Large Deviations in High Speed Communication Networks

by

Ioannis Ch. Paschalidis

Diploma, National Technical University of Athens, June 1991
S.M., Massachusetts Institute of Technology, February 1993

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1996

Author .....
.....
Department of Electrical Engineering and Computer Science
May 17, 1996

Certified by.....
.....
Dimitris Bertsimas
Professor of Operations Research
Thesis Supervisor

Certified by.....
.....
John N. Tsitsiklis
Professor of Electrical Engineering
Thesis Supervisor

Accepted by .....
.....
Frederic R. Morgenthaler
Chairman, Departmental Committee on Graduate Students

Large Deviations in High Speed Communication Networks

by

Ioannis Ch. Paschalidis

Submitted to the Department of Electrical Engineering and Computer Science
on May 17, 1996, in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Abstract

High speed communication networks accommodate various types of traffic including voice, video and real-time data. Such traffic, being sensitive to congestion phenomena (packet losses, and excessive delays), requires *Quality of Service (QoS)* guarantees. We quantify QoS by the probability of congestion, which should be extremely small, e.g. on the order of 10^{-9}. We use *Large Deviations* techniques to estimate the leading exponent of this probability in various settings.

In a single class setting, we consider an acyclic network and compute the leading exponent of congestion probabilities in each node, by decomposing the problem to a series of single node problems. In a multiclass setting, we consider a multiplexer with segregated buffers and calculate the leading exponent of congestion probabilities for each type of traffic, under various scheduling policies. We relate the problem to a deterministic optimal control problem, which we explicitly solve. Optimal state trajectories of the control problem correspond to typical congestion scenarios. Both in the single and multiclass settings our results explicitly characterize the *most likely way* that congestion occurs.

We show how to apply our performance analysis results to prevent congestion. We devise an admission control mechanism that provides loss and delay QoS guarantees which can be distinct for each type of traffic. This generalizes the notion of effective bandwidth in the multiclass case.

Finally, we propose an importance sampling technique for variance reduction in simulations of loss probability in certain single class buffers that multiplex a large number of calls. We show numerical evidence of dramatic reduction in simulation time versus direct Monte Carlo simulation.

Thesis Supervisor: Dimitris Bertsimas
Title: Professor of Operations Research

Thesis Supervisor: John N. Tsitsiklis
Title: Professor of Electrical Engineering

Acknowledgments

I wish to express my deepest appreciation to both my research advisors Prof. Dimitris Bertsimas and Prof. John Tsitsiklis. Their contribution throughout these years with ideas, support and constant encouragement was indispensable and greatly influenced this thesis. They gave me the chance to work in the stimulating environment of MIT and they created a very friendly and supportive atmosphere. I will definitely remember some of the "proof debugging" sessions with "Yanni" (John's real name) during which I was being constantly surprised by how easily he could follow and improve long and messy technical arguments. I will also remember some truly motivating meetings with Dimitri at Sloan, after which I was feeling that I could achieve anything I wanted to. I can't think of anyone who can match Dimitri's "problem posing" and Yanni's "problem solving" ability (sometimes I wonder what my role was :-)). I consider Yanni and Dimitri the "dream team" of advisors and I also greatly appreciate their care and concern for me as an individual. For all that I am deeply indebted to them.

I would like to thank Prof. Bob Gallager for serving in my thesis committee, as well as for his feedback and intuition during our very enlightening meetings. Most importantly though, for being a constant source of inspiration to me and to everyone else at LIDS.

Special thanks go to my graduate counsellor Prof. Dimitri Bertsekas for his help and support throughout these years and to Prof. Sanjoy Mitter for his keen interest in my progress. Also, lots of thanks to Dr. Irvin Schick for his help and support.

I wish to thank Dr. Debasis Mitra and Dr. Alan Weiss who arranged my visit to Bell Labs, where most of Chapter 7 took shape, for their help and support and for making me feel welcome. Alan suggested this problem to me and influenced the outcome substantially with his ideas and intuition. Lots of thanks to David Tse (now a Professor at Berkeley) for his friendship and our helpful discussions at LIDS and Bell Labs.

Many other people have contributed in various ways in this thesis, among them I would like to thank: Raj Srinivasan (for help with the random number generator),

Larry Wein, and Offer Zeitouni.

LIDS is an ideal environment, mainly due to its people, so many thanks to everyone there, especially my friends, for making it such a great place. Special thanks to the administrative team, wonderfully managed by Kathleen O' Sullivan, for taking care of every problem that came across my path.

On a more personal note, I would like to thank all my best friends (they know who they are). I can not resist mentioning, among them, Dimitris and Georgia, Angela and Bo, and Antonis and Fania, for being my "Boston family", Regina for being my older pal, and Nanno for keeping me busy inventing ways of teasing her.

More importantly, I wish to express my love to my parents Charalampos and Vasso. I owe most of what I am today to them and nothing can describe my feelings. They made my journey into life a wonderful experience. Last, but definitely not least, my love to my wife Gina for all she has done for me, but predominantly for setting out her journey to "Ithaca" with me. I do

> "pray that the road is long
> full of adventure, full of knowledge.
>
> Pray that the road is long.
> That the summer mornings are many,
> that" *we* "will enter ports seen for the first time,
> with such pleasure, with such joy !" *Constantine P. Cavafy, Ithaca*

For all that this thesis is dedicated to my parents and Gina.

The research in this thesis was supported by a Presidential Young Investigator award DDM-9158118 with matching funds from Draper Laboratory, and by the ARO under grant DAAL-03-92-G-0115.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years we have witnessed an explosion in the number of Internet users, mainly due to the development of software (World Wide Web browsers) that make Internet resources easily accessible, even to users that are not well versed in computers. Being "on-line" has become a way of life, and e-mailing as well "surfing" the net indispensable routine activities. In this new world, where connectivity is a must, new services are emerging such as interactive TV, teleconferencing, Video-On-Demand and remote access to information servers, to name a few. Moreover, educational and medical services such as remote teaching, remote medical diagnosis and even performing medical procedures remotely, have been developed or are in the development process. All these services are of a truly multimedia character, meaning that they require from the network transfers of large amounts of data, images, audio and even video. Advances in hardware (fiber optics, switching) have made available enough bandwidth to satisfy the resulting dramatic increase in bandwidth requirements. The challenge is how to manage the network resources (bandwidth) in order to support such multimedia services.

The TCP/IP protocol used in the Internet as well Ethernet, FDDI and other popular *Local Area Network (LAN)* protocols can only provide *"best effort"* service. That is, the protocol does its best to accommodate the offered load without making particular promises to the users. Congestion causes packet losses, due to buffer over-

flows, and excessive delays, thus in the event of congestion, packets may be dropped or arbitrarily delayed. The protocols achieve reliable communication by retransmitting packets that have being dropped, which generates further delays. This may be tolerable for e-mail and file transfers but results in severe degradation of the *Quality of Service (QoS)* provided to real-time services as the ones already discussed. Hence, the network should have the ability to guarantee certain QoS parameters to the user.

In the last decade a new transport protocol has been standardized, the *Asynchronous Transfer Mode (ATM)* protocol where information is traveling in tiny fixed-length packets of 53 bytes each (48 bytes information, 5 bytes header) [BG91]. ATM networks are gaining popularity and a number of organizations are currently upgrading their LANs to ATM. Many specialists believe that soon Internet (TCP/IP) will be running on top of ATM. ATM switches with capacities of up to 10 Gb/s are currently available and more powerful switches are being developed. ATM networks are a particular flavor of B-ISDNs (Broadband Integrated Services Digital Networks) that as the name indicates have the capability of providing high transmission speeds and accommodate distinct types of services (handling voice, video and data). Figure 1-1 depicts one such communication scenario.

Independently of which protocol dominates, two are the critical questions that arise from the above discussion:

1. How we quantify the notion of *Quality of Service (QoS)* and how we measure it ?

2. How we design and operate a high-speed multimedia network in order to prevent congestion and to *guarantee* QoS to real-time services ?

These questions should be addressed in a setting that preserves the fundamental features of high speed networks, as outlined above. We believe that among such features two dominate:

1. *Networking.* The high speed network is a collection of switches (nodes) and sessions go through several nodes from origin to destination. Thus, the QoS

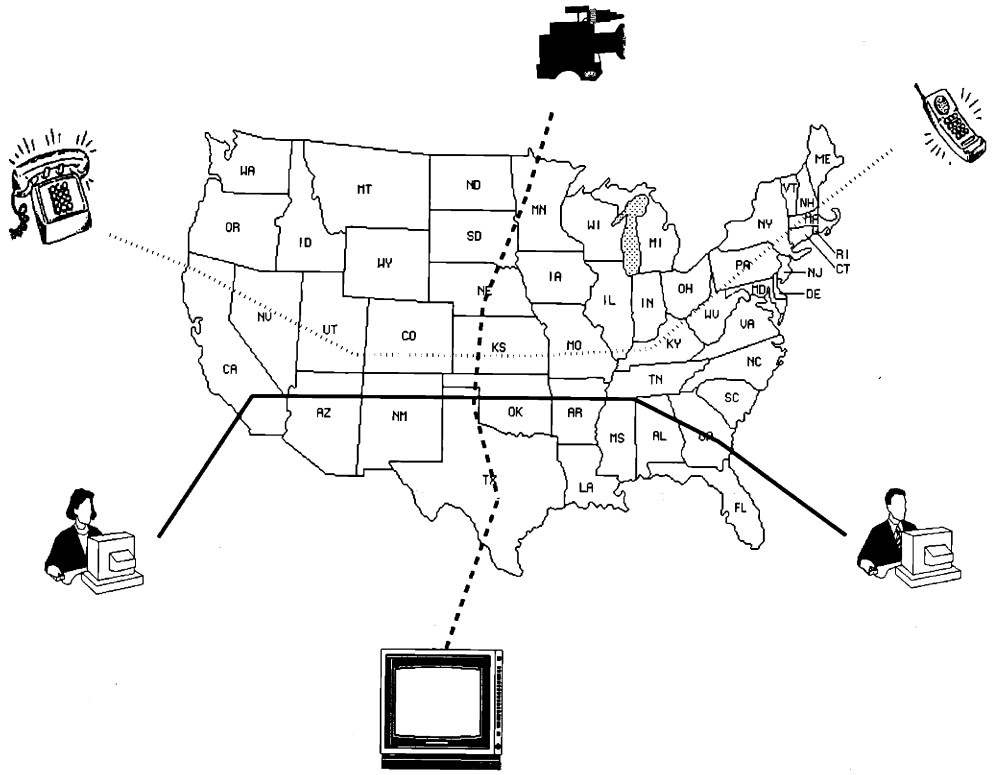

Figure 1-1: A scenario of multimedia communications via a B-ISDN network.

offered to the session by all nodes in its path has to be quantified and measured.

