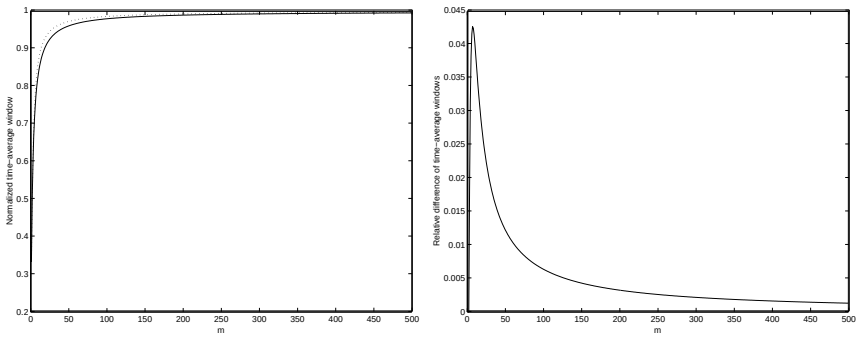


Figure 3.18: (Left) The solid line is the minimum time-average window obtained by numerically solving the problem $(\mathrm{P}_{m,\lambda,0})$ for fixed $\lambda = 1$. The dotted line is the time-average window attained under inter loss-event times fixed to λ. (Right) The same data as on the left-side, but the plot shows the relative difference of the minimum time-average window and the time-average window for constant inter-loss times.

Taking expectation on both sides, we have

$$\mathbf{E}_N^0[A(W(0-)) - A(b(W(0-)))] = \frac{1}{\lambda}.$$

Armed with the definitions of φ and χ, we can write the above display as

$$\varphi(\bar{w}) = \frac{\chi}{\lambda}.$$

By Palm inversion, $\bar{w} = \lambda/p$, hence, we can further re-write the last display as

$$\varphi(\bar{w}) = \frac{\chi}{\bar{w}p}.$$

Now, under the hypothesis that $x \to f(x)$ is decreasing, saying $\bar{w} \leq f(p)$ is equivalent to saying $p \leq f^{-1}(\bar{w})$, where f^{-1} is the inverse of f. Hence, combining the last observation with the last above identity, saying $\bar{w} \leq f(p)$ is indeed equivalent to saying

$$\varphi(\bar{w}) \geq \frac{\chi}{\bar{w}f^{-1}(\bar{w})}.$$

The rest of the proof follows straightforwardly by the hypothesis that $x \to A(x)$ is increasing.

3.7.3 Determinism is Worst-Case for AIMD

Let $\mathcal{S}_m^\lambda = (s_0, s_1, s_2, \ldots, s_{m-1})$, $m \in \mathbb{N}$, be a sequence of positive numbers, such that $\frac{1}{m}\sum_{n=0}^{m-1} s_n = 1/\lambda$, for an arbitrary finite non-null λ. Fix $0 \leq w_0 < \infty$.

Let $\bar{w}_m^{\lambda,w_0}$ be the time-average window attained under the sequence of inter loss-event times $\mathcal{S}_m^{\lambda}$. Let $\bar{w}_m^{'\lambda,w_0}$ be the time-average window attained under $\mathcal{S}_m^{'\lambda}$, a sequence of inter loss-event times in $\mathbb{R}_+^m$, each fixed to $1/\lambda$.

Consider an AIMD(α,β) control. It can be checked that

$$h_{\lambda,w_0}(s_0,s_1,\ldots,s_{m-1}) = \frac{\lambda w_0}{m}\sum_{n=0}^{m-1}\beta^n s_n + \frac{\lambda\alpha}{2m}\sum_{n=0}^{m-1}s_n^2 + \frac{\lambda\alpha}{m}\sum_{n=1}^{m-1}\sum_{k=0}^{n-1}\beta^{n-k}s_k s_n,$$

is the time-average window attained by AIMD(α,β), with the initial window w_0 and the sequence of inter loss-event times $s_1,s_2,\ldots,s_{m-1}$.

Now, imagine an *adversary* whose goal is to *minimize* the time-average window of our AIMD sender by choosing freely any sequence of the inter loss-event times subject to the only constraint that the sequence has arithmetic mean $1/\lambda$. The adversary would solve the quadratic constrained optimization problem, for a fixed m, $\lambda>0$, $w_0\geq 0$,

$\mathbf{(P_{m,\lambda,w_0})}$

$$\begin{aligned} \text{minimize} \quad & h_{\lambda,w_0}(s_0,s_1,\ldots,s_{m-1}) \\ \text{subject to} \quad & s_0\geq 0, s_1\geq 0,\ldots,s_{m-1}\geq 0 \\ & \frac{1}{m}\sum_{n=0}^{m-1}s_n = \frac{1}{\lambda}. \end{aligned} \tag{3.43}$$

We define $\bar{w}_m^{\lambda,w_0} = h_{\lambda,w_0}(s_0^*,s_1^*,\ldots,s_{m-1}^*)$, where $s_0^*,s_1^*,\ldots,s_{m-1}^*$ is a solution of the problem $(\mathrm{P}_{m,\lambda,w_0})$.

Now, let us redefine the goal of our adversary, and consider that we would like to obtain a lower bound on $\bar{w}_m^{\lambda,w_0}$ under the constraints in $(\mathrm{P}_{m,\lambda,w_0})$. An elegant way to obtain the lower bound is to consider the special case $(\mathrm{P}_{m,\lambda,0})$. In other words, we consider $(\mathrm{P}_{m,\lambda,w_0})$ for the zero initial window. Note that, for any $\lambda>0$, $w_0\geq 0$, it indeed holds

$$\bar{w}_m^{\lambda,w_0}\geq \bar{w}_m^{\lambda,0},\ \text{all } m=1,2,\ldots.$$

In general, the problem $(\mathrm{P}_{m,\lambda,w_0})$ can be solved by the method of Lagrange multipliers. For $(\mathrm{P}_{m,\lambda,0})$, it amounts to solving the system of linear equations

$$\sum_{k=0}^{n-1}\beta^{n-k}s_k + s_n + \sum_{k=n+1}^{m-1}\beta^{k-n}s_k = \frac{\gamma}{\alpha},\ n=0,1,\ldots,m-1, \tag{3.44}$$

where γ is the Lagrange multiplier. Now, one can readily note that the matrix of the system (3.44), for $0<\beta<1$, has rank m, the augmented matrix has the same rank, thus there exists unique solution that is displayed next

$$s_0 = s_{m-1} = \frac{1}{\lambda}\frac{m}{2+(1-\beta)(m-2)},\ s_k=(1-\beta)s_0,\ k=1,2,\ldots,m-1.$$

The solution is clearly non-negative, hence, the simple method of Lagrange multipliers provides the solution. By plugging this result into $h_m^{\lambda,0}(\cdot)$, we obtain

$$\bar{w}_m^{\lambda,0} = \frac{\alpha}{2\lambda} \frac{(1+\beta)m}{2\beta + (1-\beta)m}. \tag{3.45}$$

We in fact have shown the following result.

Lemma 8 *$\bar{w}_m^{\lambda,0}$ defined in (3.45) is a lower bound on the objective function in $(\mathrm{P}_{m,\lambda,w_0})$, for any positive integer m, $\lambda > 0$, $w_0 \geq 0$.*

What This Tells Us. In other words, the time-average window attained by an AIMD sender, with an arbitrary fixed initial window $w_0 \geq 0$, and any non-negative sequence of inter loss-event times with a mean $1/\lambda$ is not smaller than $\bar{w}_m^{\lambda,0}$.

Now, let us define $s_m^{\lambda} := (1/\lambda, 1/\lambda, \ldots, 1/\lambda)$ in $\mathbb{R}_+^m$, which corresponds to a sequence of constant inter loss-event times. We obtain

$$\bar{w}_m^{'\lambda,w_0} = \frac{\alpha}{2\lambda} \frac{1+\beta}{1-\beta} + \left(w_0 - \frac{\alpha\beta}{\lambda(1-\beta)} \right) \frac{1-\beta^m}{(1-\beta)m}.$$

From the above computations, we directly obtain the following result.

Lemma 9 *We have, for any finite $w_0 \geq 0$,*

$$\frac{\bar{w}_m^{'\lambda,w_0}}{\bar{w}_m^{\lambda,0}} \downarrow 1, \text{ as } m \to \infty.$$

Interpretation. The lower-bound $\bar{w}_m^{\lambda,0}$ is in the long-run attained by the sequence of inter loss-event times fixed to $1/\lambda$. The lemma tells us that in the long-run, a sequence of constant inter loss-event times is *extremal*. More precisely, it is *worst-case*, that is, in the long-run, it attains the minimum time-average window over the entire set of non-negative sequences of inter loss-event times with mean $1/\lambda$.

3.7.4 Proof of Lemma 6

The proof is based on standard tools of stochastic orderings for stochastic recursive sequences, see for instance [9], Chapter 4. Let $\leq_{icx}$ be binary relation of increasing convex ordering; for two cumulative distribution functions F and G on $\mathbb{R}$,

$$F \leq_{icx} G \Leftrightarrow \int_{\mathbb{R}} f(x)F(dx) \leq \int_{\mathbb{R}} f(x)G(dx), \text{ for any } f \in \mathcal{L},$$

where $\mathcal{L}$ is the set of all increasing convex functions. An ordering for sequences holds, if it holds component-wise.

Let $(W_0, W_1, W_2, \ldots)$ be a sequence of the embedded windows that, under the driving sequence of inter loss-event times $(S_0, S_1, S_2 \ldots)$, $\mathbf{E}[S_n] = 1/\lambda$, all $n = 0, 1, 2, \ldots$, obeys to the recurrence (3.24). Similarly, let $(W_0', W_1', W_2', \ldots)$ be a sequence of embedded windows, which obey to (3.24), but with inter loss-event times $(1/\lambda, 1/\lambda, 1/\lambda, \ldots)$. Assume also $Y_0' \leq_{icx} Y_0$ (say, $Y_0' = Y_0$).

We use a convenient transformation $Y_n = A(W_n)$, $n = 0, 1, 2, \ldots$. It indeed holds

$$(Y_0', 1/\lambda, 1/\lambda, 1/\lambda, \ldots) \leq_{icx} (Y_0, S_0, S_1, S_2, \ldots).$$

Under (A1), it follows from a known result of stochastic orderings for stochastic recursive sequences (see Section 2.2, Chapter 4, [9]) that if the the last ordering holds, then

$$(Y_1', Y_2', Y_3', \ldots) \leq_{icx} (Y_1, Y_2, Y_3, \ldots). \tag{3.46}$$

Note that by Palm inversion formula, we have

$$\mathbf{E}[W(\infty)] = \lambda \mathbf{E}[z(Y_\infty, S_\infty)], \tag{3.47}$$

where, by definition,

$$z(y, s) = \Phi(y + s) - \Phi(y),$$

and $x \to \Phi(x)$ is a primitive of A^{-1}. $W(\infty)$ is a random variable with distribution of the steady-state window. $Y_\infty := A(W_\infty)$, where W_∞ is a random variable with distribution equal to the Palm distribution of the window with respect to the point process of loss-events.

Hence, it is readily seen that $\mathbf{E}[W(\infty)] \geq \mathbf{E}[W'(\infty)]$ is equivalent to

$$\mathbf{E}[z(Y_\infty, S_\infty)] \geq \mathbf{E}[z(Y_\infty', 1/\lambda)] = z(Y_\infty', 1/\lambda). \tag{3.48}$$

Now, note, for a fixed $n = 0, 1, 2, \ldots$,

$$\begin{aligned} \mathbf{E}[z(Y_n, S_n)] &= \mathbf{E}[\Phi(Y_n + S_n) - \Phi(Y_n)] \\ &= \mathbf{E}[\mathbf{E}[\Phi(Y_n + S_n)|Y_n] - \Phi(Y_n)] \\ &\geq \mathbf{E}[\Phi(Y_n + \mathbf{E}[S_n|Y_n]) - \Phi(Y_n)] \\ &= \mathbf{E}[\Phi(Y_n + 1/\lambda) - \Phi(Y_n)] \\ &= \mathbf{E}[z(Y_n, 1/\lambda)]. \end{aligned} \tag{3.49}$$

The last inequality follows from the following property,

(P1) $x \to \Phi(x)$ is convex iff $x \to A^{-1}(x)$ is non-decreasing,

and the fact that by hypothesis $a(\cdot)$ is positive-valued, hence, $x \to A(x)$ is non-decreasing, and thus $x \to A^{-1}(x)$ is non-decreasing as well. The equality $\mathbf{E}[S_n|Y_n] = \mathbf{E}[S_n] = 1/\lambda$, used above, follows from the hypothesis that $(S_0, S_1, \ldots)$ is a sequence of independent random variables with mean $1/\lambda$.

By (A2), we have

$$z^*(y, s) \leq z(y, s) \leq (1 + \epsilon) z^*(y, s), \tag{3.50}$$

where, by definition,

$$z^*(y,s) = \int_y^{y+s} \varphi(x)dx.$$

Note that for $y \to z^*(y,s)$, a fixed $s \geq 0$, it holds

$$\frac{\partial^2 z^*(y,s)}{\partial y^2} = \varphi'(y+s) - \varphi'(y).$$

Hence, for any fixed $s \geq 0$, $y \to z^*(y,s)$ is convex iff $x \to \varphi'(x)$ is non-decreasing. The last property is implied by $x \to \varphi(x)$ convex.

From the inequalities in (3.50), the property that for any fixed $s \geq 0$, $y \to z^*(y,s)$ is convex, and the increasing convex order (3.46), it follows

$$\mathbf{E}[z(Y_n, 1/\lambda)] \geq \mathbf{E}[z^*(Y_n, 1/\lambda)] \geq \mathbf{E}[z^*(Y'_n, 1/\lambda)] \geq \frac{1}{1+\epsilon}\mathbf{E}[z(Y'_n, 1/\lambda)].$$

Lastly, recall from (3.48) that $\mathbf{E}[z(Y_\infty, S_\infty)] \geq \mathbf{E}[z(Y_\infty, 1/\lambda)]$. Combining the last with the above display, we have $\mathbf{E}[z(Y_\infty, S_\infty)] \geq \mathbf{E}[z(Y'_\infty, 1/\lambda)]$, which we already noted is equivalent to $\mathbf{E}[W(\infty)] \geq \mathbf{E}[W'(\infty)]$.

3.7.5 Proof of Lemma 7

Firstly, from the hypothesis $\bar{w}' \geq f(p')$ and the Palm inversion formula $\bar{w}' = \lambda/p'$, we conclude $\lambda \geq p'f(p')$. Hence,

$$\bar{w} = \frac{\lambda}{p} \geq \frac{p'f(p')}{p}. \tag{3.51}$$

Secondly, from the Palm inversion formulas $\bar{w} = \lambda/p$ and $\bar{w}' = \lambda/p'$, and the hypothesis $\bar{w} \geq \frac{1}{1+\epsilon}\bar{w}'$, we conclude

$$\frac{1}{1+\epsilon}p \leq p'.$$

Lastly, by (A3), we have $p'f(p') \geq \frac{1}{1+\epsilon}pf(\frac{1}{1+\epsilon}p)$. From the last inequality, and (3.51), we conclude that $\bar{w} \geq \frac{1}{1+\epsilon}f(\frac{1}{1+\epsilon}p)$.

3.7.6 Proof of Theorem 8

By the hypothesis, $1/a$ is strictly-positive, and thus $x \to A(x)$ is non-decreasing. Hence, $x \to A^{-1}(x)$ is non-decreasing as well. From the last property and the ordering (3.46), it follows

$$\mathbf{E}[W'_\infty] = A^{-1}(\mathbf{E}[Y'_\infty]) \leq A^{-1}(\mathbf{E}[Y_\infty]) = A^{-1}(\mathbf{E}[A(W_\infty)]). \tag{3.52}$$

Now recall the inequality in (A2) that reads as $\varphi(x) \leq A^{-1}(x) \leq (1+\epsilon)\varphi(x)$, all $x \geq 0$. From the former inequality we can conclude $A \leq \varphi^{-1}$. Hence,

$$A^{-1}(\mathbf{E}[A(W_n)]) \leq (1+\epsilon)\varphi(\varphi^{-1}(\mathbf{E}[W_n])) = \mathbf{E}[W_n]. \tag{3.53}$$

Conjunction of (3.52) and (3.53) completes the proof.

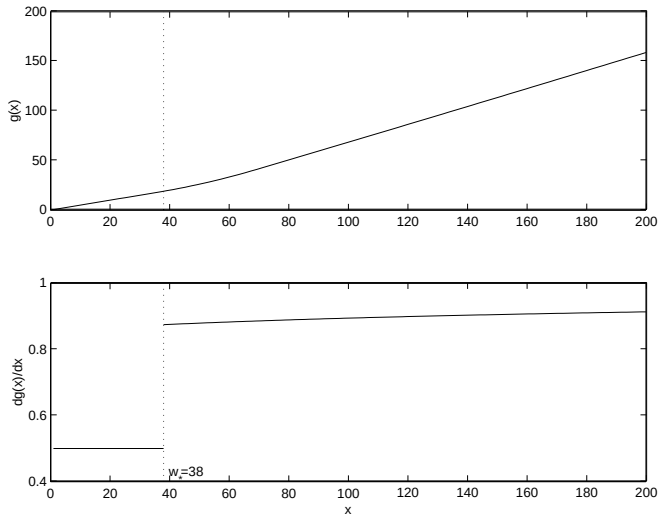


Figure 3.19: (Top) $x \to g(x)$ for Highspeed TCP, $g(x) := A(b(A^{-1}(x)))$, (Bottom) $x \to g'(x)$. The graphs demonstrate $x \to g(x)$ is convex.

3.7.7 Analytical and Numerical Confirmation of Claim 4

We only have to verify that (A1), (A2) with $\epsilon \in (0.0012)$, (A3), and $\bar{w}' \geq f(p')$, are true, then the first assertion of the claim follows from Theorem 7. The second assertion follows by carrying on the step in Remark 2.

Part 1. (A1) is true, that is, $x \to g(x)$ is convex. Here, by definition, $g(x) = A(b(A^{-1}(x)))$. We do not give a rigorous proof here, but rely on a direct numerical computation. For $x \leq w_*$, $g'(x) = 1/2$, else

$$g'(x) = \frac{b'(A^{-1}(x))a(A^{-1}(x))}{a(b(A^{-1}(x)))}.$$

$x \to g'(x)$ is non-decreasing as shown in Figure 3.19. Combining this with the fact that $x \to g(x)$ is continuous, any chord on $g(\cdot)$ lies above $g(\cdot)$, hence, $x \to g(x)$ is convex.

Part 2. (A2) is true for some $\epsilon \in (0, 0.0012)$. Let κ be such that

$$\kappa x \leq A^{-1}(x), \text{ all } x \geq 0.$$

We take

$$\kappa = \inf_{x \geq 0} \frac{A^{-1}(x)}{x}. \tag{3.54}$$

Let $x_* > 0$ be such that $\kappa x_* = A^{-1}(x^*)$. Define

$$\varphi(x) = \begin{cases} \kappa x, & x \leq x_*, \\ A^{-1}(x), & \text{else.} \end{cases}$$

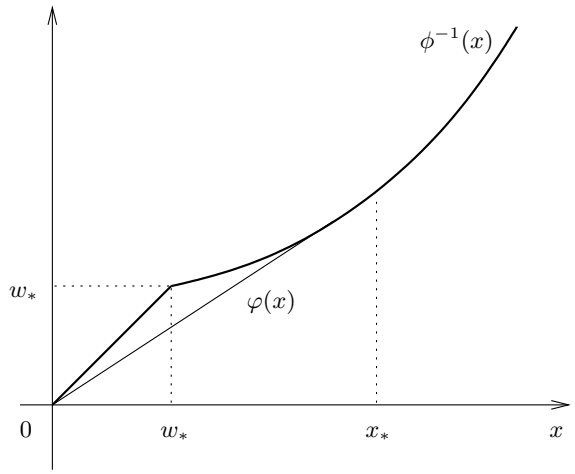


Figure 3.20: An exaggerated illustration of how for Highspeed TCP $x \to A^{-1}(x)$ deviates from its convex closure.

The function $x \to \varphi(x)$ is the convex closure of $A^{-1}(x)$, see Figure 3.20 for a pictorial illustration. Next, consider

$$A^{-1}(x) \leq \varphi(x) \sup_{y \geq 0} \frac{A^{-1}(y)}{\varphi(y)}.$$

Now, note, in our particular instance (see Figure 3.20),

$$\sup_{y \geq 0} \frac{A^{-1}(y)}{\varphi(y)} = \frac{A^{-1}(w_*)}{\kappa w_*} = \frac{1}{\kappa}.$$

From (3.54), and the substitution $x = A(y)$,

$$\frac{1}{\kappa} = \sup_{x \geq 0} \frac{x}{A^{-1}(x)} = \sup_{y \geq 0} \frac{A(y)}{y}.$$

By a direct numerical computation, we obtain $1/\kappa < 1+\epsilon$, where $0 < \epsilon < 0.0012$.
Part 3. (A3) is true. For Highspeed TCP, $pf(p) = \sqrt{3/2}p^{1/2}$, for $p \leq 0.001$, else, $pf(p) = K_1 p^{1-1/\gamma_1}$. Given the values of K_1 and γ_1, $p \to pf(p)$ is indeed non-decreasing on $[0, 0.001]$ and $[0.001, 1]$. Given also that the values of K_1 and γ_1 are set in [42] such that $p \to f(p)$ is continuous, $p \to pf(p)$ is continuous as well. Hence, $p \to pf(p)$ is non-decreasing on the entire domain $[0, 1]$.
Part 4. $\bar{w}' \geq f(p')$. See Appendix 3.7.8.
Part 5. Second assertion holds. Recall Remark 2. It is readily seen

$$\inf_{x \in [0,1]} \frac{f(\frac{1}{1+\epsilon}x)}{f(x)} = (1+\epsilon)^{-1/\gamma_1}.$$

Hence, we have

$$\bar{w} \geq (1+\epsilon)^{-(1+1/\gamma_1)} f(p).$$

By numerical computation for $\gamma_1 = 1/1.2$, $(1+\epsilon)^{-(1+1/\gamma_1)} \geq 1-\epsilon''$, for some $\epsilon'' \in (0, 0.0023)$, and the assertion follows.

3.7.8 A Numerical Confirmation of Claim 5

We solve the direct problem (3.32). Define ψ as a primitive of A. Note that for $w \in [0, w_*]$, $a(w) = 1$ and $b(w) = w/2$. Hence, $A(w) = w$, and $\psi(w) = w^2/2$, for $w \in [0, w_*]$, else

$$A(w) = A_0(w) + w_* - A_0(w_*),\ w > w_*,$$

where $A_C(x)$ is the primitive of $1/a$ with the integration constant $C \in \mathbb{R}$. A bit of elementary integral calculus reveals

$$A_0(w) = K_1^{-\frac{1}{\gamma_1}} \left[\frac{1}{c} e^{-\frac{d}{c}\frac{1-\gamma_1}{\gamma_1}} e\left(-\frac{1-\gamma_1}{\gamma_1}\left(\ln w + \frac{d}{c} \right) \right) - \frac{\gamma_1}{2(1-\gamma_1)} w^{\frac{1}{\gamma_1}-1} \right].$$

Where, we define

$$e(x) = \int \frac{e^{-x}}{x} dx = \ln x + \sum_{n=1}^{\infty} \frac{(-1)^n}{n \cdot n!} x^n.$$

Further, we calculate $w \to \psi(w)$, as

$$\psi(w) = \psi_0(w) + (w_* - A_0(w_*))(w - w_*) + \frac{1}{2} w_*^2 - \psi_0(w_*),\ w \geq w_*,$$

where, $\psi_C(\cdot)$ is the primitive of $A(\cdot)$ with the integration constant $C \in \mathbb{R}$. A bit more of elementary integral calculus yields

$$\begin{aligned} \psi_0(w) \quad = \quad & K_1^{-\frac{1}{\gamma_1}} \left\{ \frac{1}{c} e^{-\frac{d}{c\gamma_1}} \left[e^{\frac{d}{c}} w \ln\left(\frac{\gamma_1 - 1}{\gamma_1}\left(\ln w + \frac{d}{c} \right) \right) - \right. \right. \\ & \left. \left. -e\left(-\ln w - \frac{d}{c} \right) - \frac{\gamma_1}{1-\gamma_1} s\left(\frac{\gamma_1 - 1}{\gamma_1}\left(\ln w + \frac{d}{c} \right) \right) \right] - \frac{\gamma_1^2}{2(1-\gamma_1)} w^{\frac{1}{\gamma_1}} \right\}, \end{aligned}$$

where

$$s(x) = \sum_{n=1}^{\infty} \frac{(-1)^n}{n \cdot n!} \left(\frac{1-\gamma_1}{\gamma_1} \right)^{n+1} e_n\left(\frac{\gamma_1}{1-\gamma_1} x \right),$$

and $e_n(x) := \int x^n e^{-x} dx$. (In numerical solving, we use the elementary recursion $e_n(x) = -x^n e^{-x} + n \cdot e_{n-1}(x)$, $n = 1, 2, \ldots$.)

By above calculations, we have $w \to A(w)$ and $w \to \psi(w)$. Now solving the analysis problem (3.32) corresponds to, for a fixed parameter λ, solve w_0 and p that obey to

$$w_0 = b(A^{-1}(A(w_0) + 1/\lambda))$$

$$\frac{1}{p} = [A(w_0)+1/\lambda]A^{-1}(A(w_0)+1/\lambda) - w_0 A(w_0) - \psi(A^{-1}(A(w_0)+1/\lambda)) + \psi(w_0).$$

Then, by observing $\bar{w} = \frac{\lambda}{p}$, we obtain $p \to \bar{w}$, for a fixed parameter λ. We show a numerical result in Figure 3.16, which confirms Claim 5.

Chapter 4

Expedited Forwarding

In this chapter, we:

- consider a node that complies to the per-hop-behavior of Expedited Forwarding;
- under assumption that the arrival process of bits to a node is a superposition of individually regulated, and stochastically independent flows, we obtain bounds on backlog, delay, and loss that hold in probability;
- we apply our single-node performance bounds to a network of nodes.

4.1 Introduction and Outline

The per-hop behavior of Expedited Forwarding is defined as Packet Scale Rate Guarantee (PSRG) with a rate r and a latency e. PSRG is a node abstraction. There also exist other node abstractions. In Section 4.2 we review the definition of PSRG, some other node definitions and relations among them. In the specific sense given in Section 4.2, PSRG is the strongest definition; it implies other node abstractions introduced in Section 4.2. We derive our performance bounds for some of the node abstractions in Section 4.2. Given that PSRG is the strongest node abstraction, in the precise sense of Section 4.2, a property we show for a node model in Section 4.2 holds for a PSRG node as well.

We assume that the arrival process of bits to a node is a superposition of a fixed number of flows[1]. The flows are assumed to be stochastically *independent*. Each flow is *individually regulated*. We say that a flow is regulated, or constrained, by an arrival curve $\alpha(\cdot)$, if the number of bits observed on the flow during any time interval of duration t is at most $\alpha(t)$. Leaky bucket regulation corresponds to an affine function $\alpha(\cdot)$.

[1] We interchangeably use the terms *arrival process* and *flow* to mean a flow of bits. We also use the terms *Expedited-Forwarding arrival process* and *Expedited-Forwarding flow* to mean a flow of bits that belong to the traffic that uses Expedited-Forwarding service.

Existing results focus on work-conserving single queues that offer a constant service rate. However, in practice, the network nodes are often *not* work-conserving and do not offer a constant service rate for each instant of time. It turns out that many network nodes satisfy a service curve property; Parekh and Gallager [94], Le Boudec [19], Chang [24], Agrawal *et al.* [1], Le Boudec and Thiran [21]. In a deterministic context, a service curve property, with service curve $\beta(\cdot)$, means that at any time t, the total output traffic observed in $[0, t]$ is at least equal to $A(s) + \beta(t - s)$ for some s in $[0, t]$, where $A(s)$ is the number of arrival bits in $[0, s]$. Thus, it is of practical importance to derive performance bounds for a service curve node, and other node abstractions.

Why Probabilistic Bounds. One approach to dimensioning networks such as Expedited Forwarding is to use deterministic *worst-case* bounds. Some results in this vein have been obtained by Charny and Le Boudec [30] and Bennett *et al.* [14]. A worst-case bound on delay-jitter for a network of PSRG nodes, where each Expedited-Forwarding flow is regulated by a leaky-bucket at the network ingress, was shown in [30]. The bound is increasing with the maximum number of nodes a flow can traverse (hop count). The bound is finite for loads at a node smaller than a decreasing function of the hop count. In general, deterministic worst-case bounds give us a strong guarantee, but often they are a pessimistic estimate of the *average-case* performance. This motivates us to look for weaker guarantees, by using bounds that hold in probability.

Methodology. From the methodological viewpoint, a novelty of our approach is in that we systematically apply the following two steps: (1) we bound the buffer overflow event with the union of events, where an event is a deviation of a sum of random variables from its mean, (2) under the given assumptions, these random variables are independent, with bounded support, and we know an upper bound on the summation mean; these properties allow us to use Hoeffding's inequalities [56].[2] In the first step, we often make use of the sample-path results of deterministic network calculus; see, for instance, Chang [25], Le Boudec and Thiran [21], and the references therein.

Related Work. An alternative probabilistic approach for dimensioning Expedited-Forwarding networks has been proposed by Bonald, Proutière, and Roberts [17]. This approach relies on two main assumptions. The first assumption is: At the network ingress, the Expedited-Forwarding arrival process is "Better-than-Poisson." This means that the virtual waiting time distribution of a node fed with an Expedited-Forwarding arrival process is stochastically smaller than the waiting time distribution at the same node, which we would obtain if the arrival process is replaced with a Poisson process of the same intensity of bit arrivals as the original arrival process. The second assumption is a *conjecture*: The packet delay-jitter remains negligible as a flow traverses nodes in a network. This is termed "the negligible jitter" conjecture. The conjecture implies that if the Expedited-Forwarding arrival process is Better-

[2]The inequalities due to Hoeffding (1963) are for a sum of independent random variables with bounded support and a known bound on the mean of the sum. Under these assumptions, Hoeffding's inequalities follow from well-known Chernoff's inequality by some elementary convex arguments.

than-Poisson at the network ingress, it remains so at all nodes in the network. An attractive feature of the Better-than-Poisson approach is its simplicity; it requires knowing only an upper bound on the bit arrival intensity to a node, and the maximum packet length. The main problem of the approach is the "negligible-jitter" assumption; [17] provides some plausible arguments, but still, the assumption remains a conjecture. To our knowledge, a problem that still remains to be resolved is the shaping at the network ingress to make an arrival process Better-than-Poisson. Further, [17] assumes that a node offers a static non-preemptive priority for the Expedited-Forwarding traffic over other traffic. In practice, network nodes are often complex and consist of many processing elements. In this view, the specific scheduling discipline assumed in [17], may not be an exact model of a node in practice, but an approximation. This is a less fundamental problem of the approach in [17], given that performance bounds can be worked-out for more general node models, as shown in this chapter.

4.1.1 Outline of the Chapter

Steady-State Backlog. In Section 4.4 we obtain bounds on backlog for a node that offers a service curve to the aggregate arrival process. We show two sets of bounds. Our first set of bounds extends the results by Kesidis and Konstantopoulos [70] to hold for a service curve node; we use a different proof. The second set of bounds is an extension of the results of Chang, Song, and Chiu [26], which were obtained for a work-conserving constant service rate server. As a by-product, one of our bounds slightly improves [26], even for the case of a constant rate server.

Backlog at Packet Arrival Epochs. The bounds in Section 4.4 are for the steady-state backlog, that is, as seen at an arbitrary point in time. In general, this is *not* the same as observing the backlog in a node at some particular points in time, such as packet arrival instants. The backlog observed at the packet arrival instants is used in the computation of a bound on the complementary distribution of delay of a packet through the node. As a stepping stone to obtain bounds on packet delay, we first give bounds for backlog observed at packet arrival instants, see Section 4.5. This result is of a very general nature; it bounds the complementary distribution of backlog at arrival instants with the product of the stationary complementary distribution of the backlog and a pre-factor. This bound is for a node that offers any Lipschitz continuous[3] service curve, and any stationary arrival process of bits to a node, with a known intensity. The bound is an extension of "the distributional Little's law," (see Konstantopoulos and Last [73] and Konstantopoulos, Zazanis, and de Veciana [74]) to a service curve node.

Packet Delay. In Section 4.4 we show that the delay through a guaranteed rate node (defined in Section 4.2) is bounded by a delay-from-backlog bound. Upon this observation, and a few little technical adjustments, we can apply

[3] A function $f : \mathbb{R} \to \mathbb{R}$ is said to be Lipschitz continuous, if there exists a finite $c > 0$, such that $|f(x) - f(y)| \leq c|x - y|$, for all $x, y \in \mathbb{R}$.

our bounds on backlog at arrival instants to obtain a bound on delay. This is established in Section 4.6. In the same section, we also show that in some situations, a bound on the packet delay can be obtained based solely on the knowledge of the arrival curves of the arrival flows. In particular, there is no need to know the intensity of bit arrivals. In the context of Expedited Forwarding, it is desired to have bounds that require as little knowledge about the arrival processes, as possible.

Bits Lost. In the dimensioning of an Expedited-Forwarding network, one design approach is to minimize the buffer lengths in the nodes subject to the condition that the fraction of bits lost is smaller than a configured value. It is hence of interest to develop bounds on bits loss in a node, in particular, a bound on the fraction of bits lost. In Section 4.7 we show bounds on loss for a node that offers a service curve, and a stronger node abstraction, adaptive service curve, defined in Section 4.2. The upper bound on loss obtained for a node that offers an adaptive service curve is smaller than the respective bound for a node that offers a service curve. We give numerical examples that indicate the differences of the two bounds.

Network of Nodes. In Section 4.8, we show how to apply our single-node bounds to a network of nodes. We give dimensioning formulae that enable us to dimension the buffers of the nodes so that the probability of the buffer overflow or the bit loss rate is smaller than a configuration parameter. By the delay-from-backlog bound, for finite-buffer PSRG nodes, we have a bound on delay that holds with probability one. We show a numerical example that demonstrates the effectiveness of the probabilistic dimensioning over a known worst-case deterministic dimensioning.

The next section overviews the node abstractions that we consider. The knowledgeable reader may skip the next section, and move directly to new results in the remainder of this chapter.

4.2 Overview of Some Node Abstractions

Consider a node that serves packets. Let $0 = T_0 \leq T_1 \leq \cdots$ be a sequence of packet arrival instants. Let T'_n be the departure time of the n-th packet and let L_n be its length in bits. In some of our results, we assume that $\{L_n\}_n$ is a uniformly bounded sequence, and then we use the notation $L_{\max} \geq 0$, such that $L_n \leq L_{\max}$, all $n \in \mathbb{Z}_+$. We use the operators $\vee$ and $\wedge$ that act on two real numbers a and b as $a \vee b := \max\{a, b\}$ and $a \wedge b := \min\{a, b\}$.

Definition 2 (PSRG) *A node is said to offer PSRG with a rate r and a latency e, if for all $n \in \mathbb{Z}_+$, $T'_n \leq V_n + e$, where*

$$\begin{aligned} V_0 &= 0 \\ V_n &= \max\left\{T_n, \min\{V_{n-1}, T'_{n-1}\}\right\} + \tfrac{L_n}{r},\ n = 1, 2, \ldots. \end{aligned}$$

An equivalent definition [20] is that for all $j \leq n$

$$V_n \leq \left\{ e + V_j + \frac{L_{j+1} + ... + L_n}{r} \right\} \vee \bigvee_{k=j+1}^{n} \left\{ e + T_k + \frac{L_k + ... + L_n}{r} \right\}. \quad (4.1)$$

The PSRG was introduced first in the context of differentiated services Internet, as a definition of the per-hop-behavior of Expedited Forwarding service. A related node abstraction, Guaranteed Rate (GR), was introduced earlier by Goyal and Vin [50].

Definition 3 (GR) *A node is said to offer GR with a rate r and a latency e, if $T'_n \leq V_n + e$, where*

$$\begin{array}{rcl} V_0 & = & 0 \\ V_n & = & \max\left[T_n, V_{n-1}\right] + \frac{L_n}{r},\ n = 1, 2, \ldots. \end{array} \quad (4.2)$$

An equivalent definition of GR is that for all n

$$V_n \leq \bigvee_{k=1}^{n} \left\{ e + T_k + \frac{L_k + ... + L_n}{r} \right\}. \quad (4.3)$$

Both PSRG and GR capture how much a real node differs from a hypothetical minimum rate server: The GR model puts a bound on how much *later* the real node is, whereas PSRG puts a bound on how much *earlier or later* it is. The difference is important if a node multiplexes several flows into one single aggregate, thus PSRG is used in the context of aggregate scheduling, while GR is in the context of per-flow scheduling. It can easily be seen that PSRG is a stricter definition than GR, i.e., if a node is PSRG with rate r and latency e, then it is also GR with rate r and latency e (the converse is not true). See [14] for a detailed comparison of PSRG and GR, and Section 7.3.2 [21] for some examples of practical realizations of PSRG schedulers.

Both PSRG and GR are related to the concept of service curves (Sariowan [98]). We next introduce two definitions of service curves and then relate them to PSRG and GR. To that end, we need some further notations. Let $A(t)$ be the number of bits observed in $[0, t]$ at the node input. Likewise, $A^*(t)$ be the number of bits observed in $[0, t]$ at the node output.

Definition 4 (service curve) *A node is said to offer the service curve β, a positive-valued, wide-sense increasing function, if*

$$A^* \geq A \otimes \beta, \quad (4.4)$$

where $\otimes$ is the min-plus convolution[4].

[4]The min-plus convolution is defined as $A \otimes \beta(t) := \inf_{u \in [0,t]} [A(u) + \beta(t-u)]$.