2. *Multimedia.* Possibly the dominant feature of such high speed networks is that they accommodate different types of traffic, including voice, video and data, which are quite different in character (hence, a different model may be needed for each type) and have different QoS requirements from the network.

1.1 QoS measures and Literature Review

One way to provide QoS guarantees is to ensure that packets losses do not occur

and delays stay bounded below a given maximum tolerable value that depends on the particular application (type of traffic). To attain such a goal, deterministic bounds on queue lengths and delays have to be determined in the network, which requires external input traffic to be also bounded. This deterministic approach was initially proposed by Cruz [Cru91a, Cru91b]. He lets $R(t)$ represent the instantaneous rate of traffic flowing on a specific link at time t. Therefore, for $y \geq x$, the integral $\int_x^y R(t)dt$ represents the total amount of traffic transmitted in the time interval $[x, y]$. A specific regulator (leaky bucket) is applied to the incoming traffic with the property that the regulated traffic satisfies

$$\int_x^y R(t)dt \leq \sigma + \rho(y - x)$$

where σ and ρ are constants. Under regulated incoming traffic, upper bounds are provided for delays and queue lengths and the idea is extended to networks. Parekh and Gallager developed this approach further in [PG93, PG94]. They consider the *generalized processor sharing (GPS)* policy, which was also proposed in [DKS90] under the name *fair queueing*, and they obtain worst-case upper bounds on delays and queue lengths in the network from source to destination.

Using such a worst-case approach one can dimension the buffers in the network and tune the regulators in the interface between the network and the users, so that congestion phenomena (packet losses and excessive delays) do not occur. However, to calculate deterministic bounds on queue lengths and delays, the worst-case behaviour of traffic sources is assumed. This may result in substantial underutilization of the network resources, since typical traffic in communication networks is bursty and on the average it uses much less bandwidth than its peak requirements. Moreover, although such a worst-case approach, where external input traffic is regulated, provides guarantees on what is happening in the network, it does not provide concrete guarantees to the application that uses the network. For instance, losses and large delays may be incurred in the leaky bucket regulator, even though packets that finally enter the network do not suffer from congestion.

In this thesis, as many authors in the literature, we take a philosophically different standpoint. Let $D_{\max}$ be the maximum tolerable delay to deliver a packet from source

to destination. Let also d denote the delay incurred by an arbitrary packet and P_l the packet loss probability (i.e. the probability that the packet is dropped due to buffer overflow). We seek to make losses and excessive delays rare events, i.e.,

$$\mathbf{P}[d > D_{\max}] < \epsilon \tag{1.1}$$

$$P_l < \epsilon \tag{1.2}$$

where ϵ is on the order of 10^{-9}. Real-time applications can tolerate such small frequencies of congestion phenomena. Hence, we quantify QoS by the probability of excessive delays and the loss probability. We refer to $D_{\max}$ and ϵ as *QoS parameters*, since they determine how well a particular application is treated. Determining such probabilities as the ones in (1.1) and (1.2) is a highly non-trivial task, since it essentially requires finding the distributions of waiting times and queue lengths in a multiclass network of G/G/1 queues with correlated arrival processes (since it is needed to model bursty traffic) and non-exponentially distributed service times. In this light, it is natural to focus on asymptotic regimes and determine the leading exponent of such small probabilities. *Large deviations* [Buc90, DZ93b] theory will be our main analytical tool in this thesis, since it provides techniques to obtain asymptotic expressions for the tails of arbitrary distributions.

One important requirement for the above outlined approach is that a detailed statistical description of the input traffic should be available. We will use Markov modulated processes and stationary processes with mild mixing conditions to describe input traffic. Such models approximate very well voice traffic (see [MAS88]) and have been used to describe video traffic (see [EHL+94, EM93, Kel96, SW95]) with satisfactory results. In Chapter 6 we describe such models in more detail and assess their performance through experiments. Data traffic is harder to model. Recently a certain statistical behaviour termed *self-similarity* has been observed in Ethernet traffic [LTWW94, Wil95] and some complicated models have been proposed [Nor94] to explain this behaviour. Objections on the existence of self-similarity in actual

traffic have been raised in [DLO+94], where the authors argue that non-stationarity causes the observed statistical behaviour. Although the issue is not yet settled, such models, which exhibit self-similarity, are beyond the scope of this thesis. We view our results as complementary to other approaches that handle such models and to worst-case deterministic approaches. Our principal focus is voice and video, which will definitely fill up a significant portion of the total traffic.

Large deviations techniques have been applied recently to a variety of problems in communications. A nice survey can be found in [Wei95]. The problem of estimating tail probabilities of rare events in a single class queue has received extensive attention in the literature [Hui88, GH91, Kel91, KWC93, GW94, EM93, TGT95] and has been approached by two main methodologies. The first one is to use large deviations arguments. Such results were first obtained in [Hui88], [Kel91], [GH91] and later in [KWC93]. A more elaborate Markov modulated model with multiple time-scales is used in [TGT95]. The second approach is to use spectral decomposition techniques and is used in [EM93] to estimate the tail probability of the queue length in a queue with a deterministic server and Markov modulated arrival process.

The extension of these ideas to networks appears to be a rather challenging problem. Researchers have been able to obtain some bounds on the tail probabilities for delays and queue lengths in various networks models (see [Cha94b, YS93]), but it is not clear whether these bounds are tight. Recently, large deviations results for two queues in tandem, with renewal arrivals and exponential servers, were reported in [GA94]. In [dVCW93], a very interesting approach is used to obtain results for networks with deterministic servers. The departure process from a single G/D/1 queue is characterized in the large deviations regime, using a discrete time model, in an attempt to treat the whole network inductively. The main focus of [dVCW93] is to apply the large deviations results obtained to resource management for networks. It is important to point out that the departure process is a very difficult process to obtain exact results for (see for example [BN90]). However, we should note that it is not very clear to us how the large deviations result for the departure process in [dVCW93] can be applied inductively. The crux of the matter is that [dVCW93] uses

a technical result from [DZ93a] in order to obtain the large deviations behaviour of the departure process. The latter result holds under certain technical assumptions on the arrival process. Since the departure process from a queue is the arrival process in an other downstream queue in the network, one would need at this point to verify that the same technical assumptions hold for the departure process. This is not done in [dVCW93] and appears to be rather difficult.

In a multiclass setting, the asymptotic tails of the overflow probabilities for the GPS policy with deterministic service capacity are obtained in [dVK95] and [Zha95]. The latter paper raises and addresses a technical difficulty not handled in [dVK95]. Both papers use a large deviations result for the departure process from a G/D/1 queue [dVCW93]. Tail overflow probabilities for the GPS policy and deterministic service capacity were also reported in [O'C95b, CW95]. The authors in [CW95] view the problem as a control problem where control variables are the capacity that the server allocates to each buffer, as a function of the current state. This approach has some technical problems with boundaries because it requires Lipschitz continuity of the controls. In [GGG+93] the authors suggest the use of the *longest queue first (LQF)* policy in high speed networks and use a deterministic model (only the rate of each incoming stream is known) to calculate buffer sizes that guarantee no loss with certainty. In [SW95] the authors consider the LQF policy in a system with two buffers and address the question of how one queue builds up when the other is large. They consider the M/M/1 version of the system (i.e., Poisson arrivals and exponential service times).

One other issue that arises in practice is the simulation of rare event probabilities, especially in situations where a very accurate estimate should be obtained or when asymptotic results obtained through large deviations techniques are tested and verified. The problem is that simulating a system to estimate a small probability ($< 10^5$) may take a long time since a huge sample size is required. Importance sampling has been used to substantially speed up the simulation in several cases [PW89, CHJS94]. Large deviations techniques are very useful in determining the change of measure to be used in importance sampling (see [Buc90]).

1.2 Results and Contributions of the Thesis

In this thesis we address the questions raised earlier in both a networking and a multimedia (multiclass) setting. As we outlined in Section 1.1, we use the loss probability and the probability of large delays (see Eqs. (1.1) and (1.2)) to quantify QoS.

1.2.1 A network result

To motivate our network result consider the particular "bold" session in Figure 1-2. It passes through a series of switches with associated buffers dedicated only to this

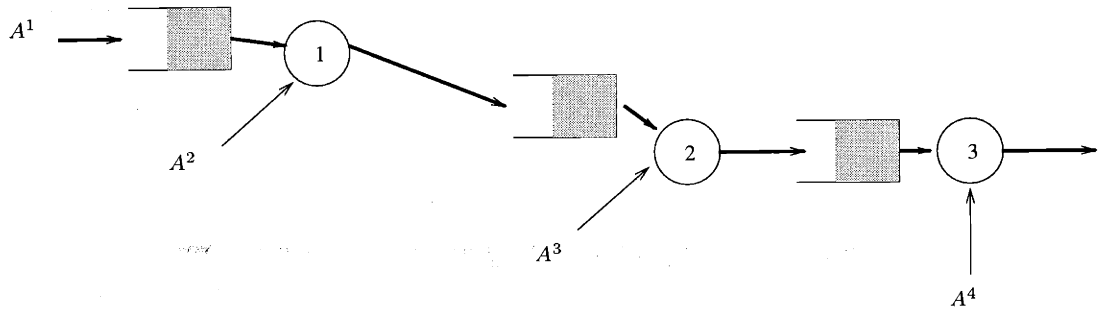

Figure 1-2: A session from origin to destination as it passes through a series of switches (nodes).

session. Even if the capacity of the switches is deterministic, due to the stochastic nature of the cross traffic in each switch, the "bold" session is facing a tandem network of G/G/1 with stochastic capacities. The objective is to determine the QoS parameters that all nodes deliver to the session. The largest loss and delay probabilities among all nodes determine the QoS delivered to the session by the network. Notice that since we are only considering the "bold" session and not the way that it interacts with other sessions we are in a single class setting and the *First Come First Serve (FCFS)* policy is a natural choice.