A special case of service curve ("strict" service curve) is when the node guarantees that for an interval of length v that is contained in a busy period, the service offered by the node is at least $\beta(v)$. Note, however, that this is not generally true; in contrast, it is the merit of this definition (like other node models mentioned in this paper) to apply to complex nodes, possibly non-work-conserving, where the concept of busy period is not relevant (consider for example a node made of a delay element followed by a server of a constant rate r, fed with a flow of constant rate $\epsilon \leq r$; the amount of service received during a busy period of duration t is ϵt which can be arbitrarily small).

Roughly speaking, a service curve and GR are equivalent, when the service curve is a rate-latency function, $\beta(t) = r(t-e)^+$.[5] A more precise statement is from [21] (Chapter 2 in the book)

Property 1 (relation of GR and service curve [21]) *A FIFO node that offers a service curve $\beta(t) = r(t-e)^+$ is GR with rate r and latency e. Conversely, a GR node with rate r and latency e offers the service curve $\beta(t) = r(t-e-\frac{L_{max}}{r})^+$.*

Note that the last property holds even for a *non-FIFO* node. Also note that the definition of GR in (4.3) is the max-plus equivalent of the min-plus definition (4.4) of service curve.

We next introduce another type of service curve, and make a connection to PSRG. This new type of service curve was termed *adaptive guarantee* (Okino [91] and Agrawal *et al.* [1]) or *adaptive service curve*, following Le Boudec and Thiran [21], Chapter 7.

Definition 5 (adaptive service curve) *A node is said to offer the adaptive service curve β if for all $t \geq 0$, and all $s \leq t$,*

$$A^*(t) \geq [A^*(s) + \beta(t-s)] \wedge \inf_{u \in [s,t]} [A(u) + \beta(t-u)]. \tag{4.5}$$

The adaptive service curve is the min-plus equivalent of the max-plus definition of PSRG in (4.1), and there is the same type of relationship between PSRG and adaptive service curve.

Property 2 (relation of PSRG and adaptive service curve [21]) *A FIFO node that offers an adaptive service curve $\beta(t) = r(t-e)^+$ is PSRG with rate r and latency e. Conversely, a PSRG node with rate r and latency e offers the service curve $\beta(t) = r(t-e-\frac{L_{\max}}{r})^+$.*

Adaptive service curve is a *stronger* property than service curve. In Figure 4.1 we visualize implications that hold among PSRG, GR, rate-latency adaptive service curve, and rate-latency service curve. We note that PSRG is the strongest node definition. If a node is PSRG, then it also verifies the other definitions. In the remainder, we utilize this observation

[5] We use the shortcut notation $(\cdot)^+ := \max[\cdot, 0]$.

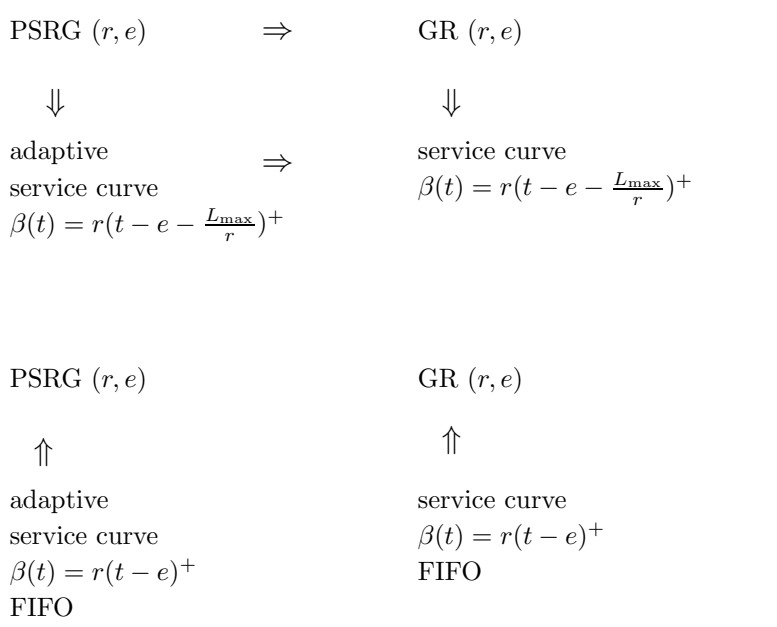


Figure 4.1: The relations among the node abstractions.

Observation 3 *A property shown to hold for any of the node models above, holds for a PSRG node.*

4.3 Assumptions and Notation

Arrival process

We assume that a node is fed by arrival processes labeled as $\mathcal{I} = \{1, 2, \ldots, I\}$. Let A_i, $i \in \mathcal{I}$, be a counting measure on some probability space $(\Omega, \mathcal{F}, \mathbf{P})$. We interpret $A_i(s,t]$ as the number of bits observed in an interval $(s,t]$ of the ith arrival process. By convention, if $s > t$, $A_i(s,t] := -A_i((t,s])$. Likewise, let $A_i^*(s,t]$ be the number of bits of the ith arrival process that depart from the node in an interval $(s,t]$. (See Figure 4.2.) Let $A(s,t] := \sum_{i=1}^{I} A_i(s,t]$ and $A^*(s,t] := \sum_{i=1}^{I} A_i^*(s,t]$, be aggregate arrival and departure processes, respectively.

In some situations, we need to distinguish between the counting processes of bits and packets. Let $\ldots < T_{-1} < T_0 \leq 0 < T_1 < \ldots$ be a point process on $\mathbb{R}$ that marks instants of packet arrivals to a node. The counting process of packet arrivals, N, is then $N(s,t] = \sum_{n \in \mathbb{Z}} \mathbf{1}_{(s,t]}(T_n)$, $s \leq t \in \mathbb{R}$. Let L_n denote the length in bits of the nth packet that arrives to a node. If we need the packet lengths to be in a bounded interval, then we use the notation $0 \leq L_{\min} \leq L_{\max}$ such that $L_{\min} \leq L_n \leq L_{\max}$, all $n \in \mathbb{Z}$.

We now introduce our assumptions:

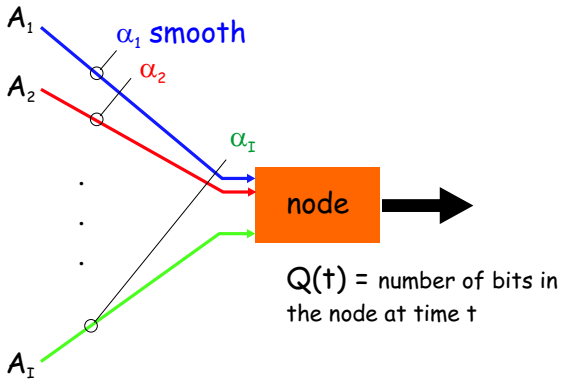


Figure 4.2: Aggregate arrival process to the node is a superposition of arrival processes (flows), assumed to be stochastically independent, and individually constrained by an arrival curve. The node offers the service to the *aggregate* arrival process, it does not discriminate among the flows. The node is assumed to offer either a service curve or an adaptive service curve or a Packet Scale Rate Guarantee or a Guaranteed Rate. $Q(t)$ is the backlog of the bits in the node at time t.

(A1) $A_1, A_2, \ldots, A_I$ are independent.

(A2) A_i has α_i as an arrival curve, i.e. for all $s, t \in \mathbb{R}$,

$$A_i(s,t] \leq \alpha_i(t-s), \text{ all } i \in \mathcal{I}, \text{ w.p.1} \tag{4.6}$$

where α_i is a non-negative, wide-sense increasing function[6] such that $\alpha_i(t) = 0$, for all $t < 0$. We assume, without loss of generality, that α_i is sub-additive, i.e. $\alpha_i(t+s) \leq \alpha_i(t) + \alpha_i(s)$ for all $t, s \in \mathbb{R}$ [19, 24, 1, 21]. When (4.6) holds, we say A_i is "α_i-smooth."

(A3) For all $i \in \mathcal{I}$ and all $s, t \in \mathbb{R}$,

$$\mathbf{E}[A_i(s,t]] \leq \bar{\alpha}_i \times (t-s), \tag{4.7}$$

where $\bar{\alpha}_i = \lim_{t\to\infty} \frac{\alpha_i(t)}{t} = \inf_{t>0} \frac{\alpha_i(t)}{t}$. The last equality is by sub-additivity of α_i (Kingman [71]).

Note that (A3) is true for arrival processes that are stationary and ergodic in the following "weak" sense, for all $s, t \in \mathbb{R}$, $i \in \mathcal{I}$,

$$\mathbf{E}[A_i(s,t]] = \rho(t-s) \text{ and } \lim_{u\to\infty} \frac{A_i(s,s+u]}{u} = \rho, \ 0 < \rho < \infty.$$

[6] We say that function $f(\cdot)$ is wide-sense increasing if $s \leq t$ always implies $f(s) \leq f(t)$. This is also called "non-decreasing".

Indeed, by stationarity $\mathbf{E}[A_i(s,t]] = \mathbf{E}[A_i(0,1]](t-s)$ and by ergodicity

$$\mathbf{E}[A_i(0,1]] = \lim_{u\to\infty} \frac{A_i(0,u]}{u} \leq \lim_{u\to\infty} \frac{\alpha_i(u)}{u} = \bar{\alpha}_i.$$

Note that the "weak sense" stationarity and ergodicity is only for the intensity of the arrival process. It is only for a subset of our results (Section 4.5), where we assume moreover

(A3bis) $A_1, A_2, \ldots, A_I$ are stationary ergodic.

Under the last assumption, we define ρ and ρ' as intensity of arrival bits and arrival packets, respectively. Formally, $\rho := \mathbf{E}[A(0,1]]$ and $\rho' := \mathbf{E}[N(0,1]]$.

Stability

(A4) There exists a sequence of points, "the construction points:"

$$\ldots < S_{-2} < S_{-1} < S_0 \leq 0 < S_1 < S_2 < \ldots$$

such that $\lim_{n\to-\infty} S_n = -\infty$ and $\lim_{n\to\infty} S_n = \infty$, and for all $n \in \mathbb{Z}$, $A(S_n, S_{n+1}] = A^*(S_n, S_{n+1}]$, w.p.1.

It follows from a known result (see, e.g., Lemma 1 in Konstantopoulos and Last [73]) that for (A4) to hold it is sufficient that
(A4a) $\sup_{t\geq 0}\{\alpha(t) - \beta(t)\} < \infty$,
(A4b) $\liminf_{t\to\infty} \{\alpha(t) - \beta(t)\} = -\infty$.

For instance, for the rate-latency service curve $\beta(t) = r\max\{t-e, 0\}$, $r, e \geq 0$, the second condition is the intuitive stability condition $\bar{\alpha} < r$. In the general case, roughly speaking, conditions (A3) and (A4) are weak stability conditions.

Node models

We slightly re-define the node models introduced in Section 4.2. This is done to make 0 an arbitrary point. Note that in the framework of Section 4.2, it is assumed that $A(-\infty, 0] = A^*(-\infty, 0] = 0$, hence, the system is empty at 0. It can be easily observed that the definitions given in this section are compatible with the respective definitions in Section 4.2.

Define $\mathcal{S}(t) = \{S_n,\ n \in \mathbb{Z} :\ S_n \leq t\}$. In (A5a)-(A5b) we re-define the node models of Section 4.2.

(A5a) PSRG. A node is said to offer PSRG with rate r and latency e, if for any $n \in \mathbb{Z}$, there exists some $m \leq n$, $m \in \mathbb{Z}$, such that $V_m \leq T_n$ and $T'_n \leq V_n + e$, where

$$V_n \quad = \quad \max\left\{T_n, \min\{V_{n-1}, T'_{n-1}\}\right\} + \tfrac{L_n}{r}.$$

(A5b) GR. A node is said to offer GR with rate r and latency e, if for any $n \in \mathbb{Z}$, there exists some $m \leq n$, $m \in \mathbb{Z}$, such that $V_m \leq T_n$ and $T'_n \leq V_n + e$, where

$$V_n = \max\{T_n, V_{n-1}\} + \tfrac{L_n}{r}. \tag{4.8}$$

(A5c) Service Curve. A node offers a service curve β to the aggregate arrival process, if with probability one, the following event holds: for all $t \in \mathbb{R}$, and any $u \in \mathcal{S}(t)$,

$$\exists s \in [u,t]: \ A^*(u,t] - A(u,s] \geq \beta(t-s).$$

(A5d) Adaptive Service Curve. A node offers an adaptive service curve β to the aggregate arrival process, if with probability one, the following event holds: for all $t \in \mathbb{R}$, and any $u \in \mathcal{S}(t)$,

$$\exists s \in [u,t]: \ [A^*(u,t] - A(u,s] \geq \beta(t-s)] \wedge \inf_{u \in [s,t]} [A(u) + \beta(t-u)].$$

Let $Q(t)$ be the number of bits in the node at an instant $t \in \mathbb{R}$. (It is the unfinished work at t; we call "backlog.")

4.3.1 Additional Notations

For two functions f and g, we define their *vertical* and *horizontal* deviation, respectively, as

$$\begin{aligned} v(f,g) &= \sup_{t \geq 0}\{f(t) - g(t)\}, \\ h(f,g) &= \sup_{t \geq 0}\{\inf\{u \geq 0: \ f(t) \leq g(t+u)\}\}. \end{aligned}$$

These are standard definitions, see e.g. [21]. Note that $v(f,g)$ is the worst-case backlog for a node that offers a service curve g to the aggregate arrival process that has f as an arrival curve. Similarly, $h(f,g)$ is the worst-case virtual delay (equal to the worst-case delay if the node would be FIFO).

We also define $\lambda_a(t) = at$, $t \geq 0$, and $\lambda_a(t) = 0$, $t < 0$, $a \in \mathbb{R}$. Let $\bar{\alpha} = \sum_{i=1}^{I} \bar{\alpha}_i$ and $\alpha = \sum_{i=1}^{I} \alpha_i$.

4.3.2 The Bounding Method

Our bounds are derived following a two-step method, which we describe next. Consider $\mathcal{A}$, an event whose probability we want to bound. The bounding method consists of:
Step 1. construction of a containment, $\mathcal{A} \subseteq \bigcup_i \mathcal{A}_i$;
Step 2. bounding the probability of $\mathcal{A}_i$.

In our context, $\{\mathcal{A}_i\}_i$ are "the overflow events," for example, $\mathcal{A}_i = \{A(t_i,t] \geq \beta(t-t_i) + b\}$, for some fixed $t_i \leq t$, $b \geq 0$. In the last step, in order to bound $\mathbf{P}[\mathcal{A}_i]$, we often use Hoeffding's inequalities [56].

4.4 Backlog Bounds

In this section we consider a node that offers a service curve β to the aggregate arrival process. We give bounds on the complementary distribution of the backlog $Q(0)$, where 0 is an arbitrary instant.

We assume that the node has a buffer that is sufficient for *loss-free* operation. Then, indeed, $Q(t) = A(u,t] - A^*(u,t]$, $t \in \mathbb{R}$, for some $u \in \mathcal{S}(t)$. From (A5c), it follows that,

$$Q(t) \leq \sup_{-\infty < s \leq t} \{A(s,t] - \beta(t-s)\},\ t \in \mathbb{R}. \tag{4.9}$$

We give two sets of the bounds. We discuss later how they relate to some previous work.

4.4.1 A Set of Backlog Bounds

The first theorem gives us a bound when all the arrival curves are identical.

Theorem 10 (homogeneous case) *Assume (A1)–(A5c), $v(\alpha,\beta), h(\alpha,\beta) < \infty$, and $\alpha_i = \alpha_1$, for all $i \in \mathcal{I}$. Then, for $\bar{\alpha}h(\alpha,\beta) < q < v(\alpha,\beta)$,*

$$\mathbf{P}[Q(0) > q] \leq \exp\left(-Ig(q,\alpha,\beta)\right),$$

where

$$g(q,\alpha,\beta) = \frac{q}{v(\alpha,\beta)} \ln \frac{q}{\bar{\alpha}h(\alpha,\beta)} - \left(1 - \frac{q}{v(\alpha,\beta)}\right) \ln \frac{v(\alpha,\beta) - \bar{\alpha}h(\alpha,\beta)}{v(\alpha,\beta) - q}.$$

The theorem gives us a bound for $q \in (\bar{\alpha}h(\alpha,\beta), v(\alpha,\beta))$. Else, for $q \leq \bar{\alpha}h(\alpha,\beta)$, use a trivial upper-bound $\mathbf{P}[Q(0) > q] \leq 1$, else for $q \geq v(\alpha,\beta)$, $\mathbf{P}[Q(0) > q] = 0$.

Next, we give a bound that applies to arbitrary arrival curves (not necessarily identical). This bound is less sharp than in Theorem 10, which is discussed shortly. Let

$$\mathcal{G} = \{(\gamma_1, \gamma_2, \ldots, \gamma_I) \in \mathbb{R}_+^I :\ \forall i \in \mathcal{I},\ v(\alpha_i, \gamma_i\beta), h(\alpha_i, \gamma_i\beta) < \infty,\ \sum_{i=1}^{I} \gamma_i = 1\}$$

Theorem 11 (heterogeneous case) *Assume (A1)–(A3) and (A5c). Assume that for each $i \in \mathcal{I}$, (A4) holds for a virtual node that offers a service curve $\gamma_i\beta$, which is fed with the arrival process A_i. Then, for any $\underline{\gamma} \in \mathcal{G}$, and $\sum_{i=1}^{I} \bar{\alpha}_i h(\alpha_i, \gamma_i\beta) < q < v(\alpha,\beta)$,*

$$\mathbf{P}[Q(0) > q] \leq \exp\left(-F(\underline{\gamma})\right), \tag{4.10}$$

where

$$F(\underline{\gamma}) = \frac{2(q - \sum_{i=1}^{I} \bar{\alpha}_i h(\alpha_i, \gamma_i\beta))^2}{\sum_{i=1}^{I} v(\alpha_i, \gamma_i\beta)^2}.$$

We next give another set of bounds on the backlog.

4.4.2 Another Set of Backlog Bounds

We need an additional assumption:

(A6) there exists a finite τ such that for all $s \geq \tau$, $\beta(s) \geq \alpha(s)$.

(A6) is a stronger form of (A4-b), which holds in practice (for example, but not only, when α is concave and β is convex) when the natural stability conditions are met. Notice that τ replaces, in the context of service curve, the concept of an upper bound on the duration of a busy period, which is useful only for work-conserving servers.

Let, for any $K \in \mathbb{N}$ and $t \geq 0$, $\mathcal{T}_K(t)$ be the set of partitions of $[-t, 0]$ in K intervals; in other words

$$\mathcal{T}_K(t) = \{(t_0, t_1, ..., t_K) : \ -t = t_0 \leq t_1 \leq ... \leq t_K = 0\}.$$

If time would be discrete, we would require that the partition $\mathcal{T}_K(t)$ is uniform, that is, $t_k = -kt/K$, $k = 0, \ldots, K$.

We first carry out the first step of the bounding method in Section 4.3.2, and then display the main result of this section.

Lemma 10 *Under (A2), (A5c), and (A6), for any $q \geq 0$, it holds*

$$\{Q(0) > q\} \subseteq \{ \sup_{0 \leq s \leq \tau} \{A(-s, 0] - \beta(s)\} > q\}.$$

Lemma 11 *We have, for any $K \in \mathbb{N}$, $\underline{t} \in \mathcal{T}_K(\tau)$, and $q \geq 0$,*

$$\{ \sup_{0 \leq s \leq \tau} \{A(-s, 0] - \beta(s)\} > q\} \subseteq \bigcup_{k \in \{0,1,...,K-1\}} \{A(-t_{k+1}, 0] > q + \beta(t_k)\}. \quad (4.11)$$

Conjunction of the last two lemmas implies

$$\{Q(0) > q\} \subseteq \bigcup_{k \in \{0,1,...,K-1\}} \{A(-t_{k+1}, 0] > q + \beta(t_k)\}. \quad (4.12)$$

The last is a containment of the "buffer overflow event" into a union of the "overflow events."

Theorem 12 (homogeneous case) *Assume (A1)–(A4), (A5c), (A6), and $\alpha_i = \alpha_1$, all $i \in \mathcal{I}$. Let $c \geq 0$ be a constant such that $\rho \leq c$ (in particular, we can take $c = \bar{\alpha}$). Then, for any $K \in \mathbb{N}$ and $\underline{t} \in \mathcal{T}_K(\tau)$,*

$$\mathbf{P}[Q(0) > q] \leq \sum_{k=0}^{K-1} \exp\left(-Ig(t_k, t_{k+1})\right), \quad (4.13)$$

where, for $q > \alpha(v) - \beta(u)$, $g(u, v) = +\infty$, else for $q < cv - \beta(u)$, $g(u, v) = 0$, else

$$g(u, v) = \frac{\beta(u) + q}{\alpha(v)} \ln \frac{\beta(u) + q}{cv} + \left(1 - \frac{\beta(u) + q}{\alpha(v)}\right) \ln \frac{\alpha(v) - \beta(u) - q}{\alpha(v) - cv}.$$

Application of Hoeffding's Inequalities. The last result is obtained by using the union bound on the containment in (4.12). The resulting bound is a sum of probabilities of the "overflow events." Recall our hypotheses on the arrival process. We assumed the arrival process is a superposition of mutually independent arrival processes, each individually constrained with an arrival curve. Knowing that an arrival process is constrained with a sub-additive arrival curve, implies knowing an upper bound on the intensity of the arrival process. Note that probability of an "overflow event" is exactly the complementary distribution of a sum of independent random variables with bounded support and a known upper bound on their expectations. This is a set of assumptions under which we can apply Hoeffding's inequalities [56].

We provide another bound in Theorem 12, which is less tight, but holds for the heterogeneous case. The proof follows closely the last theorem—the only difference the use of a different Hoeffding's bound.

Theorem 13 (heterogeneous case) *Assume (A1)–(A6). Let c be a constant $c \geq 0$ such that $\rho \leq c$. Then, for any $K \in \mathbb{N}$ and any $\underline{t} \in \mathcal{T}_K(\tau)$, for $q < v(\alpha, \beta)$,*

$$\mathbf{P}[Q(0) > q] \leq \sum_{k=0}^{K-1} \exp\left(-g(t_k, t_{k+1})\right), \tag{4.14}$$

where

$$g(u, v) = \frac{2((\beta(u) + q - cv)^+)^2}{\sum_{i=1}^{I} \alpha_i(v)^2}.$$

We can derive an additional bound for the heterogeneous case that requires only aggregate information about the arrival curves. The next result is obtained by using a new containment, a union of "the overflow events," similar to the one used by Massoulié and Buson [37] for leaky-bucket constrained processes. Note that the next result is obtained under a stronger assumption (A3bis).

Theorem 14 (heterogeneous case) *Assume (A1), (A2), (A3bis), (A4), (A5c), (A6). Let $c_i \geq 0$ be a constant such that $\rho_i \leq c_i$, $i \in \mathcal{I}$, $c := \sum_{i=1}^{I} c_i$. Then, the same bound as in (4.14) holds, with*

$$g(u, v) = \frac{((\beta(u) + q - cv)^+)^2}{2\sum_{i=1}^{I} v(\alpha_i, \lambda_{c_i})^2}.$$

A Better Bound for the Heterogeneous Case

By using the union bound on the containment in (4.12), the "buffer overflow event" is bounded with a sum of the following probabilities

$$P[A(-t_{k+1}, 0] > q + \beta(t_k)],\ k = 0, 1, \ldots, K-1.$$

Of course, if we know two upper bounds on a probability in the last sequence of probabilities, then the minimum of these two bounds is also an upper bound.

We exploit this simple fact in the next theorem. We use bounds on the "overflow events" used in Theorem 13 and Theorem 14. We find later in our numerical examples that the resulting bound in some cases yields a significant improvement.

Theorem 15 *Assume (A1), (A2), (A3bis), (A4), (A5c), (A6). Then, the same bound as in (4.14) holds, with*

$$g(u,v) = \frac{2[(q+\beta(u)-\rho v)^+]^2}{[\sum_{i=1}^I \alpha_i(v)^2] \wedge [4\sum_{i=1}^I v(\alpha_i, \lambda_{\rho_i})^2]}.$$

The last bound is obviously better than the bounds in Theorem 13 and Theorem 14.

Four Backlog Bounds for Leaky-bucket Regulated Arrivals

In the context of Expedited Forwarding, it is desirable to know bounds that are based on some *global knowledge* about the arrival processes, and require knowledge of only a *few* global parameters. In this section we give bounds on the complementary distribution of the steady-state backlog for a particular case when arrival curves are *affine* functions. In practice, this would be realized by *leaky-bucket* regulators. The bounds are in terms of some global knowledge about the arrival curves, as will be made specific shortly.

Let $\alpha_i(t) = \rho_i t + \sigma_i$, $t \geq 0$, else $\alpha_i(t) = 0$. $\rho_i, \sigma_i \in \mathbb{R}_+$, $i \in \mathcal{I}$. We can think of ρ_i as an upper bound on the intensity of bit arrivals from the flow i and σ_i as an upper bound on the "burstiness" of the same flow.

Theorem 16 *Consider a node that offers a service curve β. Assume $\alpha_i(t) = \rho_i t + \sigma_i$, $i \in \mathcal{I}$. Then, under (A1)-(A4), (A6), and $\rho < \hat{\beta}$,*

$$\begin{aligned}
& \mathbf{P}[Q(0) > q] \\
\leq\ & \sum_{k=0}^{K-1} \exp\left(-\frac{2[(q+\beta(t_k)-\rho t_{k+1})^+]^2}{[\sum_{i=1}^I (\rho_i t_{k+1}+\sigma_i)^2] \wedge (4\sum_{i=1}^I \sigma_i^2)}\right) && (4.15)\\
\leq\ & \sum_{k=0}^{K-1} \exp\left(-\frac{2[(q+\beta(t_k)-\rho t_{k+1})^+]^2}{(\sqrt{\sum_{i=1}^I \rho_i^2 t_{k+1}} + \sqrt{\sum_{i=1}^I \sigma_i^2})^2 \wedge (4\sum_{i=1}^I \sigma_i^2)}\right) && (4.16)\\
\leq\ & \sum_{k=0}^{K-1} \exp\left(-\frac{2[(q+\beta(t_k)-\rho t_{k+1})^+]^2}{(\rho t_{k+1} + \sqrt{\sum_{i=1}^I \sigma_i^2})^2 \wedge (4\sum_{i=1}^I \sigma_i^2)}\right), && (4.17)\\
\leq\ & \sum_{k=0}^{K-1} \exp\left(-\frac{2[(q+\beta(t_k)-\rho t_{k+1})^+]^2}{(\rho t_{k+1}+\sigma)^2 \wedge (4\sigma^2)}\right), && (4.18)
\end{aligned}$$

for a $K \in \mathbb{N}$, and a sequence $0 = t_0 \leq t_1 \leq \cdots \leq t_K = \tau$.

Discussion. We discuss the bounds (4.15)-(4.18). Consider the parameters (P1) $\sum_{i=1}^{I} \rho_i$, (P2) $\sum_{i=1}^{I} \sigma_i^2$, (P3) $\sum_{i=1}^{I} \rho_i^2$, (P4) $\sum_{i=1}^{I} \rho_i \sigma_i$, (P5) $\sum_{i=1}^{I} \sigma_i$. Note that the parameters require only a global knowledge about the leaky-bucket regulators. We in fact need upper-bounds on:

(P1) the aggregate intensity of bit arrivals;

(P2) the "variability" of the burstiness parameters;

(P3) the "variability" of intensities of individual arrival processes;

(P4) the "correlation" of the intensities and burstiness parameters;

(P5) the aggregate burstiness.

Now, note that the bounds require only upper bounds on the parameters (P1)-(P5) as follows: (4.15) (P1)-(P4), (4.16) (P1)-(P3), (4.17) (P1)-(P2), (4.18) (P1). The bound (4.18) would be in most cases over-pessimistic. It is noteworthy that the bound (4.17) requires only two global parameters: (P1) and (P5). This bound is indeed less sharp than (4.15) and (4.16), but it may be of merit in practice, for instance with networks such as Expedited Forwarding, where the dimensioning would need to be based on a few aggregate parameters about the arrival curves. (For instance, [15] considers a deterministic worst-case dimensioning of an Expedited-Forwarding network, assuming only bounds on (P1) and (P5) are known.) It is natural to ask how much we lose in terms of sharpness of the bounds as we know fewer global parameters. We explore this numerically in Section 4.9.

4.4.3 Discussion and Related work

Extension of Kesidis and Konstantopoulos's Bound [70, 69]. Theorem 10 in Section 4.4.1 generalizes a bound by Kesidis and Konstantopoulos [70, 69] that is for a work-conserving constant rate server with arrival curves that are combination of two leaky-buckets, as is commonplace with ATM and in the Internet. Theorem 10 in Section 4.4.1 extends their result to a node that offers an arbitrary service curve, and to any arrival curve constraints. Our proof is different; it is simpler, even for the original case considered in [70]. We can apply Theorem 10 to the original case in [70] by letting $\alpha_1(t) = \min(\pi_1 t, \bar{\alpha}_1 t + \sigma_1)$ and $\beta(t) = ct$. It can be found in the proof of Theorem 10 that our bound is obtained by computing $\sup_{\theta>0} F(\theta)$, where in the foregoing special case,

$$F(\theta) = \theta q - I \ln \left(1 - \frac{\bar{\alpha}_1}{c} + \frac{\bar{\alpha}_1}{c} e^{\theta \frac{\pi_1 - c}{\pi_1 - \bar{\alpha}_1} \sigma_1}\right).$$

This is exactly the function in Theorem 1 of [70]. This shows that we do have an extension of that result.[7]

[7] In fact, [70] proves a tighter bound than that of Theorem 1 [70], but which is not expressible in a closed-form (see discussion in Sec. III [70]).

Extension of Chang, Song, and Chiu's Bound [26]. Our bound in Theorem 12 extends a bound in [26] to hold for a service curve node under mild assumptions on the arrival and service curve. Extending [26] to a service curve is a simple adjustment. However, by the virtue of stochastic comparisons and Hoeffding's inequalities we were able to obtain new bounds for arbitrary, not necessarily equal, arrival curves. We also slightly improve the bound in [26], even for the original case, by using an *under-sampling* argument. Incidentally, this makes the bound valid in continuous time, whereas [26] assumes time to be discrete. If time would be discrete, and we let $\beta(s) = c(s+1)$, $K = t$, $t_k = k$, then our Theorem 12 gives the same bound as [26]. However, even for the original scenario in [26], we have a slight improvement: if τ is large (which may happen simply because our time unit is very small), we expect the bound in [26] to be large, because it relies on the union bound. We expect to have a better bound by allowing K to be smaller than τ (under-sampling). This is verified in Section 4.9. Note that the theorem implies that for any $K \in \mathbb{N}$ and $\underline{t} \in \mathcal{T}_K(\tau)$, the right hand-side in (4.13) *is* a bound. This allows us to take the infimum over all possible partitions of $[0, \tau]$.

Both [70] and [26] give explicit results under the assumption that all the arrival curves are equal and leave the general case as an optimization problem to solve. For both cases, we give simple formulas that apply in general (Theorems 11 and 13). Of course, the bounds that hold in general also apply for identical arrival curves, but they are not as tight; this is inherited from Hoeffding's inequalities.

We also derive a variant for the heterogeneous case in Theorem 14, by combining the proof of Theorem 13 with a bounding technique similar to that used by Massoulié and Buson [37]. The bound in Theorem 13 (as with Theorem 11) requires knowing the arrival curves of all the flows. In contrast, Theorem 14 requires a *partial* knowledge about the arrival curves; it suffices to know bounds on the aggregate burstiness and the long-run aggregate bit intensity. The bound is less tight than Theorem 13, but may be more useful in the context of differentiated services, where only aggregate information may be available.

Many-Sources Asymptotics. All the bounds on the backlog obtained in this section are for a multiplex of an arbitrary number of arrival flows. The second set of the backlog bounds in Section 4.4.2 are obtained by bounding the buffer overflow event at an arbitrary point in time with a union of the "arrival overflow events." This results in bounds that involve a sum of probabilities of the "arrival overflow events" at different timescales. One may also consider some limit cases. It is common to consider the many-sources scaling (see Likhanov and Mazumdar [82]), whereas the buffer capacity and the service rate of a node are set proportionaly to the number of arrival flows. Under this scaling, there often exists typical time-scale of overflow. In other words, a single summation element in the bounds of Section 4.4.2 is dominant; other elements are of the order $O(e^{-I\epsilon})$, $\epsilon > 0$. This property would hold for most of our bounds, but not all; an exception is the bound (4.18). We do not give a substantiate asymptotic versions of our bounds under the many-sources scaling. The interested reader may find some details in [109].

4.5 Bounds on Backlog as Seen at Bit and Packet Arrival Instants

Our goal in this section is to obtain upper bounds on the complementary distribution of the backlog as seen by either *bit* or *packet* arrivals. Before we state the main result of this section, we introduce an additional assumption:

(A9) There exists a $\hat{\beta} < \infty$ such that $\beta(t) - \beta(s) \leq \hat{\beta} \times (t - s)$, all $s \leq t$.

In other words, we assume β is a Lipschitz continuous function. In particular, we can set

$$\hat{\beta} = \sup_{t,s \geq 0} \frac{\beta(t+s) - \beta(s)}{s}. \tag{4.19}$$

Let $\tilde{Q}(t)$ be the backlog at time t of a greedy shaper that offers the service curve β. We use the notation $\tilde{Q}(t-) := \lim_{s \uparrow t} \tilde{Q}(s)$.

Theorem 17 *Consider a node that offers a service curve β. Assume A is an arrival process that verifies (A3bis) and $\rho < \hat{\beta}$. Then, for any measurable function $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, it holds*

$$\mathbf{E}_A^0[\psi(\tilde{Q}(0-))] \leq \frac{\hat{\beta}}{\rho} \mathbf{E}[\psi(\tilde{Q}(0))]. \tag{4.20}$$

The theorem is related to that of Konstantopoulos and Last [73]. They showed that *strict equality* holds in (4.20) for a work-conserving single server with a constant service rate, and any measurable function $\psi : \mathbb{R}_+ \to \mathbb{R}$. This equality is known as the "distributional version of Little's law" [73, 74].

An immediate corollary of the theorem is

Corollary 6 *Under the assumptions of Theorem 17,*

$$\mathbf{P}_A^0[\tilde{Q}(0-) > q] \leq \frac{\hat{\beta}}{\rho} \mathbf{P}[\tilde{Q}(0) > q], \ q > 0.$$

Note that the Palm distribution $\mathbf{P}_A^0$ is with respect to the point process of *bit arrivals*. If we want a bound for the backlog with respect to $\mathbf{P}_N^0$, the Palm distribution with respect to the point process of *packet arrivals*, then, several more steps are required to arrive at the desired result.

By definition of Palm expectation [9], for a measurable function $\varphi : \mathbb{R} \to \mathbb{R}_+$,

$$\mathbf{E}_A^0[\varphi(0)] = \frac{1}{\rho} \mathbf{E}[\int_{(0,1]} \varphi(s) A(ds)], \tag{4.21}$$

and

$$\mathbf{E}_N^0[\varphi(0)] = \frac{1}{\rho'} \mathbf{E}[\int_{(0,1]} \varphi(s) N(ds)]. \tag{4.22}$$

The last two displays clearly show that, in general, the Palm expectations with respect to the bit and packet arrivals are *not* the same. The two are the same

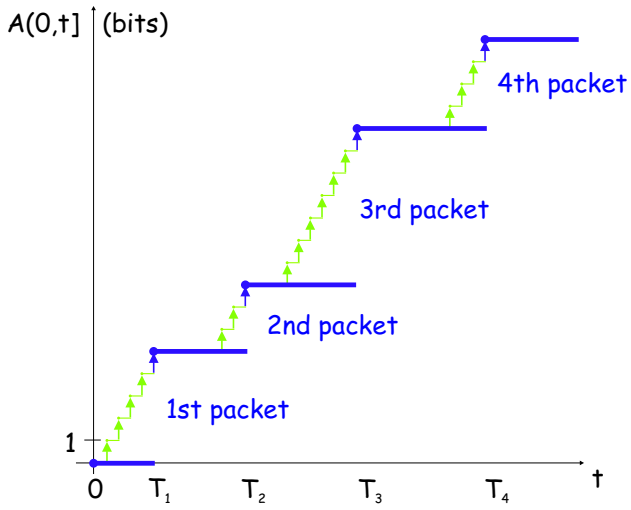


Figure 4.3: The *sampling bias* due to variable-length packets. The Palm expectation of some state of the node (e.g. backlog) with respect to the bit arrivals corresponds to per-arrival-bit averaging of the state variable. The average is as seen by an arbitrary bit arrival. If we pick an arbitrary bit in the flow of the arrival bits, then it is more likely that we lend into *a large* packet. (In the picture, if we pick an arrival bit uniformly at random, then it is the most likely that we pick a bit of the third packet.) Now the length of an arrival packet may be *non-independent* to the state of the node. On the other hand, if we pick up an arbitrary arrival packet, then with *equal likelihood* we would pick a large or a small packet. It is clear that the two viewpoints are not the same.

in some particular cases; for instance for fixed-length packets. A more intuitive argument is given in the caption of Figure 4.3.

Indeed, we have

$$L_{\min}N(s,t] \le A(s,t] \le L_{\max}N(s,t], \text{ any } s \le t. \tag{4.23}$$

Applying the last inequalities into (4.21) and (4.22), we obtain

$$L_{\min}\mathbf{E}_N^0[\varphi(0)] \le \frac{\rho}{\rho'}\mathbf{E}_A^0[\varphi(0)] \le L_{\max}\mathbf{E}_N^0[\varphi(0)]. \tag{4.24}$$

By Palm inversion formula applied to the stochastic intensity of the bit arrivals, $\rho = \rho'\mathbf{E}_N^0[L_0]$. This identity and the left inequality in (4.24) imply the next inequality

$$\mathbf{E}_N^0[\varphi(0)] \le \frac{\mathbf{E}_N^0[L_0]}{L_{\min}}\mathbf{E}_A^0[\varphi(0)] \le \frac{L_{\max}}{L_{\min}}\mathbf{E}_A^0[\varphi(0)]. \tag{4.25}$$

The second inequality is obvious.