In Chapter 2, we consider a network model that generalizes the tandem network faced by the "bold" session. The results of that chapter were reported in [BPT94]. In particular, we consider a *single class, acyclic network* of G/G/1 queues. Customers arrive to the network in a number of independent streams and are treated uniformly. Different streams may share a queue and the first-come first-serve (FCFS) policy is implemented. A constant fraction p_{ij} of customers departing queue i, is routed to queue j and a fraction p_{i0} leaves the network. The aim is to derive large deviations results for the waiting time and the queue length observed by an arbitrary customer at different queues of the network. To this end, we initially seek to characterize the large deviations behaviour of the *aggregate arrival process* in each node. Our results are self-contained in the sense that we do not need the technical results of [DZ93a]. Instead, we impose certain assumptions on the external arrival processes and we characterize the large deviations behaviour of all the processes resulting from various operations in the network. For the network model that we are considering, these operations are *passing-through-a-queue* (the process resulting from this operation being the departure process), *superposition* of independent processes, and *deterministic splitting* of a process to a number of processes. We prove that the assumptions imposed on the external arrival processes are *preserved* by these operations, and thus we are able to apply these results inductively to obtain large deviations results for the aggregate arrival process in each node. As a by-product of our analysis we also obtain large deviations results for the internal traffic in the network. For a single queue, in isolation, we characterize the large deviations behaviour of the waiting time incurred by a typical customer and, by using ideas from distributional laws (see [BN95, BM92]), the large deviations behaviour of the queue length observed by a typical customer. Finally, we *combine* the large deviations behaviour of the aggregate arrival process in each node of the network with the results for a single queue to obtain the large deviations behaviour of the waiting time and queue length in each node.

Our approach provides particular insight on how these large deviations occur, by concretely characterizing the most likely path that leads to them. Characterizations of most likely paths were also obtained for the single queue case in [Asm82], [Ana88]

and [DZ93a]. Results similar to ours were independently obtained in [Cha94a], [CZ95] and [O'C95a]. In [Cha94a] the author obtained the large deviations behaviour for a network model of G/D/1 queues when the external arrival processes are bounded. In [CZ95] the authors obtain the large deviations behaviour of the departure process of a G/G/1 queue, in isolation.

It is interesting to note, that in order to obtain the large deviations behaviour of the superposition operation we prove a general result that connects the *stationary distribution* (i.e., as it is seen at a random time) and the *Palm distribution* (i.e., as it is seen by a typical customer) of a point process in the large deviations regime. This result could be of independent interest.

1.2.2 Multiclass Performance Analysis Results

To address the questions raised in the beginning of this chapter in a multimedia environment we consider a multiclass multiplexer (one node), with segregated buffers for each type of traffic. We seek to determine the QoS parameters that the switch delivers to each type of traffic under specific scheduling policies for sharing bandwidth. That is, the aim is to obtain loss and delay probabilities for each type of traffic.

In Chapters 3 and 5, we consider the *generalized processor sharing policy (GPS)* (loss and delay, respectively). According to this policy each type i of traffic receives a guaranteed constant ϕ_i fraction of the node's capacity. In Chapter 4, we consider the loss probability under the *generalized longest queue first policy (GLQF)*, which as the name indicates is a generalization of the LQF policy. In particular, there is a threshold level, β, and the server allocates all of its capacity to the first buffer, if the ratio of the queue length in the second buffer versus the queue length in the first buffer is below the threshold, otherwise it allocates all of its capacity to the second buffer. For $\beta = 1$ we have the LQF policy. The LQF policy can be viewed as an attempt to reduce the variance of delay between different types of traffic. Both the GPS and the GLQF policies are parametric policies and for specific values of the parameters reduce to strict priority policies. Thus, the performance of strict priority policies is

obtained as a corollary of our results (approximate results for priority policies are reported in [EM94]).

In Chapter 4, we compare the loss probability characteristics of the GLQF and the GPS policy and find that the first outperforms the second. However, this may be happening at the expense of greater delay. Though, since delay is due to long queues, it is intuitive that the GLQF policy tries to balance (with a β "bias") the delay of the two traffic streams. In any case, if only loss probability guarantees are needed, our results clearly suggest the use of the GLQF policy instead of the GPS.

Regarding the analysis technique, in the standard *large deviations* methodology we provide a lower and a matching (up to first degree in the exponent) upper bound on the buffer overflow and delay probabilities. We prove that congestion occurs in one of two *most likely* ways (modes of overflow) and we explicitly and in detail characterize these modes. We address the case of multiplexing two different traffic streams; for the general case of N streams our lower bound approach (which also determines the modes of overflow) can be easily extended. It should be noted, however, that there is an exponential explosion of the number of overflow modes (there are 2^{N-1} modes). Proving a tight upper bound for the case of N streams is still an open problem. Our results have important implications in traffic management of high-speed networks. They extend (in the GPS case) the deterministic, worst-case analysis of [PG93] to the case where a detailed statistical model of the input traffic is available.

More importantly, we provide an *optimal control formulation* of the problem, both in the GPS and the GLQF case. Our formulation is different from the one in [CW95] and does not fall into problems with the boundaries of the state-space. In particular, the exponent of the congestion probability is the optimal value of a corresponding control problem, which we explicitly solve. Optimal state-trajectories of the control problem correspond to the most likely modes of congestion; from the solution of the control problem we obtain a detailed characterization of these modes. This formulation, as will be apparent later, motivates the selection of congestion scenarios that are used to obtain the lower bound, a selection which is sort of arbitrary in most of the existing literature. This optimal control formulation is general enough

to include any scheduling policy. The only thing that changes with the policy is the dynamics of the system. Optimal control formulations are also used in [SW95] for large deviations results for jump Markov processes.

Our results consider the case of stochastic node capacity (in contrast to [dVK95, Zha95, O'C95b, CW95]). This makes it possible to treat more complicated service disciplines. Consider for example the case where we have a deterministic server and three types of traffic with dedicated buffers. We give priority to the first stream and use the GPS policy for the remaining streams. These two remaining streams face a server with stochastic capacity, a model of which can be obtained using the model for the arrival process of the first stream. Stochastic capacity significantly alters the way congestion occurs. To see this notice that in deriving their results [dVK95] and [Zha95] use the departure process from a G/D/1 queue. The large deviations behaviour of the departure process is different with deterministic and stochastic service capacity as it is pointed out in Chapter 2 and [BPT94, CZ95].

1.2.3 Admission Control

In Chapter 6, we apply the performance analysis results outlined in Section 1.2.2 to prevent congestion through admission control.

More specifically we consider the same model of the multiclass switch with segregated buffers for each type of traffic and under the GPS policy we

1. Dimension the buffers based on the maximum tolerable delay characteristics of the corresponding types of traffic, and

2. Devise an admission control algorithm on a call basis that guarantees to each type of traffic the required QoS parameters.

Among the main advantages of the proposed algorithm are

- It allows for the use of separate models for each type of traffic.

- Delivers to each type of traffic the required QoS parameters, which may be type-dependent.

- Takes into account both packet loss and delay in measuring congestion. To see the importance of this feature consider a switch with two buffers and high priority to the first buffer. Notice that guaranteeing the loss probability in both buffers is not sufficient for acceptable QoS since traffic in the second buffer may experience arbitrarily large delay.

Our admission control algorithm can be run on-line and is based on the calculation of an *admission region* that contains the set of loads under which QoS is guaranteed to all types of traffic. Such an admission region can be calculated off-line, if detailed statistics for the incoming traffic are available. If such statistics are not available the the admission control mechanism should be coupled with an on-line estimator.

Also, in Chapter 6, we report a couple of experiments with traffic models and actual traffic which assess the performance of our approach. These experiments show a substantial gain from statistical multiplexing versus more naive worst-case based admission control.

1.2.4 Quick Simulation

Lastly, in Chapter 7, we consider a particular single class buffer which accommodates a large number of calls and estimate the loss probability through simulation. Since, as we discussed earlier, direct Monte-Carlo simulation takes a large amount of CPU time, we apply importance sampling techniques. Based on the large deviations result in [BD94] we infer a *change of measure* that speeds-up the simulation dramatically.

1.2.5 Main Contributions

We next summarize the main contributions of the thesis:

1. We provide a rigorous large deviations analysis of single class acyclic networks of G/G/1 queues that enable us to determine the QoS that a particular session acquires from the network. In the course of the analysis we characterize the internal traffic in the network, in the large deviations regime, and prove a result that relates *Palm* and *stationary* probabilities in the same regime.

2. We introduce a deterministic optimal control approach to analyze the performance of the GPS and the GLQF policy in multiclass multiplexers. The optimal control approach yields

   (a) A tight lower bound on the dominant exponent of congestion probabilities, and

   (b) The *most likely way* that congestion occurs, in the sense that the optimal trajectories of the control problem correspond to the maximum probability scenario (sample path) of congestion.

3. We devise an *admission control mechanism* in multiclass switches that allows the use of distinct source models per type of traffic and provides *loss* and *delay* guarantees that can be type-dependent.

4. We use an *importance sampling* technique (introducing an appropriate *change of measure*) for obtaining the loss probability in a particular single class buffer through simulation. We show numerical evidence that the simulation speeds up dramatically when compared with direct Monte Carlo simulation.

1.3 Background Material

In this section we review some basic results on Large Deviations Theory that will be used in the rest of the thesis. An introductory reference to this theory and some of its important applications is the book by Bucklew [Buc90]. A very rigorous treatment can be found in the book by Dembo and Zeitouni [DZ93b]. Finally, the book by Shwartz and Weiss [SW95] specializes in Large Deviations for jump-Markov processes

and contains numerous applications of the theory to various problems, especially in communications.