Let $\varphi(t) = \psi(\tilde{Q}(t-))$, $t \in \mathbb{R}$. Now, from the left inequality in (4.24), and Theorem 17, we obtain the following adaptation of Theorem 17 to the Palm expectation with respect to *packet* arrivals.

Theorem 18 *Under the same setting as in Theorem 17, we have*

$$\mathbf{E}_N^0[\psi(\tilde{Q}(0-))] \leq \frac{\mathbf{E}_N^0[L_0]}{L_{\min}} \frac{\hat{\beta}}{\rho} \mathbf{E}[\psi(\tilde{Q}(0-))].$$

Comments. If we do not know a better upper bound on $\mathbf{E}_N^0[L_0]$, then we can always use $L_{\max}$ in the inequality of the theorem. Note that the difference with respect to Theorem 17 is the additional pre-factor $\mathbf{E}_N^0[L_0]/L_{\min}$. If the packet lengths were equal, then the inequality in Theorem 18 would boil down to that of Theorem 17.

We next give a particular inequality. Take $\psi(x) = \mathbf{1}_{(q,\infty)}(x)$. Then from the last inequality and Theorem 17, we can make the following statement.

Corollary 7 *Under the assumptions of Theorem 17 and arrival process with packet lengths in* $[L_{\min}, L_{\max}]$*, it holds*

$$\mathbf{P}_N^0[\tilde{Q}(0-) > q] \leq \frac{\mathbf{E}_N^0[L_0]}{L_{\min}} \frac{\hat{\beta}}{\rho} \mathbf{P}[\tilde{Q}(0) > q].$$

We obtained a bound on the complementary distribution of the backlog observed at bit and packet arrival instants in terms of the steady-state complementary distribution of the backlog. We have bounds for the latter in Section 4.4.

4.5.1 When the Intensity of the Bit Arrivals is Unknown

Note that the bound of Theorem 4.20 requires knowing the intensity of bit arrivals ρ. It is of interest to have a bound that would require knowing only the arrival curves of the arrival processes that constitute the aggregate arrival process. We show next that in some situations, we can achieve this.

Assume there exists a function $(x, \rho) \to f(x, \rho)$ such that

$$\mathbf{P}[\tilde{Q}(0) > q] \leq f(q, \rho),$$

where ρ is the intensity of bit arrivals. Assume, for a fixed $q \geq 0$, $x \to \frac{1}{x} f(q, x)$ is a non-decreasing function on an interval $[0, c]$ on which $f(q, x) \leq 1$. If $\bar{\alpha} \leq c$, then

$$\mathbf{P}_A^0[\tilde{Q}(0-) > q] \leq \frac{\hat{\beta}}{\bar{\alpha}} f(q, \bar{\alpha}). \tag{4.26}$$

What This Tells Us. Under the foregoing assumptions, we can bound the complementary distribution of backlog as seen by bit arrivals in terms of a bound that is solely in terms of the arrival curves; we do not need to know the intensity of bit arrivals ρ; it is sufficient to know $\bar{\alpha}$. This has merit in situations when we

do *not* know exactly the value of the intensity of bit arrivals to a node, but we may well know arrival curves of the arrival flows, hence, we would know $\bar{\alpha}$.

Application to Theorem 12. We will see next that the bound in Theorem 12 admits a bound on the complementary distribution of the backlog at bit arrival instants in terms of α *only*. Admit the setting of Theorem 12. We have

$$f(q,\rho) = \sum_{k=0}^{K-1} \exp(-Ig(s_k, s_{k+1})),$$

where $g(s_k, s_{k+1})$ depends on ρ as defined in Theorem 12. In this case, for $x \to f(q,x)$, a fixed $q \geq 0$, to be non-decreasing, it is sufficient that, $\rho \to h_k(\rho)$ defined as

$$h_k(\rho) = -\ln \rho - I\frac{\beta(s_k)+q}{\alpha(s_{k+1})} - I\left(1 - \frac{\beta(s_k)+q}{\alpha(s_{k+1})}\right)\ln\frac{\alpha(s_{k+1})-\beta(s_k)-q}{\alpha(s_{k+1})-\rho s_{k+1}},$$

is a non-decreasing function for $0 \leq \rho \leq \frac{\beta(s_k)+q}{\alpha(s_{k+1})}$, all $k = 0, 1, \ldots, K-1$. After a little calculus, we note $h_k'(\rho) \geq 0$ is equivalent to

$$\rho \leq \frac{I}{I-1}\frac{\beta(s_k)+q}{s_{k+1}},$$

which is indeed true on the interval that we consider $0 \leq \rho \leq \frac{\beta(s_k)+q}{s_{k+1}}$. We know that $\rho \leq \bar{\alpha}$, hence, the last inequality is indeed implied by

$$\bar{\alpha} \leq \frac{\beta(s_k)+q}{s_{k+1}}.$$

In summary, we showed the following

Proposition 13 *Under the setting of Theorem 12, with (A3) replaced by (A3bis),* $\mathbf{P}_A^0[\tilde{Q}(0-) > q]$ *(resp.* $\mathbf{P}_N^0[\tilde{Q}(0-) > q]$*) is bounded by the right-hand side in (4.13), multiplied with the pre-factor* $\hat{\beta}/\bar{\alpha}$ *(resp.* $\mathbf{E}_N^0[L_0]\hat{\beta}/(\bar{\alpha}L_{\min})$*).*

4.5.2 Bounds on Backlog as Seen at Arrival Instants of a Sub-Aggregate

In the previous section, we considered bounds on backlog as seen by bit arrivals of the *whole* aggregate arrival process to a node. In some cases, we may be interested in the backlog as seen by a *subset* of the arrival flows, or a single flow, as a particular case. Let A^0 and A^1 be an arbitrary split of the arrival bits A into two disjoint sets of arrival flows. $A^1(s,t] = A(s,t] - A^0(s,t]$, $s \leq t$. In particular, in order to consider one arrival flow, we would set $A^0(s,t] = A_i(s,t]$, for a $i \in \mathcal{I}$. (See Figure 4.4 for an illustration.) We use the notation $\rho^0 = \mathbf{E}[A^0(0,1]]$. Indeed,

$$\rho\mathbf{P}_A^0[\tilde{Q}(0-) > q] = \rho^0\mathbf{P}_{A^0}^0[\tilde{Q}(0-) > q] + (\rho - \rho^0)\mathbf{P}_{A^1}^0[\tilde{Q}(0-) > q].$$

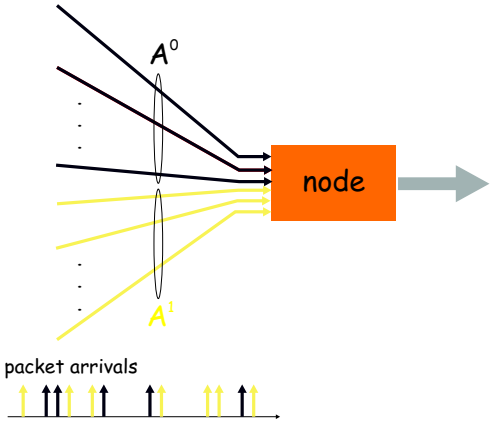


Figure 4.4: Observing the system at bit arrival instants of a subset of the aggregate arrival process. A^0 is the counting process of bits of the subset.

By Theorem 17, we have

$$\mathbf{P}^0_{A^0}[\tilde{Q}(0-) > q] \leq \frac{\hat{\beta}}{\rho^0}\mathbf{P}[\tilde{Q}(0) > q] - \left(\frac{\rho}{\rho^0} - 1\right)\mathbf{P}^0_{A^1}[\tilde{Q}(0-) > q].$$

The probability $\mathbf{P}^0_{A^1}[\tilde{Q}(0-) > q]$ may be non-trivial to compute. To have the inequality in the above last display, we need a lower bound on $\mathbf{P}^0_{A^1}[\tilde{Q}(0-) > q]$. A trivial lower bound is 0, which yields

$$\mathbf{P}^0_{A^0}[\tilde{Q}(0-) > q] \leq \frac{\hat{\beta}}{\rho^0}\mathbf{P}[\tilde{Q}(0) > q].$$

A better bound can be in fact obtained. However, under the foregoing assumptions on the arrival processes, this better bound involves an element which we are not able to bound. However, the bound may be of interest under some other assumptions on the arrival process; for this reason, we show it in Section 4.11.4.

4.6 Bound on Packet Delay

We give a bound on delay for a GR node, defined in (A5b), Section 4.3. Let $D_n = T'_n - T_n$ be sojourn time in the node of the packet labeled n. We call D_n *delay* of the packet n. From the definition of GR (A5b), note

Lemma 12 *Consider a GR node with a rate r and a latency e. A sequence of packet delays $\{D_n\}_{n\in\mathbb{Z}}$ through the node is bounded as*

$$D_n \leq \frac{\tilde{Q}(T_n)}{r} + e, \ n \in \mathbb{Z}, \tag{4.27}$$

where $\tilde{Q}(t) := \sup_{s\leq t}\{A(s,t] - r(t-s)\}$, $t \in \mathbb{R}$.

In other words, for a GR node there holds a "delay-from-backlog" bound. The delay-from-backlog is a property known to hold for PSRG nodes.

Theorem 19 (Theorem III.1 [20]) *Consider a node offering the PSRG with rate r and latency e, not necessarily FIFO. Call $Q(t)$ the backlog at time t. All packets that are in the system at time t will leave the system no later than at time $t + Q(t)/r + e$.*

It is known that the delay-from-backlog property *cannot* hold for a node that offers a rate-latency service curve; see Section IV.E [20] and Chapter 7, Section 7.2.1 in [21]. From the relation of GR and rate-latency service curve (see Figure 4.1), it follows that for a GR node does not hold the delay-from-backlog property. However, Lemma 12 tells us that the delay-from-backlog property *does hold* for a GR node with rate r and latency e, but with respect to the backlog of a *virtual system*; a work-conserving single-server with constant service rate r, which is fed with the same arrival process A as the original GR system.

From (4.27), we have, for $d \geq 0$,

$$\mathbf{P}[D_0 > d] \leq \mathbf{P}_N^0[\tilde{Q}(0) > r(d-e)]. \tag{4.28}$$

The right-hand side is the complementary distribution of the backlog at packet arrival instants. We can obtain a bound in terms of the complementary distribution of the steady-state backlog by the result of Theorem 7, with an intermediate step as shown next.

Indeed, for $T_0 = 0$ (a packet arrival at 0), we have $\tilde{Q}(0) = \tilde{Q}(0-) + L_0$. Note $L_0 \leq L_{\max}$, and hence $\tilde{Q}(0) \leq \tilde{Q}(0-) + L_{\max}$. This gives us

$$\mathbf{P}[D_0 > d] \leq \mathbf{P}_N^0 \left[\tilde{Q}(0-) > r\left(d - e - \frac{L_{\max}}{r}\right)\right]. \tag{4.29}$$

The right-hand side can be directly bounded by the result found in Theorem 7 in Section 4.5, and a bound on the steady-state backlog, for instance, one of those found in Section 4.4.

4.6.1 Discussion

Consider the inequality in (4.28). Note that the right-hand side is probability of the buffer overflow with respect to $\mathbf{P}_N^0$, the Palm distribution. In some situations we may not know a bound on the probability of the buffer overflow at packet arrival instants. However, we may well know a bound on the steady-state probability of the buffer overflow, that is, with respect to $\mathbf{P}$. This gives a merit to a bound on delay obtained as a conjunction of (4.29) and Theorem 7. Notice that the obtained bound is under the assumption that the arrival process to the node is stationary ergodic, with non-null finite packet lengths.

4.7 Bounds on Arrival Bits Lost

In this section we consider a node whose buffer capacity is not sufficient to ensure loss-free operation. This can happen if the buffer capacity q is smaller than $v(\alpha, \beta)$, which is the vertical deviation of the arrival and service curve, and it is a sufficient buffer capacity to ensure no losses. Let $L(s,t]$ be the number of bits lost in an interval $(s,t] \in \mathbb{R}$. Locally to this section, we assume that time is *discrete*, that is, a time instant $t \in \mathbb{Z}$.

We use additional notation. Fix $x \geq 0$. If A is constrained with an arrival curve α, then define

$$\tau(x) = \max\{v \in \mathbb{Z}_+ | \alpha(v) \geq \beta(v) + q + x\}, \tag{4.30}$$

else, $\tau(x) = \infty$.

In particular, for $\alpha(t) = \rho t + \sigma$ and $\beta(t) = r(t-e)^+$, $\rho < r$, we have $v(\alpha, \beta) = \sigma + \rho e$ and

$$\tau(x) = \frac{v(\alpha, \beta) - q - x}{r - \rho} + e.$$

4.7.1 A Bound on Loss for a Service Curve Node

Theorem 20 *Consider a node that offers a service curve β to the arrival process A. Assume the node offers a buffer of capacity $q < v(\alpha, \beta)$. Then, for an arbitrary instant 0, for a $x > 0$,*

$$\mathbf{P}[L(-1,0] \geq x] \leq \sum_{s=-\tau(x)}^{-1} \mathbf{P}[A(s,0] \geq \beta(-s) + q + x], \tag{4.31}$$

else, $\mathbf{P}[L(-1,0] \geq 0] = 1$.

4.7.2 A Bound on Loss for an Adaptive Service Curve Node

Theorem 21 *Consider a node that offers an adaptive service curve β to the arrival process A. Assume the node offers a buffer of capacity $q < v(\alpha, \beta)$. Then, for an arbitrary instant 0, for a $x > 0$,*

$$\mathbf{P}[L(-1,0] \geq x] \leq \sum_{s=-\tau(x)}^{-1} \mathbf{P}[A(s,0] \geq \beta(-s) + q + x] \wedge B(s), \tag{4.32}$$

else, $\mathbf{P}[L(-1,0] \geq 0] = 1$. *Here*

$$B(s) = \min_{u \in [s,-1]} \mathbf{P}[A(u,0] \geq \beta(-u) + x].$$

Note that the above bounds assume neither stationarity nor ergodicity of the arrival process. They are based on a sample-path bound that can be established on the amount of loss in terms of an infimum of the difference of the arrival counting process and the service curve; see proofs of Theorem 20 and Theorem 21.

To obtain a bound on the fraction of bits lost, we assume the arrival process is stationary ergodic, that is (A3bis). Indeed, $\mathbf{E}[L(-1,0]]$ is the expected number of bits lost in a unit time interval, that is, the intensity of bits lost in real-time. By the telescopic formula,

$$\mathbf{E}[L(-1,0]] = \int_0^\infty \mathbf{P}[L(-1,0] > x]dx.$$

Having established bounds on the amount of the bits lost in Theorem 20 and Theorem 21, we in principle have a bound on the intensity of the bits lost. The only thing left is to compute the integral, which we expect computationally would not be a problem, at least for some bounds of Hoeffding that we may use in bounding the probabilities in the sums of Theorem 20 and Theorem 21. Nevertheless, we can obtain a simpler expression, and avoid computing an integral. In the proofs of the last two theorems we showed

$$L(-1,0] \leq (\tilde{Q}(0) - q)^+,$$

where $\tilde{Q}$ is the backlog of a virtual system with the arrival process A and departure process $A^* = A \otimes \beta$ (greedy shaper [21]). Indeed,

$$\mathbf{E}[L(-1,0]] \leq \mathbf{E}[(\tilde{Q}(0) - q)^+] = (\mathbf{E}[\tilde{Q}(0)|\tilde{Q}(0) > q] - q)\mathbf{P}[\tilde{Q}(0) > q].$$

We know that $\tilde{Q}(0) \leq v(\alpha,\beta)$, and hence, $\mathbf{E}[\tilde{Q}(0)|\tilde{Q}(0) > q] \leq v(\alpha,\beta)$. Putting the pieces together, we showed

$$\mathbf{E}[L(-1,0]] \leq [v(\alpha,\beta) - q]^+ \mathbf{P}[\tilde{Q}(0) > q]. \tag{4.33}$$

We can now directly apply the results of Section 4.4 to bound the right-hand side in the last inequality.

Fraction of the Bits Lost. In practice, one may be interested in the fraction of bits that are lost in the number of arrival bits observed over a long time interval. Under the assumption that the arrival process is stationary ergodic and the system is stable, the last defined quantity observed over an infinite time interval is equal to ℓ, the probability that an arrival bit is lost. By Palm inversion formula, we have

$$\ell = \frac{\mathbf{E}[L(-1,0]]}{\rho}.$$

By the same arguments as in (4.5.1), we can obtain a bound on ℓ solely based on the knowledge of the arrival curves of the arrival processes. Combining the last display with (4.33), we obtain

$$\ell \leq \frac{[v(\alpha,\beta) - q]^+}{\rho} \mathbf{P}[\tilde{Q}(0) > q]. \tag{4.34}$$

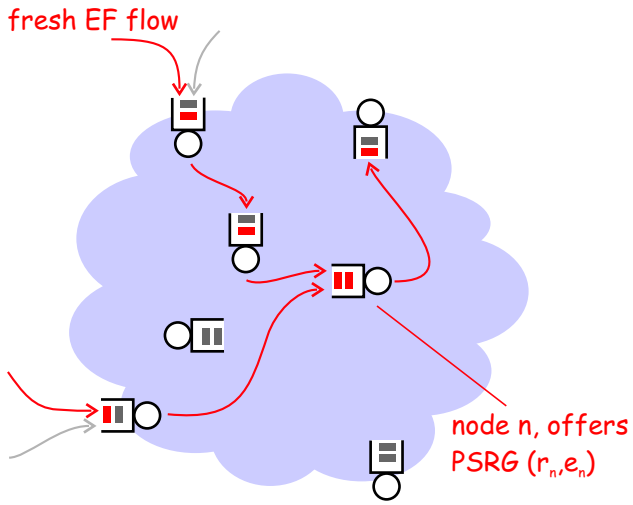


Figure 4.5: Network of PSRG nodes. We assume that Expedited-Forwarding flows are independent and individually regulated at the network ingress. We do assume *neither* independence *nor* regulation of the flows for a node in the network. Each flow is assume to traverse at most h hops; other than this, no particular assumptions are made on the routing of the flows.

4.7.3 Discussion

We have seen that for an adaptive service curve node (Theorem 21) we were able to obtain a better bound on the number of bits lost than for a service curve node (Theorem 20). In Section 4.9.2, we give a numerical example that demonstrates a situation when the bound of Theorem 21 is substantially smaller than the bound of Theorem 20.

4.8 Network of Nodes

In this section we consider a network of PSRG nodes; see Figure 4.5 for an illustration. A difficulty with the network case, where each node offers service to the *aggregate* arrival process (not per-flow service) is that, in general, it is unreasonable to assume that individual arrival processes to any node in the network are stochastically independent. Indeed, individual arrival processes to a node may share some upstream nodes to *this* node, and may become non-independent, even the flows may have been independent at the network ingress. We believe that for some networks it is reasonable to assume that arrival processes are stochastically independent at the network ingress. This is an assumption that we take in the rest of this section. We bound an arrival process of bits to a node with the number of bits of this arrival process as observed at the network

ingress ("fresh arrivals") as explained shortly. This approach was perhaps first used by Kurose [76]. The bounds found in this section are *exact* probabilistic bounds, and *not* some large-deviations asymptotics, see e.g. Wischik [114].

Consider a network of nodes. Suppose a node labeled with n offers a PSRG with a rate r_n and a latency e_n, and a buffer of capacity q_n. Let $\mathcal{I}_n$ be a set of labels of the arrival processes to the node n. Let $A_i^n(s,t]$ be the number of bits that arrive in $(s,t]$ for an arrival processes i to the node n, $i \in \mathcal{I}_n$. Let $\mathcal{N}(i)$ be the ingress node of the arrival process i. Then, $A_i^{\mathcal{N}(i)}(s,t]$ is the number of bits of the arrival process i that arrive at the network ingress ("the fresh traffic"). Let $Q_n(t)$ be the backlog of the node n at an instant t.

We assume that the delay-jitter to any node in the network is bounded with a finite constant Δ. By the delay-from-backlog property of PSRG nodes, such a constant indeed exists for a network of PSRG nodes with finite buffers. The delay-from-backlog property of a PSRG node with a rate r_n and a latency e_n implies that if the backlog in the node is bounded with q, then the delay through this node is bounded with $q/r_n + e_n$. We assume that each Expedited-Forwarding flow in the network traverses at most h nodes, $h \geq 1$. Then, we can define

$$\Delta = (h-1) \max_n \left\{ \frac{q_n}{r_n} + e_n \right\}.$$

It is easy to show that we can bound the arrival bits of an arrival process i to a node n, $i \in \mathcal{I}_n$, by its "fresh traffic." This can be done as follows

$$A_i^n(s,t] \leq A_i^{\mathcal{N}(i)}(s-\Delta,t],\ s \leq t. \tag{4.35}$$

We have seen earlier in this chapter that some of our bounds, the bounds on backlog in Section 4.4.2 and bounds on the bit loss rate in Section 4.7, are a sum of probabilities of the "arrival overflow events." For the given set of the bounds, we have to consider the following probabilities, for a node n, $0 \leq t_1 < t_2 < \ldots$,

$$\mathbf{P}[A^n(-t_{k+1},0] > r_n \left(t_k - e_n - \frac{L_{\max}}{r_n} \right)^+ + c],\ c \in \mathbb{R}_+ \tag{4.36}$$

Here, by definition

$$A^n(s,t] = \sum_{i \in \mathcal{I}(n)} A_i^n(s,t].$$

From (4.35) and (4.36), we can write, for a $c \in \mathbb{R}_+$

$$\begin{aligned} & \mathbf{P}[A^n(-t_{k+1},0] > r_n \left(t_k - e_n - \frac{L_{\max}}{r_n} \right)^+ + c] \\ \leq\ & \mathbf{P}[\sum_{i \in \mathcal{I}(n)} A_i^{\mathcal{N}(i)}(-t_{k+1}-\Delta,0] > r_n \left(t_k - e_n - \frac{L_{\max}}{r_n} \right)^+ + c]. \end{aligned} \tag{4.37}$$

It can be easily checked that the sequence $0 \leq t_1 < t_2 < \ldots$ can be restricted to $0 \leq t_1 < t_2 < \ldots < t_K = \tau_{\Delta,c}$, for a $K \in \mathbb{N}$, where $\tau_{\Delta,c}$ is a positive real

number such that

$$f_n(s+\Delta) \leq r_n(s-e_n)^+ + c, \text{ all } s \geq \tau_{\Delta,c}. \tag{4.38}$$

Here $f_n(\cdot)$ is an arrival curve of the *fresh traffic* of the Expedited-Forwarding arrival processes to the node n.

Hoeffding's Inequalities. In order to apply Hoeffding's inequalities[8] to bound the probability of the event in (4.37), so that the bound is less than one, for a $c \in \mathbb{R}_+$, it is necessary to have

$$r_n \left(t_k - e_n - \frac{L_{\max}}{r_n} \right)^+ + c > \sum_{i \in \mathcal{I}(n)} \rho_i (t_{k+1} + \Delta), \text{ all } k = 0, 2, \ldots, K-1.$$

Recall that ρ_i is the intensity of bit arrivals of the flow i. The last inequality can be re-written as

$$\frac{\sum_{i \in \mathcal{I}(n)} \rho_i}{r_n} < \min_{k \in \{0,1,\ldots,K-1\}} \left\{ \frac{\left(t_k - e_n - \frac{L_{\max}}{r_n} \right)^+ + \frac{c}{r_n}}{t_{k+1} + \Delta} \right\}. \tag{4.39}$$

The last inequality tells us that some bounds on backlog or loss for a node in the network would be non-trivial (< 1) for some loads of the node, not all. We show this for a particular case, which is of interest for dimensioning an Expedited-Forwarding network.

Dimensioning an Expedited-Forwarding Network. Assume the same setting as before. Assume in addition that arrival processes are regulated by leaky-buckets at the network ingress, that is an arrival process i at the network ingress has $\alpha_i(t) = \bar{\alpha}_i t + \sigma_i$ as an arrival curve. Assume we have the following "global knowledge" about the network: We know some positive-valued constants a, b, e such that, for any node n in the network,

$$\begin{aligned} \frac{\sum_{i \in \mathcal{I}(n)} \bar{\alpha}_i}{r_n} &\leq a \\ \frac{\sum_{i \in \mathcal{I}(n)} \sigma_i}{r_n} &\leq b \\ e_n + \frac{L_{\max}}{r_n} &\leq e \end{aligned}$$

This is precisely the set of assumptions made by Bennett *et al.* [15].

If we know a constant d such that $q_n/r_n \leq d$, any node n, then, we can define $\Delta = (h-1)(d+e)$. Then, also, for $\tau_{c,\Delta}$ in (4.38), we can set $\tau_{c,\Delta} = \tau$, where

$$\tau = \frac{b + ahe + a(h-1)d}{1-a}. \tag{4.40}$$

[8] In the way it is done for all the bounds in Section 4.4.2, and would also be used to bound the probabilities in the bounds for the arrival bits lost in Section 4.7.

Bound on the Load of EF Traffic. Assume that the buffer lengths are such that $q_n = dr_n$, for a node n. Consider (4.39), for $c = q_n$. Recall that in applying Hoeffding's inequalities, we can choose any sequence $0 = t_0 < t_1 < t_2 < \ldots t_K = \tau$ (see Section 4.4.2). We restrict $(t_1, t_2, \ldots, t_{K-1})$ to be any increasing sequence on the interval $[0, \tau]$ such that $t_{k+1} - t_k \leq g$, all $k = 1, 2, \ldots, K$, g a fixed number on $(0, \tau]$. Then, the following inequality implies (4.39)

$$a < \inf_{s \in [0,\tau]} \frac{(s-e)^+ + d}{s + (h-1)(d+e) + g}.$$

It can be checked that the infimum is attained for $s = e$. Hence, we can re-write the last display as

$$a < \frac{1}{(h-1) + \frac{he+g}{d}}. \tag{4.41}$$

Compare the last condition with the related condition of a worst-case deterministic approach in [15], $a < 1/(h-1)$. (4.41) is stronger.

Bounding the Probability of Buffer Overflow. We dimension the buffer of a node in the network such that the probability of the buffer overflow is not larger than a fixed $0 \leq \epsilon < 1$. Suppose we know an additional global parameter, a positive constant f such that

$$\frac{\sum_{i \in \mathcal{I}(n)} \sigma_i^2}{r_n^2} \leq f, \text{ any node } n.$$

We use a bound that follows from (4.17), for the fresh arrivals as described above. We have

$$\mathbf{P}[Q_n(0) > q_n] \leq \sum_{k=0}^{K-1} \exp\left(- \frac{[(\frac{q_n}{r_n} + (t_k - e)^+ - at_{k+1} - a\Delta)^+]^2}{2f} \right).$$

We require that the right-hand side is not larger than ϵ. The last is true, if for a fixed sequence of non-negative reals $(\epsilon_0, \epsilon_1, \ldots, \epsilon_{K-1})$ such that $\sum_{k=0}^{K-1} \epsilon_k = \epsilon$,

$$\left(d + (t_k - e)^+ - at_{k+1} - a\Delta\right) \vee 0 \geq \sqrt{\ln \epsilon_k^{-2} f}, \text{ all } k = 0, 1, \ldots, K-1.$$

Assume $a < 1/(h-1)$. Then, the last condition can be re-written as, for all $k = 0, 1, \ldots, K-1$,

$$d \geq \frac{1}{1-(h-1)a} \left(a(h-1)e + at_{k+1} - (t_k - e)^+ + \sqrt{\ln \epsilon_k^{-2} f} \right). \tag{4.42}$$

Take $\epsilon_k = \epsilon/K$, all $k = 0, 1, \ldots, K-1$. It is easily seen that (4.42) is implied by

$$d \geq \frac{1}{1-(h-1)a} \left(ahe + g + \sqrt{2(\ln K - \ln \epsilon) f} \right).$$

Now, for a fixed $K \in \mathbb{N}$, the right-hand side is minimized by the sequence $(\tau/K, 2\tau/K, \ldots, (K-1)\tau/K)$, over all increasing sequences $(t_1, t_2, \ldots, t_{K-1})$ such that $0 = t_0 < t_1$ and $t_{K-1} < t_K = \tau$. Thus, $g = \tau/K$. We have

$$d \geq \frac{1}{1-(h-1)a} \left(ahe + a\frac{\tau}{K} + \sqrt{2(\ln K - \ln \epsilon) f} \right).$$

Substituting the definition of τ (4.40) into the last display, and a simple re-arrangement, we can state;

Lemma 13 *Assume $a < 1/(h-1)$. Let $q_n = d' r_n$, all n,*

$$d' = \min_{K \in \mathcal{K}} \frac{a(b+ahe) + (1-a)[ahe + \sqrt{2(\ln K - \ln \epsilon) f}]K}{(1-a)[1-(h-1)a]K - a^2(h-1)}, \tag{4.43}$$

where $\mathcal{K} = \{k \in \mathbb{N} : \ k > a^2(h-1)/[1-(h-1)a]/(1-a)\}$. Then, the probability of the buffer overflow, at any node, is not larger than ϵ.

Assume that we fixed buffer lengths of the nodes in the network such that $q_n = d' r_n$, all n. From the delay-from-backlog property of PSRG nodes, we have that, a bound on the end-to-end delay jitter is, for $a < 1/(h-1)$, $h(d'+e)$.

In practice, one may be more interested in a bound on the bit loss rate than a bound on the probability of buffer overflow. We can indeed obtain the former from the latter, by (4.34). To that end, we need a bound on the vertical deviation of an arrival curve and a service curve to a node in the network. An arrival curve for arrival process to a node n is $f_n(t) = ar_n t + br_n + ar_n\Delta$, $t \geq 0$. The node n offers adaptive service curve $\beta_n(t) = r_n(t-e)^+$. We have

$$\frac{v(f_n, \beta_n)}{r_n} = b + a\Delta + ae.$$

From the delay-from-backlog property of PSRG nodes, and the fact that $v(f_n, \beta_n)$ is a bound on the maximum backlog at node n, we have that the delay at any node in the network is bounded by $\max_m\{v(f_m, \beta_m)/r_m\} + e$. Hence, $\Delta \leq (h-1)[\max_m\{v(f_m, \beta_m)/r_m\} + e]$. Combining this with the last above display, we have

$$\frac{v(f_n, \beta_n)}{r_n} \leq b + ahe + a(h-1) \max_m \left\{ \frac{v(f_m, \beta_m)}{r_m} \right\}, \text{ all } n.$$

The last can be re-written as, for $a < 1/(h-1)$,

$$\max_m \left\{ \frac{v(f_m, \beta_m)}{r_m} \right\} \leq \frac{b+ahe}{1-(h-1)a}. \tag{4.44}$$

Define the function

$$h(x) := \frac{1}{x} \sum_{k=0}^{K-1} \exp\left(-\frac{[(d' + (\tau k/K - e)^+ - x\tau(k+1)/K - x\Delta)^+]^2}{2f} \right).$$

Assume that

(H) $h(x) \leq h(a)$, for all $x \in (0, a]$.

In view of Lemma 13, (4.44), and (4.33), under (H), we have

$$\ell \leq \frac{[b - d' + a(h-1)d' + ahe]^+}{a[1-(h-1)a]}\epsilon. \tag{4.45}$$

The last inequality gives us a bound on the bit loss rate, with buffers dimensioned such that the probability of buffer overflow is bounded by ϵ for all nodes in the network.

Bounding the Bit Loss Rate. Suppose now we want to dimension buffers in a network such that $\ell \leq \epsilon$, for a fixed $0 \leq \epsilon < 1$. We can achieve this by adapting the results obtained earlier. From (4.34) and (4.44), we have that the bit loss rate ℓ_n at the node n is bounded as

$$\ell_n \leq \frac{b + ahe}{[1-(h-1)a](\rho'_n/r_n)}\mathbf{P}[Q_n(0) > q_n],$$

where ρ'_n is the intensity of bit arrivals to the node n.

Lemma 14 *Assume (H). In (4.43), replace ϵ with*

$$\epsilon\frac{a[1-(h-1)a]}{b+ahe}.$$

Then, under the same setting as in Lemma 13, we have that the bit loss rate, at any node, is not larger than ϵ.

Discussion. Compare our bound on the end-to-end delay jitter, hd', with a worst-case deterministic bound in [15] (Theorem V.1), which says: For $a < 1/(h-1)$, a bound on the end-to-end delay-jitter is

$$\frac{h}{1-(h-1)a}(b+e). \tag{4.46}$$

Our bound can be significantly smaller than (4.46); we confirm this by a numerical example in Section 4.9.3.

We expect our bound on the end-to-end delay-jitter, $h(d'+e)$, to become smaller under the many-sources asymptotics, defined by the scaling $q_n = I_n q_n^0$, $r_n = I_n r_n^0$, for some fixed $r_n^0 > 0$, $q_n^0 > 0$, and assuming that σ_i's are fixed. I_n is the number of arrival processes to the node n. Then, asymptotically, we would have $b \sim O(1)$, whereas $f \sim O(1/I)$.

4.9 Numerical Examples

4.9.1 Examples for Backlog Bounds in Section 4.4

We show numerical examples for leaky-bucket constrained arrival flows; for the ith flow $\alpha_i(t) = \bar{\alpha}_i t + \sigma_i$, $\bar{\alpha}_i > 0$, $\sigma_i \geq 0$. We assume packets lengths are fixed to

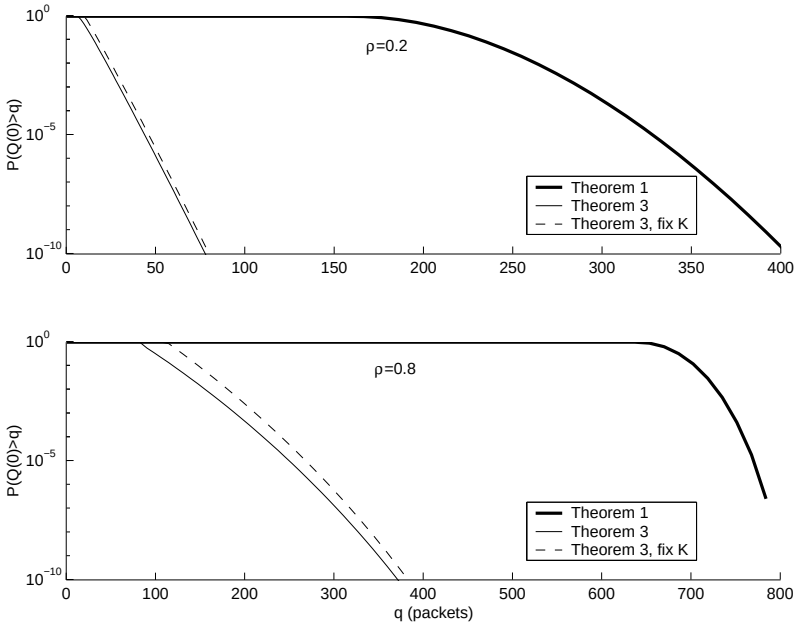


Figure 4.6: Bounds of Theorems 10 and 12 for the homogeneous setting with $I = 100$ arrival flows. The graphs are given for the loads: (Top) $\alpha = 0.2$, and (Bottom) $\alpha = 0.8$. The bound of Theorem 12 is computed for a uniform partition of $[0, \tau]$, for the optimum K, and K fixed to $\lceil \tau/e \rceil$.

$L = 1500$ bytes. We consider a node that offers the rate-latency service curve $\beta(t) = r(t-e)^+$, with rate $r = 150$ Mb/s, and latency $e = L/r$.

In our computations we partition the interval $[0, \tau]$ uniformly into K subintervals with $t_k = k\tau/K$, for $k = 0, 1, \ldots, K$. Then, we find $K \in \mathbb{N}$ that minimizes the bound on backlog.

Numerical Comparison of Bounds of Theorem 10 and Theorem 12

We set $\bar{\alpha}_1 = \rho r/I$ and $\sigma_1 = 8L$, where $0 < \rho < 1$ is the load of the node. We fix the number of arrival flows to $I = 100$, and the total load to $\rho = 0.2$ and 0.8. See Figure 4.6. We observe:

- the bound of Theorem 12 is much better than the bound of Theorem 10;
- minimizing the bound of Theorem 12 over the partitions of $[0, \tau]$ may yield a significant sharpening.

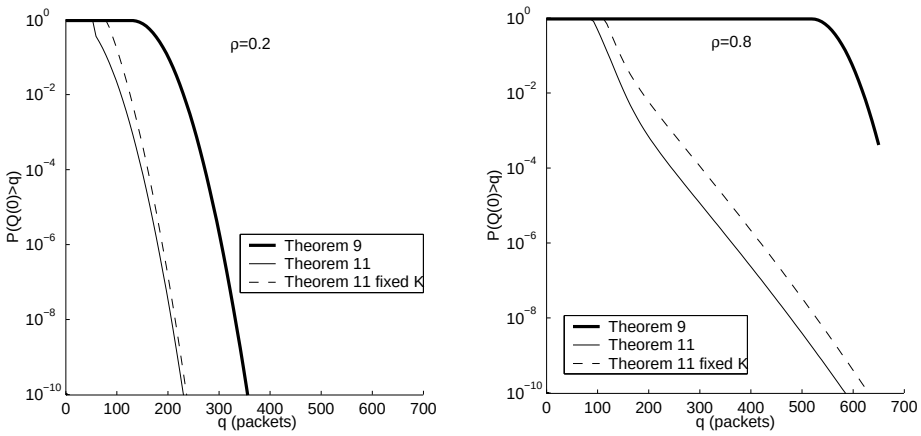


Figure 4.7: Bounds of Theorems 11 and 13 for the heterogeneous case of two classes of the input flows each consisting of 50 flows. The graphs are for the loads (Left) $\alpha = 0.2$ and (Right) $\alpha = 0.8$. Bound of Theorem 13 is for uniform partition of $[0, \tau]$; for the optimum K and K fixed to $\lceil \tau/e \rceil$.

Numerical Comparison of Bounds of Theorem 11 and 13

We fix the arrival flows to belong to two classes, each consisting of I_1 and I_2 flows, respectively. We fix $\bar{\alpha}_1 = 2\bar{\alpha}_2$, $\sigma_1 = 8L$, and $\sigma_2 = 5L$. (Here the subscript 1 (resp. 2) refers to the first (resp. second) class flow.) We set the number of flows of each class to be the same $I_1 = I_2 = 50$, and we fix the node loads to $\rho = 0.2$, 0.8. From Figure 4.7, we note:

- the bound of Theorem 13 is much better than the bound of Theorem 13;
- minimizing the bound of Theorem 13 over the partitions of $[0, \tau]$ may yield a significant sharpening.

Effect of the Node Latency

We show the numerical values of the bound of Theorem 12, for a node that offers the rate-latency service curve defined in this section. We take the latency from the discrete set of values $e = 0$, 4, 8 multiples of L/r. The case $e = 0$ is a special case; it corresponds to approximation of the node with a work-conserving, constant service rate server. See Figure 4.8. We observe:

- an approximation of the node with a constant service rate server can be over-optimistic. The discrepancy is larger for a smaller load ($\rho = 0.2$). In some cases, the discrepancy is about one order of magnitude.

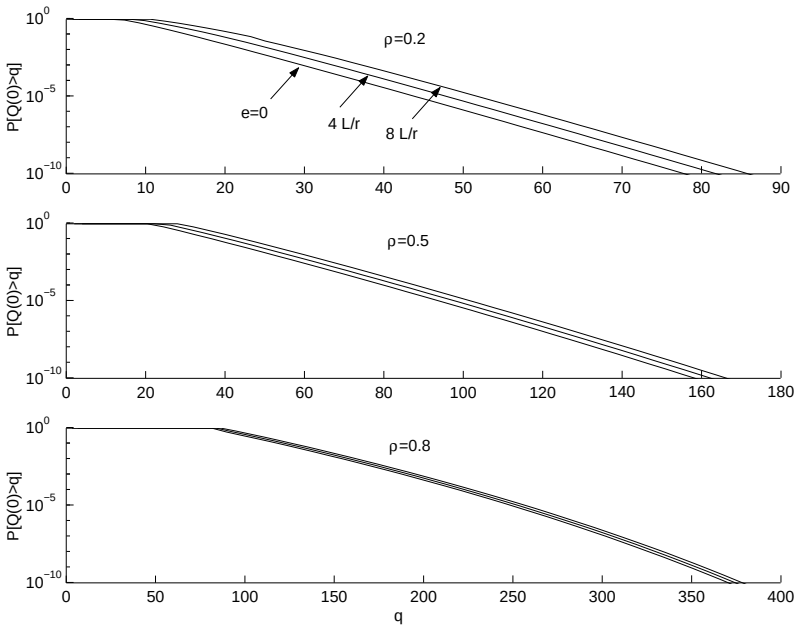


Figure 4.8: Bound of Theorem 12 for the homogeneous setting of $I = 100$ arrival flows. The node latencies are $e = 0$, 4, and 8 multiples of L/r. The graphs are given for the loads (Top) $\rho = 0.3$, (Middle) 0.5, and (Bottom) 0.8.

Bound of Theorem 13 Against the Bound of Theorem 12 in a Homogeneous Setting

From the results of Hoeffding [56], which we use in our derivations, we know that in a homogeneous setting, the bound of Theorem 13 is worse than the bound of Theorem 12. We substantiate this with a numerical example, where we also plot the bound of Theorem 14. From Figure 4.9, we observe:

- for a light to moderate load, the bound of Theorem 13 is significantly conservative with respect to the bound of Theorem 12;
- for a high load, the bound of Theorem 13 is near to the bound of Theorem 12, except for the buffer beyond a certain value, when it becomes more conservative.

We did some further experiments in order to evaluate the sharpness of the bounds that hold for arbitrary arrival curves, by comparing with respective bounds obtained under the assumption that arrival curves are the same. We assumed the flows are constrained with affine arrival curves, which corresponds to leaky-bucket regulation. We also compared with the first bound in Theorem 16. We do not show the numerical results here, for the benefit of space, but refer the interested reader to [109] (Figure 3 therein). The observations are:

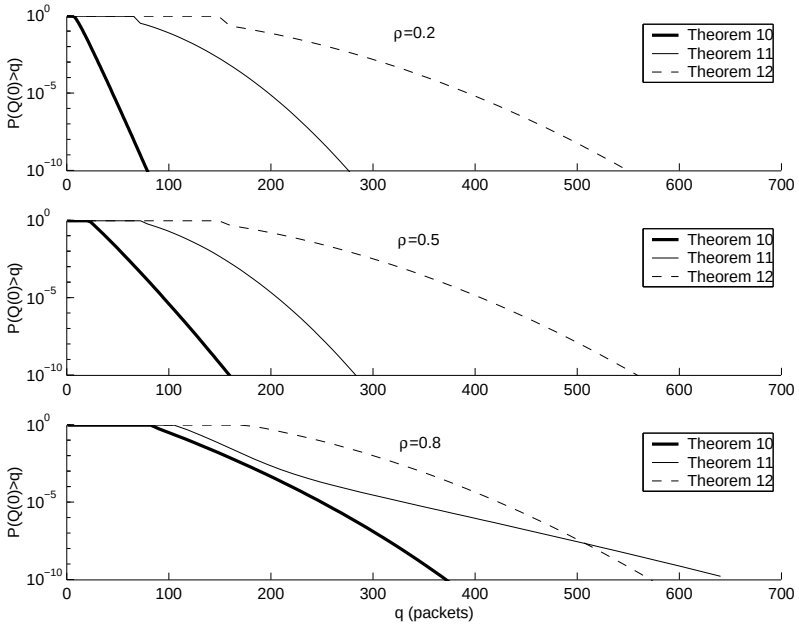


Figure 4.9: Bounds of Theorems 12, 13, and 14, for the homogeneous setting. The graphs are given for the loads $\rho = 0.2$, 0.5, and 0.8, top to bottom, respectively.

- the observations made from Figure 4.9 remain to hold;
- the qualitative observations made for the bound in Theorem 12 apply to the first bound of Theorem 16, except for a high load the latter bound is not conservative, in some cases it is even sharper than the bound of Theorem 12.

Comparison of Bounds of Theorem 16, Theorem 13, and Theorem 14

We compare by a numerical example the three bounds in Theorem 16 , the bound of Theorem 13, and the bound of Theorem 14. Recall that the first bound in Theorem 16 is always at least as good as the bounds of Theorem 13 and Theorem 14. The aim of the numerical example of this section is to evaluate the improvement of the first bound in Theorem 16 over the bounds in Theorem 13 and Theorem 14. Another aim is to assess the loss in sharpness of the second and third bound in Theorem 16 with respect to the first bound in the same theorem.

We consider a numerical example for a homogeneous and a heterogeneous setting, as defined in the caption of Figure 4.11. We observe:

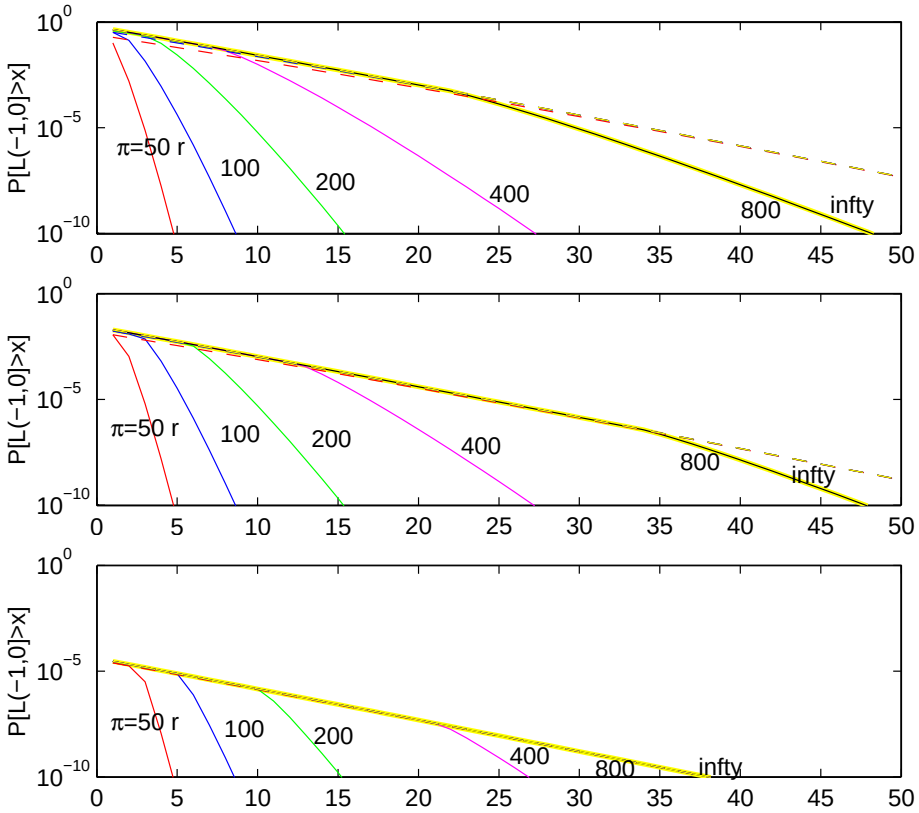


Figure 4.10: Bounds on the complementary distribution of the bit loss. The buffer capacity is set as: (Top) $q = 10L$, (Middle) $20L$, and (Bottom) $40L$. Packet lengths are fixed to $L = 1500$ bytes. The solid lines depict the bound of Theorem 21. The dashed lines depict the bound of Theorem 20. The bounds are plotted for different values of the peak-rate constraint π.

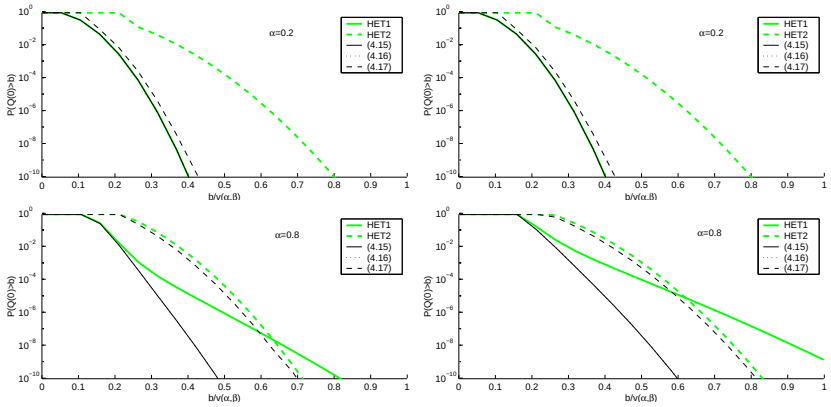


Figure 4.11: Bounds on backlog of Theorem 16, Theorem 13, and Theorem 14. The bounds are labeled as (HET1) Theorem 13, (HET2) Theorem 14, other bounds of Theorem 4.11.2. (Left) homogeneous setting, (Right) heterogeneous setting. The arrival curves are $\alpha_i(t) = \rho_i + \sigma_i$. The heterogeneous setting is for $\rho_1\rho_2 = 1/4$, $\sigma_1 = 2L$, and $\sigma_2 = 8L$. The node offers the rate-latency service curve with rate $r = 150$ Mb/s and latency $e = L/r$. $L = 1500$ bytes.

- the first bound of Theorem 16 yields a significant improvement over the bounds of Theorem 13 and Theorem 14. This is pronounced for a high load of the node;
- the second bound in Theorem 16 is near the first bound of the same theorem;
- the third bound in Theorem 16 is comparable with the first bound of the same theorem, for a light load. For a high load, it is conservative.

Comparison with the Better-than-Poisson Proposal

At the beginning of this chapter (see Section 4.1) we referred to an alternative proposal for dimensioning Expedited-Forwarding networks; the Better-than-Poisson [17]. In this section, we compare the first bound of Theorem 16 and the bound of Theorem 12, with the bound of [17]. The example is for a work-conserving node that offers a constant service rate r, we set $\beta(t) = rt$, $r = 150$ Mb/s. We consider a homogeneous setting of the arrival curves. The arrival curve of the aggregate arrival process to the node is fixed to $\alpha(t) = \rho t + \sigma$, $\rho = \alpha r$, $\sigma = 500L$, $L = 1500$ bytes. The parameter α is the load of the node, which is varied. An individual arrival flow in the aggregate is assumed constrained with the arrival curve $(\rho t + \sigma)/I$, where, recall I is the total number of the flows in the aggregate. We fix I to 100 and 500. Notice that for a work-conserving constant service rate node that we consider in our example, the

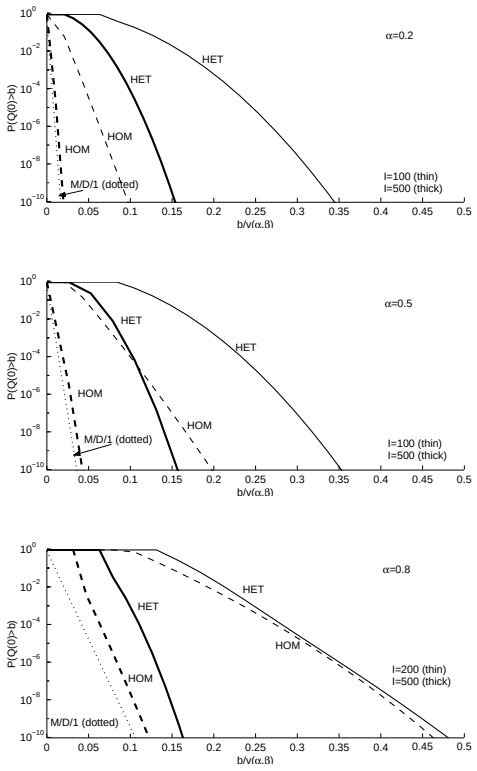


Figure 4.12: Comparison with the Better-than-Poisson proposal. (HET) the solid lines depict the first bound in Theorem 16, (HOM) the dashed lines are for the bound in Theorem 12, (M/D/1) the dotted lines depict the "Better than Poisson" prediction. The backlog bounds are a work-conserving constant service rate node, $\beta(t) = rt$. The load of the node is varied as (Top) $\alpha = 0.2$ (Middle), 0.5 (Bottom) 0.8.

complementary distribution of the backlog for the Better-than-Poisson proposal is exactly of an $M/D/1$ queue. We use the asymptotic expression in [17]. From Figure 4.12, we observe:

- for a non-large number of the arrival flows, the bound of the Better-than-Poisson proposal predicts smaller probability of the buffer overflow. The Better-than-Poisson proposal may be over-optimistic in this case;
- for a large number of arrival flows, our bounds become closer to the prediction of the Better-than-Poisson proposal.

The last observation would not come as a surprise. Note that for the given example we scale the intensity of bit arrivals and burstiness of individual arrival flows with the number of arrival flows as $O(1/I)$. The buffer capacity and the

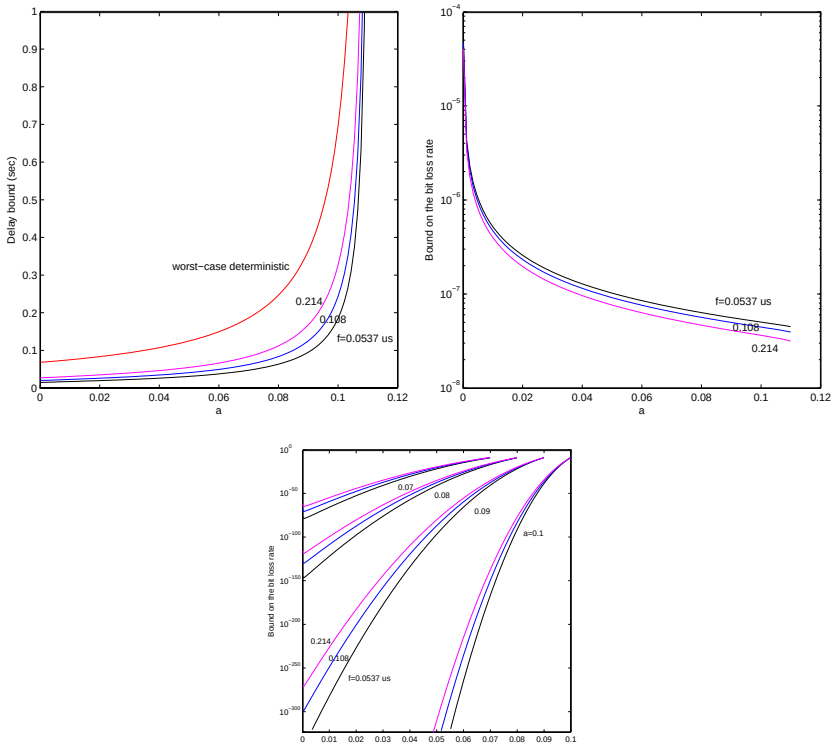


Figure 4.13: (Top, left) Bound on the end-to-end delay-jitter versus the bound on EF load at a node, (Top, right) bound on the bit loss rate, (Bottom) $h(\cdot)$ of the hypothesis (H) multiplied with $(b + ahe)/(1 - (h - 1)a)$. The results are for a network with each EF flow traversing at most ten hops ($h = 10$). $b = 6.4$ ms, $e = 0.384$ ms. Probability of the buffer overflow at a node is bounded with $\epsilon = 10^{-6}$.

service rate are fixed. This corresponds to the many-sources scaling, which would be obtained with arrival curves of individual flows fixed, and the buffer capacity and the service rate scaled as $O(I)$. It is readily observed that, under the many-sources scaling, our backlog bounds decrease exponentially with I.

4.9.2 Examples for Bounds on Loss in Section 4.7

We consider a node that offers either the adaptive service curve or service curve, $\beta(t) = rt$, $r = 150$ Mb/s. The buffer capacity of the node is denoted as q and is varied. The arrival process to the node is a superposition of $I = 100$ flows. Packet lengths are fixed to $L = 1500$ bytes. We assume all flows have the same arrival curve, $\alpha_i(t) = \min\{\pi t, \rho_1 t + \sigma_1\}$, where $\rho_1 = \alpha r/I$, $\sigma_1 = 8L$. π is a bound on the peak-rate, and is varied. α is a bound on the load of the node,

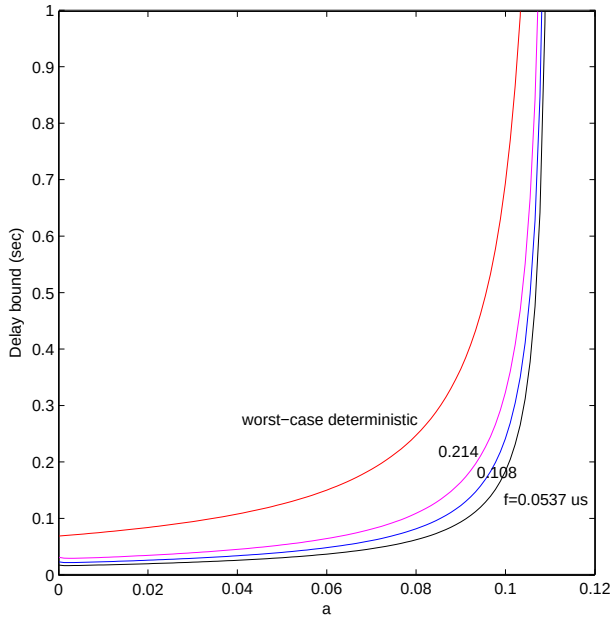


Figure 4.14: Bound on the end-to-end delay-jitter for the same scenario as in Figure 4.13, but with the bit loss rate bounded with $\epsilon = 10^{-6}$.

which we fix to 0.2. ρ_1 is an upper bound on the intensity of the arrival bits. σ_1 is the burstiness parameter. Time is discrete. Each time slot is of length L/r, a packet transmission time by the node.

We compare the bounds on loss of Theorem 20 and Theorem 21, which are for a node that offers a service curve, and an adaptive service curve, respectively. From Figure 4.10, we observe:

- the bound of Theorem 20 remains fairly insensitive to the value of π;
- in contrast, for the bound of Theorem 21, the smaller π is, the smaller the bound is.

Note that $B(s)$ in the bound of Theorem 21, Equation (4.32), accounts for arrival overflow over small timescales. It is exactly for these small timescales where the peak-rate constraint has a role; it decrease the term $B(s)$. It is to be expected from (4.32), the larger q is, the smaller the impact of $B(s)$ is. This is confirmed by the numerical results, see Figure 4.10.

4.9.3 Network of Nodes

In this section we give numerical examples for our bounds on the end-to-end delay-jitter in Section 4.8. In particular, we demonstrate that our bound on the

end-to-end delay-jitter can be significantly more optimistic than a previously-known worst-case deterministic bound [15].

We consider an Expedited-Forwarding network that obeys to $h = 10$, $b = 6.4$ ms, $e = 0.384$ ms, we vary $f = 0.214, 0.108, 0.0537$ μs, and a. An example family of networks that obeys to the given parameters is with each node in the network with a rate $r_0 I$, $r_0 = 0.78125$ Mb/s, $\sigma_i = 640$ B, where $I = 200, 400, 800$ is the number of the Expedited-Forwarding arrival processes to a node, for the respective values of the parameter f. Note that the given setting corresponds to the rates $156.25, 312.5, 625$ Mb/s, respectively. The setting is exactly the many-sources scaling. The latency of 0.384 ms, for a node with the rate 156.25 Mb/s, is equivalent to the transmission time of 5 packets of length 1500 B.

We first bound the probability of the buffer overflow by $\epsilon = 10^{-6}$, as given by Lemma 13, Section 4.8. See the numerical values for bounds on the end-to-end delay-jitter, in Figure 4.13 (Left). We observe:

- our bounds are significantly more optimistic than the worst-case deterministic bound (4.46);
- the smaller the f, the smaller the bound. The difference can be significant.

In Figure 4.13 (Middle) we show the bound on the bit loss rate (4.45). Figure 4.13 (Right) verifies the hypothesis (H) for some values of a.

Lastly, we bound the bit loss rate by $\epsilon = 10^{-6}$, as given in Lemma 14, Section 4.8. See Figure 4.14. We observe qualitatively the same results as earlier. The bounds on the end-to-end delay-jitter do not quantitatively differ much from the bounds seen in Figure 4.13 (Left).

4.10 Conclusions

- We obtained performance bounds that hold in probability for some node abstractions that contain the per-hop-behavior of the Expedited Forwarding networks.

- We obtained bounds on backlog for a node that offers a service curve. Some of these results generalize some bounds obtained by Kesidis and Konstantopoulos [70] and Chang, Song, and Chiu [26]. Our numerical computations indicate that the set of bounds that generalizes [70] is less sharp than the set that generalizes [26]. Hence, the latter set of the bounds should be preferred in practice.

- We found that all the bounds under the given assumptions on the arrival processes can be obtained by using known inequalities due to Hoeffding (1963) [56]. In particular, this implies that the bounds in [70] and [26] follow from Hoeffding's inequalities. Having realized that Hoeffding's inequalities can be directly used in the present context, enabled us to make use of some of the known inequalities, and obtain *new* bounds of interest; in particular, for the case of arbitrary arrival curves of the arrival processes to a node.

- We showed an exact bound on delay for a guaranteed rate node. The bound on delay holds under the assumption that the arrival process of bits to a node is stationary ergodic, we know the minimum packet length, and we know an upper bound on the expected packet length (a trivial bound that can be used is the maximum packet length).

- We derived bounds on the complementary distribution of the amount of bits lost on an arbitrary time interval, for a lossy node that offers a service curve and a node that offers an adaptive service curve. Under the latter, stronger node model, we obtained a better bound than under the former node model. We demonstrate this through numerical examples. Having a bound on the complementary distribution of the amount of the bits lost, on an arbitrary interval of time, enables us to compute a bound on the bit loss rate, that is, the fraction of arrival bits lost.

- We stress that all the bounds showed in this chapter are *exact*, they are *not* some asymptotic limits.

- We compared our bounds with the bounds of the Better-than-Poisson proposal, an alternative proposal for a probabilistic dimensioning of Expedited-Forwarding networks [17]. The results indicate that our backlog bounds become asymptotically close to the prediction of the Better-than-Poisson proposal, under a commonly used many-sources scaling. In situations where a node is fed with a few independent, individually regulated flows, our bounds may be significantly larger than the Better-than-Poisson prediction. In such situations, the Better-than-Poisson approach may be

over-optimistic. Note that the validity of our bounds is immune to the number of flows that constitute an arrival process to a node.

- We applied our single-node performance bounds to a network of nodes that conform to the per-hop-behavior of Expedited-Forwarding. We assume that Expedited Forwarding flows are stochastically independent, individually regulated at the network *ingress*, and each flow can traverse at most some configured number of nodes. We assumed to know bounds on some global network parameters, the same as found in [15], plus a bound on one additional global parameter. Our bounds are non-trivial under the same constraint on the load of the Expedited-Forwarding traffic to a node as in the worst-case deterministic result of [15]. This is a limitation, however, we do *not* know of an *exact* result that would hold without this constraint. Our probabilistic dimensioning consists in dimensioning the buffers of nodes such that the bit loss rate at any node is smaller than a configured value. With finite buffers of the nodes, we have a bound on the end-to-end delay-jitter that holds with probability one. Our numerical computations demonstrate the *gain* of our probabilistic over the worst-case deterministic dimensioning.

- As a by-product of our work, we found that a bound on delay for guaranteed rate nodes, obtained by Goyal, Lam, and Vin [48, 49] is incorrect. We gave a fix to the problem in [110].

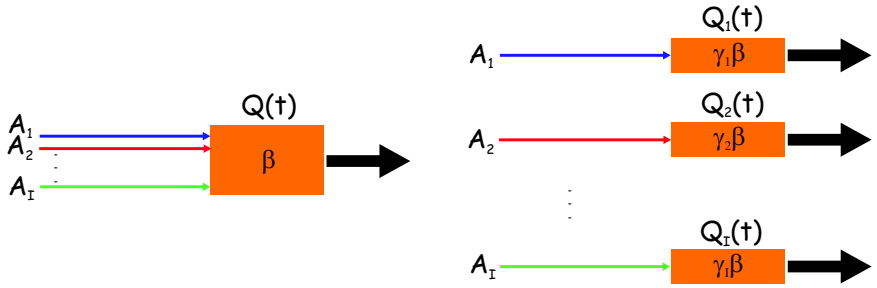


Figure 4.15: Bounding by the backlogs of the virtual seggregated system. The original system (left drawing) is a node that offers the service curve β to the aggregate of the arrival processes $A_1, A_2, \ldots, A_I$. The virtual system (right drawing) is a system of greedy shapers where the ith greedy shaper is with the service curve $\gamma_i\beta$, where γ_i is a positive-valued real number such that $\sum_{i=1}^{I} \gamma_i = 1$. Call $Q(t)$ the backlog at time t in the original system. Call $Q_i(t)$ the backlog in the ith greedy shaper in the virtual system. We have that for all t, $Q(t) \leq \sum_{i=1}^{I} Q_i(t)$.

4.11 Proofs

4.11.1 Proofs for Section 4.4.1

The next two proofs follow the lines of the bounding method introduced in Section 4.3.2. In the present context, the notation in Section 4.3.2 specializes to $\mathcal{A} = \{Q(0) > q\}$, $\mathcal{A}_i = \{Q_i(0) > q\}$, $q \geq 0$. For $i \in \mathcal{I}$, $Q_i(t)$, $t \in \mathbb{R}$, is a virtual queue introduced shortly. The second step of the bounding method is carried out by using Hoeffding's inequalities.

Proof of Theorem 10

Define, for $i \in \mathcal{I}$, $\gamma_i > 0$,

$$Q_i(t) = \sup_{-\infty < s \leq t} \{A_i(s,t] - \gamma_i\beta(t-s)\},\ t \in \mathbb{R}.$$

From (4.9), for any $(\gamma_1, \gamma_2, \ldots, \gamma_I) \in \mathbb{R}_+^I$ such that $\sum_{i=1}^{I} \gamma_i = 1$, we have, $Q(t) \leq \sum_{i=1}^{I} Q_i(t)$, $t \in \mathbb{R}$. (See Figure 4.15.) Hence,

$$\mathbf{P}[Q(0) > q] \leq \mathbf{P}[\sum_{i=1}^{I} Q_i(0) > q]. \tag{4.47}$$

Note the following properties;

(i) For any $t \in \mathbb{R}$,

$$Q_1(t), Q_2(t), \ldots, Q_I(t) \text{ are independent.} \tag{4.48}$$

(ii) For any $t \in \mathbb{R}$, all $i \in \mathcal{I}$,

$$0 \leq Q_i(t) \leq v(\alpha_i, \gamma_i \beta). \tag{4.49}$$

(iii) For any $t \in \mathbb{R}$,

$$\mathbf{E}[Q(t)] \leq \bar{\alpha} h(\alpha, \beta). \tag{4.50}$$

The property (i) follows from (A1). The property (ii) follows from (A2) and the definition of vertical deviation. We prove the property (iii) next. To that end, define, for any $t \in \mathbb{R}$,

$$V(t) = \inf\{v \in [0, s] : \ s \in \mathcal{S}(t), \ A^*(s, t] \geq A(s, t - v]\}.$$

Note that $V(t)$ is the virtual delay (sojourn time) of a bit that departs at instant t. If the node would be FIFO, then $V(t)$ is the delay of a bit that departs at t. It can easily be shown that for any $t \in \mathbb{R}$, $V(t) \leq h(\alpha, \beta)$.

Now, note, for any $t \in \mathbb{R}$, some $s \in \mathcal{S}(t)$, it holds

$$\begin{aligned} Q(t) &= A(s, t] - A^*(s, t] \\ &\leq A(s, t] - A(s, t - V(t)] \\ &\leq A(s, t] - A(s, t - h(\alpha, \beta)] = A(t - h(\alpha, \beta), t]. \end{aligned}$$

Taking expectation on both sides of the above display, and then combining with (A3), we recover (4.50). That is, we proved the property (iii).

Further, let $\gamma_i = 1/I$. By (4.47)-(4.50) and using (4.5) in the proof of Hoeffding's inequality (Theorem 1, [56]), we obtain that for any $\theta > 0$,

$$\mathbf{P}[Q(0) > q] \leq e^{-\theta q} \left(1 - \frac{\mathbf{E}[Q_1(0)]}{v(\alpha_1, \beta/I)} + \frac{\mathbf{E}[Q_1(0)]}{v(\alpha_1, \beta/I)} e^{\theta v(\alpha_1, \beta/I)}\right)^I.$$

The right-hand side in the last inequality is increasing with $\mathbf{E}[Q_1(0)]$. Now, by (4.50) applied to $Q_i(0)$, we obtain $\mathbf{E}[Q_i(0)] \leq \bar{\alpha}_i h(\alpha_i, \gamma_i \beta) = \bar{\alpha}_1 h(\alpha_1, \beta/I)$. It is easy to observe $h(\alpha_1, \beta/I) = h(\alpha, \beta)$ and $v(\alpha_1, \beta/I) = v(\alpha, \beta)/I$. We showed

$$\mathbf{P}[Q(0) > q] \leq \exp\left(-\sup_{\theta > 0} F(\theta)\right),$$

where

$$F(\theta) = q\theta - I \ln\left(1 - \bar{\alpha} \frac{h(\alpha, \beta)}{v(\alpha, \beta)} + \bar{\alpha} \frac{h(\alpha, \beta)}{v(\alpha, \beta)} e^{\theta \frac{v(\alpha, \beta)}{I}}\right). \tag{4.51}$$

Computing $\sup_{\theta > 0} F(\theta)$ yields the desired result.

Remark. Note that we could immediately apply Hoeffding's inequality (Theorem 1, [56]) to (4.47)-(4.50). However, the last part of the proof is given for the sake of a comparison with [70], which is done in the text.

Proof of Theorem 11

The proof builds upon the proof of Theorem 10. Given (4.47)-(4.50), the problem is equivalent to deriving an upper bound on the complementary distribution (4.47) of a sum of independent *non-uniformly* bounded random variables. From Hoeffding's inequality (Theorem 2, [56]) it follows that, for any $\underline{\gamma} \in \mathcal{G}$, and $q > \sum_{i=1}^{I} \mathbf{E}[Q_i(0)]$,

$$\mathbf{P}[Q(0) > q] \leq \exp\left(-\frac{2(q - \sum_{i=1}^{I} \mathbf{E}[Q_i(0)])^2}{\sum_{i=1}^{I} v(\alpha_i, \gamma_i\beta)^2}\right).$$

The right-hand side is increasing with $\sum_{i=1}^{I} \mathbf{E}[Q_i(0)]$, hence we can replace it with its upper bound $\sum_{i=1}^{I} \bar{\alpha}_i h(\alpha_i, \gamma_i\beta)$ and still have a bound. This recovers the inequality in (4.10), hence, it completes the proof.

4.11.2 Proofs for Section 4.4.2

The proofs of the above theorems require two lemmas, which we prove first, and then give proofs of the theorems. The proofs again follow the two steps in the bounding method of Section 4.3.2. Herein $\mathcal{A} = \{Q(0) > q\}$, $q > 0$. An intermediate step is Lemma 10, which establishes $\mathcal{A} \subseteq \mathcal{B}$, where $\mathcal{B} = \{Y(0) > q\}$, with $Y(t)$, $t \in \mathbb{R}$, defined below. Then, Lemma 11 shows $\mathcal{B} \subseteq \bigcup_i \mathcal{A}_i$, for $\mathcal{A}_i = \{A(-t_{i+1}, 0] > q + \beta(t_i)\}$, for a sequence $t_0 \leq t_1 \leq \ldots \leq t_K$, $K \in \mathbb{N}$, defined shortly.

Proof of Lemma 10

From (4.9), for an arbitrary time instant 0,

$$Q(0) \leq \max\{X(0), Y(0)\},$$

where

$$\begin{aligned} X(t) &:= \sup_{-\infty < s \leq t-\tau} \{A(s,t] - \beta(t-s)\}, \\ Y(t) &:= \sup_{t-\tau < s \leq t} \{A(s,t] - \beta(t-s)\}. \end{aligned}$$

From (A2), $X(0) \leq \sup_{-\infty < s \leq -\tau} \{\alpha(-s) - \beta(-s)\}$. Now, assume τ satisfies (A6), then we conclude $X(0) \leq 0$. It follows, $Q(0) \leq \max\{0, Y(0)\}$. Hence, for any $q \geq 0$,

$$\{Q(0) > q\} \subseteq \{\max\{0, Y(0)\} > q\} = \{Y(0) > q\},$$

which by definition of $Y(0)$ recovers the assertion of the lemma.

Proof of Lemma 11

Fix $K \in \mathbb{N}$ and and a sequence $0 = t_0 \leq t_1 \leq \ldots \leq t_K = \tau$. Assume τ satisfies (A6). Note, for any s such that $t_k \leq s < t_{k+1}$,

$$A(-\tau, -s] \geq A(-\tau, -t_{k+1}] \text{ and } \beta(s) \geq \beta(t_k).$$

Hence,

$$\begin{aligned} \sup_{0 \leq s \leq \tau} \{A(-s, 0] - \beta(s)\} &= \max_{k \in \{0,\ldots,K-1\}} \{ \sup_{t_k \leq s \leq t_{k+1}} \{A(-s, 0] - \beta(s)\}\} \\ &\leq \max_{k \in \{0,\ldots,K-1\}} \{A(-t_{k+1}, 0] - \beta(t_k)\}. \end{aligned} \tag{4.52}$$

Let $\mathcal{A} := \{\sup_{0 \leq s \leq \tau} \{A(-s, 0] - \beta(s)\} > q\}$, for a fixed $q \geq 0$. By the inequality above, we obtain

$$\begin{aligned} \mathcal{A} &\subseteq \{ \max_{k \in \{0,\ldots,K-1\}} \{A(-t_{k+1}, 0] - \beta(t_k)\} > q\} \\ &= \bigcup_{k \in \{0,\ldots,K-1\}} \{A(-t_{k+1}, 0] > q + \beta(t_k)\}. \end{aligned} \tag{4.53}$$

This completes the proof.

Proof of Theorem 12

By the hypothesis of the theorem, (A2), for any $s, t \in \mathbb{R}$, and any $i \in \mathcal{I}$, $A_i(s, t]$ is uniformly bounded with $\alpha_1(t - s)$. Thus, the kth summation term in (4.11) is the complementary distribution of a sum of independent uniformly bounded random variables. By Hoeffding's inequality (Theorem 1, [56]), the assumption (A3), $\mathbf{E}[A_i(-t_{k+1}, 0]] \leq \bar{\alpha}_i t_{k+1}$, the kth summation term in (4.11) is upper-bounded by $\exp(-Ig(t_k, t_{k+1}))$, for $q > \bar{\alpha} t_{k+1} - \beta(t_k)$. This proves the result.

Proof of Theorem 13

By Hoeffding's inequality (Theorem 2, [56]), the kth summation term in (4.11) is upper-bounded with

$$\exp\left(-\frac{2(q + \beta(t_k) - \mathbf{E}[A(-t_{k+1}, 0]])^2}{\sum_{i=1}^{I} \alpha_i^2(t_{k+1})} \right).$$

For $q > \mathbf{E}[A(-t_{k+1}, 0]] - \beta(t_k)$, the last display is wide-sense increasing with $\mathbf{E}[A(-t_{k+1}, 0]]$. From (4.7), $\mathbf{E}[A(-t_{k+1}, 0]] \leq \bar{\alpha} t_{k+1}$, thus for $q > \bar{\alpha} t_{k+1} - \beta(t_k)$ we can replace $\mathbf{E}[A(-t_{k+1}, 0]]$ with $\bar{\alpha} t_{k+1}$ and still have an upper bound.