The theory of Large Deviations is concerned with the estimation of rare event probabilities. Consider for instance a sequence of iid random variables X_i, $i \geq 1$, with mean $\mathbf{E}[X_1] = m$. The strong law of large numbers asserts that $\frac{\sum_{i=1}^n X_i}{n}$ converges to m, as $n \to \infty$, w.p.1. Thus, for large n the event $\sum_{i=1}^n X_i > na$, where $a > m$, (or $\sum_{i=1}^n X_i < na$, for $a < m$) is a rare event. In particular, its probability behaves as $e^{-nr(a)}$, as $n \to \infty$, where the function $r(\cdot)$ determines the rate at which the probability of this event is dropping. Cramér's theorem [Cra38] determines $r(\cdot)$, and is considered the first Large Deviations statement. Although, Cramér's theorem applies to iid random variables it has been extended by Gärtner and Ellis to cover dependent processes. We will next state the Gärtner-Ellis Theorem (see [Buc90] and [DZ93b]) which establishes a *Large Deviations Principle (LDP)* for random variables.

Consider a sequence $\{S_1, S_2, \dots\}$ of random variables, with values in $\mathbb{R}$ and define

$$\Lambda_n(\theta) \triangleq \frac{1}{n} \log \mathbf{E}[e^{\theta S_n}]. \tag{1.3}$$

For the applications that we have in mind, S_n is a partial sum process. Namely, $S_n = \sum_{i=1}^n X_i$, where X_i, $i \geq 1$, are identically distributed, possibly dependent random variables.

Assumption A

1. *The limit*

$$\Lambda(\theta) \triangleq \lim_{n\to\infty} \Lambda_n(\theta) = \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{\theta S_n}] \tag{1.4}$$

exists for all θ, where $\pm\infty$ are allowed both as elements of the sequence $\Lambda_n(\theta)$ and as limit points.

2. *The origin is in the interior of the domain $D_\Lambda \triangleq \{\theta \mid \Lambda(\theta) < \infty\}$ of $\Lambda(\theta)$.*

3. *$\Lambda(\theta)$ is differentiable in the interior of D_Λ and the derivative tends to infinity as θ approaches the boundary of D_Λ.*

4. $\Lambda(\theta)$ *is lower semicontinuous, i.e.,* $\liminf_{\theta_n \to \theta} \Lambda(\theta_n) \geq \Lambda(\theta)$, *for all* θ.

Theorem 1.3.1 ***(Gärtner-Ellis)*** *Under Assumption A, the following inequalities hold*

Upper Bound: *For every closed set* F

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left[\frac{S_n}{n} \in F\right] \leq -\inf_{a\in F} \Lambda^*(a). \tag{1.5}$$

Lower Bound: *For every open set* G

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\left[\frac{S_n}{n} \in G\right] \geq -\inf_{a\in G} \Lambda^*(a), \tag{1.6}$$

where

$$\Lambda^*(a) \triangleq \sup_{\theta}(\theta a - \Lambda(\theta)). \tag{1.7}$$

We say that $\{S_n\}$ satisfies a LDP with *good rate function* $\Lambda^*(\cdot)$. The term "good" refers to the fact that the level sets $\{a \mid \Lambda^*(a) \leq k\}$ are compact for all $k < \infty$, which is a consequence of Assumption A (see [DZ93b] for a proof).

It is important to note that $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$ are convex duals (Legendre transforms of each other). Namely, along with (1.7), it also holds

$$\Lambda(\theta) = \sup_{a}(\theta a - \Lambda^*(a)). \tag{1.8}$$

The Gärtner-Ellis Theorem intuitively asserts that for large enough n and for small $\epsilon > 0$,

$$\mathbf{P}[S_n \in (na - n\epsilon, na + n\epsilon)] \sim e^{-n\Lambda^*(a)}.$$

In this thesis, and in particular in Chapter 2, we are mostly estimating tail probabilities of the form $\mathbf{P}[S_n \leq na]$ or $\mathbf{P}[S_n \geq na]$. We therefore define large deviations rate functions associated with such tail probabilities.

Consider the case where $S_n = \sum_{i=1}^{n} X_i$, the random variables X_i, $i \geq 1$, being identically distributed, and let $m = \mathbf{E}[X_1]$. It is easily shown (see [DZ93b]) that $\Lambda^*(m) = 0$. Let us now define

$$\Lambda^{*+}(a) \triangleq \begin{cases} \Lambda^*(a) & \text{if } a > m \\ 0 & \text{if } a \leq m \end{cases} \tag{1.9}$$

and

$$\Lambda^{*-}(a) \triangleq \begin{cases} \Lambda^*(a) & \text{if } a < m \\ 0 & \text{if } a \geq m. \end{cases} \tag{1.10}$$

Notice that $\Lambda^{*+}(a)$ is non-decreasing and $\Lambda^{*-}(a)$ non-increasing functions of a, respectively. The convex duals of these functions are

$$\Lambda^{+}(\theta) \triangleq \begin{cases} \Lambda(\theta) & \text{if } \theta \geq 0 \\ +\infty & \text{if } \theta < 0 \end{cases} \tag{1.11}$$

and

$$\Lambda^{-}(\theta) \triangleq \begin{cases} \Lambda(\theta) & \text{if } \theta \leq 0 \\ +\infty & \text{if } \theta > 0 \end{cases} \tag{1.12}$$

respectively.

Using the Gärtner-Ellis Theorem we can now state

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_n \leq na] = -\sup_{\theta}(\theta a - \Lambda^{-}(\theta)) = -\Lambda^{*-}(a) \tag{1.13}$$

and

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_n \geq na] = -\sup_{\theta}(\theta a - \Lambda^{+}(\theta)) = -\Lambda^{*+}(a). \tag{1.14}$$

1.3.1 Sample Path Large Deviations

A stronger concept than the LDP for the partial sum random *variable* $S_n \in \mathbb{R}$ that we introduced in the previous section, is the LDP for the partial sum *process*

$$S_n(t) = \frac{1}{n} \sum_{i=1}^{\lfloor nt \rfloor} X_i, \qquad t \in [0,1].$$

Note that the random variable $S_n = \sum_{i=1}^{n} X_i$ corresponds to the terminal value (at $t = 1$) of the process $S_n(t)$, $t \in [0,1]$.

In a key paper [DZ93a], the authors establish an LDP for the process $S_n(\cdot)$ in $D[0,1]$ (the space of right continuous functions with left limits). In particular, under certain mild mixing conditions on the stationary sequence $\{X_i;\ i \geq 1\}$ they first establish that the following two assumptions are satisfied.

Assumption B

Fix $m \in \mathbb{N}$ and $0 = t_0 < t_1 < \ldots < t_m \leq 1$, setting $Z_n = (S_n(t_1), S_n(t_2) - S_n(t_1), \ldots, S_n(t_m) - S_n(t_{m-1}))$. Then, $\{Z_n\}$ satisfies an LDP in $\mathbb{R}^m$ with the good rate function

$$\Lambda_m^*(z) = \sum_{i=1}^{m} (t_i - t_{i-1}) \Lambda^* \left(\frac{z_i}{t_i - t_{i-1}} \right),$$

where $z = (z_1, \ldots, z_m)$, and $\Lambda^*(\cdot)$ is the convex good rate function associated with the LDP of $S_n(1)$.

Assumption C

For all non-negative $\gamma, r < \infty$

$$g_r(\gamma) = \sup_{k,m \in \mathbb{N}, k \in [0, rm]} \frac{1}{m} \log \mathbf{E}[e^{\gamma |\sum_{i=k+1}^{k+m} X_i|}] < \infty,$$

and $A = \sup_\gamma \limsup_{r \to \infty} \frac{1}{r} g_r(\gamma) < \infty$.

Under these assumptions they prove the following theorem.

Theorem 1.3.2 ***(Sample path LDP, Dembo and Zajic [DZ93a])*** *Assuming*

that Assumptions B and C hold, the sequence of partial sums $\{S_n(\cdot)\}$ *satisfies the LDP in* $D[0,1]$ *with the convex good rate function*

$$I_\infty(x(\cdot)) = \begin{cases} \int_0^1 \Lambda^*(\dot{x})\ dt & \text{if } x \in \mathcal{AC}^0 \\ \infty & \text{otherwise.} \end{cases}$$

We use $\mathcal{AC}^0$ to denote the set of maps $x : [0,1] \to \mathbb{R}$ that are absolutely continuous and satisfy $x(0) = 0$. Dotted variables denote derivatives. The space $D[0,1]$ is equipped with the metric topology induced by

$$d_\infty(y(\cdot), z(\cdot)) = \sup_{t\in[0,1]} |y(t) - z(t)|.$$

The integral in the above theorem can viewed as the cost that the process $S_n(\cdot)$ incurs to follow the path $x(\cdot)$. In the simpler case when dependencies are not present (i.e., $S_i = \sum_{j=1}^i X_j$, where X_i's are iid) the above theorem was first proved by Mogulskii (see [DZ93b]). In [DZ93a] and [Cha94a] it is proved that Assumptions B and C (and therefore Theorem 1.3.2) are satisfied by processes that are commonly used in modeling the input traffic to communication networks, that is, renewal processes, Markov modulated processes with some uniformity assumptions on the stationary distribution (see [DZ93a, Section 4]) and correlated stationary processes with mild mixing conditions.

In this thesis we will focus to such dependent processes and base our analysis on the sample path LDP. We will therefore assume that arrival and service processes satisfy a statement similar to Thm. 1.3.2, the exact assumption to be specified on a case by case basis. The above sample path LDP in discrete time takes the form of the following assumption.