Proof of Theorem 14

Let $t \in \mathbb{R}$ and $\epsilon > 0$,

$$\tilde{Q}_i^\epsilon(t) := \sup_{-\infty < s \le t} \{A_i(s,t] - (1+\epsilon)\bar{\alpha}_i(t-s)\},\ i \in \mathcal{I},$$

and $Z_i^\epsilon(t) := \tilde{Q}_i^\epsilon(0) - \tilde{Q}_i^\epsilon(-t)$. Note, by (A3bis) and (A2), $\mathbf{E}[A_i(0,1]] < (1+\epsilon)\bar{\alpha}_i$ and thus $\tilde{Q}_i^\epsilon$ is stable. The last implies $\mathbf{E}[Z_i^\epsilon(t)] = 0$, any $t \in \mathbb{R}$. Note, from (A2),

$$-v(\alpha_i, \lambda_{\bar{\alpha}_i}) \le Z_i^\epsilon(t) \le v(\alpha_i, \lambda_{\bar{\alpha}_i}),\ t \in \mathbb{R}.$$

Now, observe, for any $s, t \in \mathbb{R}$, $s \le t$,

$$\tilde{Q}_i^\epsilon(t) - \tilde{Q}_i^\epsilon(s) \ge A_i(s,t] - (1+\epsilon)\bar{\alpha}_i(t-s).$$

Hence, for $t \in \mathbb{R}$, $Z_i^\epsilon(t) \ge A_i(-t,0] - (1+\epsilon)\bar{\alpha}_i t$, and thus

$$\begin{aligned} \mathbf{P}[A(-t,0] - (1+\epsilon)\bar{\alpha}t > z] &\le \mathbf{P}[\sum_{i=1}^I Z_i^\epsilon(t) > z] \\ &\le \exp\left(-\frac{z^2}{2\sum_{i=1}^I v(\alpha_i, \lambda_{\bar{\alpha}_i})^2}\right). \end{aligned} \quad (4.54)$$

The last inequality is by a Hoeffding's inequality (Theorem 2, [56]) applied to the sum of independent random variables, each with a bounded support and zero-mean.

Finally, from (4.54) and a simple variable substitution, we have, for any $u, v \ge 0$, $q \ge (1+\epsilon)\bar{\alpha}(v) - \beta(u)$,

$$\mathbf{P}[A(-v,0] > q + \beta(u)] \le \exp\left(-\frac{(q+\beta(u) - (1+\epsilon)\bar{\alpha}v)^2}{2\sum_{i=1}^I v(\alpha_i, \lambda_{\bar{\alpha}_i})^2}\right).$$

By continuity of the right-hand side, we can let $\epsilon \to 0$, and then combining with the lemmas 10 and 11, we complete the proof.

Proof of Theorem 16

The inequality in (4.15) is a corollary of Theorem 15 for leaky-bucket regulated inputs. The second inequality is obtained by upper-bounding the first term in the minimum operation in (4.15) as follows

$$\begin{aligned} \sum_{i=1}^I (\rho_i s_{k+1} + \sigma_i)^2 &= \sum_{i=1}^I \rho_i^2 s_{k+1}^2 + 2\sum_{i=1}^I \rho_i \sigma_i s_{k+1} + \sum_{i=1}^I \sigma_i^2 \\ &\le \sum_{i=1}^I \rho_i^2 s_{k+1}^2 + 2\sqrt{\sum_{i=1}^I \rho_i^2}\sqrt{\sum_{i=1}^I \sigma_i^2} s_{k+1} + \sum_{i=1}^I \sigma_i^2 \end{aligned}$$

$$= \left(\sqrt{\sum_{i=1}^{I} \rho_i^2} s_{k+1} + \sqrt{\sum_{i=1}^{I} \sigma_i^2} \right)^2 .$$

The last inequality is by Cauchy-Schwartz's inequality.

The inequality (4.17) follows from (4.16) by a trivial inequality $\sum_{i=1}^{I} \rho_i^2 \leq \rho^2$. Similarly, (4.18) is obtained from (4.17) by $\sum_{i=1}^{I} \sigma_i^2 \leq \sigma^2$. This completes the proof of the theorem.

4.11.3 Proofs for Section 4.5

Let $\tilde{A}^* = A \otimes \beta$. $A \otimes \beta$ is called the min-plus convolution of A and β, defined by $(A \otimes \beta)(t) = \inf_{u \in [0,t]} \{A(t-u) + \beta(u)\}$. By [26, 111], the infimum is obtained for $u \in [0, \tau]$, thus $\tilde{Q}(t)$ defined by the right-hand side in Equation (4.9), satisfies $\tilde{Q}(t) = A(t) - \tilde{A}^*(t)$.

We now state and prove a preparatory lemma, and then continue with the proof of the theorem. Recall the definition of $\hat{\beta}$ in (4.19).

Lemma 15 *We have*

$$\tilde{A}^*(t, t+u] \leq u\hat{\beta}.$$

Proof 4 *Define* $\gamma(u) = u\hat{\beta}\mathbf{1}_{\{u \geq 0\}}$. *It follows from the definition of* $\hat{\beta}$ *that, for all* $0 \leq s \leq t$,

$$\beta(t-s) + \gamma(s) \geq \beta(t).$$

Thus

$$\beta \otimes \gamma \geq \beta.$$

It follows that

$$\tilde{A}^* = A \otimes \beta \leq A \otimes (\beta \otimes \gamma) = (A \otimes \beta) \otimes \gamma = \tilde{A}^* \otimes \gamma.$$

Coming back to the definition of $\otimes$ *we find that*

$$\tilde{A}^*(t, t+u] \leq u\hat{\beta}.$$

Let $\varphi : \mathbb{R} \to \mathbb{R}_+$ be a wide-sense increasing function. Let $\varphi' = \psi$. We can write the evolution equation, for $\tilde{Q}(t)$, $t \in \mathbb{R}_+$,

$$\varphi(\tilde{Q}(t)) - \varphi(\tilde{Q}(0)) = \int_0^t \varphi'(\tilde{Q}(s-))\tilde{Q}(ds), \tag{4.55}$$

where $\tilde{Q}(ds) = A(ds) - \tilde{A}^*(ds)$. It follows from Lemma 15,

$$\int_0^t \psi(\tilde{Q}(s))\tilde{A}^*(ds) \leq \hat{\beta} \int_0^t \psi(\tilde{Q}(s-))ds. \tag{4.56}$$

Combining (4.55) with (4.56) we obtain

$$\varphi(\tilde{Q}(t)) - \varphi(\tilde{Q}(0)) \geq \int_0^t \psi(\tilde{Q}(s-))A(ds) - \hat{\beta}\int_0^t \psi(\tilde{Q}(s-))ds.$$

Taking expectation at both sides, by the stationarity hypothesis (A3bis), we obtain

$$0 \geq \rho t \mathbf{E}_A^0[\varphi'(\tilde{Q}(0-))] - \hat{\beta} t \mathbf{E}[\varphi'(\tilde{Q}(0))].$$

The Palm expectation is by Campbell's formula [9]. Replacing φ' with ψ we prove (4.20).

4.11.4 A Better Bound on Backlog as Seen by Bit Arrivals

Admit the setting of Section 4.5.2. Here we show a bound on the backlog as seen by a subset of the bit arrivals. Indeed, $\tilde{Q}(0-) \geq A(-u,0) - \beta(u)$, all $u \geq 0$. This yields, for any $u \geq 0$,

$$\mathbf{P}_{A^1}^0[\tilde{Q}(0-) > q] \geq \mathbf{P}_{A^1}^0[A(-u,0) > \beta(u) + q] \geq \mathbf{P}_{A^1}^0[A^0(-u,0) > \beta(u) + q].$$

By definition of the Palm probability,

$$\mathbf{P}_{A^1}^0[A^0(-u,0) > \beta(u) + q] = \frac{1}{\rho - \rho^0}\mathbf{E}[\int_{(0,1]} \mathbf{1}_{\{A^0(-u+s,s) > \beta(u)+q\}} A^1(ds)].$$

From the last identity, by independence of A^0 and A^1, we obtain

$$\mathbf{P}_{A^1}^0[A^0(-u,0) > \beta(u) + q] = \mathbf{P}[A^0(-u+s,s) > \beta(u) + q].$$

Putting the pieces together, we have

$$\mathbf{P}_{A^0}^0[\tilde{Q}(0-) > q] \leq \frac{\hat{\beta}}{\rho^0}\mathbf{P}[\tilde{Q}(0) > q] - \left(\frac{\rho}{\rho^0} - 1\right)\sup_{u \geq 0} \mathbf{P}[A^0(-u,0) > \beta(u) + q].$$

The new element on the right-hand side cannot be bounded by assuming only that the arrival processes to the node are with *upper* bounds on their increments. However, the bound may be of interest under some other assumptions on the arrival processes.

4.11.5 Proofs for Section 4.7

Proof of Theorem 20

Note, $L(-1,0] = L(-1,0]\mathbf{1}_{Q(0)=q} \leq L(s,0]\mathbf{1}_{Q(0)=q}$, for all $s \leq 0$. It follows from the definition of a service curve β, that there exists some $s \in [S(0),0]$ such that

$$L(s,0] \leq A(s,0] - \beta(-s) - Q(0).$$

If $s = 0$, then $Q(0) \leq -\beta(0) \leq 0$, and thus $L(-1,0] = 0$. As a result, we can write, for some $s \in [S(0),-1]$,

$$\begin{aligned} L(-1,0] &\leq [A(s,0] - \beta(-s) - q]\mathbf{1}_{Q(0)=q} \\ &\leq [A(s,0] - \beta(-s) - q] \vee 0. \end{aligned}$$

Then, for a $x > 0$,

$$\{L(-1,0] \geq x\} \subset \bigcup_{s\in[-\tau(x),-1]} \{A(s,0] \geq \beta(-s) + q + x\}. \tag{4.57}$$

Above we reduce the union of events from $[S(0),-1]$ to $[-\tau(x),-1]$. We are allowed to do so by the definition of $\tau(x)$ (4.30). (4.31) is obtained from (4.57) by the union bound.

Proof of Theorem 21

The proof follows closely a related result by Cruz and Liu [33], which is for a work-conserving constant service rate server. The definition of an adaptive service curve β is equivalent to: for an arbitrary instant 0, all $s \in [S(0),0]$, either

$$A^*(s,0] \geq \beta(-s), \tag{4.58}$$

or

$$\exists\, u \in [s,0]:\ A^*(S(0),0] \geq A'(S(0),u] + \beta(-u). \tag{4.59}$$

Note that the inequality in (4.58) can be written as

$$A(s,0] \geq \beta(-s) + Q(0) - Q(s) + L(s,0], \tag{4.60}$$

and, the inequality in (4.59) as:

$$A(u,0] \geq \beta(-u) + Q(0) + L(u,0]. \tag{4.61}$$

In the remainder, fix a $x > 0$. Note,

$$\{L(-1,0] \geq x\} \subset \{Q(0) = q\}. \tag{4.62}$$

Define

$$V(t) = \max[v \in [S(t),t] | A^*(S(t),t] \geq A'(S(t),v] + \beta(-v)]. \tag{4.63}$$

In other words, $V(0)$ is the largest integer in $[S(0),0]$ such that the inequality in (4.59) is true. $V(0)$ is well defined. To check this, impose as a hypothesis $A^*(S(0),0] - A'(S(0),u] < \beta(-u)$, all $u \in [S(0),0]$, that is (4.59) is not true. Then, it must be that (4.58) is true, that is $A^*(s,0] \geq \beta(-s)$, all $s \in [S(0),0]$. Pick $s = S(0)$, then $A^*(S(0),0] \geq \beta(-S(0))$, which contradicts the hypothesis.

From (4.61) and (4.63),

$$L(-1,0] \leq L(-V(0),0] \leq A(V(0),0] - \beta(-V(0)) - Q(0).$$

Thus, by (4.62),

$$\{L(-1,0] \geq x\} \subset \{A(V(0),0] \geq \beta(-V(0)) + q + x\}\,. \tag{4.64}$$

From the definition of $V(\cdot)$ (4.63), for $s \in [S(0),0]$,

$$\begin{aligned}
\{V(0)=s\} \quad &\Leftrightarrow \quad \{A^*(S(0),0] - A'(S(0),s] \geq \beta(-s)\} \cap \\
&\qquad \bigcap_{u\in[s+1,0]} \{A^*(S(0),0] - A'(S(0),u] < \beta(-u)\} \\
&\subset \quad \{A^*(S(0),0] - A'(S(0),s] \geq \beta(-s)\} \cap \\
&\qquad \bigcap_{u\in[s,0]} \{A^*(u,0] \geq \beta(-u)\} \\
&\subset \quad \{A^*(S(0),0] - A'(S(0),s] \geq \beta(-s)\} \cap \\
&\qquad \bigcap_{u\in[s,-1]} \{A^*(u,0] \geq \beta(-u)\} \\
&\Leftrightarrow \quad \{A(s,0] \geq \beta(-s) + Q(0) + L(s,0]\} \cap \\
&\qquad \bigcap_{u\in[s,-1]} \{A(u,0] \geq \beta(-u) + Q(0) - Q(u) + L(u,0]\} \qquad (4.65)
\end{aligned}$$

The first containment in the last display is by definition of β ((4.58) and (4.59)): it must hold, for all $u \in [V(0),0]$, $A^*(u,0] \geq \beta(-u)$. Note that whenever (4.59) is verified with $u = s$, then (4.58) holds as well.

From the final containment in (4.65), (4.64), and (4.62), we conclude, for $s \in [S(0),0]$,

$$\begin{aligned}
\{L(-1,0] \geq x, V(0)=s\} \quad &\subset \quad \{A(s,0] \geq \beta(-s) + q + x\} \cap \\
&\qquad \bigcap_{u\in[s,-1]} \{A(u,0] \geq \beta(-u) + x\}\,.
\end{aligned}$$

Next, one can easily check $\{V(0)=0\} \subset \{Q(0)=0\}$, and thus together with (4.62), $\{L(-1,0] \geq x, V(0)=0\} = \emptyset$. Hence,

$$\begin{aligned}
&\{L(-1,0] \geq x\} \\
\subset \quad &\bigcup_{s\in[S(0),-1]} \{A(s,0] \geq \beta(-s) + q + x\} \bigcap_{u\in[s,-1]} \{A(u,0] \geq \beta(-u) + x\}\,.
\end{aligned}$$

Finally, (4.32) follows by the union bound and the intersection bound $\mathbf{P}[\bigcap_n A_n] \leq \min_n \mathbf{P}[A_n]$.

Chapter 5

Input-Queued Switch

In this chapter, we:

- obtain bounds on the latency for a decomposition-based input-queued switch.

5.1 Introduction and Outline

We consider a switch with a given number of input and output ports. We assume that at any point in time, a transmission from an input port can be to at most one output port, and an output port can receive from at most one input port. In practice, these constraints are imposed by the *crossbar*; see Figure 5.1 for a standard graphical representation of a crossbar switch. The problem is to schedule transmissions between input/output port pairs of the switch, subject to the crossbar constraints. We assume that the crossbar connectivity is scheduled with the rate equal to the line rate of an input port of the switch. The last means that we consider an *input-queued* switch. The core routers in the Internet are commonly built as input-queued switches. A desirable feature of input-queued switches is that the scheduling rate is equal to the line rate of *one* input port. Some other architectures require the scheduling rate to be larger than the line rate of one input port. For instance, an output-queued switch may require the scheduling rate to be equal to the *sum* of line rates of *all* input ports. Apparently, this may be prohibitively expensive with large line rates of the input ports, at this moment, the state-of-the-art is the line rates in the order of tens of giga bits-per-second.

We consider a particular class of the switch scheduling algorithms we call the *decomposition-based* algorithms. With decomposition-based algorithms, we are given a matrix of the rates, which have to be offered across the input/output port pairs in the *long-run*. If a schedule is such that for an input/output port pair ij, the number of the offered slots, in the long-run, is at least the value given by the rate matrix, say r_{ij}, the schedule offers the *rate guarantee*. The schedule is said to offer the *latency guarantee*, with a latency e, if for an input/output

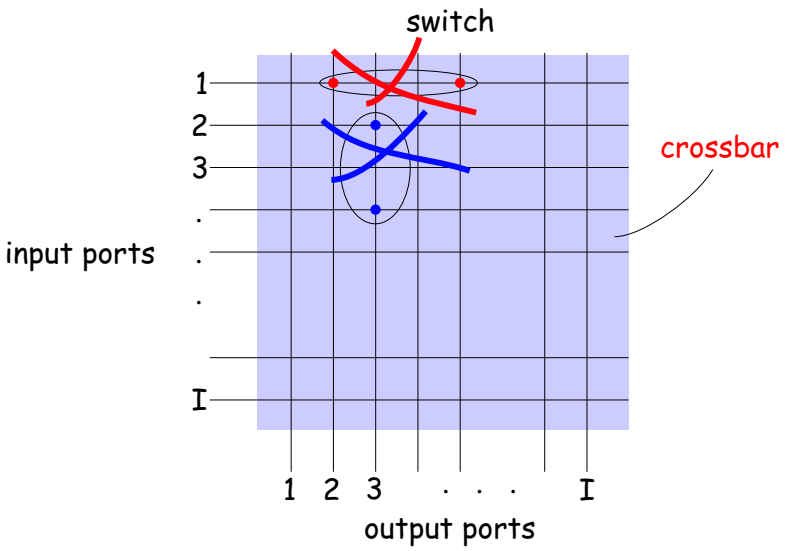


Figure 5.1: Switch *crossbar* constraints: At all instants, an input port can transmit to at most one output port and an output port can receive from at most one input port.

port pair ij, the number of the offered slots, over any interval of length m, is at least $r_{ij} \max\{m - e, 0\}$. In other words, the schedule offers the latency e, if for all input/output port pairs ij, from any time slot, the next slot offered to the input/output port pair ij arrives *no later* than in the next e slots. One may also introduce some other definitions of the latency guarantee, as we do in Section 5.2.1.

For the case of high-speed optical switches, it is reasonable to assume that the arrival rates are known to us because such switches are likely to be deployed in the core of networks where traffic engineering using Multi-Protocol Label Switching (MPLS) is becoming more prevalent. Switches that support MPLS must be able to provide *rate guarantees* for certain input-output pairs. For each MPLS path that passes through input i and output j on a switch, the switch will be required to *reserve* bandwidth for the path between these two ports. Another justification that the input rates are known can be found with Expedited Forwarding (EF) in the context of differentiated services [38]. There it is commonplace to assume the network is engineered such that the load of EF traffic at each node is bounded by some configured value. It is noteworthy that for a node to support EF, it needs to conform to a rigorous definition of the per-hop-behavior, namely, Packet Scale Rate Guarantee with a rate r and a latency e; see Chapter 4, Section 4.2. Our work can be viewed as calculating what the value of the latency e would be for an input-queued switch.

Another approach for service provisioning with input-queued switches is em-

ulation of output-queuing (for which the scheduling can be realized by known low-jitter single-server fair schedulers). It is shown by Charny *et al.* [29] and Chuang *et al.* [32] that it is sufficient to have a speed-up of 2 in configuration of the switch fabric to exactly emulate output-queueing behavior. However, with some high-speed switches the speed-up may be less than 2.

The total amount of bandwidth that needs to be reserved for engineered traffic will depend on the mix between engineered traffic and best effort traffic. (Note that the input-output rates for best effort traffic can change over time.) In this chapter we consider *non-idle* scheduling of the *entire* bandwidth of the switch for engineered traffic. However, this does not mean that the entire bandwidth is actually assigned to the engineered traffic.

One challenge is to schedule transmissions of an input-queued switch so as the rate and latency guarantees are offered between input/output port pairs of the switch. Our prime goal is the construction of *low-latency* decomposition-based schedules for *reserved* traffic.

5.1.1 Outline of the Chapter

In Section 5.2 we define the problem, introduce our notation and assumptions, and outline our results. Section 5.3 gives some preliminary results that are used in the remainder of this chapter. The main results of the chapter are given in Section 5.4, where we define our four schedulers and obtain latencies for them. Section 5.5 shows numerical values of some of our latency bounds for specific rate matrices and compares our results with some related work. We give our conclusions in Section 5.6.

5.2 Assumptions and Notations

We consider an $I \times I$ switch. Let ρ_{ij} be the bandwidth that needs to be reserved between input i and output j, normalized by the link rate. Let M be the matrix whose ij entry is ρ_{ij}. We refer to M as the *rate matrix*. In this chapter, we mostly consider the case in which M is a *doubly stochastic* matrix, i.e. $\sum_i \rho_{ij} = 1$ and $\sum_j \rho_{ij} = 1$. This corresponds to the case in which the entire bandwidth of the switch is reserved. However, we shall sometimes consider the sub-stochastic case in which we only have $\sum_i \rho_{ij} \leq 1$ and $\sum_j \rho_{ij} \leq 1$. In this case the residual bandwidth of the switch could be used by best effort traffic.

By standard results of Birkhoff and von Neumann (see e.g. [27]) we can decompose the matrix M into a convex combination of permutation matrices,

$$M = \sum_{k=1}^{K} \varphi_k M_k,$$

where $K \leq I^2 - 2I + 2$. Here, M_k is a permutation matrix (a $0-1$ matrix with exactly one "1" in each row and column), φ_k is the *rate* of matrix M_k and $\sum_{k=1}^{K} \varphi_k = 1$. Let S_{ij} be the set of matrices in the decomposition that have a

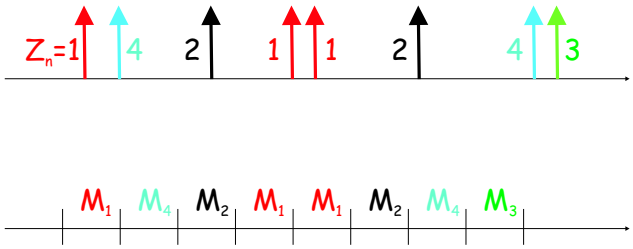


Figure 5.2: (Top) The token process. T_n is the time at which the nth token appears. Z_n is the type of the nth token, i.e. if the nth token corresponds to permutation matrix M_k then $Z_n = k$. (Bottom) The corresponding schedule.

1 in the ij position. Then $\rho_{ij} = \sum_{k \in S_{ij}} \varphi_k$. Our aim is to create a schedule in which exactly one of the permutation matrices is scheduled in each time slot. Input-output pair ij is served whenever a matrix from the set S_{ij} is scheduled. Hence we require that a matrix from S_{ij} is scheduled approximately once every $1/\rho_{ij}$ time slots.

As an example, suppose that,

$$M = \begin{pmatrix} 1/6 & 5/6 & 0 \\ 1/2 & 1/6 & 1/3 \\ 1/3 & 0 & 2/3 \end{pmatrix}.$$

Then,

$$M = \frac{1}{2}M_1 + \frac{1}{3}M_2 + \frac{1}{6}M_3,$$

where,

$$M_1 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, M_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, M_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and a possible schedule is,

$$M_1, M_2, M_1, M_2, M_1, M_3, M_1, M_2, M_1, M_2, M_1, M_3, \ldots$$

The class of schedulers that we consider can be formulated by the following unifying framework. We first place *tokens* for each matrix M_k in continuous time. We schedule matrix M_k in time slot n if the nth token to appear corresponds to matrix M_k. (See Figure 5.2).

More formally, we associate with M_k a *counting process* N_k defined on $\mathbb{R}_+$. For an interval $\mathcal{I} \subseteq \mathbb{R}_+$, $N_k\mathcal{I}$ equals the number of tokens for M_k that land in interval $\mathcal{I}$. We require that N_k has *intensity* φ_k, i.e. $\lim_{t\to\infty} N_k[0,t)/t = \varphi_k$.

We define the superposition process $N\mathcal{I} = \sum_{k=1}^{K} N_k\mathcal{I}$. Let $\{T_n\}_{n\geq 0}$ be the point processes on $\mathbb{R}_+$ defined by the counting process N. Next, let $\{Z_n\}_{n\geq 0}$

be a sequence of marks such that $Z_n = k$ iff the n-th point of the superposition point process, T_n, belongs to N_k. Let $N_{ij}\mathcal{I}$ be the number of tokens for input-output pair ij that land in the interval $\mathcal{I}$, i.e. $N_{ij}\mathcal{I} = \sum_{k \in S_{ij}} N_k\mathcal{I}$. Conversely, let $N_{\overline{ij}}\mathcal{I}$ be the number of tokens that land in $\mathcal{I}$, but are not for input-output pair ij. Indeed, $N_{ij}\mathcal{I} + N_{\overline{ij}}\mathcal{I} = N\mathcal{I}$, for any $\mathcal{I} \in \mathbb{R}_+$.

The schedule is given by the sequence $\{Z_n\}_{n\geq 0}$. If for any given n, $Z_n = k$, then the matrix M_k is scheduled in the nth slot. We say the nth token is of type k. Notice that by taking $\{Z_n\}_{n\geq 0}$ we in fact construct a *non-idle* schedule. A key feature of this schedule is,

Observation 4 *The total number of slots in which input-output pair ij can be served during the time slots $n, n+1, \ldots, n+m-1$ is equal to $N_{ij}[T_n, T_{n+m})$.*

5.2.1 Rate-Latency Service Characterization

We give different characterizations of the service offered to an arbitrary input-output pair ij. Informally speaking, we would like $N_{ij}[T_n, T_{n+m})$ to be close to $\rho_{ij}m$. The following is the simplest, but weakest, service characterization: for any $n \geq 0$ and $m > 0$, and some fixed $E_1^{ij} \geq 0$,

$$\{N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E_1^{ij})\}. \tag{5.1}$$

If the event (5.1) holds with probability $1-\epsilon$, $\epsilon \geq 0$, a probabilistic interpretation of the service offered is: for a fixed m one picks an arbitrary slot n, then, the number of slots offered to the input-output pair ij in the next m slots is at least $\rho_{ij}(m - E_1^{ij})$ with probability $1 - \epsilon$.

A natural extension of the above characterization is to require that for a $n \geq 0$ and some fixed $E_2^{ij} \geq 0$,

$$\{\forall_{m>0} : \; N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E_2^{ij})\}. \tag{5.2}$$

The strongest guarantee is offered by requiring, for some fixed $E_3^{ij} \geq 0$,

$$\{\forall_{n\geq 0}\forall_{m>0} : \; N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E_3^{ij})\}. \tag{5.3}$$

If this event holds, then we lower bound the service offered to input-output pair ij over *any* interval of time slots. It follows from a known property that $f_{ij}(m) := \rho_{ij} \max[m - E_3^{ij}, 0]$ is a strict minimum service curve offered to the ijth pair, see Proposition 1.3.6, Section 1.3.2, [21]. The service curve is "rate-latency" with a rate ρ_{ij} and a latency E_3^{ij}. See Figure 5.3 for an illustration.

The service characterizations introduced so far bound how much the service offered is *behind* the service that would be offered by an idealistic fluid system (which would serve $\rho_{ij}m$ bits in m slots). Thus, these service characterizations bound the *lateness* of the scheduler. Analogous characterizations can be established to bound the *earliness*; one only needs to reverse the inequalities in the above definitions, and replace minus with plus in the rate-latency functions. Small earliness of the schedule is desirable to reduce burstiness at the output of the switch.

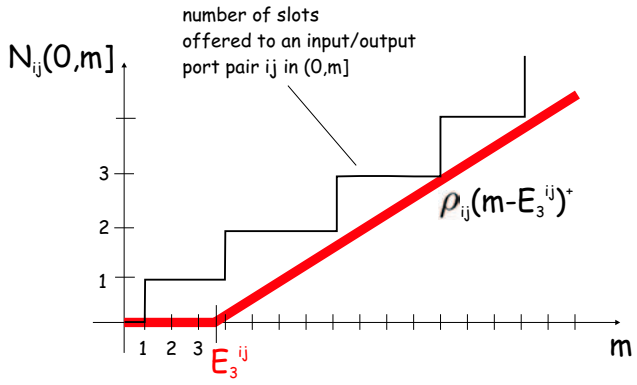


Figure 5.3: The rate-latency function $f_{ij}(m) = \rho_{ij}(m - E_3^{ij})$ is a lower bound on the number of the offered slots to the input/output port ij over any interval of m slots.

5.2.2 Why the Rate-Latency Service Characterization

Consider an input-output port pair ij. Let $A_{ij}(n)$ be the number of the bits that arrive in the interval of the slots $[0, n]$ at input port i and are destined to the output port j. By the result known for variable-capacity nodes (see [21], Sec. 1.3.2, also Sec. 4.3.2), we know that the number of bits in $[0, n]$ observed at the output port j that originate from i (we denote as $A^*_{ij}(n)$) satisfies,

$$A^*_{ij}(n) = \min_{1 \le m \le n} [A_{ij}(m) + N_{ij}[T_m, T_n)].$$

In particular, suppose that the arrivals are (σ_{ij}, ρ_{ij})-*bounded*, i.e., $A_{ij}(n) - A_{ij}(m) \le \rho_{ij}(n - m) + \sigma_{ij}$, for all $m \le n$. Then, the following is a classical network calculus result,

Fact 1 *The backlog of ij packets waiting for service at the switch is at most $\sigma_{ij} + \rho_{ij}E_3^{ij}$. If FIFO scheduling is used* within *the aggregate of ij packets then the maximum delay for these packets is at most $\sigma_{ij}/\rho_{ij} + E_3^{ij}$.*

If $\sigma_{ij} = 0$ (i.e. the arrivals are bounded by an idealized fluid of the rate ρ_{ij}) then the packet delay is bounded by E_3^{ij} (assuming FIFO scheduling within the aggregate). However, in a perfect schedule for ij, a matrix in S_{ij} would appear *exactly once* every $1/\rho_{ij}$ time slots. In this case the packet delay would be $1/\rho_{ij}$. Hence, if $\sigma_{ij} = 0$ we have,

$$\frac{\text{worst case packet delay}}{\text{optimal packet delay}} \le \rho_{ij}E_3^{ij}.$$

For these reasons our objective is to keep E_3^{ij} small.

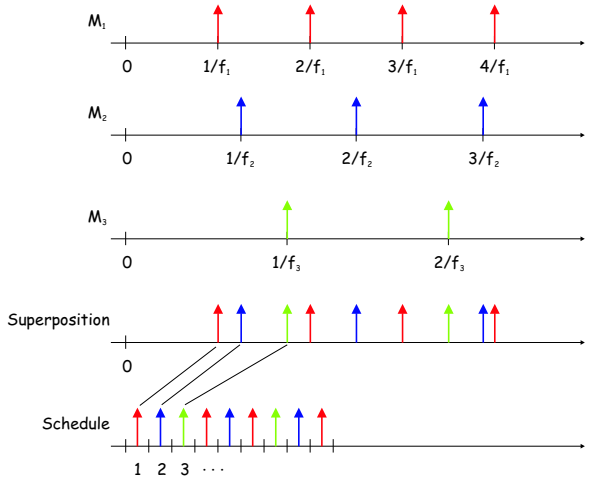


Figure 5.4: The construction of the deterministic schedule by Chang *et al.* [27]. The example is for three permutation matrices, with the intensity of the ith permutation matrix equal to f_i.

5.2.3 Our Results

Our algorithms will be divided into two types, *frame-based* and *non-frame-based*. Suppose that for some fixed integers ℓ_k and L, $\varphi_k = \ell_k/L$, for all $k = 1, 2, \ldots, K$. (Note that this is always possible if the φ_k are rational.) We can compute a schedule for the interval $[0, L)$ that contains exactly ℓ_k occurrences of the permutation matrix M_k and then simply repeat this schedule for all subsequent intervals of length L. We call such a schedule a *frame-based* schedule of length L. Notice that the frame length L and the number of permutation matrices K are related as $K = L/\bar{\ell}$, where $\bar{\ell}$ is the arithmetic mean of ℓ_k, $k = 1, \ldots, K$. Since $\ell_k \geq 1$ for all k it follows that $L \geq K$, with equality iff $\ell_k = 1$ for all k.

If the schedule is not periodic in the above way then we say that it is *non-frame-based*. For a non-frame-based schedule we have to define it explicitly in the entire interval $[0, \infty)$.

In [27], Chang *et al.* propose a non-frame-based algorithm in which the permutation matrices are scheduled according to a PGPS system (Parekh and Gallager [94]) that is fed with its own departures. In our setting, this corresponds to placing the nth token for matrix M_k at time n/φ_k. More formally, for each $k = 1, \ldots, K$,

$$N_k[0, t) = \sum_{n>0} \mathbf{1}_{[0,t)} \left(\frac{n}{\varphi_k} \right).$$

An example is shown in Figure 5.4.

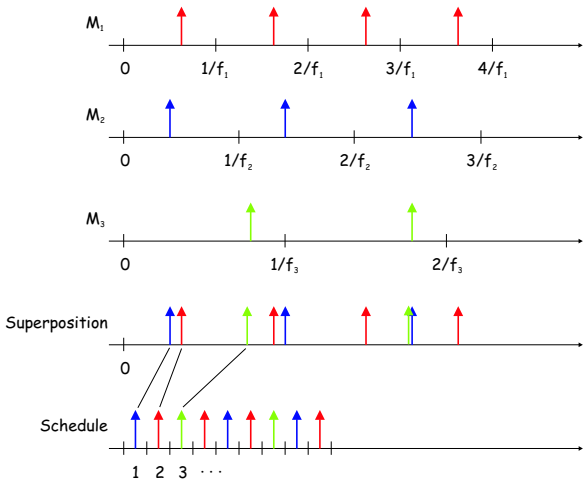


Figure 5.5: The random-phase periodic competition scheduler. The setting of the example matches that of Figure 5.4. The token process of the ith permutation matrix is periodic, with period $1/f_i$, and random phase, which is drawn independently of the phases of the token processes of other permutation matrices, uniformly at random on the interval $(0, 1/f_i]$.

Chang *et al.* [27] show that for this PGPS algorithm,[1]

$$E_3^{ij} \leq \min\left\{\frac{K}{\rho_{ij}}, \frac{|S_{ij}|}{\rho_{ij}} + (K-1)\right\}. \tag{5.4}$$

The aim of our work is to show that by using probabilistic techniques, it is possible to obtain algorithms with better bounds.
Our results are as follows.

1. We begin in Section 5.4.1 with an extremely simple frame-based scheduler in which the tokens for the permutation matrices in a frame are randomly permuted. We call this the *Random Permutation* scheduler. We require that (5.3) holds with probability $1-\epsilon$ and we show that

$$E_3^{ij} \sim \sqrt{A_\epsilon \left(\frac{1}{\rho_{ij}} - 1\right) L}, \text{ large } L \tag{5.5}$$

where A_ϵ is a positive constant specified later. For the latency E_2^{ij}, the same expression holds, with $A_\epsilon = \frac{1}{2}\ln\epsilon^{-1}$.

[1] We remark that this bound (5.4) for PGPS can be almost tight. Consider an example in which $\phi_1 = 1/2$, $\phi_k = 1/2(K-1)$ for $k = 2, \ldots, K$ and $S_{ij} = \{1\}$. PGPS will schedule matrix M_1 in the first $K-1$ slots and matrices $M_2, \ldots, M_K$ in the next $K-1$ slots. Then $E_3^{ij} = K-1$ whereas the bound (5.4) for input-output pair ij equals $\min\{2K, 2+(K-1)\} = K+1$.

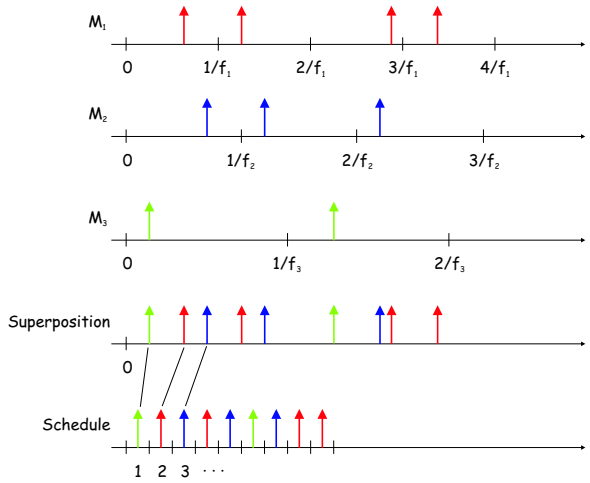


Figure 5.6: The random-distortion periodic competition scheduler. The setting of the example matches that of Figure 5.4. For the ith permutation matrix, the tokens are placed independently, uniformly at random on the intervals $((n-1)/f_i, n/f_i]$, $n = 1, 2, \ldots$. In other words, the tokens of a deterministic periodic sequence of tokens, as found in Figure 5.4, are randomly distored on the periods in which they fall in.

2. In Section 5.4.2 we present a *deterministic* frame-based algorithm. We require that (5.3) holds with probability 1 and we show that,

$$E_3^{ij} \leq \frac{|S_{ij}|}{\rho_{ij}} + (2 + \sqrt{2K \ln(2L+1)}). \tag{5.6}$$

We derive this algorithm from a randomized algorithm in which the nth token for matrix M_k is placed at time $U_k + n/\varphi_k$ where U_k is chosen uniformly at random in $[0, 1/\varphi_k)$. We call this the *Random-Phase Periodic Competition* scheduler. See Figure 5.5 for an illustration.

We show how to *de-randomize* this scheduler to obtain a deterministic algorithm using the method of conditional probabilities (Motwani and Raghavan [89]). In Section 5.5 we show that in many scenarios, (5.6) is significantly smaller than (5.4), largely due to the presence of the square-root in (5.6).

3. In Section 5.4.4 we present a deterministic non-frame-based algorithm. We require that (5.3) holds with probability 1 and we show that,

$$E_3^{ij} \leq \frac{1}{\rho_{ij}} \sqrt{2|S_{ij}| \ln D} + (2 + \sqrt{2K \ln D}), \tag{5.7}$$

where $D = 1 + (4(2I^2 + 2)/\min_k \varphi_k)$. This algorithm is derived from a randomized algorithm in which the nth token for matrix M_k is placed

uniformly at random in the interval $[(n-1)/\varphi_k, n/\varphi_k)$. We call this the *Random-Distortion Periodic Competition* scheduler. An illustration is shown in Figure 5.6.

By using the method of conditional probabilities we are able to *derandomize* this scheduler although the analysis is more complex than it was for the random-phase scheduler since we now have to consider the entire interval $[0, \infty)$ instead of a finite frame. In Section 5.5 we show that in many scenarios, (5.7) is significantly smaller than (5.4), largely due to the presence of the square-roots in (5.7).

4. In Section 5.4.7 we analyze a non-frame-based scheduler in which the tokens for matrix M_k are placed according to a Poisson process. We call this the *Poisson Competition* scheduler. For this scheduler only we assume that the load ρ on each input and output is strictly less than 1. We show using the Brownian approximation (e.g. Whitt [113]) that,

$$E_2^{ij} \approx \frac{1}{2} \ln \epsilon^{-1} \frac{\rho}{1-\rho} \left(\frac{1-\rho}{\rho_{ij}} + \rho \right). \tag{5.8}$$

The latency E_3^{ij} does not make much sense for this scheduler since the event in (5.3) would fail with probability 1 as we require that the inequality in (5.3) holds for all n.

5.2.4 Complexity

For all four of our algorithms the vast majority of the computations can be performed *offline*. For the frame-based schedulers, namely the Random Permutation scheduler and the Random-Phase Periodic Competition Scheduler, the schedule for an entire frame can be computed offline. During the online operation of the scheduler, all we have to do is keep repeating the stored schedule. Hence the complexity is low.

For the non-frame-based schedulers we can still use an offline computation to decompose the rate matrix M into permutation matrices. All that remains to do is to place the tokens online and then find the earliest token that has not yet been scheduled. The latter operation is elementary for any scheduler. For the Random-Distortion Periodic Competition scheduler the tokens are placed periodically and then shifted by a random amount.[2] For the Poisson Competition scheduler we place the tokens by running a Poisson process for each permutation matrix. Hence for both non-frame-based schedulers the token placement algorithm is inexpensive.

5.2.5 Contrasting with a Single Server Polling

We remark that our problem is significantly different from the single server polling problem (e.g. see Takagi [103] and references therein) in which a sin-

[2] Derandomization introduces some additional complexity. For each token we must calculate a set of expectations for each possible placement of the token.

gle server has to poll a set of clients at predetermined frequencies. Note that in our problem it is not sufficient for each matrix M_k to be served at evenly spaced intervals of $1/\varphi_k$ slots. This is because input-output pair ij is served whenever a matrix in S_{ij} is served. If $k, \ell \in S_{ij}$ and M_k and M_ℓ are served close together then the service to ij is bursty even though each permutation matrix might receive smooth service. Note however that we cannot, in isolation, change the schedule to improve service for one particular input-output pair since each permutation matrix is a member of S_{ij} for I different pairs ij.

5.2.6 Previous Work

As mentioned earlier, papers that analyze schedulers based on decomposing the rate matrix include Chang, Song, and Chiu [27, 28], Li and Ansari [81]. Analyses of MWM-type schedulers can be found in, for example, in McKeown, Anantharam, and Walrand [86], Mekkittikul and McKeown [87], Tassiulas and Ephremides [104], Dai and Prabhakar [34], and Leonardi *et al.* [80]. Some frame-based schedulers were presented by Hung and Kesidis [59], and Lee and Lam [78]. If the switch fabric has an internal speedup of 2, then it is known that it can emulate output-queued switches (in which there is no contention at the inputs), see Chuang *et al.* [32], Charny *et al.* [29], and Stoica and Zhang [100]. In [102], Tabatabaee, Georgiadis, and Tassiulas, present an algorithm whose aim is to "track" an idealized fluid policy.

If the switch is sufficiently underloaded, then tight delay bounds can be achieved. Giles and Hajek [47] show that if the total load on any input or output is at most one quarter of the link rate, then it is possible to serve each ij pair at least once every $1/\rho_{ij}$ steps.

5.3 Preliminary Analysis

Consider the event (5.1), which can be re-written as

$$\{\exists_{t>0}\exists_{s<t} : N_{ij}[s, s+t) \geq \rho_{ij}(m - E_1^{ij}), N[0, s+t) = n+m-1, N[0, s) = n-1\}.$$

Unfortunately, it is hard in general to calculate the probability of the above event because of the dependency of the counting processes. It is however feasible for the case of point processes with independent increments. An example of this special case is the Poisson Competition scheduler that we analyze in Section 5.4.7.

In the remainder of the section, we define a subevent of (5.1) whose probability is easier to bound in the general case. We let $G_{n,m}$ be the *good* event, defined as

$$G_{n,m} = \{N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m - E)\}.$$

Let $\Delta_1, \Delta_2 \in \mathbb{Z}_+$ and $\Delta_3, \Delta_4 \in \mathbb{R}_+$ satisfy,

$$\Delta_1 + \Delta_2 + \frac{\Delta_3 + \Delta_4}{\rho_{ij}} \leq E,$$

where $E = E_1^{ij}, E_2^{ij}$ or E_3^{ij}, depending on our calculation. Let

$$t = n + m - \Delta_1,\ s = n + \Delta_2.$$

Note that s and t are integers.

From Lemma 19, Lemma 20 and (5.17), it follows

$$\begin{aligned} G_{n,m} \supseteq & \{N[0,t) < t + \Delta_1\} \cap \{N[0,s) \geq s - \Delta_2\} \cap \\ & \cap \{N_{ij}[s,t) \geq \rho_{ij}(t-s) - (\Delta_3 + \Delta_4)\}, \end{aligned}$$

and,

$$\begin{aligned} G_{n,m} \supseteq & \{N[0,t) < t + \Delta_1\} \cap \{N[0,s) \geq s - \Delta_2\} \cap \\ & \cap \{N_{ij}[0,t) \geq \rho_{ij} t - \Delta_3\} \cap \{N_{ij}[0,s) \leq \rho_{ij} s + \Delta_4\}. \end{aligned}$$

If we are interested in calculating E_1^{ij} then we only need to focus on some fixed n and m. However, if we are interested in E_3^{ij} then we need to know whether $G_{n,m}$ is true for *all* n, m. For the latter case we have,

$$\begin{aligned} \bigcap_{n,m} G_{n,m} \supseteq & \bigcap_t \{N[0,t) < t + \Delta_1\} \cap \bigcap_s \{N[0,s) \geq s - \Delta_2\} \cap \\ & \bigcap_{s,t} \{N_{ij}[s,t) \geq \rho_{ij}(t-s) - (\Delta_3 + \Delta_4)\}. \end{aligned} \tag{5.9}$$

and,

$$\begin{aligned} \bigcap_{n,m} G_{n,m} \supseteq & \bigcap_t \{N[0,t) < t + \Delta_1\} \cap \bigcap_s \{N[0,s) \geq s - \Delta_2\} \cap \\ & \bigcap_t \{N_{ij}[0,t) \geq \rho_{ij} t - \Delta_3\} \cap \bigcap_s \{N_{ij}[0,s) \leq \rho_{ij} s + \Delta_4\} \end{aligned} \tag{5.10}$$

We note that since s and t are integers we only need to take the intersection over a discrete set of events.

5.4 Four Schedulers

5.4.1 Random Permutation

We consider a frame-based scheduler that schedules the permutation matrices in a frame at random order. Formally, let $z = (z_1, z_2, \ldots, z_L)$ be a sequence of token types such that in z there exist ℓ_k tokens of type k, $k = 1, \ldots, K$. Let $\pi = (\pi(1), \pi(2), \ldots, \pi(L))$ be a random permutation of the elements $(1, 2, \ldots, L)$. The schedule in a frame is defined as

$$Z_n = z_{\pi(n)},\ n = 1, \ldots, L.$$

The schedule is extended for $n > L$ by the periodic extension of the frame fixed above, so that the schedule is

$$z_{\pi(1)}, z_{\pi(2)}, \ldots, z_{\pi(L)}, z_{\pi(1)}, z_{\pi(2)}, \ldots, z_{\pi(L)}, \ldots$$

Note that the scheduler as defined above can be formulated in the framework of the token point processes. We would first place ℓ_k points uniformly at random on the interval $[0, 1)$. Then, the counting process of tokens N_k would be defined as the periodic extension of the points placed in $[0, 1)$.

We first discuss some elementary properties of the scheduler, and then show the main result that concerns latencies of the scheduler. By a standard combinatorial argument[3] we obtain, for $l = 0, 1, \ldots, \min[\ell_{ij}, m]$,

$$\mathbf{P}[N_{ij}[T_n, T_{n+m}) = l] = \frac{\left(\begin{array}{c} m \\ l \end{array} \right) \left(\begin{array}{c} L-m \\ \ell_{ij}-l \end{array} \right)}{\left(\begin{array}{c} L \\ \ell_{ij} \end{array} \right)},$$

where $\ell_{ij} = \sum_{k \in S_{ij}} \ell_k$. This is hypergeometric distribution. The above explicit expression would enable us to compute the latency E_1^{ij} defined by (5.1). The variance of $N_{ij}[T_n, T_{n+m})$ is (see Appendix 5.7.2)

$$\sigma_{ij}^2(m) = \frac{L^2}{L-1} \rho_{ij}(1-\rho_{ij}) \frac{m}{L} \left(1 - \frac{m}{L}\right). \tag{5.11}$$

Note that the variance forms a bridge, see Figure 5.7. Indeed,

$$\sigma_{ij}^2(m) \sim L\rho_{ij}(1-\rho_{ij}) \frac{m}{L} \left(1 - \frac{m}{L}\right), \text{ large } L.$$

We now show the main results of this section: asymptotic expressions for the latencies E_2^{ij} and E_3^{ij}. We first show the expression for the latency E_2^{ij}.

Theorem 22 *It holds*

$$\frac{E_2^{ij}}{\sqrt{L}} \to \sqrt{\frac{1}{2} \ln \frac{1}{\epsilon} \left(\frac{1}{\rho_{ij}} - 1 \right)}, \ L \to \infty.$$

The expression for the latency E_3^{ij} is similar as shown next.

Theorem 23 *It holds*

$$\frac{E_3^{ij}}{\sqrt{L}} \to \sqrt{A_\epsilon \left(\frac{1}{\rho_{ij}} - 1 \right)}, \ L \to \infty \tag{5.12}$$

[3]Consider an urn containing L balls with ℓ_{ij} balls labeled with "1" and the rest of the balls labeled with "0." We draw at random m balls without replacement and look at how many of the m balls are labeled with "1."

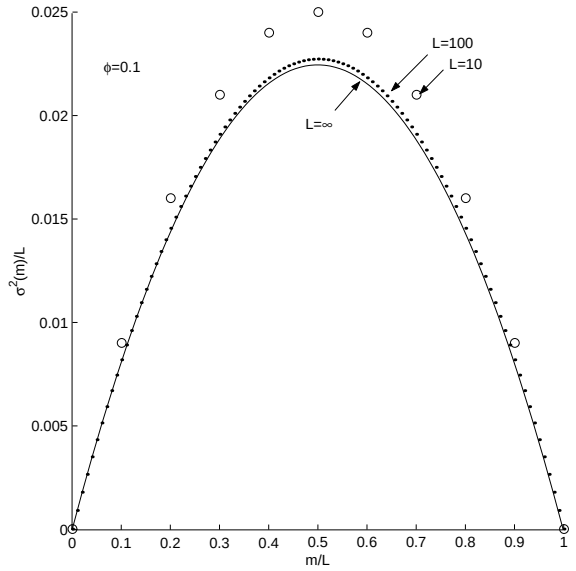


Figure 5.7: Normalized variance of $N_{ij}[T_n, T_{n+m})$ for varying frame size L, and $\rho_{ij} = \phi = 1/10$.

where A_ϵ is the positive solution of

$$\sum_{\ell=1}^{\infty} \left(4\ell^2 A_\epsilon - 1\right) e^{-2\ell^2 A_\epsilon} = \frac{1}{2}\epsilon.$$

Comments. Note that both latencies in the last two theorems are asymptotically $\sqrt{L}$ up to a multiplicative constant. Compare this with the deterministic worst-case bound on latency, $(1-\rho_{ij})L$. In Figure 5.8, we plot values for E_2^{ij} and E_3^{ij} obtained empirically, together with the above limits, for different values of L. We observe that analytical results compare well with the empirical companions. Note that the above stated latencies are for a fixed input/ouput port pair ij. If we want that the latencies E_2^{ij} and E_3^{ij} hold for *all* input/output port pairs ij with probability $1-\epsilon$, then we can easily adjust our results as follows. First, consider the latencies E_2^{ij}. We require

$$\mathbf{P}[\{\exists\ ij, m > 0 :\ N_{ij}[T_0, T_m) < \rho_{ij}(m - E_2^{ij})\}] \leq \epsilon.$$

By the union bound, the last inequality is implied by: for all input/output port pairs ij, it holds

$$\mathbf{P}[\{\exists\ m > 0 :\ N_{ij}[T_0, T_m) < \rho_{ij}(m - E_2^{ij})\}] \leq \frac{\epsilon}{I^2}.$$

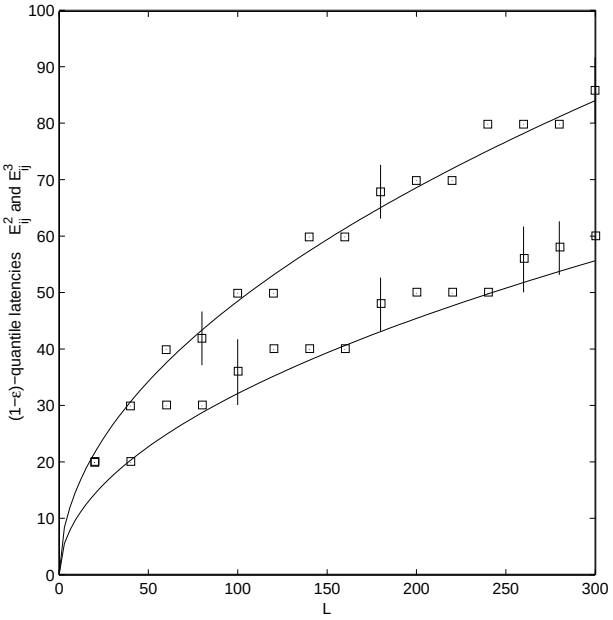


Figure 5.8: Latency of the Random Permutation Scheduler that hold with probability $1-\epsilon$, $\epsilon = 0.1$. (Bottom curve) E_2^{ij} and (Top curve) E_3^{ij}. The empirical quantiles (squares) are computed from 5 independent samples each of 500 samples of random permutations. The empirical quantiles are shown as averages over the 5 samples along with 0.95-confidence intervals. The solid curves are analytical results of Theorem 22 and Theorem 23.

Note that the left-hand side is exactly the probability that we considered earlier. Hence, the statement of Theorem 22 that would hold for all ij is

$$E_2^{ij} \sim \sqrt{\left(\ln I + \frac{1}{2}\ln\frac{1}{\epsilon}\right)\left(\frac{1}{\rho_{ij}} - 1\right)L}, \text{ large } L.$$

Similarly, we can adjust the statement of Theorem 23 to hold for all input/output port pairs ij, by replacing ϵ with ϵ/I^2 in the equation that defines A_ϵ.

5.4.2 Random-Phase Periodic Competition

Let $(U_1, U_2, \ldots, U_K)$ be a collection of independent uniformly distributed random variables on $[0,1]$. We define the scheduler as follows, for each $k = 1, \ldots, K$,

$$N_k[0,t) = \sum_{n>0} \mathbf{1}_{[0,t)}\left(\frac{n-1}{\varphi_k} + \frac{1}{\varphi_k}U_k\right), \ t \geq 0.$$

Thus the tokens for matrix M_k form a periodic stream of period $1/\varphi_k$ with random phase shift U_k/φ_k. Note that we can write

$$N_k[0,t) = \lfloor \varphi_k t \rfloor + 1_{U_k < \varphi_k t - \lfloor \varphi_k t \rfloor}. \tag{5.13}$$

We assume that the φ_k have the property that we can define a frame-based scheduler; see Section 5.2.3. Therefore we only need to concentrate on the time interval L. For any interval $[s,t)$, $N_k[s,t) \geq \varphi_k(t-s) - 1$. This implies,

$$N_{ij}[s,t) \geq \sum_{k \in S_{ij}} (\varphi_k(t-s) - 1) \Rightarrow N_{ij}[s,t) \geq \rho_{ij}(t-s) - |S_{ij}|.$$

We follow the method of Section 5.3 and set $t = n + m - \Delta_1$, $s = n + \Delta_2$, where Δ_1, Δ_2 are defined below. For any permutation matrix M_k, let $a_k = \lfloor \varphi_k t \rfloor$. Then $N_k[0,t) = a_k + X_k$ where X_k is a binary random variable with mean $\varphi_k t - a_k$. Let $\mu = \mathbf{E}[\sum_k X_k] = \sum_k (\varphi_k t - a_k) = t - \sum_k a_k$. We have,

$$\begin{aligned} N[0,t) \geq t + \Delta_1 &\Leftrightarrow \sum_k (a_k + X_k) \geq t + \Delta_1 \\ &\Leftrightarrow \sum_k X_k \geq t + \Delta_1 - \sum_k a_k \\ &\Leftrightarrow \sum_k X_k \geq \mu + \Delta_1. \end{aligned}$$

Therefore, by a Hoeffding's inequality (Theorem 2, [56]), $\mathbf{P}[N[0,t) \geq t + \Delta_1] \leq \exp(-2(\Delta_1)^2/K)$. Similarly, $\mathbf{P}[N[0,s) < s - \Delta_2] \leq \exp(-2(\Delta_2)^2/K)$.

Let,

$$\begin{aligned} \Delta_1 &= \Delta_2 = \left\lceil \sqrt{(K/2) \cdot \ln(2L+1)} \right\rceil, \\ \Delta_3 &= \Delta_4 = |S_{ij}|/2, \\ E_3^{ij} &= (|S_{ij}|/\rho_{ij}) + (2 + \sqrt{2K \ln(2L+1)}). \end{aligned}$$

Then, by the above Hoeffding bounds and the containment (5.9) from Section 5.3 we have,

$$\begin{aligned} & \mathbf{P}[\bigcap_{ij} \bigcap_{nm} \{N_{ij}[T_n, T_m) \geq \rho_{ij}(m - E_3^{ij})\}] \\ \geq\ & 1 - \sum_{t=1}^{L} \mathbf{P}[N[0,t) \geq t + \Delta_1] - \sum_{s=1}^{L} \mathbf{P}[N[0,s) < s - \Delta_2] \\ \geq\ & 1 - \sum_{t=1}^{L} \exp(\frac{-2(\Delta_1)^2}{K}) - \sum_{s=1}^{L} \exp(\frac{-2(\Delta_2)^2}{K}) \\ \geq\ & 1 - \frac{2L}{2L+1}. \end{aligned} \tag{5.14}$$

Note that since we are considering a finite frame we only need sum over $s, t \in \{1, \ldots, L\}$.

Hence with probability $1/(2L+1)$, the rate-latency condition (5.3) holds *for all* ij. We now show how to *de-randomize* the algorithm so that condition (5.3) holds with probability 1.

5.4.3 De-randomization

We use the method of conditional probabilities (see e.g. [89]) that is motivated by the following lemma.

Lemma 16 *Let $\sigma_1, \ldots, \sigma_{n_1}$ be a collection of events, let $Y_1, \ldots, Y_{n_2}$ be a collection of random variables and let $X_{i,1}, \ldots, X_{i,n_2}$, $i = 1, 2, \ldots, n_1$ be another collection of random variables such that for some non-random functions $f_{ik}(\cdot)$,*

$$\mathbf{P}[\sigma_i | Y_1 = z_1, \ldots, Y_v = z_v] \leq \mathbf{E}[\prod_{k=1}^{n_2} f_{ik}(X_{ik}) | Y_1 = z_1, \ldots, Y_v = z_v], \qquad (5.15)$$

for any $v, z_1, \ldots, z_v$. Then there exists a set of fixed *values $y_1, \ldots, y_{n_2}$ such that,*

$$\sum_i \mathbf{P}[\sigma_i | Y_1 = y_1, \ldots, Y_{n_2} = y_{n_2}] \leq \sum_i \mathbf{E}[\prod_{k=1}^{n_2} f_{ik}(X_{ik})].$$

In particular, if $\sum_i \mathbf{E}[\prod_{k=1}^{n_2} f_{ik}(X_{ik})] < 1$ and σ_i is completely determined by $Y_1, \ldots, Y_{n_2}$ then,

$$\sum_i \mathbf{P}[\sigma_i | Y_1 = y_1, \ldots, Y_{n_2} = y_{n_2}] = 0.$$

Comments. Note that if, in addition, $Y_1, \ldots, Y_{n_2}$ are mutually independent and $X_{ik} = \varphi_{ik}(Y_k)$ for some non-random function φ_{ik}, then,

$$\sum_i \mathbf{E}[\prod_{k=1}^{n_2} f_{ik}(X_{ik}) | Y_1 = y_1, \ldots, Y_v = y_v] =$$
$$= \sum_i \prod_{k=1}^{n_2} \mathbf{E}[f_{ik}(X_{ik}) | Y_1 = y_1, \ldots, Y_v = y_v],$$

In order to compute y_v given $y_1, \ldots, y_{v-1}$ we can minimize the above function with respect to y_v over the support of Y_v.

Application. To apply the last lemma in our setting we take Y_k to be the random phase shift U_k; $X_{ik} = \phi_{ik}(U_k)$, for $\phi_{ik}(x) := \mathbf{1}_{x < \varphi_k i - \lfloor \varphi_k i \rfloor}$, and σ_i to be an event of the form $\{N[0,t) \geq t + \Delta_1\}$ or $\{N[0,s) < s - \Delta_2\}$. Note that X_{ik} is a binary random variable, it is a non-random function of U_k. The functions $f_{ik}(\cdot)$ are defined by Chernoff's inequality,

$$\mathbf{P}[N[0,t) \geq t + \Delta_1] \leq e^{-\theta(t+\Delta_1)} \mathbf{E}[\prod_{k=1}^{K} e^{\theta N_k[0,t)}], \qquad (5.16)$$

where

$$\begin{aligned}\theta &= \ln \frac{b(1-a)}{a(1-b)},\\ a &= \frac{1}{K}(t - \sum_{k=1}^{K} \lfloor \varphi_k t \rfloor),\\ b &= \frac{1}{K}(t + \Delta_1 - \sum_{k=1}^{K} \lfloor \varphi_k t \rfloor).\end{aligned}$$

A similar inequality holds for $\mathbf{P}[N[0,s) < s - \Delta_2]$.

In the derivation of (5.14) we showed that,

$$\sum_{t=1}^{L} \mathbf{P}[N[0,t) \geq t + \Delta_1] + \sum_{s=1}^{L} \mathbf{P}[N[0,s) < s - \Delta_2] \leq \frac{2L}{2L+1},$$

using Hoeffding bounds [56] that are not less than the right-hand side in (5.16). Hence by Lemma 16 there exist fixed values $u_1, \ldots, u_K$ for the initial phase shifts such that $N[0,t) < t + \Delta_1$ for all t and $N[0,s) \geq s - \Delta_2$ for all s.

The one complication that arises in the calculation of the u_k is that the U_k are continuous random variables, they do not take discrete values. However, as U_k is varied between 0 and 1, $N_k[0,t) - \lfloor \varphi_k t \rfloor$ changes from 0 to 1 at one discrete point. Hence it is sufficient to consider only $L+1$ values of U_k. The right-hand side of (5.16) may be computed in time polynomial in K and L, even if some of the phase shifts have already been fixed. Hence, we can fix the value of U_k in time polynomial in K and L.

Theorem 24 *The resulting deterministic scheduler satisfies (5.3) with,*

$$E_3^{ij} = \frac{|S_{ij}|}{\rho_{ij}} + (2 + \sqrt{2K \ln(2L+1)}).$$

5.4.4 Random-Distortion Periodic Competition

Let U_{kn}, $k = 1, \ldots, K$, $n \in \mathbb{Z}_+$, be a collection of independent uniformly distributed random variables on $[0,1]$. We define the scheduler as follows; for each $k = 1, \ldots, K$,

$$N_k[0,t) = n + \mathbf{1}_{U_{kn} < \varphi_k t - n},$$

where $n = \lfloor \varphi_k t \rfloor$. Another interpretation is: the nth point of the kth token type is placed uniformly at random in the interval $[n/\varphi_k, (n+1)/\varphi_k)$.

We make use of the containment (5.10) from Section 5.3. We apply Hoeffding bounds in a similar manner to the previous subsection to obtain.

$$\begin{aligned}\mathbf{P}[N[0,t) \geq t + \Delta_1] &\leq \exp(-2(\Delta_1)^2/K),\\ \mathbf{P}[N[0,s) < s - \Delta_2] &\leq \exp(-2(\Delta_2)^2/K),\\ \mathbf{P}[N_{ij}[0,t) \leq \rho_{ij} t - \Delta_3] &\leq \exp(-2(\Delta_3)^2/|S_{ij}|),\\ \mathbf{P}[N_{ij}[0,s) \geq \rho_{ij} s + \Delta_4] &\leq \exp(-2(\Delta_4)^2/|S_{ij}|).\end{aligned}$$

Let,

$$\begin{aligned}
\Gamma_1(t) &= \mathbf{P}[N[0,t) \geq t + \Delta_1] + \sum_{ij} \mathbf{P}[N_{ij}[0,t) < \rho_{ij} t - \Delta_3], \\
\Gamma_2(s) &= \mathbf{P}[N[0,s) < s - \Delta_2] + \sum_{ij} \mathbf{P}[N_{ij}[0,s) > \rho_{ij} s + \Delta_4], \\
D &= 1 + 4\frac{2I^2+2}{\min_k \varphi_k}, \\
\Delta_1 = \Delta_2 &= \left\lceil \sqrt{(K/2) \cdot \ln D} \right\rceil, \\
\Delta_3 = \Delta_4 &= \sqrt{(|S_{ij}|/2) \cdot \ln D}, \\
E_3^{ij} &= \Delta_1 + \Delta_2 + \frac{\Delta_3 + \Delta_4}{\rho_{ij}}.
\end{aligned}$$

For fixed s and t we have,

$$\Gamma_1(t) + \Gamma_2(s) \leq \frac{2I^2+2}{D}.$$

Note that we cannot apply a union bound over s and t as we did in the previous subsection because s and t range over the entire interval $[0, \infty)$. However, note that if $\Gamma_1(t) = \Gamma_2(s) = 0$ for all s, t then from (5.10) we know that (5.3) holds for all i, j with probability 1. Hence we focus on de-randomizing the algorithm.

5.4.5 De-randomization

Instead of placing the nth token for matrix M_k at random into the interval $[n/\varphi_k, (n+1)/\varphi_k)$, we now wish to place it deterministically. Let $P = \lceil 2/\min_k \varphi_k \rceil$. We divide time into intervals of length P, namely, $[0, P), [P, 2P), \ldots$. Let A^ω be the set of tokens that fall into the interval $[\omega P, (\omega+1)P)$ with probability 1, i.e. the nth token for matrix M_k is in A^ω if and only if $[n/\varphi_k, (n+1)/\varphi_k) \subseteq [\omega P, (\omega+1)P)$. Let B^ω be the set of tokens that are not in $A^{\omega'}$ for any ω' and that fall into the interval $[(\omega - \frac{1}{2})P, (\omega + \frac{1}{2})P)$ with probability 1. We have chosen P sufficiently large so that all tokens are in $A^\omega \cup B^\omega$ for some ω.

Suppose inductively that for $\omega' < \omega$ we have fixed the positions of all the tokens in $A^{\omega'} \cup B^{\omega'}$. Since none of the tokens that have already been fixed affect the interval $[\omega P, (\omega+1)P)$, our previous analysis implies,

$$\sum_{\omega P}^{(\omega+1)P} \Gamma_1(t) + \sum_{\omega P}^{(\omega+1)P} \Gamma_2(s) \leq \frac{P(2I^2+2)}{D}.$$

By applying the method of conditional probabilities in a similar manner to Section 5.4.2, we can fix the positions of tokens in A^ω one after the other so

that we still have,

$$\sum_{\omega P}^{(\omega+1)P} \Gamma_1(t) + \sum_{\omega P}^{(\omega+1)P} \Gamma_2(s) \leq \frac{P(2I^2+2)}{D}.$$

Here, the constituent probabilities of $\Gamma_1(t)$ and $\Gamma_2(t)$ are now conditioned on the fact that the tokens in A^ω are fixed. We obtain,

$$\begin{aligned}\sum_{(\omega-\frac{1}{2})P}^{(\omega+\frac{1}{2})P} \Gamma_1(t) + \sum_{(\omega-\frac{1}{2})P}^{(\omega+\frac{1}{2})P} \Gamma_2(s) &\leq \sum_{(\omega-1)P}^{(\omega+1)P} \Gamma_1(t) + \sum_{(\omega-1)P}^{(\omega+1)P} \Gamma_2(s) \\ &\leq \frac{2P(2I^2+2)}{D} < 1.\end{aligned}$$

By the method of conditional probabilities we can fix the positions of tokens in B^ω so that we still have,

$$\sum_{(\omega-\frac{1}{2})P}^{(\omega+\frac{1}{2})P} \Gamma_1(t) + \sum_{(\omega-\frac{1}{2})P}^{(\omega+\frac{1}{2})P} \Gamma_2(s) \leq \frac{2P(2I^2+2)}{D} < 1.$$

All tokens in $A^\omega \cup B^\omega$ are now fixed and so we have a deterministic schedule up to time $(\omega + \frac{1}{2})P$. Recall that $\Gamma_1(t)$ and $\Gamma_2(s)$ are sums of probabilities. Therefore $\Gamma_1(t) = \Gamma_2(s) = 0$ for all $s, t \in [(\omega - \frac{1}{2})P, (\omega + \frac{1}{2})P)$. This process can be repeated indefinitely.

Theorem 25 *The resulting deterministic scheduler satisfies (5.3) with,*

$$E_3^{ij} = \frac{1}{\rho_{ij}} \sqrt{2|S_{ij}| \ln D} + (2 + \sqrt{2K \ln D}).$$

5.4.6 Adaptation to the Substochastic Case

For the previous three schedulers, we have assumed that the rate matrix M is doubly stochastic, i.e. $\sum_i \rho_{ij} = 1$ and $\sum_j \rho_{ij} = 1$. For the case in which M is only substochastic, i.e. $\sum_i \rho_{ij} \leq 1$ and $\sum_j \rho_{ij} \leq 1$, it is known by a result of von Neumann (see e.g. [27]) that there exists a matrix M' with ij entry ρ'_{ij} such that $\rho_{ij} \leq \rho'_{ij}$ for all ij and M' is doubly stochastic. In this case, we can apply all the results of this paper to the matrix M' to obtain latencies E_1^{ij}, E_2^{ij} and E_3^{ij}. Note that the ij traffic might not be able to use all the service it is offered. In this case the residual bandwidth can be used for best-effort traffic.

5.4.7 Poisson Competition

For our final scheduler we require that the load on each input and output is strictly less than 1. Let N_k be Poisson with intensity φ_k, all $k = 1, \ldots, K$. Then the following holds.

Lemma 17 *For any ij, and $n, m \geq 0$, $l = 1, 2, \ldots, m$,*

$$\mathbf{P}[N_{ij}[T_n, T_{n+m}) = l] = \left(\begin{array}{c} m \\ l \end{array} \right) \rho_{ij}^l (1 - \rho_{ij})^{m-l}.$$

The above result may be obvious to many. We note that $(T_n, Z_n)_{n \geq 0}$ is a marked point process with independent, identically distributed marks, where $Z_n = k$ with probability φ_k. Our naming of this scheduler is inspired by the Poisson competition theorem (Theorem 1.3, Chapter 8 Bremaud [23]).

We continue further by observing the following queueing interpretation of the latencies defined in (5.2) and (5.3). Locally to this section, assume $\sum_{k=1}^{K} \varphi_k < 1$; we impose this condition to ensure stability. Moreover, for a fixed ij, let $\rho < 1$ be such that $\sum_{k \ni S_{ij}} \varphi_k = \rho(1 - \rho_{ij})$. We also assume that the counting processes N_k are extended to $\mathbb{R}$, the whole real line. Then, it is not difficult to observe that (5.2) is equivalent to

$$\{V_{\overline{ij}}(0) \leq \rho_{ij} E_2^{ij}\},$$

where $V_{\overline{ij}}(n)$, $n = 0, \pm 1, \pm 2, \ldots$, is *the unfinished work* of a slotted single server queueing system with infinite buffer capacity, service rate $(1 - \rho_{ij})$ and an arrival process that is 0 or 1 with the probability of an arrival equal to $\rho(1 - \rho_{ij})$. The above observation follows immediately by Reich's formula,

$$V_{\overline{ij}}(n) = \max_{m \geq 1} [N_{\overline{ij}}[T_n, T_{n+m}) - (1 - \rho_{ij})m].$$

From Lemma 17, it follows that the unfinished work is that of a Geo/D/1 queue. The distribution of the unfinished work of a Geo/D/1 queue is known in closed form (Gravey, Louvion, and Boyer [51]), which in our context amounts to

$$\mathbf{P}[V_{\overline{ij}}(0) \leq v] = \frac{1 - qD}{(1 - q)^{v+1}} \sum_{l=0}^{j} [q(1 - q)^{D-1}]^l (-1)^l \left(\begin{array}{c} v - (D-1)l \\ l \end{array} \right),$$

where j is the integer such that $jD \leq v \leq (j+1)D - 1$, $q := \rho(1 - \rho_{ij})$, and $D := 1/(1 - \rho_{ij})$ is implicitly assumed to be an integer. (If D would be a real, then we can redefine $D := \lceil 1/(1 - \rho_{ij}) \rceil$ to obtain a lower bound, provided that $\rho D < 1$.)

Our last expression would enable us at least in theory to exactly compute the latency E_2^{ij} in (5.2). The following heuristic argument gives us an approximation that brings us some insight about the latency. By appealing to the Brownian approximation (see Whitt [113], Sec. 5.7, Equation (7.16)) we claim

$$\mathbf{P}[V_{\overline{ij}}(0) \leq \rho_{ij} E_2^{ij}] \approx 1 - e^{-2 \frac{1-\rho}{\rho(1-\rho(1-\rho_{ij}))} \rho_{ij} E_2^{ij}}.$$

Hence, we have

$$E_2^{ij} \approx \frac{1}{2} \ln \epsilon^{-1} \frac{\rho}{1 - \rho} \left(\frac{1 - \rho}{\rho_{ij}} + \rho \right).$$

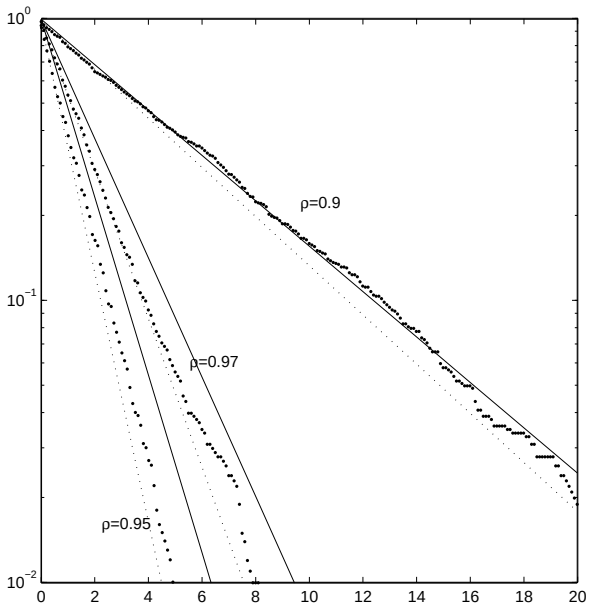


Figure 5.9: Complementary distribution of $V_{\overline{ij}}$: (dots) empirical estimates, (dotted line) M/D/1, (solid line) Brownian approximation. $V_{\overline{ij}}$ is estimated by averaging over 1000 random samples of length 10000. $\rho_{ij} = 0.1$.

Another approximation can be obtained by considering the M/D/1 queue, which is a continuous time analogue[4] of Geo/D/1. A simple exponential approximation is known for M/D/1 ([39], Equation 6.1.6, Section 6.1.2). In Figure 5.9 we show a numerical comparison of the approximations mentioned above with their empirical companions. We observe that the above approximation for E_2^{ij} should be good in the heavy-traffic limit as $\rho \to 1$. It is perhaps interesting to observe that in the heavy-traffic limit E_2^{ij} becomes insensitive to ρ_{ij}.

As mentioned at the beginning of this chapter, the latency E_3^{ij} does not make much sense for this scheduler since the event in (5.3) would fail with probability 1 as we require that the inequality in (5.3) holds for all n.

5.5 Numerical Results

In this section we evaluate some of our bounds for specific rate matrices. Recall that the best possible latency for input-output pair ij is $1/\rho_{ij}$. Hence the ratio between the latency provided by the scheduler, E_3^{ij}, and the best possible latency is $\rho_{ij} E_3^{ij}$. For this reason we define $\max_{ij} \rho_{ij} E_3^{ij}$ to be the figure of merit for a

[4] A notable difference is that with Geo/D/1, in contrast to M/D/1, the number of arrivals over any interval of length m is bounded by m.