Assumption D

For all $m \in \mathbb{N}$, *for every* $\epsilon_1, \epsilon_2 > 0$ *and for every scalars* $a_0, \ldots, a_{m-1}$, *there exists*

$M > 0$ such that for all $n \geq M$ and all $k_0, \ldots, k_m$ with $1 = k_0 \leq k_1 \leq \cdots \leq k_m = n$,

$$e^{-\left(n\epsilon_2 + \sum_{i=0}^{m-1}(k_{i+1}-k_i)\Lambda^*(a_i)\right)} \leq \mathbf{P}[|S_{k_{i+1}} - S_{k_i} - (k_{i+1} - k_i)a_i| \leq \epsilon_1 n,\ i = 0, \ldots, m-1] \leq e^{\left(n\epsilon_2 - \sum_{i=0}^{m-1}(k_{i+1}-k_i)\Lambda^*(a_i)\right)}. \tag{1.15}$$

Intuitively, Assumption D asserts that for the partial sum process $\{S_i,\ i = 1, \ldots, n\}$ to reach the improbable level $S_n \sim \sum_{i=0}^{m-1}(k_{i+1} - k_i)a_i$ it is constrained in an ϵ_1-tube around the "polygonal" path constructed with linear segments of slopes $a_0, \ldots, a_{m-1}$.

In [Cha94a] a uniform bounding condition is given under which Thm. 1.3.2 is true. In other words, this condition implies Assumptions B and C. It is verified that the condition is satisfied by renewal, Markov-modulated and stationary processes with mild mixing conditions. In particular, the following result is proved.

Theorem 1.3.3 ***([Cha94a])*** *Suppose that $\{X_i;\ i \geq 1\} \in \mathbb{R}^d$ is adapted to filtration $\mathcal{F}_i$. If for all $\gamma \in \mathbb{R}$,*

$$\sup_{k,m} \frac{1}{m} \log \mathbf{E}[e^{\gamma \sum_{i=k+1}^{k+m} |X_i|}] < \infty,$$

and for all $\theta \in \mathbb{R}^d$, $k, m \geq 0$, there is a differentiable function $\Lambda(\theta) < \infty$ and a function $0 \leq \Gamma(\theta) < \infty$ independent of k, m, such that

$$m\Lambda(\theta) - \Gamma(\theta) \leq \log \mathbf{E}[e^{\theta \cdot \sum_{i=k+1}^{k+m} X_i} \mid \mathcal{F}_k] \leq \Lambda(\theta)m + \Gamma(\theta), \qquad a.s.,$$

then the process $\{S_i;\ i \geq 1\}$ satisfies a sample path LDP (Thm. 1.3.2).

Using this uniform bounding condition it is not hard to verify (see [Cha94a] for a proof) that the following assumption is satisfied. This assumption can be viewed as the "convex dual analog" of Assumption D.

Assumption E

For all $m \in \mathbb{N}$ there exists $M > 0$ and a function $0 \leq \Gamma(y) < \infty$, for all $y > 0$, such

that for all $n \geq M$ *and all* $k_0, \ldots, k_m$ *with* $1 = k_0 \leq k_1 \leq \cdots \leq k_m = n$,

$$\mathbf{E}[e^{\theta \cdot Z}] \leq \exp\{\sum_{j=1}^{m}[(k_j - k_{j-1})\Lambda(\theta_j) + \Gamma(\theta_j)]\}, \tag{1.16}$$

where $\theta = (\theta_1, \ldots, \theta_m)$ *and* $Z = (S_{k_0}, S_{k_2} - S_{k_1}, \ldots, S_{k_m} - S_{k_{m-1}})$.

Finally, on a notational remark, in the rest of the thesis we will be denoting by $S_{i,j}^X \triangleq \sum_{k=i}^{j} X_k$, $i \leq j$, the partial sums of the random sequence $\{X_i;\ i \in \mathbb{Z}\}$. We will be also denoting by $\Lambda_X(\cdot)$ and $\Lambda_X^*(\cdot)$ the limiting log-moment generating function and the large deviations rate function (see eqs. (1.4) and (1.7) for definitions), respectively, of the process X.

Chapter 2

Acyclic Single Class Networks

Consider a single class, acyclic network of G/G/1 queues. Customers arrive to the network in a number of independent streams and are treated uniformly by the network. Different streams may share a queue and the first-come first-serve (FCFS) policy is implemented. A constant fraction p_{ij} of customers departing a queue i, is routed to queue j and a fraction p_{i0} leaves the network. In this chapter we derive large deviations results for the waiting time and the queue length observed by an arbitrary customer at different queues of the network.

As we outlined in the introduction, we use a decomposition approach that essentially treats the problem as a series of single queue problems. The main task is to prove an LDP for the aggregate arrival process in each queue of the network. We observe that for the particular network model that we are considering the external traffic is transformed by three "filtering" operations as it slips through the network. The first operation is to pass through a queue (passing-through-a-queue), the second to split at a router (routing) and the third to get superimposed (superposition) as it reaches a node. We impose an LDP on the external arrival processes, along with some mild technical assumptions. We analyze all three "filtering" operations in isolation and determine an LDP for their output. We prove that all these operations preserve the assumptions imposed on the external arrival processes. Thus, by using induction rigorously we determine an LDP for the aggregate arrival process in each queue of

the network. The last step is to invoke single queue results for the waiting time and the queue length and determine the tail of the waiting time and the queue length in all the queues of the network.

Regarding the structure of this chapter, we start in Section 2.1 by presenting the network model that we are considering and establish our notation. In Section 2.2 we treat the single queue case. This section is comprised by two subsections. In Subsection 2.2.1 we review the existing result for the large deviations behaviour of the waiting time and we completely characterize the most likely path along which the waiting time takes large values. In Subsection 2.2.2, using an idea from distributional laws we obtain the tail probability of the queue length. In Section 2.3 we derive the large deviations behaviour of the departure process (passing-through-queue operation) from a G/G/1 queue. Particular attention is given to the way that such a deviation occurs. In Subsection 2.3.1, some special cases are studied. Namely, we apply the result for the departure process of a G/G/1 queue to a G/D/1 queue and an M/M/1 queue. For the latter case, Burke's Theorem is verified in the large deviations regime. In Sections 2.4 and 2.5 we study the large deviations behaviour of the processes resulting from the following operations: superposition of independent processes, and deterministic splitting of a process to a number of processes, respectively. In Subsection 2.4.1 we prove a result that connects the Palm and the stationary distribution of a point process in the large deviations regime. This result is used in the rest of Section 2.4 to derive the large deviations behaviour of the superposition process. In Section 2.6, we treat, as an example, a network consisting of two queues in tandem. We characterize the way that the waiting time in the second queue reaches large values and we include some numerical results.

2.1 The Network Model

In this section, we formally define the network model of which we will derive the large deviations behaviour. Moreover, we establish the notation that we will be using and state a set of assumptions on the arrival and service processes.

Consider a *directed acyclic graph (dag)* with J nodes. For reasons that will become soon apparent, we assume that any two directed paths do not meet in more than one node. Each node of the graph is equipped with an infinite buffer and a single server. Customers enter the network in a number of independent streams $A^1, A^2, \ldots, A^J$. In particular, A^i is the stream of customers that enter the network at node i. Customers are treated uniformly by the network, i.e., the network is assumed to be *single class*. Let $\mathbb{Z}$ denote the set of integers. By $A_i^j,\ i \in \mathbb{Z}$, we denote the interarrival time of the ith customer in the jth stream (the interval between the arrival epochs of the $(i-1)$st and the ith customer). By $B_i^j,\ i \in \mathbb{Z}$, we denote the service time of the ith customer in the jth node. We assume that for each arriving stream j the process $\{A_i^j,\ i \in \mathbb{Z}\}$, is stationary, and $A_i^j,\ i \in \mathbb{Z}$, are possibly dependent random variables. Moreover, for each node j, the service times $B_i^j,\ i \in \mathbb{Z}$, are iid random variables. We also assume that interarrival and service times at a specific node are mutually independent and that service times at different nodes are independent.

Independent streams may share a queue and the FCFS policy is implemented. A fraction $p_{ij_1}, p_{ij_2}, \ldots$ of customers departing node i, which is connected to nodes $j_1, j_2, \ldots$, are routed to these nodes, respectively, and a fraction p_{i0} leaves the network. The exact way that the routing is performed is not of importance in the large deviations regime. Roughly, out of every $1/p_{ij}$ customers leaving node i, the routing mechanism sends one to node j. Figure 2-1 depicts an example of the class of networks considered. Such a network is intended to model packet-switched communication networks.

We denote by $W^1, W^2, \ldots, W^J$ and $L^1, L^2, \ldots, L^J$ the steady-state waiting times and queue lengths, incurred by a typical customer at nodes $1, 2, \ldots, J$ of the network, respectively. For each node j, W_n^j (resp. L_n^j) denotes the waiting time incurred (resp. queue length observed) by the nth customer. We assume that the process $\{(W_n^j, L_n^j);\ n \in \mathbb{Z}, j = 1, \ldots, J\}$ is stationary.

In this chapter, we derive large deviations results for the steady-state waiting times $W^1, W^2, \ldots, W^J$, and the corresponding queue lengths $L^1, L^2, \ldots, L^J$, incurred at nodes $1, 2, \ldots, J$ of the network, respectively (as these random variables are seen