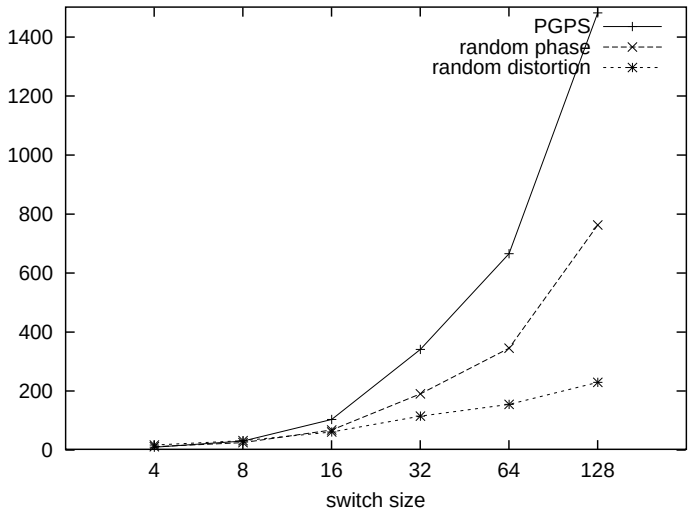


Figure 5.10: The value of $\max_{ij} \rho_{ij} E_3^{ij}$ for switches of varying size.

scheduler.

We evaluate the bounds (5.6) and (5.7) for the deterministic algorithms derived from Random-Phase Periodic Competition and Random-Distortion Periodic Competition. We compare them with the bound (5.4) for PGPS. We would like to use matrices drawn uniformly from the set of doubly-stochastic matrices. However, we do not know a method to generate such a matrix uniformly. Hence we use the following method to generate our example matrices. We start with a uniform matrix in which all entries are equal to I/L where $L = I \times I$. We then repeatedly choose parameters i_1, i_2, j_1, j_2 and δ uniformly at random such that $\delta \leq \min\{\rho_{i_1 j_1}, \rho_{i_2 j_2}\}$. We subtract δ from $\rho_{i_1 j_1}$ and $\rho_{i_2 j_2}$ and we add δ to $\rho_{i_1 j_2}$ and $\rho_{i_2 j_1}$. We carry out this operation 100000 times. Note that it preserves the doubly stochastic nature of the matrix. We also ensure that all entries of the rate matrix are integer multiples of $1/L$. Hence we can define frame-based schedulers with frame-length L.

In Figure 5.10 we plot the value of $\max_{ij} \rho_{ij} E_3^{ij}$ for different values of I, the switch size. We see that except for extremely small switches, the bound for the Random-Distortion scheduler is smaller than the bound for the Random-Phase scheduler which is in turn smaller than the bound for PGPS.

In Figure 5.11 we examine how $\rho_{ij} E_3^{ij}$ varies for different pairs ij. In particular we examine a 64×64 matrix for which $K = 2423$. For each value of x we plot the fraction of ij pairs for which $\rho_{ij} E_3^{ij} \leq x$. We see that the bound (5.6) for the Random-Phase based algorithm is consistently smaller than the bound (5.4) for PGPS. The bound (5.7) for the Random-Distortion based algorithm has a smaller range than the other two bounds. There are fewer pairs ij with large

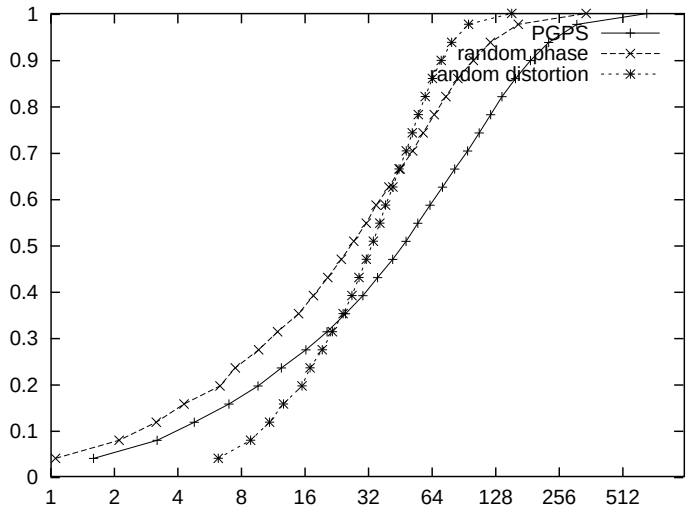


Figure 5.11: Fraction of ij pairs for which $\rho_{ij}E_3^{ij} \leq x$. The matrix has $I = 64$, $L = 4096$ and $K = 2423$.

values of $\rho_{ij}E_3^{ij}$ but there are also fewer pairs ij with small values of $\rho_{ij}E_3^{ij}$. The reason for the latter phenomenon is that the bound (5.7) is typically larger than the bounds, (5.4), (5.6) when the value of $|S_{ij}|$ is small.

We remark that we cannot directly compare the expressions (5.5) and (5.8) for the Random Permutation and Poisson Competition schedulers with the bounds (5.4), (5.6) and (5.7) considered in this section. This is because the expression (5.5) is a limit and the expression (5.8) is for E_2^{ij}, not E_3^{ij}.

5.6 Conclusions

- We obtained bounds on latency for four schedulers that we define for a decomposition-based input-queued switch.
- For two of our schedulers (Random-permutation and Poisson-competition) the bounds on latencies that we obtained hold in probability. For the former, we have asymptotic expressions for large frame lengths of the schedule, which are simple and provide some insight.
- For the other two of our schedulers (Random-phase periodic competition and Random-distortion periodic competition) the bounds on latencies are obtained that hold with probability 1. It is demonstrated that for some numerical examples, our bounds are in many cases tighter than the bound in [27].
- To our knowledge, our approach of the point processes to construct a schedule for the input-queued switch problem is *novel*. The same approach may lead to the construction of new schedulers and perhaps better bounds.

5.6.1 Possible Directions of Future Work

- It is possible that our analysis can be refined so that we would obtain sharper results.
- The question remains: What is *the best possible* latency for a switch load larger than 1/4? (when the load is not larger than 1/4, a schedule exists that schedules an input/output port pair ij, at least once every $1/\rho_{ij}$ slots [47]).

5.7 Proofs

5.7.1 Lemmas and their Proofs

Lemma 18 *Suppose that* $N_{ij}[s,t) \geq \rho_{ij}(t-s)-(\Delta_3+\Delta_4)$ *and* $[s,t) \subseteq [T_n, T_{n+m})$. *Then,* $N_{ij}[T_n, T_{n+m}) \geq \rho_{ij}(m-E)$.[5]

Proof 5 *We have,*

$$\begin{aligned} N_{ij}[T_n, T_{n+m}) &\geq N_{ij}[s,t) \\ &\geq \rho_{ij}(t-s) - (\Delta_3+\Delta_4) \\ &= \rho_{ij}((n+m-\Delta_1) - (n+\Delta_2) - (\Delta_3+\Delta_4)/\rho_{ij}) \\ &= \rho_{ij}(m-(\Delta_1+\Delta_2) - (\Delta_3+\Delta_4)/\rho_{ij}) \\ &\geq \rho_{ij}(m-E). \end{aligned}$$

The first two inequalities come from the assumptions of the lemma. The first equality comes from the definitions of s *and* t. *The final inequality comes from our constraint on the* Δ*'s.*

From the definition of $G_{n,m}$, Lemma 18 implies,

$$G_{n,m} \supseteq \{N_{ij}[s,t) \geq \rho_{ij}(t-s) - (\Delta_3+\Delta_4)\} \cap \{[s,t) \subseteq [T_n, T_{n+m})\}. \quad (5.17)$$

By the above results we can focus on the quantity $N_{ij}[s,t)$ and the relationship between the intervals $[s,t)$ and $[T_n, T_{n+m})$. However, for the random processes we consider, each interval $[s,t)$ will be dependent on too many other intervals. The way to solve this problem is to concentrate on intervals that have one of their endpoints fixed. For this purpose we must refine our results.

Lemma 19 *If* $N[0,t) < t+\Delta_1$ *and* $N[0,s) \geq s-\Delta_2$ *then* $[s,t) \subseteq [T_n, T_{n+m})$.

Proof 6 *Since* $t = n+m-\Delta_1$, $N[0,t) < t+\Delta_1 \Rightarrow [0,t) \subseteq [0, T_{n+m})$. *Similarly, since* $s = n+\Delta_2$, $N[0,s) \geq s-\Delta_2 \Rightarrow [0,s) \supseteq [0,T_n)$. *Therefore,* $[s,t) \subseteq [0,T_{n+m}) \backslash [0,T_n) = [T_n, T_{n+m})$.

Lemma 20 *If* $N_{ij}[0,t) \geq \rho_{ij}t - \Delta_3$ *and* $N_{ij}[0,s) \leq \rho_{ij}s + \Delta_4$ *then* $N_{ij}[s,t) \geq \rho_{ij}(t-s) - (\Delta_3+\Delta_4)$.

Proof 7 *We have,* $N_{ij}[s,t) = N_{ij}[0,t) - N_{ij}[0,s) \geq (\rho_{ij}t - \Delta_3) - (\rho_{ij}s + \Delta_4) = \rho_{ij}(t-s) - (\Delta_3+\Delta_4)$.

[5]Note that we are interested in the values of m such that $m \geq E$. This implies $m \geq \Delta_1+\Delta_2$, which is equivalent to $s \leq t$. For $m < E$ the inequalities in (5.1), (5.2), (5.3) do indeed hold.

5.7.2 Proof of Equation (5.11)

We have to obtain the variance of $N[1,m]$, a hypergeometric random variable,

$$\mathbf{P}[N[1,m]=n]=\frac{\left(\begin{array}{c} m \\ n \end{array}\right)\left(\begin{array}{c} L-m \\ k-n \end{array}\right)}{\left(\begin{array}{c} L \\ k \end{array}\right)},\ n=0,1,\ldots,n\wedge m,$$

where $m=1,2,\ldots,L$. By definition, $N[1,s]=0$, for $s\leq 0$. We suspect this would be a known result. Nevertheless, we give our original proof of (5.11). The proof is based on the known mean of $N[1,m]$, that is

$$\mathbf{E}[N[1,m]]=\frac{mk}{L}.$$

Note

$$N[1,m]=N[1,m-1]+\mathbf{1}_{N(m-1,m]=1}.$$

From the last identity, we have

$$N[1,m]^2=N[1,m-1]^2+(2N[1,m-1]+1)\mathbf{1}_{N(m-1,m]=1}.$$

Define $f(m)=\mathbf{E}[N[1,m]^2]$, the second moment of $N[1,m]$, and $g(m)=\mathbf{E}[(2N[1,m-1]+1)\mathbf{1}_{N(m-1,m]=1}]=(2\mathbf{E}[N[1,m-1]|N(m-1,m]=1]+1)\mathbf{P}[N(m-1,m]=1]$. Armed with this new notation, taking the expectation in the last above display, we have

$$f(m)=f(m-1)+g(m).$$

Hence,

$$f(m)=\sum_{j=1}^{m}g(j). \tag{5.18}$$

It remains to compute $g(j)$, $j=1,2,\ldots,m$. Note that, for $n=0,1,\ldots,n\wedge(m-1)$,

$$\mathbf{P}[N[1,m-1]=n|N(m-1,m]=1]=\frac{\left(\begin{array}{c} m-1 \\ n \end{array}\right)\left(\begin{array}{c} L-1-(m-1) \\ k-1-n \end{array}\right)}{\left(\begin{array}{c} L-1 \\ k-1 \end{array}\right)}.$$

The last follows by a probabilistic argument. Consider an urn of L balls, among which k balls are labeled with "1", and other balls are labeled with "0." Now, interpret $N[1,m]$ as number of balls labeled with "1," among m balls drawn at random from the urn without replacement. The condition $N(m-1,m]=1$ means that we remove a ball labled "1" from the urn, and then under this condition, $N[1,m-1]$ is again the urn problem without replacement, but with $L-1$ balls, among which $k-1$ labeled "1".

This implies that $N[1, m-1]$, given that $N(m-1, m] = 1$, has a hypeometric distribution. Hence,

$$\mathbf{E}[N[1, m-1]|N(m-1, m] = 1] = \frac{(k-1)(m-1)}{L-1}.$$

We compute

$$g(j) = \frac{k}{L(L-1)}(2(k-1)j - 2k + L + 1).$$

Then, from (5.18), we obtain

$$f(m) = \frac{k}{L(L-1)}(m^2(k-1) + [L-k]m).$$

Finally, the variance of $N[1, m]$ is equal to $f(m) - k^2m^2/L^2$, so that after a few straightforward calculations, we obtain the variance is equal to

$$\frac{L^2}{L-1}\frac{k}{L}\left(1 - \frac{k}{L}\right)\frac{m}{L}\left(1 - \frac{m}{L}\right).$$

This with some appropriate substitutions recovers (5.11).

5.7.3 Proof of Proposition 22

Note that by periodicity of the counting process N_{ij}, (5.2) is equivalent to

$$\begin{aligned} & \left\{ \max_{1 \le m \le L} [\rho_{ij} m - N_{ij}[T_n, T_{n+m})] \le \rho_{ij} E_2^{ij} \right\} \\ \Leftrightarrow \quad & \left\{ \max_{1 \le m \le L} [N_{\overline{ij}}[T_n, T_{n+m}) - (1-\rho_{ij})m] \le \rho_{ij} E_2^{ij} \right\}. \end{aligned}$$

Now, it is a standard result (e.g. Theorem 1, Section 6.3.7 [2]) that as $L \to \infty$ we have the convergence in distribution

$$\frac{1}{\sqrt{L}} \max_{1 \le m \le L} [N_{\overline{ij}}[T_n, T_{n+m}) - (1-\rho_{ij})m] \Rightarrow \sqrt{\rho_{ij}(1-\rho_{ij})} \sup_{0 \le t \le 1} B_0(t),$$

where B_0 is Brownian bridge, a Gaussian process with $\mathbf{E}[B_0(t)] = 0$ and $\mathbf{cov}[B_0(s)B_0(t)] = s(t-s)$, $0 \le s \le t \le 1$. Another definition of Brownian bridge is given by $B_0(t) = B(t) - tB(1)$, $t \in [0, 1]$, where $B(t)$, $t \ge 0$, is standard Brownian motion. Hence, Brownian bridge is a Brownian motion conditioned on hitting 0 at $t = 1$.

An exact expression for the complementary distribution of the maximum of Brownian bridge is known (Doob [36]),

$$\mathbf{P}[\sup_{0 \le t \le 1} B_0(t) > b] = e^{-2b^2}.$$

From the above convergence and equating the last limit distribution with ϵ, we obtain the stated result.

5.7.4 Proof of Proposition 23

Note that (5.3) is equivalent to

$$\{\max_{n\geq 0, m>0}[N_{\overline{ij}}[T_n, T_{n+m}) - (1-\rho_{ij})m] \leq \rho_{ij}E_3^{ij}\}.$$

From the periodicity of $N_{\overline{ij}}$, it follows

$$\begin{aligned} Y &:= \max_{n\geq 0, m>0}[N_{\overline{ij}}[T_n, T_{n+m}) - (1-\rho_{ij})m] \\ &= \max_{1\leq k\leq m\leq 2L}[N_{\overline{ij}}[T_{k-1}, T_{m-1}) - (1-\rho_{ij})(m-k)] \\ &= \max_{1\leq k\leq 2L} X_k - \min_{1\leq k\leq 2L} X_k, \end{aligned}$$

where $X_k := N_{\overline{ij}}[T_0, T_{k-1}) - (1-\rho_{ij})k$, $k = 1, 2, \ldots$. Now, similarly as in the proof of Proposition 22, we conclude that, as $L \to \infty$, $L^{-1/2}Y \Rightarrow \sqrt{\rho_{ij}(1-\rho_{ij})}W$, where $W =_d \sup_{0\leq t\leq 1} B_0(t) - \inf_{0\leq t\leq 1} B_0(t)$, the range of the Brownian bridge. It is known that the range of the Brownian bridge is equal in distribution to the maximum Brownian excursion (see Vervaat [106] and [53]). The Brownian excursion can be represented in terms of standard Brownian motion B as $(Z(t))_{t\in[0,1]} =_d ((\tau_+ - \tau_-)^{-1/2}|B((1-t)\tau_+ + t\tau_-)|)_{t\in[0,1]}$, where τ_- is the last zero of B before 1 and τ_+ the first zero after 1. It is known that ([72] Theorem 5.2.10), for $z > 0$,

$$\mathbf{P}[\sup_{0\leq t\leq 1} Z(t) > z] = 2\sum_{\ell=1}^{\infty}(4\ell^2 z^2 - 1)e^{-2\ell^2 z^2}.$$

Now let E_3^{ij} be equal to the right-hand side in (5.12) for some $A_\epsilon > 0$. It follows from the above convergence in distribution that, as $L \to \infty$,

$$\mathbf{P}[\max_{n\geq 0, m>0}[N_{\overline{ij}}[T_n, T_{n+m}) - (1-\rho_{ij})m] > \rho_{ij}E_3^{ij}] \to 2\sum_{\ell=1}^{\infty}\left(4\ell^2 A_\epsilon - 1\right)e^{-2\ell^2 A_\epsilon}.$$

Lastly, we equate the limit in the last display to ϵ, for a fixed $\epsilon \geq 0$, which completes the proof.

5.7.5 Proof of Lemma 16

Suppose inductively that we have already chosen $y_1, \ldots, y_{v-1}$ such that,

$$\begin{aligned} &\sum_i \mathbf{E}[\prod_{j=1}^{n_2} f_{ij}(X_{ij}) | Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}] \\ \leq\ &\sum_i \mathbf{E}[\prod_{j=1}^{n_2} f_{ij}(X_{ij})]. \end{aligned}$$

This can trivially be done for $v = 1$. Then,

$$\begin{aligned} & \sum_y \sum_i \mathbf{P}[Y_v = y | Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}] \cdot \\ & \cdot \mathbf{E}[\prod_{j=1}^{n_2} f_{ij}(X_{ij}) | Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y] \\ = & \sum_i \mathbf{E}[\prod_{j=1}^{n_2} f_{ij}(X_{ij}) | Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}] \leq \\ \leq & \sum_i \mathbf{E}[\prod_{j=1}^{n_2} f_{ij}(X_{ij})]. \end{aligned}$$

Since $\sum_y \mathbf{P}[Y_v = y | Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}] = 1$, there exists a fixed value y_v such that,

$$\begin{aligned} & \sum_i \mathbf{E}[\prod_{j=1}^{n_2} f_{ij}(X_{ij}) | Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y_v] \\ \leq & \sum_i \mathbf{E}[\prod_{j=1}^{n_2} f_{ij}(X_{ij})]. \end{aligned}$$

Then, by the inequality in (5.15),

$$\begin{aligned} & \sum_i \mathbf{P}[\sigma_i | Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y_v] \\ \leq & \sum_i \mathbf{E}[\prod_{j=1}^{n_2} f_{ij}(X_{ij}) | Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y_v] \\ \leq & \sum_i \mathbf{E}[\prod_{j=1}^{n_2} f_{ij}(X_{ij})]. \end{aligned}$$

The proof follows by induction.

Comment. To find y_v we minimize,

$$\begin{aligned} & \sum_i \mathbf{E}[\prod_{j=1}^{n_2} f_{ij}(X_{ij}) | Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y] = \\ & = \sum_i \prod_{j=1}^{n_2} \mathbf{E}[f_{ij}(X_{ij}) | Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y], \end{aligned}$$

over all possible values of y. We can exchange the expectation with the product due to the independence of the X_{ij}. Hence we only need to be able to calculate the $\mathbf{E}[f_{ij}(X_{ij}) | Y_1 = y_1, \ldots, Y_{v-1} = y_{v-1}, Y_v = y]$ in isolation. This is feasible in all our applications of Lemma 16.

5.7.6 Proof of Lemma 17

A counting process N on $\mathbb{R}$ is said to be Poisson with intensity λ if for any two disjoint intervals $\mathcal{I}$ and $\mathcal{J}$ on $\mathbb{R}$, $N\mathcal{I}$ and $N\mathcal{J}$ are independent, and in addition, for any $\mathcal{I}$ on $\mathbb{R}$,

$$\mathbf{P}[N\mathcal{I} = m] = \frac{(\lambda|\mathcal{I}|)^m}{m!} e^{-\lambda|\mathcal{I}|}, \ m = 0, 1, \ldots .$$

It is a known result that if N_k, $k = 1, 2, \ldots, K$ are Poisson counting processes with respective finite intensities φ_k $k = 1, 2, \ldots, K$, then $N_{ij}\mathcal{I} = \sum_{k \in S_{ij}} N_k\mathcal{I}$, for any ij and $\mathcal{I} \in \mathbb{R}$, is Poisson with intensity $\rho_{ij} = \sum_{k \in S_{ij}} \varphi_k$.

Hence, we can write

$$\begin{aligned}
& \mathbf{P}[N_{ij}[T_n, T_{n+m}) = l] \\
= \ & \mathbf{P}[N_{ij}\{T_n\} = 0, N_{ij}(T_n, T_{n+m}) = l] + \mathbf{P}[N_{ij}\{T_n\} = 1, N_{ij}(T_n, T_{n+m}) = l - 1] \\
= \ & (1 - \rho_{ij})\mathbf{P}[N_{ij}(T_n, T_{n+m}) = l] + \rho_{ij}\mathbf{P}[N_{ij}(T_n, T_{n+m}) = l - 1]. \qquad (5.19)
\end{aligned}$$

We exercise a simple calculus, for any $l = 0, 1, \ldots, m - 1$,

$$\begin{aligned}
& \mathbf{P}[N_{ij}(T_n, T_{n+m}) = l] \\
= \ & \int_0^\infty \mathbf{P}[N_{ij}(T_n, T_n + t) = l, N(T_n, T_n + t) = m - 1]dt \\
= \ & \int_0^\infty \mathbf{P}[N_{ij}(T_n, T_n + t) = l, N_{\overline{ij}}(T_n, T_n + t) = m - 1 - l]dt \\
= \ & \int_0^\infty \mathbf{P}[N_{ij}(0, t) = l, N_{\overline{ij}}(0, t) = m - 1 - l]dt \\
= \ & \int_0^\infty \mathbf{P}[N_{ij}(0, t) = l]\mathbf{P}[N_{\overline{ij}}(0, t) = m - 1 - l]dt \\
= \ & \frac{\rho_{ij}^k(1 - \rho_{ij})^{m-1-l}}{k!(m - 1 - l)!} \int_0^\infty t^{m-1}e^{-t}dt \\
= \ & \frac{(m - 1)!}{(l - 1)!(m - l)!}\rho_{ij}^l(1 - \rho_{ij})^{m-1-l}.
\end{aligned}$$

The second equality is obtained by $N\mathcal{I} = N_{ij}\mathcal{I} + N_{\overline{ij}}\mathcal{I}$, any $\mathcal{I} \subset \mathbb{R}$; the third equality is by independence and stationarity of the increments of N_{ij} and $N_{\overline{ij}}$; the forth equality follows by the independence of N_{ij} and $N_{\overline{ij}}$; in the fifth equality we utilize the fact that for any fixed $\mathcal{I} \in \mathbb{R}$, $N_{ij}\mathcal{I}$ and $N_{\overline{ij}}\mathcal{I}$ are Poisson random variables; and finally, the last equality follows from $\int_0^\infty t^m e^{-t}dt = m!$, for m an integer.

The statement of the lemma follows by plugging the resulting identity in the last above display into (5.19).

Bibliography

[1] R. Agrawal, R. L. Cruz, C. Okino, and R. Rajan. Performance Bounds for Flow Control Protocols. *IEEE/ACM Trans. on Networking*, 7(3):310–323, June 1999.

[2] Jaroslav Hájek, Zbyněk Šidák, and Pranab K. Sen. *Theory of Rank Tests*. Academic Press, 1999.

[3] Werner Almesberger. Linux Network Traffic Control–Implementation Overview. Technical report, EPFL, April 1999.

[4] Eitan Altman, Konstantin Avrachenkov, and Chadi Barakat. A stochastic model of TCP/IP with stationary random losses. In *Proc. of the Sigcomm'00*, pages 231–242, 2000.

[5] Eitan Altman, Chadi Barakat, and Victor M. Ramos R. Analysis of AIMD protocols over paths with variable delay, July 2003. preprint.

[6] Eitan Altman, Chadi Barakat, and Victor M. Ramos Ramos. Analysis of AIMD Protocols over Paths with Variable Delay. In *To Appear in Proc. of IEEE INFOCOM'2004*, 2004.

[7] Oskar Andereasson. http://ipsysctl-tutorial.frozentux.net/ipsysctl-tutorial.html, 2003.

[8] Matthew Andrews and Milan Vojnović. Scheduling Reserved Traffic in Input-Queued Switches: New Delay Bounds via Probabilistic Techniques. *IEEE Journal on Selected Areas of Communications*, 21(4):595–605, May 2003.

[9] Francois Baccelli and Pierre Brémaud. *Elements of Queueing Theory*, volume 26. Applications of Mathematics, Springer-Verlag, 1991.

[10] Deepak Bansal and Hari Balakrishnan. Binomial Congestion Control Algorithms. In *Proc. of the IEEE Infocom'01*, volume 2, pages 631–640, 2001.

[11] Deepak Bansal, Hari Balakrishnan, Sally Floyd, and Scott Shenker. Dynamic behavior of slowly-responsive congestion control algorithms. In *Proc. of ACM Sigcomm'01*, San Diego, California, USA, August 2001.

[12] Chadi Barakat, October 2002. private communication.

[13] Chadi Barakat, February 2003. private communication.

[14] J. Bennett, K. Benson, A. Charny, W. Courtney, and J.-Y. Le Boudec. Delay jitter bounds and packet scale rate guarantee for expedited forwarding. In *Proc. of IEEE INFOCOM'2001*, March 2001.

[15] J. Bennett, K. Benson, A. Charny, W. Courtney, and J.-Y. Le Boudec. Delay jitter bounds and packet scale rate guarantee for expedited forwarding. *IEEE/ACM Trans. on Networking*, 10(4):529–540, August 2002.

[16] G. Birkhoff. Tres observaciones sobre el algebra lineal. *Univ. Nac. Tucumán Rev. Ser. A5*, pages 147 – 150, 1946.

[17] T. Bonald, A. Proutiére, and J. Roberts. Statistical performance guarantees for streaming flows using expedited forwarding. In *Proc. of IEEE INFOCOM'2001*, March 2001.

[18] A. A. Borovkov. *Ergodicity and Stability of Stochastic Processes*. John Wiley & Sons, Wiley Series in Probability and Statistics, 1998.

[19] Jean-Yves Le Boudec. Application of network calculus to guaranteed service network. *IEEE Trans. on Information Theory*, 44:1087–1096, May 1998.

[20] Jean-Yves Le Boudec and Anna Charny. Packet scale rate guarantee for non-fifo nodes. In *Proc. of IEEE Infocom 2002*, New York, USA, June 2002.

[21] Jean-Yves Le Boudec and Patrick Thiran. *Network Calculus*. Springer-Verlag, 2001. (also available on-line at http://ica1www.epfl.ch/PS_files/NetCal.htm).

[22] C. Boutremans and J.-Y. Le Boudec. Adaptive delay aware error control for internet telephony. In *Proc. of 2nd IP-Telephony Workshop*, pages 81–92, Columbia University, New York, April 2001.

[23] Pierre Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues*. Springer, 1999.

[24] C. S. Chang. On deterministic traffic regulation and service guarantee: A systematic approach by filtering. *IEEE/ACM Trans. on Networking*, 44:1096–1107, August 1998.

[25] Cheng-Shang Chang. *Performance Guarantees in Communication Networks*. Springer-Verlag, 2000.

[26] Cheng-Shang Chang, Wheyming Song, and Yuh ming Chiu. On the Performance of Multiplexing Independent Regulated Inputs. In *Proc. of Sigmetrics 2001*, Massachusetts, USA, May 2001.

[27] C.S. Chang, W.J. Chen, and H.Y. Huang. On service guarantees for input buffered crossbar switches: A capacity decomposition approach by Birkhoff and von Neumann. In *Proc. of IEEE IWQoS*, London, UK, 1999.

[28] C.S. Chang, W.J. Chen, and H.Y. Huang. Birkhoff-von Neumann Input Buffered Crossbar Switches. In *Proc. of IEEE INFOCOM'00*, Tel-Aviv, Israel, March 2000.

[29] A. Charny, P. Krishna, N. Patel, and R. Simcoe. Algorithms for Providing Bandwidth and Delay Guarantees in Input-Buffered Crossbars with Speedup. In *Proc. of IEEE IWQoS*, Napa, CA, 1998.

[30] Anna Charny and Jean-Yves Le Boudec. Delay Bounds in a Network With Aggregate Scheduling. In *First International Workshop on Quality of future Internet Services*, September 2000.

[31] D. Chiu and R. Jain. Analysis of the Increase and Decrease Algorithms for Congestion Avoidance in Computer Networks. *Computer Networks and ISDN Systems*, 17:1–14, June 1989.

[32] S. Chuang, A. Goel, N. McKeown, and B. Prabhakar. Matching Output Queueing with Combined Input and Output Queueing. *IEEE Journal on Selected Areas in Communications*, 17(6):1030–1039, 1999.

[33] R. L. Cruz and Hai-Ning Liu. Single Server Queues with Loss: A Formulation. In *Proc. 1993 CISS, Johns Hopkins University*, March 1993.

[34] J. Dai and B. Prabhakar. The Throughput of Data Switches with and without Speedup. In *Proc. of IEEE INFOCOM '00*, pages 556 – 564, Tel Aviv, Israel, March 2000.

[35] B. Davie, A. Charny, F. Baker, J. Bennett, K. Benson, J.-Y. Le Boudec, A. Chiu, W. Courtney, S. Davari, V. Firoiu, C. Kalmanek, K. K. Ramakrishnan, and D. Stiliadis. An Expedited Forwarding PHB, March 2002.

[36] J. L. Doob. Heuristic Approach to the Kolmogorov-Smirnov Theorems. *Ann. Math. Statist.*, 20:293–403, 1949.

[37] Laurent Massoulié and Anthony Busson. Stochastic majorization of aggregates of leaky bucket-constrained traffic streams. preprint, http://www.research.microsoft.com/users/lmassoul/, 2000.

[38] B. Davie (Editor), A. Charny, F. Baker, J. Bennett, K. Benson, J.-Y. Le Boudec, A. Chiu, W. Courtney, S. Davari, V. Firoiu, C. Kalmanek, K. K. Ramakrishnan, and D. Stiliadis. An Expedited Forwarding PHB, April 2001.

[39] J. W. Roberts (Editor). *COST 224: Performance evaluation and design of multiservice networks.* Commision of the European Communities, 1991.

[40] Bert Hubert et al. Linux advanced routing & traffic control howto. http://www.linux.org/docs/ldp/howto/Adv-Routing-HOWTO.

[41] Sally Floyd. Connections with Multiple Congested Gateways in Packet-Switched Networks Part 1: One-way Traffic. *Computer Communication Review (also available from: ftp://ftp.ee.lbl.gov/papers/gates1.ps.Z)*, 21(5):30–47, October 1991.

[42] Sally Floyd. HighSpeed TCP for Large Congestion Windows, August 2002. Internet Engineering Task Force, INTERNET-DRAFT, draft-floyd-tcp-highspeed-01.txt.

[43] Sally Floyd and Kevin Fall. Promoting the Use of End-to-End Congestion Control in the Internet. *IEEE/ACM Trans. on Networking*, 7(4):458–472, August 1999.

[44] Sally Floyd, Mark Handley, and Jitendra Padhye. A Comparison of Equation-Based and AIMD Congestion Control. In *Workshop on the modeling of flow and congestion control mechamisms*, Ecole Normale Superieure, Paris, http://www.ens.fr/∼mistral/tcp2.html, September 2000.

[45] Sally Floyd, Mark Handley, Jitendra Padhye, and Jörg Widmer. Equation-Based Congestion Control for Unicast Applications. In *Proc. of ACM Sigcomm'00*, pages 43–56, 2000.

[46] Sally Floyd and Van Jacobson. Random Early Detection Gateways for Congestion Avoidance. *IEEE/ACM Trans. on Networking*, 1(4):397–413, August 1993.

[47] J. Giles and B. Hajek. Scheduling multirate periodic traffic in a packet switch. In *1997 Conference on Information Sciences and Systems at John Hopkins University*, 1997.

[48] Pawan Goyal, Simon S. Lam, and Harrick M. Vin. Determining end-to-end delay bounds in heterogeneous networks. In *Proc. 5th Int. Workshop Network and Operating System Support for Digital Audio and Video*, pages 287–298, April 1995.

[49] Pawan Goyal, Simon S. Lam, and Harrick M. Vin. Determining end-to-end delay bounds in heterogeneous networks. *Springer Multimedia Systems*, 5:157–163, 1997.

[50] Pawan Goyal and Harrick M. Vin. Generalized guaranteed rate scheduling algorithms: A framework. *ACM/IEEE Trans. on Networking*, 5(4):561–571, 1997.

[51] Annie Gravey, Jean-Raymond Louvion, and Pierre Boyer. On the Geo/D/1 and Geo/D/1/n queues. *Performance Evaluation, North-Holland*, 11:117–125, 1990.

[52] Fabrice Guillemin, Philippe Robert, and Bert Zwart. Aimd algorithms and exponential functionals. *To Appear in The Annals of Applied Probability*, 2002.

[53] Bruce Hajek. A queue with periodic arrivals and constant service rate. *in Probability, Statistics and Optimization – a Tribute to Peter Whittle (F.P. Kelly ed., John Wiley and Sons*, pages 147–158, 1994.

[54] Mark Handley, Jitendra Padhye, Sally Floyd, and Jörg Widmer. TCP Friendly Rate Control (TFRC) Protocol Specification, IETF INTERNET-DRAFT, Janunary 2003. RFC 3448, ftp://ftp.isi.edu/in-notes/rfc3448.txt.

[55] Martin Hasler. Nonlinear systems. Lecture Notes, EPFL, 2000.

[56] Wassily Hoeffding. Probability Inequalities For Sums of Bounded Random Variables. *American Statistical Association Journal*, pages 13–30, March 1963.

[57] Bert Hubert. Linux advanced routing & traffic control, manpages, http://lartc.org/manpages, 2003.

[58] Pierre A. Humblet. Determinism minimizes waiting time in queues. Technical Report LIDS-P-1207, MIT, May 1982.

[59] A. Hung, G. Kesidis, and N. McKeown. ATM input-buffered switches with the guaranteed rate property. In *Proc. of IEEE ISCC*, pages 331–335, 1998.

[60] Paul Hurley, Jean-Yves Le Boudec, and Patrick Thiran. A note on the fairness of additive increase and multiplicative decrease. Technical Report 98/296, Institute for Computer Applications, Swiss Federal Institute of Technology at Lausanne, October 1998.

[61] Paul Hurley, Jean-Yves Le Boudec, and Patrick Thiran. A Note on the Fairness of Additive Increase and Multiplicative Decrease. In *Proceedings of ITC-16*, Edinburgh, UK, June 1999.

[62] V. Jacobson. ftp://ftp.ee.lbl.gov/traceroute.tar.z, 1989.

[63] Van Jacobson and Michael J. Karels. Congestion Avoidance and Control. In *Proc. of the ACM SIGCOMM'88*, pages 314–329, Stanford, August 1988.

[64] Shudong Jin, Liang Guo, Ibrahim Matta, and Azer Bestavros. A Spectrum of TCP-Friendly Window-Based Congestion Control Algorithms. *IEEE/ACM Trans. on Networking*, pages 341–355, June 2003.

[65] k claffy, Greg Miller, and Kevin Thompson. The nature of the beast: recent traffic measurements from an Internet backbone. In *Inet98*, Geneva, http://www.caida.org/outreach/papers/1998/Inet98/, July 1998.

[66] J. B. Keller. Inverse problems. *Am. Math. Mon.*, 83:107–118, 1976.

[67] F. P. Kelly. Mathematical modelling of the Internet. In *Forth International Congress on Industrial and Applied Mathematics*, Edinburgh, Scotland, July 1999. (also available from http://www.statslab.cam.uc.uk/ frank).

[68] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan. Rate control for communication networks: Shadow prices, proportional fairness and stability. *Journal of the Operational Research Society*, 49, 1998. (also available from http://www.statslab.cam.uc.uk/ frank).

[69] George Kesidis and Takis Konstantopoulos. Extremal Traffic and Worst-Case Performance for Queues with Shaped Arrivals. *Analysis of Communication Networks: Call Centers, Traffic and Performance (edt. D. R. McDonald and S. R. E. Turner), Fields Institute Communications/AMS, ISBN 0-8218-1991-7*, 2000.

[70] George Kesidis and Takis Konstantopoulos. Worst-Case Performance of a Buffer with Independent Shaped Arrival Processes. *IEEE Communication Letters*, 2000.

[71] J. F. C. Kingman. Subadditive processes. In *Ecole d'été de probabilité de Saint-Flour*, volume Lecture Notes in Mathematics (539), pages 165–223. Springer Verlag, 1976.

[72] Frank B. Knight. *Essentials of Brownian Motion and Diffusion*, volume 18. Mathematical Surveys, American Mathematical Society, 1981.

[73] T. Konstantopoulos and G. Last. On the dynamics and performance of stochastic fluid systems. *Journal of Applied Probability*, 37:652–667, 2000.

[74] T. Konstantopoulos, M. Zazanis, and G. de Veciana. Conservation laws and reflection mappings with an application to multiclass mean value analysis for stochastic fluid queues. *Stoch. Proc. Appl.*, 65(1):139–146, 1997.

[75] Srisankar Kunniyur and R. Srikant. End-to-End Congestion Control Schemes: Utility Functions, Random Losses and ECN Marks. In *Proc. of the IEEE INFOCOM'2000*, Tel-Aviv, Israel, March 2000.

[76] J. Kurose. On computing per-session performance bounds in high-speed multi-hop computer networks. In *Proc. of ACM Sigmetrics and Performance '92*, Newport, Rhode Island, June 1992.

[77] Harold J. Kushner and G. George Yin. *Stochastic Approximations Algorithms and Applications*. Springer-Verlag, 1997.

[78] T. Lee and C. Lam. Path switching–A quasi-static routing scheme for large scale ATM packet switches. *IEEE Journal on Selected Areas of Communications*, 15:914 – 924, 1997.