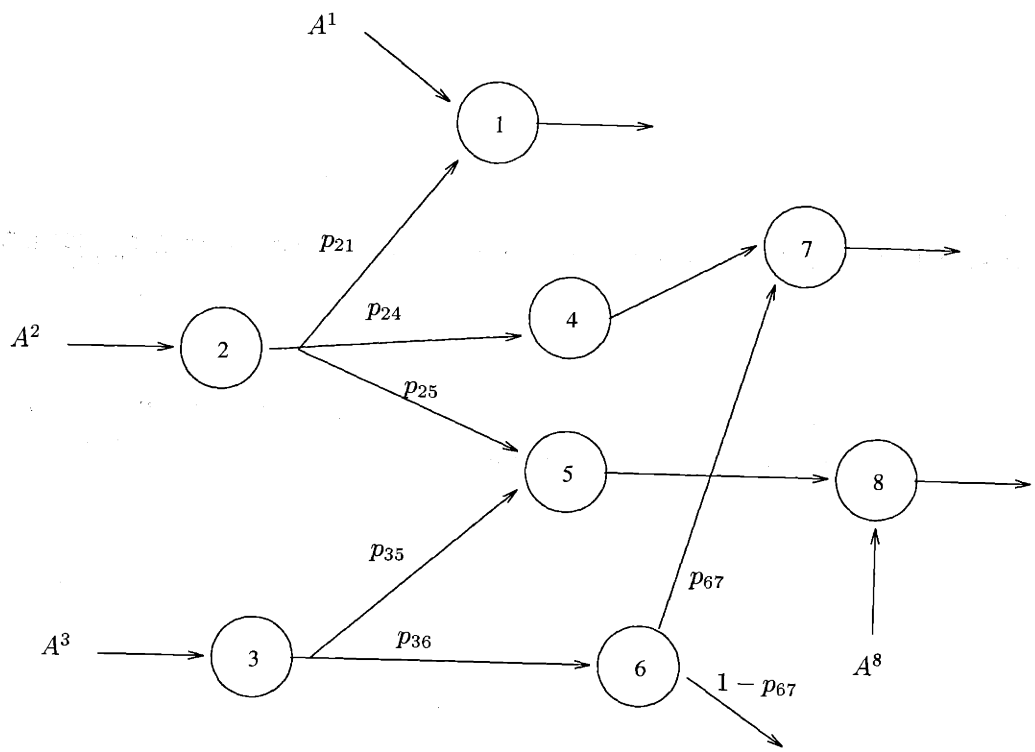

Figure 2-1: A network example.

by a typical customer). Our strategy is first to obtain large deviations results for the steady-state waiting time and the corresponding queue length in a single G/G/1 queue. Then it suffices to derive a LDP for the partial sum of the aggregate arrival process in each queue of the network and apply the result for the single queue case. It is important to note that by the definition of the network all the streams sharing the same queue are independent. Therefore, from the model description, it is apparent that it suffices to obtain LDP's for the processes resulting from the following operations

1. Passing-through-a-queue (the process resulting from this operation being the departure process).

2. Superposition of independent streams.

3. Deterministic splitting of a stream to a number of streams.

Let $\{A_i,\ i \in \mathbb{Z}\}$ be an arbitrary external arrival process and $\{B_i,\ i \in \mathbb{Z}\}$ be an arbitrary service process. Hereafter, we will be using the notation $S^X_{i,j} \triangleq \sum_{k=i}^{j} X_k;\ i \leq j$ for the partial sums of the random sequence $\{X_i;\ i \in \mathbb{Z}\}$ along with the convention $S^X_{i,j} \triangleq 0;\ i > j$.

Assumption F

1. *The sequence of partial sums* $\{S^A_{1,n};\ n \geq 1\}$ *satisfies*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S^A_{1,n} \leq na] = -\Lambda^{*-}_A(a), \tag{2.1}$$

where

$$\Lambda^-_A(\theta) \triangleq \begin{cases} \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{\theta S^A_{1,n}}] & \text{if } \theta \leq 0 \\ +\infty & \text{if } \theta > 0 \end{cases} \tag{2.2}$$

and

$$\Lambda^{*-}_A(a) \triangleq \sup_\theta (\theta a - \Lambda^-_A(\theta)). \tag{2.3}$$

We will say that $\{S^A_{1,n};\ n \geq 1\}$ *satisfies an* one sided LDP.

2. *The sequence of partial sums* $\{S^B_{1,n};\ n \geq 1\}$ *satisfies the requirements of the Gärtner-Ellis theorem (i.e., Assumption A) with limiting log-moment generating function*

$$\Lambda_B(\theta) \triangleq \lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{\theta S^B_{1,n}}], \tag{2.4}$$

which is finite for all $\theta > 0$, *and large deviations rate function*

$$\Lambda^*_B(a) \triangleq \sup_\theta (\theta a - \Lambda_B(\theta)). \tag{2.5}$$

Assumption G

1. *For every* $\epsilon_1, \epsilon_2, a > 0$, *there exists* M_A *such that for all* $n \geq M_A$

$$e^{-n(\Lambda_A^{*-}(a)+\epsilon_2)} \leq \mathbf{P}[S_{1,i}^A - ia \leq \epsilon_1 n,\ i = 1, \ldots, n]. \tag{2.6}$$

2. *For every* $\epsilon_1, \epsilon_2, a > 0$, *there exists* M_B *such that for all* $n \geq M_B$

$$e^{-n(\Lambda_B^{*-}(a)+\epsilon_2)} \leq \mathbf{P}[S_{i,j}^B - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n], \tag{2.7}$$

and

$$e^{-n(\Lambda_B^{*+}(a)+\epsilon_2)} \leq \mathbf{P}[S_{i,j}^B - (j-i+1)a \geq -\epsilon_1 n,\ 1 \leq i \leq j \leq n]. \tag{2.8}$$

We consider external arrival and service processes that satisfy Assumptions F and G. We will show that these assumptions are satisfied by the processes resulting from the three operations mentioned above. In this way, our approach provides a *calculus of acyclic networks* since we will be able to determine the large deviations behaviour of each individual queue inductively.

Assumption F provides a LDP for the arrival and service processes. Based on these LDP's we will derive LDP's for all the processes of interest in the network. Note that only the tail probability of the external arrival processes corresponding to "many arrivals" is characterized by Assumption F. We will prove that in order to estimate probabilities of large waiting times and long delays, as we do in this chapter, only such a tail probability of the aggregate arrival process in each queue of the network is needed.

Assumption G is needed in order to derive a LDP for the departure process of a G/G/1 queue. It intuitively asserts that besides the LDP for the partial sum random *variable* $S_{1,n}$, we also have a LDP for the partial sum *process* $\{S_{1,i},\ i = 1, \ldots, n\}$ for the arrivals and $\{S_{i,j},\ 1 \leq i \leq j \leq n\}$ for the service times. In other words, (2.6) and (2.7) guarantee that the partial sum process follows a path that never overshoots the straight line of slope a, in order to reach an improbable level $S_{1,n} \leq na$. A

similar interpretation can be given to (2.8). Mild mixing conditions on the arrival and service processes suffice to guarantee Assumption G. A thorough treatment is given in [DZ93a]. In the Appendix A we provide some conditions under which Assumption G is satisfied based on the results of [DZ93a].

Assumptions F and G are satisfied by processes that are used to model external arrival and services in communications networks, such as renewal processes, stationary processes with mild mixing conditions, as well as Markov-modulated processes with some uniformity assumptions on the stationary distribution (see [DZ93a, Section 4]).

2.2 Large Deviations of a G/G/1 Queue

In this section, we establish a LDP for the Palm distributions of the steady-state waiting time and queue length (i.e., as these random variables are seen by a typical customer), in a G/G/1 queue with stationary arrivals and service times.

The setting is the same as in Section 2.1. We denote by $\{A_i,\ i \in \mathbb{Z}\}$ the stationary aggregate arrival process to the queue and we assume that it satisfies Assumption F.1. We also denote by $\{B_i,\ i \in \mathbb{Z}\}$ the stationary service process and we assume that it satisfies Assumption F.2. For this section, the independence assumption for the service times can be relaxed. For stability purposes, we further assume $\mathbf{E}[A] > \mathbf{E}[B]$, where A (resp. B) denotes a typical interarrival (resp. service) time.

2.2.1 Large Deviations of the Waiting Time

Let us first characterize the steady-state waiting time, W, incurred by a typical customer. By W_n we denote the waiting time of the nth customer. The condition $\mathbf{E}[A] > \mathbf{E}[B]$ is necessary [1] for the existence and the uniqueness of a stationary process (see [Wal88]). From the Lindley equation, the waiting time of the 0th customer, at

[1] for sufficiency ergodicity is also needed.

steady-state, is given by

$$W_0 = [W_{-1} + B_{-1} - A_0]^+ \triangleq \max[W_{-1} + B_{-1} - A_0, 0] = \max_{i \geq 0}[S^B_{-i-1,-1} - S^A_{-i,0}, 0]. \tag{2.9}$$

The intuitive meaning of this relation is the following: For a particular sample path, if i^* is the optimum i, then the customer with label $-i^* - 1$ is the one who initializes the busy period in which the 0th customer is served.

The next theorem establishes a LDP for W_0. This result is not new. The proof is almost identical with the proof in [dVW93, Thm 3.1], where a discrete time model is used, and is therefore omitted. An upper bound on the tail probability, of the steady-state waiting time, was first obtained by Kingman [Kin70].

Theorem 2.2.1 *The tail of the Palm distribution of the steady-state waiting time, W, in a FCFS G/G/1 queue with arrivals and service times satisfying Assumption F is characterized by*

$$\lim_{U \to \infty} \frac{1}{U} \log \mathbf{P}[W \geq U] = \theta^*, \tag{2.10}$$

where $\theta^* < 0$ is the smallest root of the equation

$$\Lambda_A(\theta) + \Lambda_B(-\theta) = 0. \tag{2.11}$$

Remarks : Intuitively, Theorem 2.2.1 asserts that for large enough U, we can state

$$\mathbf{P}[W \geq U] \sim e^{\theta^* U}, \qquad \text{where } \theta^* < 0 \text{ is such that} \qquad \Lambda_A(\theta^*) + \Lambda_B(-\theta^*) = 0. \tag{2.12}$$

Note that θ^* exists as an extended real number since $\mathbf{E}[A] > \mathbf{E}[B]$ and the functions $\Lambda_A(\cdot), \Lambda_B(\cdot)$ are convex [2]. Figure 2-2 depicts the function $\Lambda_A(\theta) + \Lambda_B(-\theta)$ and the root θ^*. If $\Lambda_A(\theta) + \Lambda_B(-\theta) < 0$ for all $\theta < 0$ we use the convention $\theta^* = \infty$.