[79] Simon Leinen. UDP vs. TCP distribution, 2001. http://www.postel.org/pipermail/end2end-interest/2001-February/000215; http://www.postel.org/pipermail/end2end-interest/2001-March/000218.

[80] E. Leonardi, M. Mellia, F. Neri, and M. Ajmone Marsan. Bounds on average delays and queue size averages and variances in input-queued cell based switch. In *Proc. of IEEE Infocom 2001*, Anchorage, AK, 2001.

[81] S. Li and N. Ansari. Input-queued switching with QoS guarantees. In *Proc. of IEEE INFOCOM '99*, pages 1152 – 1159, New York, NY, March 1999.

[82] Nikolay Likhanov and Ravi R. Mazumdar. Cell loss asymptotics in buffers fed with a large number of independent stationary sources. *Journal of Applied Probability*, 36:86–96, 1999.

[83] Page maintaned by Jörg Widmer. Implementation of the TCP-Friendly Congestion Control Protocol (TFRC), February 2000. http://www.icir.org/tfrc.

[84] Matt Mathis and Raghu Reddy. Enabling High Peformance Data Transfers. http://www.psc.edu/networking/perf_tune.html.

[85] Matthew Mathis, Jeffrey Semke, Jamshid Mahdavi, and Teunis Ott. The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm. *Computer Communication Review*, 27(3), July 1997.

[86] N. W. McKeown, V. Anantharam, and J. Walrand. Achieving 100% Throughput in an Input-Queued Switch. In *Proc. of IEEE INFOCOM*, 1996.

[87] A. Mekkittikul and N. W. McKeown. A practical algorithm to achieve 100% throughput in input-queued switches. In *Proc. of IEEE INFOCOM*, 1998.

[88] Vishal Misra, Wei-Bo Gong, and Don Towsley. A Fluid-based Analysis of a Network of AQM Routers Supporting TCP Flows with an Application to RED. In *SIGCOMM'2000*, Stockholm, Sweden, August/September 2000.

[89] R. Motwani and P. Raghavan. *Randomized algorithms*. Cambridge University Press, 1995.

[90] NIST Net Emulation Package. http://is2.antd.nist.gov/itg/nistnet, 2002.

[91] Clayton M. Okino. A Framework for Performance Guarantees in Communication Networks. Ph.D. Dissertation, UCSD, 1998.

[92] Teunis Ott, J. H. B. Kemperman, and Matt Mathis. The Stationary Behavior of Ideal TCP Congestion Avoidance. Technical Report In progress, August 1996, Bellcore, ftp://ftp.bellcore.com/pub/tjo/TCPwindow.ps, August 1996.

[93] Jitendra Padhye, Victor Firoiu, Don Towsley, and Jim Kurose. Modeling TCP Reno Performance: A Simple Model and its Empirical Validation. *IEEE/ACM Trans. on Networking*, 8(2):133–145, 2000.

[94] A. K. Parekh and R. G. Gallager. A generalized processor sharing approach to flow control in integrated services networks: The single node case. *IEEE/ACM Trans. on Networking*, 1-3:344–357, June 1993.

[95] Vern Paxson. End-to-End Internet Packet Dynamics. *IEEE/ACM Trans. on Networking*, pages 277–292, June 1999.

[96] K. K. Ramakrishnan and Raj Jain. A Binary Feedback Scheme for Congestion Avoidance in Computer Networks with a Connectionless Network Layer. In *Proc. of ACM SIGCOMM'88*, pages 303–313, 1988.

[97] Reza Rejaie, Mark Handley, and Deborah Estrin. RAP: An End-to-end Rate-based Congestion Control Mechanism for Realtime Strems in the Internet. In *IEEE Infocom 1999*, New York, USA, 1999.

[98] H. Sariowan. *A Service Curve Approach to Performance Guarantees in Integrated Service Networks*. PhD thesis, UCSD, 1996.

[99] W. Richard Stevens. *TCP/IP Illustration Volume 1: The Protocols*. Addison Wesley, 1994.

[100] I. Stoica and H. Zhang. Exact Emulation of an Output Queueing Switch by a Combined Input Qutput Queueing Switch. In *Proc. of IEEE IWQoS*, Napa, CA, 1998.

[101] Rockafellar R. T. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[102] V. Tabatabaee, L. Georgiadis, and L. Tassiulas. QoS Provisioning and Tracking Fluid Policies in Input Queuing Switches. *IEEE/ACM Trans. on Networking*, 9(5), October 2001.

[103] H. Takagi. Analysis and application of polling models. In G. Haring, C. Lindemann, and M. Reiser, editors, *Performance Evaluation: Origins and Directions, Lecture Notes in Computer Science 1769*, pages 423–442, 2000.

[104] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 37(12):1936 – 1948, December 1992.

[105] The ATM Forum Technical Committee. Traffic Managment Specification Version 4.0. Technical Report af-tm-0056.000, ATM Forum, April 1996.

[106] Wim Vervaat. A Relation Between Brownian Bridge and Brownian Excursion. *The Annals of Probability*, 7(1):143–149, 1979.

[107] Milan Vojnović and Jean-Yves Le Boudec. Global TCP Modeling: the Limit Mean ODE and its Convergence. In *Workshop on the Modeling of Flow and Congestion Control Mechamisms*, Ecole Normale Superieure, Paris, http://www.ens.fr/∼mistral/tcp2.html, September 2000.

[108] Milan Vojnović and Jean-Yves Le Boudec. Some Observations on Equation-based Rate Control. In *Proc. of ITC-17*, pages 173–184, Salvador, Bahia, Brazil, December 2001.

[109] Milan Vojnović and Jean-Yves Le Boudec. Stochastic Analysis of Some Expedited Forwarding Networks. In *Proc. of IEEE Infocom 2002*, New York, USA, June 2002.

[110] Milan Vojnović and Jean-Yves Le Boudec. Stochastic Bound on Delay for Guaranteed Rate Nodes. *IEEE Communications Letters*, 6(10):449–451, October 2002.

[111] Milan Vojnović and Jean-Yves Le Boudec. Bounds for Independent Regulated Inputs Multiplexed in a Service Curve Network Element. *IEEE Trans. on Communications*, 51(5):735–740, May 2003.

[112] J. von Neumann. A certain zero-sum two-person game equivalent to the optimal assignment problem. *Contributions to the Theory of Games*, 2:5 – 12, 1953.

[113] Ward Whitt. *Stochastic-Processes Limits*. Springer, 2002.

[114] Damon Wischik. The output of a switch, or, effective bandwidths for networks. *Queueing Systems*, 32:383–396, 1999.

[115] Yin Zhang, Nick Duffield, Vern Paxson, and Scott Shenker. On the Constancy of Internet Path Properties. In *Proc. of ACM Sigcomm Internet Measurement Workshop*, November 2001.

Appendix A

Details about our Experiments

The purpose of this appendix is:

- to give a concise description of the configuration of our experiments;
- to justify why we have taken a particular design of an experiment;
- to highlight a few caveats that we encountered.

A.1 Basic Software Components

A.1.1 TFRC

We used the experimental TFRC code available at [83]. The original code does not conform to some parts of TFRC specification [54]. We made adjustments described as follows.

Adjusting the Feedback Control Data

The feedback conveyed from the receiver to the sender was the send rate computed at the receiver. This required the samples of the round-trip times, which are measured at the sender, to be sent to the receiver. The moving-average of the round-trip time samples and the estimate of the loss-event rate were used to compute the send rate. The feedback, as just described, does not conform to TFRC specification [54]; by the specification, the feedback is an estimate of the loss-event rate. Adjusting the original code was a little tweak.

Biased Estimation of the Round-Trip Time

One issue in the original code, which we believe is worth mentioning, and we do not consider as a correct design, is the fact that the exponentially-weighted

smoothing of the round-trip times is executed per each packet arrival at the receiver. This in fact corresponds to on-line estimating the event-average of the round-trip time at packet arrival instances at the receiver. However, we know that the event-average of the round-trip time that acts in TCP throughput formulae is the event-average by sampling the round-trip time once in a round-trip time round. It is more accurate to run an estimator of the round-trip time per each report received at the sender, given that reports are roughly sent per each smoothed round-trip time. Our estimation of the round-trip time was made at the sender as described above, as such, it conforms to the specification [54], see Section 4.3 therein.

Biased Estimation of the Loss-event Rate: Under-prediction by Over-counting

We discovered that the loss-event intervals at the TFRC receiver of the original experimental code are over-counted. TFRC detects loss-events by emulating triple duplicated acknowledgments of TCP. When the receiver counts three out-of-order packets, a loss-event is declared. A loss-event interval is computed as the number of packets sent since the last loss-event, with the third out-of-order packet not counted. However, we discovered that these out-of-order packets were roughly counted twice in the computation of a loss-event interval. The over-counting would be exactly equal to 2, if the network would guarantee no packet disordering. Note that over-counting of the loss-event interval means *underestimating* the loss-event rate. This bias is not of a fundamental nature. The relative error is of course small if the loss-event rate is small (large loss-event interval). However, we found that the relative error becomes significant and discernible for larger loss-event rates. This is demonstrated later. After we discovered this bug, we re-ran some of our experiments. A part of the experiments were not repeated, however, we account for the bias due to the over-counting bug in the analysis of the results.

Other Adjustments of the Code

We made some further adjustments to the original TFRC experimental code:

- we used the basic moving-average estimator of the expected loss-event interval. This, in the code, corresponds to `LOSS_V1`;
- we disabled a heuristic control law that is based on delay;
- we disabled the gradual increase of the send rate;
- in some of our experiments, we disabled the control law of the comprehensive control.

A.1.2 TCP

We used a system implementation of a given host. The data was sent over a fixed duration of an experiment by executing `write` persistently on a TCP socket.

A.1.3 Tools that we Used

We used:

- a utility to analyze TCP from a tcpdump file, which was generously given to us by Chadi Barakat [12]. This utility infers loss-events, the round-trip time, window, and some other TCP-specific control variables, from a tcpdump file that contains TCP packets captured in both forward and backward direction of a connection. The utility was originally written for TCP NewReno. We made a few small adjustments of the code, including: accommodating TCP Sack/Fack, making the code to account for the wrapping of TCP segment sequence numbers, detecting undone loss-events;

- some Perl scripts were designed to automate the measurement process. We took the scripts from [83] as an initial design template.

A.2 Caveats

We have to bear in mind that TFRC, as used in our experiments, is a user space program. In contrast, TCP is a kernel space program. TFRC runs on a UDP socket. To the best of our knowledge, a TFRC kernel implementation did not exist at the time we were performing our experiments. Although we tried to get an in-depth understanding of how TFRC and TCP packets are treated inside the kernel (in particular, Linux kernel) we lack a through understanding.

A.2.1 A Back-pressure for TCP, not for UDP

We observed that an experiment with several TFRC and TCP senders running on a single host may not be a good design choice. Our first observation was that TCP sees a much larger loss-event rate. After careful examination of TCP congestion window dynamics, we observed that in some cases the window was halved, even though no packets had been retransmitted. This means that TCP reduced its congestion window as if there would be a loss-event, even in the absence of packet losses. We explored both TCP Linux and FreeBSD Linux code, and found indications that there exists a back-pressure mechanism (so called, congestion indication of the local device, in Linux code). There does not seem to exist such a mechanism for UDP. Hence, running a TFRC and a TCP sender on a single host may give a preferential treatment to TFRC due to the back-pressure mechanism. This observation led us to design our experiments

such that test TCP and TFRC senders are run on two separate hosts, with no other experimental traffic on these hosts.

A.2.2 Enabling TCP Window Scaling under Linux

The TCP congestion window is constrained by the value advertised by the receiver. TCP window scaling is an option that enables a receiver to advertise receive window larger than $2^{16} = 65535$ bytes. Without this option, the TCP congestion window would be upper bounded by 65535. This upper bound can be effective for Internet channels with sufficiently large bandwidth-delay product. A necessary condition to enable TCP window scaling is to set up a flag in

```
/proc/sys/net/ipv4/tcp_window_scaling
```

which is set by default. However, this is not a sufficient condition. The hard limits on the send and receive TCP socket buffer lengths in bytes are, respectively, specified in

```
/proc/sys/net/core/tcp_wmem
/proc/sys/net/core/tcp_rmem
```

which are by default set to 65535 bytes. These values have to be enlarged to make TCP window scaling actually work. Some handy guidelines on setting these and some other configuration parameters specific to various TCP implementations are nicely compiled in [84]; also, see [7] for definitions of `ipsysctl` configuration parameters.

A.3 Lab experiments

A.3.1 Configuring a Queue Discipline with `tc`

We used the Linux DiffServ architecture, which is a part of the Linux kernel network implementation since kernel 2.2. It is controlled by the `tc` command (see, e.g., [40, 3, 57]). This command allows to set a broad range of traffic shaping and filtering options. For our experiments, we only used this tool to configure either DropTail or RED [46] queue discipline. A command for setting up a packet-mode DropTail queue discipline to an interface `ethX` with buffer length `L` packets is as follows

```
tc qdisc add dev ethX root pfifo limit Lp
```

The `pfifo` queue discipline serves the arrival packets transparently. A byte-mode DropTail queue can be setup similarly; one would only need to replace `pfifo` with `bfifo`, and Lp with Lb, to create a byte-mode DropTail with buffer length L bytes. We aimed to configure RED queue discipline to roughly match those in our ns-2 simulations; the setting is displayed in Table A.1, see also Figure A.3 for the definitions of the parameters. However, we were not able to

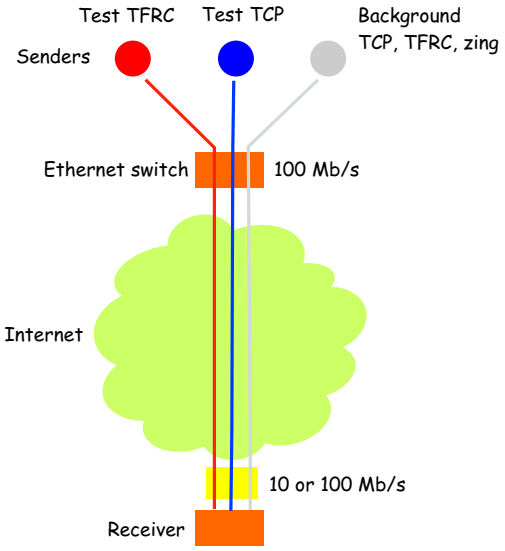


Figure A.1: Configuration of our Internet experiments. The background traffic consists of an equal number of TCP and TFRC connections, and two `zing` flows. A `zing` was configured to produce a Poisson stream of packets with intensity 50 packets per second (packet size equal to 256 B).

configure an *exact* match, because the *gentle*[1] variant of RED was not implemented in the Linux kernel. We consulted some engineering guidelines provided in http://www.icir.org/floyd/REDparameters.txt. We note that RED was configured in Linux with two specific configurations: The average queue length is measured in bytes and the packet dropping is done in packet-mode. With the `tc` utility, we set RED as, $q_m = m$, $q_M = M$, limit=L, $p_M = p$, and service rate C Mb/s,

```
tc qdisc add dev ethX root red limit L ...
... min m max M avpkt a burst b ...
... probability p bandwith cMbit
```

Other configuration parameters are defined as follows. Parameter a is the average packet length in bytes. The parameter burst is defined in the original RED reference [46]; it corresponds to the maximum number of packets that can arrive in a single batch, given that before the arrival the queue is empty, and the moving-average queue estimator is equal to 0, such that the resulting moving-average queue length computed on the given sequence of packets in the batch is not larger than the minimum queue threshold, q_m. The time constant in the moving-average estimator, w, is assumed to take values on the set $\mathcal{W} = \{1/2^k,$

[1]The "gentle" variant of RED refers to a specific dropping function, other than displayed in Figure A.3; see http://www.icir.org/floyd/REDfunc.txt

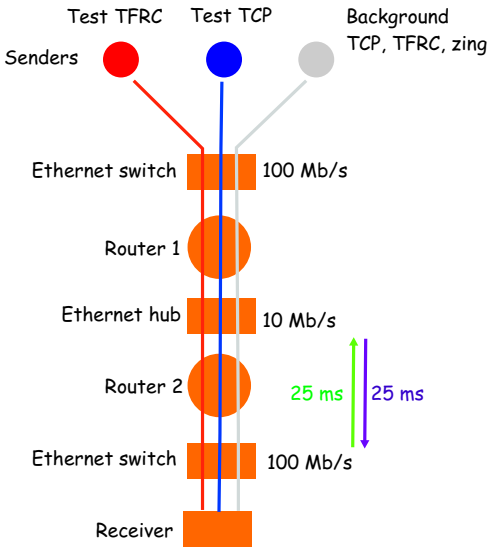


Figure A.2: Configuration of our lab experiments. All end-hosts and the two routers are Linux hosts (kernel 2.4.19). The configuration is designed such that router 1 would be a bottleneck. On the router 1 we configure queueing disciplines at the outgoing interface to the router 2. On the router 2, we run NISTNet [90], a wide-area network emulator, which we only use to add a fixed delay through the router 2 in both directions from the senders to the receiver and back.

$k = 1, 2, \ldots, 32\}$. Given q_m, a, and b, w is computed as the largest value in the set $\mathcal{W}$ such that

$$b + 1 - \frac{q_m}{a} < \frac{1 - (1 - w)^b}{w}$$

Some care needs to be exercised when setting a RED configuration, as there exist a number of constraints on the feasible values

$$\begin{aligned} limit &\geq q_M \\ q_M &> 2 * q_m \\ a &< q_m \\ b + 1 - q_m/a &\geq 1 \end{aligned}$$

In addition,

$$\frac{p_M}{q_M - q_m} = \frac{1}{2^j}$$

for some positive integer j, named `Plog` in the Linux implementation.

```
tc qdisc add dev eth1 root red ...
... limit 156250 min 15625 max 78125 ...
```

	q_m	q_M	limit	w	p_M
RED	$3/20B$	$5/4B$	$5/2B$	0.002	1/10
DropTail	n/a	n/a	$3/2B$	n/a	n/a

Table A.1: RED parameters; see Figure A.3 for definitions. B is the bandwidth-delay product in bytes. q_m and q_M are the minimum and maximum averaged queue thresholds in bytes. The parameter "limit" is the maximum value of the queue in bytes. w is the queue averaging constant defined on $(0, 1]$. p_M is the dropping/marking probability at q_M.

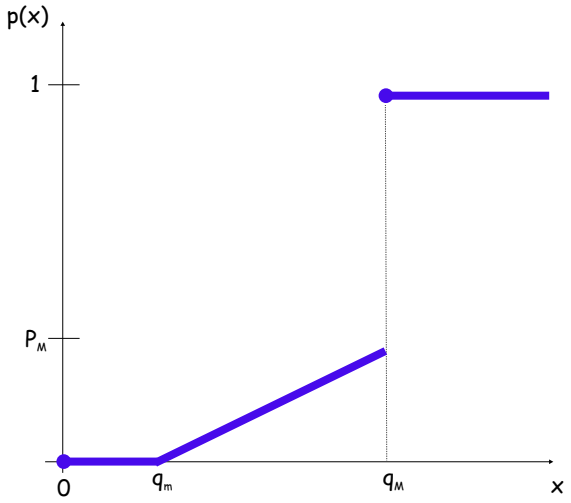


Figure A.3: The RED dropping function as used in our experiments.

```
... avpkt 500 burst 187 probability ...
... 0.1 bandwidth 10mbit
```

This results in the averaging time constant $w = \frac{1}{2^{20}} \simeq 0.002$.

A queue discipline can be removed by executing

```
tc qdisc del dev ethX root
```

A.4 Supplemental Experimental Results

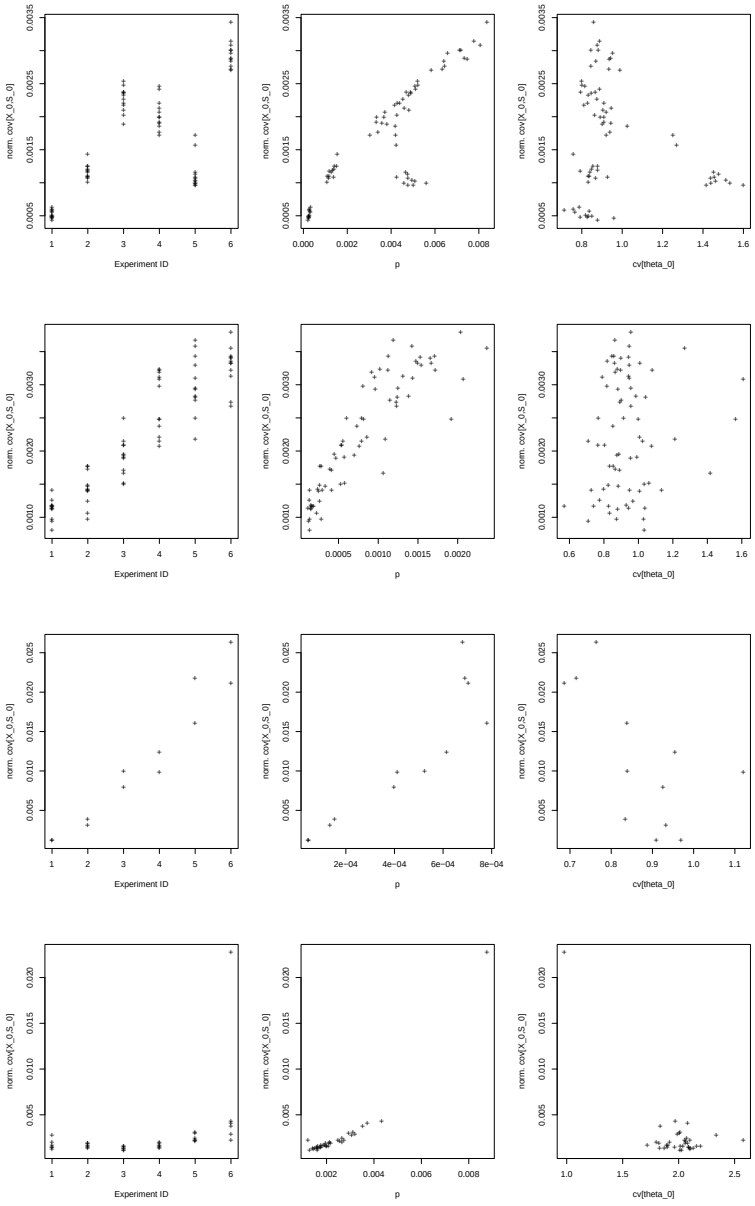


Figure A.4: Internet experiments. $\mathbf{cov}_N^0[X_0, S_0]/(\mathbf{E}_N^0[X_0]\mathbf{E}_N^0[S_0])$ for (Top to Bottom) INRIA, UMASS, KTH, UMELB.

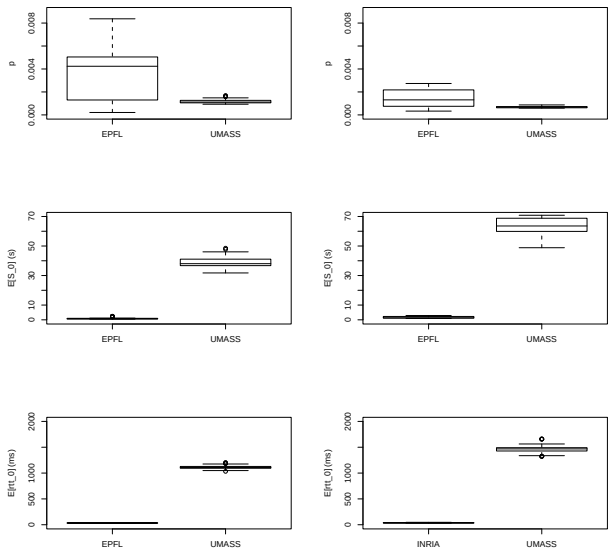


Figure A.5: Cable modem: Boxplots of empirical estimates of the loss-event rate, average inter loss-event time, average round-trip time sampled once in a round-trip time round. (Left) TFRC, (Right) TCP.

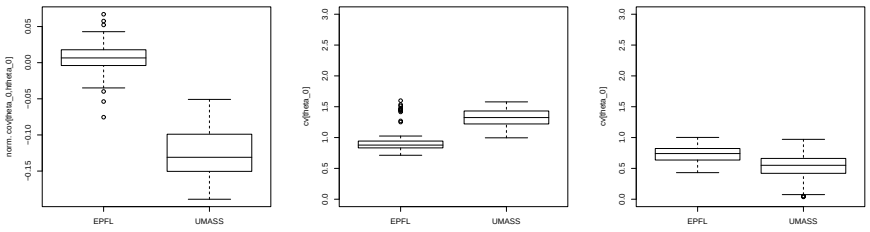


Figure A.6: Cable modem: (Left) the normalized covariance $\mathbf{cov}_N^0[\theta_0, \hat{\theta}_0]p^2$ for TFRC, (Middle) the coefficient of the variation $\mathbf{cv}_N^0[\theta_0]$ for TFRC, (Right) the coefficient of the variation $\mathbf{cv}_N^0[\theta_0]$ for TCP.

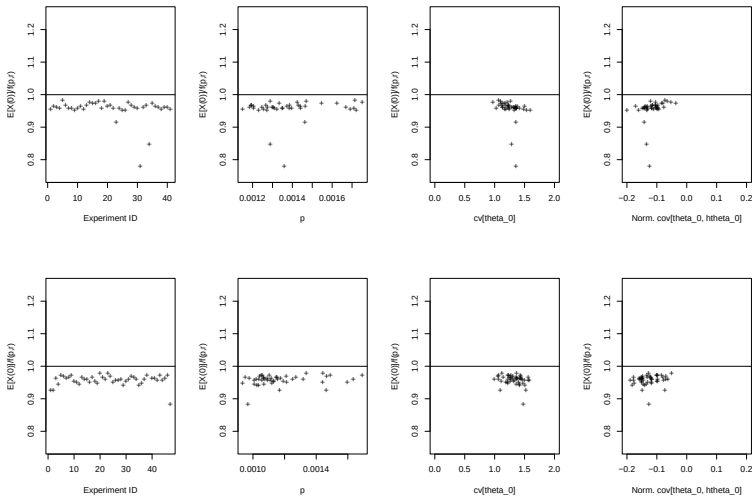


Figure A.7: Cable modem: Is TFRC conservative? (Top to Bottom) EPFL, UMASS. In all the cases, TFRC is conservative, but moderately. Note that in the experiments, the loss-event rate is small—by Claim 1, a moderate conservativeness is expected.

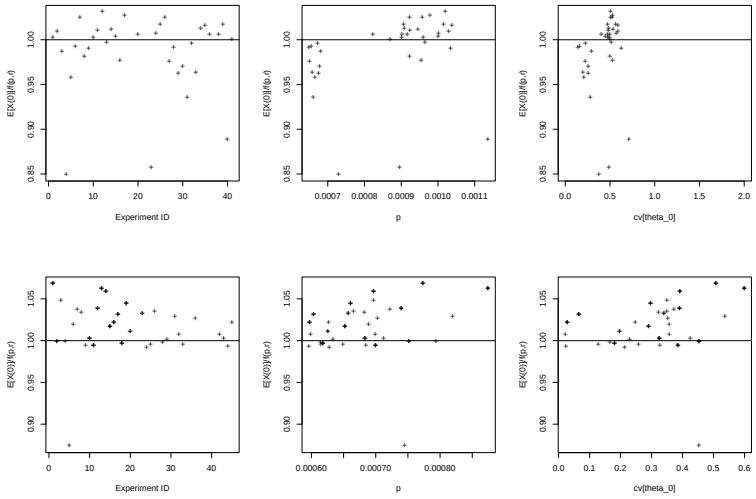
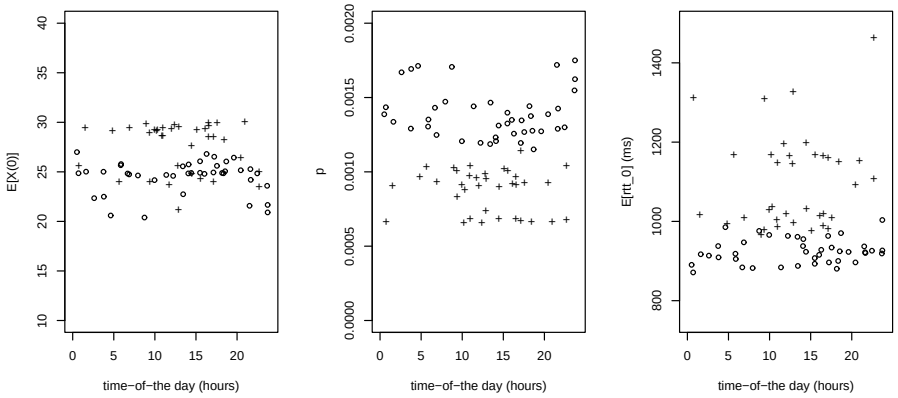
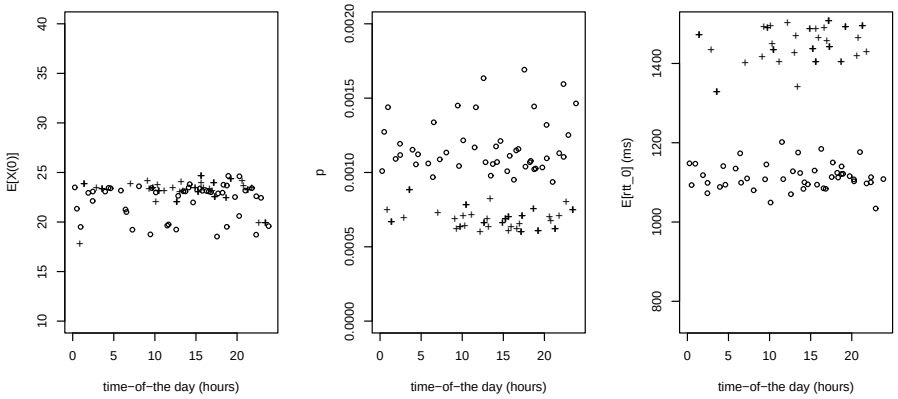


Figure A.8: Cable modem: Does TCP conform to its formula? (Top to Bottom) EPFL, UMASS. In most cases, TCP overshoots the value predicted by PFTK formula, but slightly.



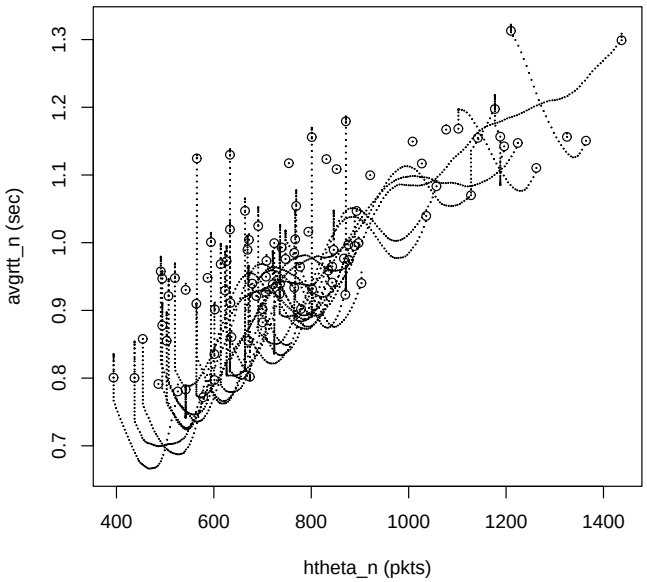
EPFL



UMASS

Figure A.9: Cable modem: TFRC throughput against TCP throughput. (Top) EPFL (Bottom) UMASS. In the plots, "circles" are for TFRC, "plus signs" are for TCP.

EPFL



UMASS

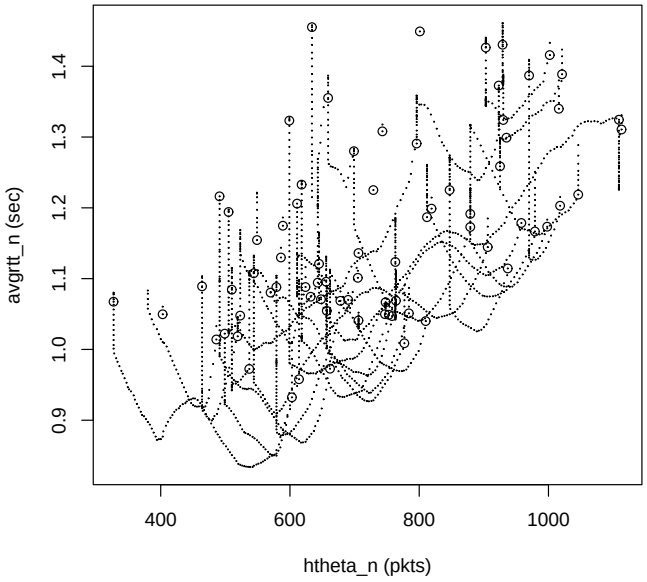


Figure A.10: LAN to cable modem: the scatter-plot of $\hat{\theta}_n(T'_n)$ against $\hat{R}(T'_n)$, T'_n is the arrival instant of the nth report at the sender. The top plot is for the sender at EPFL, the bottom plot is the sender at UMASS. The circles depict loss-events. Just after a loss-event, the estimator of the loss-event interval remains constant for some time (vertically aligned dots emanating from a circle). Then, the loss-event interval estimator begins to increase, due to the comprehensive control. For some interval of time, after a loss-event, the round-trip time estimator decreases, and then typically increases, but not always. This increase of the round-trip time should be due to the increase of the send rate, which amounts to increase of the queueing delay. A global positive correlation is evident.

List of Figures

List of Tables

Publications

[1] M. Andrews and M. Vojnović, "Scheduling Reserved Traffic in Input-Queued Switches: New Delay Bounds via Probabilistic Techniques," *IEEE Journal on Selected Areas in Communications, Special Issue on High-Performance Optical/Electronic Switches/Routers for High-Speed Internet*, Vol. 21, No. 4, May 2003, pp 595–605.

[2] M. Andrews and M. Vojnović, "Scheduling Reserved Traffic in Input-Queued Switches: New Delay Bounds via Probabilistic Techniques," *to appear in Proc. of IEEE INFOCOM 2003*, San Francisco, CA, April 1–3, 2003.

[3] M. Vojnović and J.-Y. Le Boudec, "Bounds for Independent Regulated Inputs Multiplexed in a Service Curve Network Element," *IEEE Trans. on Communications*, Vol. 51, No. 5, May 2003, pp 735–740.

[4] M. Vojnović and J.-Y. Le Boudec, "Stochastic Bound on Delay for Guaranteed Rate Nodes," *IEEE Communications Letters*, Vol. 6, No. 10, October 2002, pp. 449-451.

[5] M. Vojnović, J.-Y. Le Boudec, D. Towsley, and V. Misra, "A Note on the Stochastic Bias of Some Increase-Decrease Congestion Controls: High-Speed TCP Case Study," *invited paper to PFLDNet 2003 (1st Int'l Workshop on Protocols for Fast Long-Distance Networks)*, CERN, Geneve, Switzerland, February 3–4, 2003; chairs: J.-P. Martin-Flatin and S. Low, http://datatag.web.cern.ch/datatag/pfldnet2003.

[6] M. Vojnović and J.-Y. Le Boudec, "On the Long-Run Behavior of Equation-Based Rate Control," *in Proc. of ACM SIGCOMM 2002*, Pittsburgh, PA, August 19-23, 2002, pp 103–116.

[7] M. Vojnović and J.-Y. Le Boudec, "Elements of Probabilistic Network Calculus for Packet Scale Rate Guarantee Nodes," *in Proc. of MTNS'02 (the 15th Int'l Symp. on Mathematical Theory of Networks and Systems)*; invited session on Computer Networks; session chaired by Martin Haeengi; University of Notre Dame, South Bend, IN, August 12–16, 2002.

[8] M. Vojnović and J.-Y. Le Boudec, "Stochastic Analysis of Some Expedited Forwarding Networks," *in Proc. of IEEE INFOCOM 2002*, New-York, NY, June, 2002.

[9] M. Vojnović and J.-Y. Le Boudec, "Bounds for Independent Regulated Inputs Multiplexed in a Service Curve Network Element," *in Proc. of IEEE GLOBECOM 2001*, San Antonio, TX, November, 2001, pp 1857–1861.

[10] M. Vojnović and J.-Y. Le Boudec, "Some Observations on Equation-Based Rate Control," *in Proc. of ITC-17*, Salvador da Bahia, Brazil, December, 2001, pp 173–184, Best Student Paper Award.

[11] M. Vojnović, J.-Y. Le Boudec, and C. Boutremans, "Global Fairness of the Additive-increase and Multiplicative-decrease with Heterogeneous Round-trip times," *in Proc. of IEEE INFOCOM 2000*, Tel-Aviv, Israel, March, 2000, pp 1303–1312.

[12] M. Vojnović and J.-Y. Le Boudec, "How (Un)Fair are the ABR Binary Schemes, Actually?," *in Proc. of IEEE SoftCOM 1999*, Split-Rijeka-Trieste-Venice, Croatia/Italy, October 1999, pp 411–424.

Curriculum Vitae

Milan Vojnović was graduated from the University of Split, Croatia, in 1995, with a B.Sc. degree in electrical engineering. He earned his M.Sc. degree in electrical engineering in 1998, with first class honors, also from the University of Split. He attended the graduate school in communication systems, EPFL, in 1998/1999. He was awarded a three-year EPFL Ph.D. fellowship. He began his Ph.D. work at EPFL in 1999 with the Laboratory for Computer Communications and Applications, now within the School of Computer and Communication Sciences. He undertook an internship in August-October 2001 with the Mathematics Research Center, Bell Laboratories, Lucent Technologies, Murray Hill, NJ.

He was awarded best student paper award at ITC-17, for his work on the analysis of equation-based rate control, co-authored with Prof. Jean-Yves Le Boudec.