[2] This is proven under the conditions of Assumption F in [DZ93b, Lemma 2.3.9]

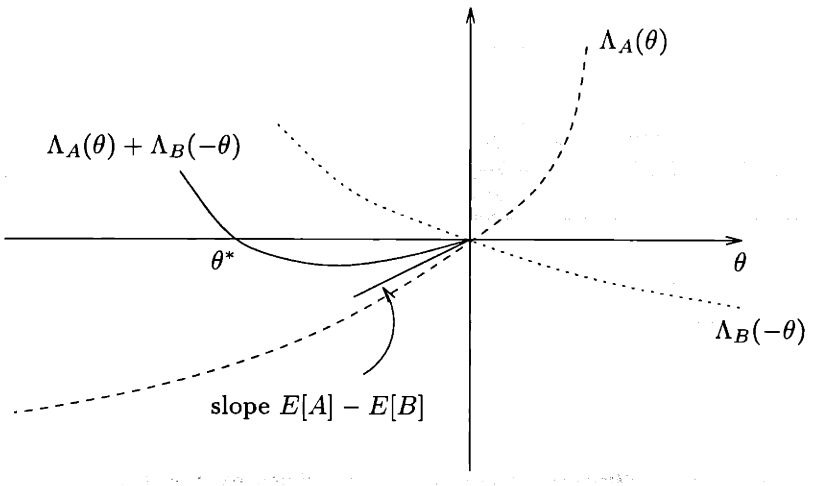

Figure 2-2: The root of $\Lambda_A(\theta) + \Lambda_B(-\theta) = 0$.

It is instructive to characterize the most likely "path" along which the large deviation of the waiting time occurs. Such a characterization can also provide an alternative proof of Thm. 2.2.1. Let $a > 0$ and $x_1, x_2 \in \mathbb{R}^+$, such that $x_2 - x_1 = a$. Using Eq. (2.9), we have

$$\begin{aligned} \mathbf{P}[W_0 \geq (i+1)a] &\geq \mathbf{P}[S^B_{-i-1,-1} - S^A_{-i,0} \geq (i+1)a] \\ &\geq \mathbf{P}[S^A_{-i,0} \leq (i+1)x_1]\mathbf{P}[S^B_{-i-1,-1} \geq (i+1)x_2] \\ &\geq e^{-(i+1)[\Lambda^{*-}_A(x_1) + \Lambda^{*+}_B(x_2) + \epsilon]}, \end{aligned} \tag{2.13}$$

where the last inequality makes use of Assumption F and holds for any $\epsilon > 0$ and for large i.

Setting $U = (i+1)a$, we obtain

$$\mathbf{P}[W_0 \geq U] \geq \exp\left\{-U \inf_{a>0} \frac{1}{a} \inf_{x_2 - x_1 = a} [\Lambda^{*-}_A(x_1) + \Lambda^{*+}_B(x_2)] - U\epsilon\right\}. \tag{2.14}$$

Let $a^* > 0$ be a solution to the above optimization problem. Thus, for large U,

and by taking $\epsilon \to 0$ in (2.14), we obtain

$$\mathbf{P}[W_0 \geq U] \geq \exp\left\{-U\frac{\inf_{x_2-x_1=a^*}[\Lambda_A^{*-}(x_1)+\Lambda_B^{*+}(x_2)]}{a^*}\right\}. \tag{2.15}$$

The tightness of this bound can be proven by obtaining a matching (i.e., with the same exponent) upper bound; the proof is omitted. Let i^* be defined by the equation $i^*+1 = U/a^*$. Let also x_1^* and x_2^* solve the optimization problem in (2.15). Consider a scenario where customers $(-i^*-1), \ldots, -1, 0$ arrive at an empirical arrival rate of $\frac{1}{x_1^*}$ and customers $(-i^*-1), \ldots, -1$ are served with an empirical service rate of $\frac{1}{x_2^*}$. Such a scenario, which is depicted in Figure 2-3, has probability comparable to the right hand side of (2.15) and is therefore a most likely way for the large deviation of the waiting time to occur.

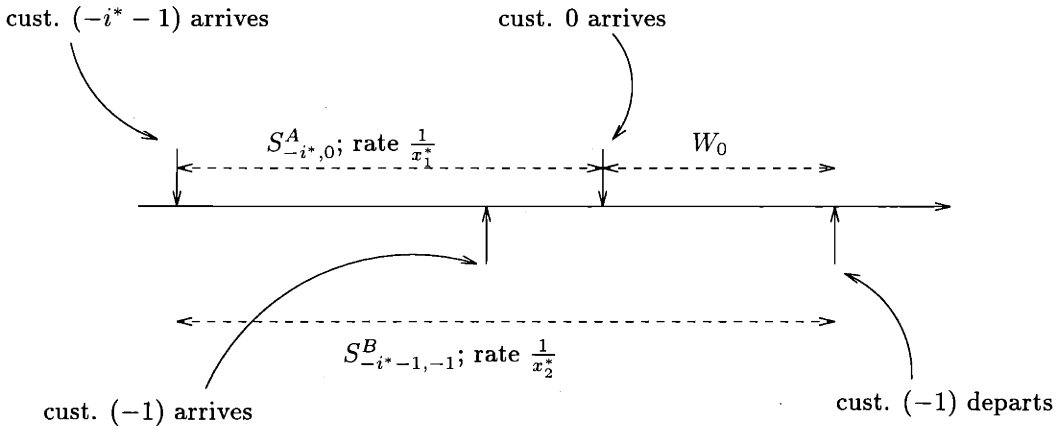

Figure 2-3: The optimal path for large deviations in the waiting time.

2.2.2 Large Deviations of the Queue Length

In this subsection, we present a LDP for the steady-state queue length in a G/G/1 queue, as seen by a typical customer (Palm distribution). To accomplish that we

use the main argument used in deriving distributional laws; that is, a probabilistic relation between the waiting time and the queue length. A detailed discussion of distributional laws and their applications can be found in [BN95, BM92]. It is important to note that distributional laws have been proven there only for renewal arrival and service processes. However in the large deviations setting, we are able to relax the renewal assumption and state a result that holds even for correlated arrival and service processes.

Let us now characterize the steady-state queue length L seen by a typical customer (not including herself) upon arrival (this is sometimes denoted by L^- in the literature). The goal is to estimate $\mathbf{P}[L \geq n]$. Let us denote by L_n the queue length observed by the nth customer. As in Section 2.1, we assume that the process $\{(L_n, W_n);\ n \in \mathbb{Z}\}$ is stationary. The main idea, in order to establish a relation between the waiting time and the queue length, is to look backwards in time from the arrival epoch of the nth customer. Figure 2-4 depicts the situation. We denote

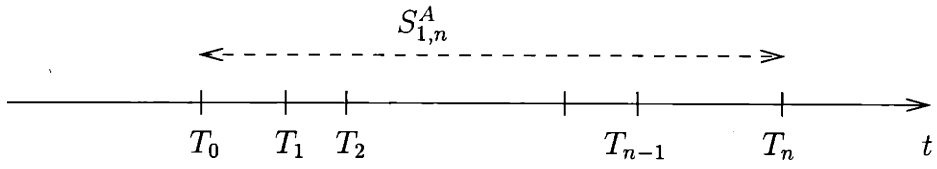

Figure 2-4: The system at time T_n.

with $T_0, T_1, \ldots$ the arrival epochs of customers $0, 1, \ldots$, respectively. Recall that W_n and B_n denote the waiting and the service time of the nth customer, respectively.

The main observation is the following: In order for the queue length right before T_n to be at least n, the 0th customer should be in the system at that time. Namely,

$$\mathbf{P}[L_n \geq n] = \mathbf{P}[W_0 + B_0 \geq S^A_{1,n}] \tag{2.16}$$

and by using (2.9) we obtain

$$\begin{aligned}\mathbf{P}[L_n \geq n] &= \mathbf{P}[\max_{i\geq 0}[S^B_{-i-1,0} - S^A_{-i,n}, -S^A_{1,n}] \geq 0] = \\ &= \mathbf{P}[\max_{i\geq -1}[S^B_{-i-1,0} - S^A_{-i,n}] \geq 0]. \end{aligned} \tag{2.17}$$

The next theorem establishes a LDP for L_n. We will need a technical lemma which we prove next.

Lemma 2.2.2 *Under Assumption F, and for $\theta < 0$, satisfying $\Lambda_A(\theta) + \Lambda_B(-\theta) < 0$, it holds*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{-\theta \max_{i\geq -1}[S^B_{-i-1,0} - S^A_{-i,n}]}] \leq \Lambda_A(\theta). \tag{2.18}$$

Proof : We have

$$\mathbf{E}[e^{-\theta \max_{i\geq -1}[S^B_{-i-1,0} - S^A_{-i,n}]}] \leq \sum_{i\geq -1} \mathbf{E}[e^{-\theta S^B_{-i-1,0}}]\mathbf{E}[e^{\theta S^A_{-i,n}}].$$

From (2.4) it can be seen that there exists $j > 0$ such that for all $i > j$ and all $\epsilon > 0$ it holds

$$\mathbf{E}[e^{-\theta S^B_{-i-1,0}}] \leq e^{(i+2)(\Lambda_B(-\theta)+\epsilon)}. \tag{2.19}$$

Also from (2.2), we have that for $\theta < 0$ there exists N such that for all $n > N$, all $i \geq -1$ and all $\epsilon > 0$

$$\mathbf{E}[e^{\theta S^A_{-i,n}}] \leq e^{(n+i+1)(\Lambda_A(\theta)+\epsilon)}. \tag{2.20}$$

Fix now some $\theta < 0$ satisfying $\Lambda_A(\theta) + \Lambda_B(-\theta) < 0$ and some $\epsilon > 0$ such that $\Lambda_A(\theta) + \Lambda_B(-\theta) + 2\epsilon < 0$. Note that the existence of such a θ is guaranteed by the

condition $\mathbf{E}[A] > \mathbf{E}[B]$ (see Figure 2-2). We then have that for all $n > N$

$$\begin{aligned}\mathbf{E}[e^{-\theta \max_{i\geq -1}[S^B_{-i-1,0}-S^A_{-i,n}]}] \leq & \sum_{i=-1}^{j} \mathbf{E}[e^{-\theta S^B_{-i-1,0}}]\mathbf{E}[e^{\theta S^A_{-i,n}}] + \sum_{i>j} \mathbf{E}[e^{-\theta S^B_{-i-1,0}}]\mathbf{E}[e^{\theta S^A_{-i,n}}] \\ \leq & e^{n(\Lambda_A(\theta)+\epsilon)} \sum_{i=-1}^{j} \mathbf{E}[e^{-\theta S^B_{-i-1,0}}]e^{(i+2)(\Lambda_A(\theta)+\epsilon)} \\ & + e^{n(\Lambda_A(\theta)+\epsilon)} \sum_{i>j} e^{2\Lambda_B(-\theta)+\Lambda_A(\theta)+3\epsilon} e^{i(\Lambda_B(-\theta)+\Lambda_A(\theta)+2\epsilon)} \\ \leq & K(\theta, j, \epsilon) e^{n(\Lambda_A(\theta)+\epsilon)}, \end{aligned} \tag{2.21}$$

where $K(\theta, j, \epsilon)$ is some constant depending on θ, j and ϵ but not on n. To see that, notice that in the last inequality above we use the fact that the first sum is finite, and the infinite geometric series in the second sum converges to a constant independent of n. From Eq. (2.21) we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{-\theta \max_{i\geq -1}[S^B_{-i-1,0}-S^A_{-i,n}]}] \leq \Lambda_A(\theta) + \epsilon. \tag{2.22}$$

Since this is true for all small enough $\epsilon > 0$, the result follows.

■

Theorem 2.2.3 *The tail of the Palm distribution of the steady-state queue length, L, in a FCFS G/G/1 queue with arrivals and service times satisfying Assumption F is characterized by*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[L \geq n] = \Lambda_A(\theta^*), \tag{2.23}$$

where $\theta^* < 0$ is the smallest root of the equation

$$\Lambda_A(\theta) + \Lambda_B(-\theta) = 0. \tag{2.24}$$

Proof : Due to stationarity, it suffices to characterize the tail distribution of L_n. For

an upper bound define

$$G_n \stackrel{\triangle}{=} \max_{i \geq -1} [S^B_{-i-1,0} - S^A_{-i,n}]. \tag{2.25}$$

Using the Markov inequality, we obtain

$$\mathbf{P}[L_n \geq n] = \mathbf{P}[G_n \geq 0] \leq \mathbf{E}[e^{-\theta G_n}],$$

for $\theta < 0$. Taking the limit as $n \to \infty$, using Lemma 2.2.2, and optimizing over θ to get the best bound we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[L_n \geq n] \leq \inf_{\{\theta |\ \Lambda_A(\theta)+\Lambda_B(-\theta)<0\}} [\Lambda_A(\theta)] = \Lambda_A(\theta^*), \tag{2.26}$$

where the last equality is justified by Figure 2-2.

For a lower bound, set $i = \delta n$ for $\delta \geq 0$ (δn is assumed integer), and notice that

$$\begin{aligned} \mathbf{P}[L_{n-1} \geq n] &= \mathbf{P}[G_n \geq 0] \\ &\geq \sup_{\delta \geq 0} \mathbf{P}[S^B_{-\delta n-1,0} - S^A_{-\delta n,n} \geq 0]. \end{aligned}$$

The limiting log-moment generating function of $S^B_{-\delta n-1,0} - S^A_{-\delta n,n}$ is

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{-\theta(S^B_{-\delta n-1,0} - S^A_{-\delta n,n})}] = \delta \Lambda_B(-\theta) + (1+\delta)\Lambda_A(\theta)$$

and by using Assumption F we obtain

$$\begin{aligned} \liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}[L_n \geq n] &\geq \sup_{\delta \geq 0} (-\sup_{\theta} [-\delta(\Lambda^-_A(\theta) + \Lambda^+_B(-\theta)) - \Lambda^-_A(\theta)]) \\ &= \sup_{\delta \geq 0} \inf_{\theta} [\delta(\Lambda^-_A(\theta) + \Lambda^+_B(-\theta)) + \Lambda^-_A(\theta)] \\ &= \inf_{\{\theta |\ \Lambda^-_A(\theta)+\Lambda^+_B(-\theta)<0\}} [\Lambda^-_A(\theta)] \\ &= \Lambda^-_A(\theta^*) = \Lambda_A(\theta^*), \end{aligned} \tag{2.27}$$

where the second equality follows by dualizing the constraint $\Lambda_A^-(\theta) + \Lambda_B^+(-\theta) < 0$. The lower bound in (2.27) along with (2.26) proves (2.23).

■

Remark : Intuitively, Theorem 2.2.3 asserts that for large enough n, we can state

$$\mathbf{P}[L \geq n] \sim e^{n\Lambda_A(\theta^*)}, \qquad \text{where } \theta^* < 0 \text{ such that} \qquad \Lambda_A(\theta^*) + \Lambda_B(-\theta^*) = 0. \tag{2.28}$$

2.3 The Departure Process of a G/GI/1 queue

In this section we obtain a LDP for the process resulting from the passing-through-a-queue operation of our network model. That is, we establish a LDP for the steady-state departure process of a G/GI/1 queue, as seen by a typical departing customer. We denote by D_i, $i \in \mathbb{Z}$, the inter-departure time of the ith customer (the interval between the departure epochs of the $(i-1)$st and the ith customer). As in Section 2.1 we assume that the interarrival times process $\{A_i,\ i \in \mathbb{Z}\}$ is stationary, and A_i are possibly dependent random variables. The service times B_i are independent and identically distributed (iid) random variables. The arrival and service processes are also assumed to satisfy Assumptions F and G. As explained in Section 2.1, we will prove that the departure process satisfies Assumptions F and G when the arrival and service processes do.

We denote by $S_{1,n}^D \stackrel{\Delta}{=} \sum_{i=1}^n D_i$, the partial sum of the departure process. The objective of this section is to prove a LDP for $S_{1,n}^D$. The inter-departure times can be expressed as follows

$$D_i = B_i + I_i, \tag{2.29}$$

where B_i denotes the service time of the ith customer and I_i the idling period of the system that ended with the arrival of the ith customer ($I_i = 0$ if the ith customer finds the system busy upon arrival). By using the Lindley equation one can obtain

an expression for I_i and after some algebra derive an expression for $S^D_{1,n}$ in terms of the partial sums for the arrival and the service process. Using such an expression one can prove a LDP for $S^D_{1,n}$. Here we follow a more intuitive approach. We derive an upper bound and a matching lower bound on $\mathbf{P}[S^D_{1,n} \leq na]$ based on sample path arguments. To that effect, we explicitly characterize the most likely path leading to the large deviation of the departure process. The next proposition establishes an upper bound for the tail probability of $S^D_{1,n}$.

Proposition 2.3.1 (Upper Bound) *Under Assumption F, the partial sum $S^D_{1,n}$ of the departure process of a G/GI/1 queue under FCFS satisfies*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S^D_{1,n} \leq na] \leq -\Lambda^{*-}_D(a), \tag{2.30}$$

where

$$\Lambda^{*-}_D(a) \stackrel{\Delta}{=} \Lambda^{*-}_B(a) + \Lambda^{*-}_\Gamma(a) \tag{2.31}$$

and

$$\Lambda^{*-}_\Gamma(a) \stackrel{\Delta}{=} \sup_{\{\theta | \Lambda_A(\theta)+\Lambda_B(-\theta)<0\}} [\theta a - \Lambda^-_A(\theta)]. \tag{2.32}$$

Proof : Since $D_i \geq B_i$ for all i we obtain

$$S^D_{1,n} \geq S^B_{1,n}. \tag{2.33}$$

Consider some $j \leq 1$ and let $(j-1)$ be the customer who initializes the busy period in which the 0th customer is served. Let t be the time that the $(j-1)$st customer arrived, t' the time that the $(j-1)$st customer departed, and t'' the time that the nth customer departed. Figure 2-5 depicts the situation. Note that

$$B_{j-1} + S^D_{j,n} \geq S^A_{j,n}. \tag{2.34}$$

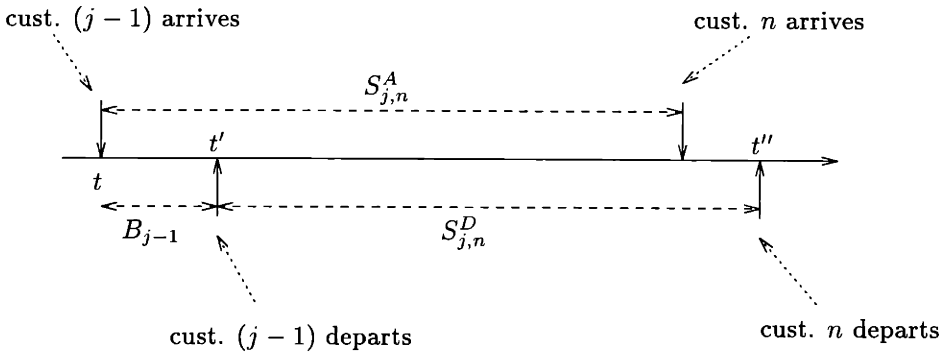

Figure 2-5: Deriving an upper bound on $\mathbf{P}[S^D_{1,n} \leq na]$.

Since the system is busy from the arrival of the $(j-1)$st customer until the departure of customer 0, we have

$$S^D_{j,0} = S^B_{j,0}. \tag{2.35}$$

Therefore, from (2.35) and (2.34) we have

$$S^D_{1,n} = S^D_{j,n} - S^D_{j,0} \geq S^A_{j,n} - B_{j-1} - S^B_{j,0} = S^A_{j,n} - S^B_{j-1,0}. \tag{2.36}$$

Now, from (2.33) and (2.36) we obtain

$$\begin{aligned} \mathbf{P}[S^D_{1,n} \leq na] &\leq \mathbf{P}[S^B_{1,n} \leq na,\ \exists\ j \leq 1 \text{ s.t. } S^A_{j,n} - S^B_{j-1,0} \leq na] \\ &= \mathbf{P}[S^B_{1,n} \leq na]\mathbf{P}[\min_{j\leq 1}[S^A_{j,n} - S^B_{j-1,0}] \leq na], \end{aligned} \tag{2.37}$$

since the service times B_i are assumed to be independent and independent of the arrival process. Since $\min_{j\leq 1}[S^A_{j,n} - S^B_{j-1,0}] = -\max_{j\leq 1}[S^B_{j-1,0} - S^A_{j,n}]$, we use Lemma 2.2.2 to obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}[e^{\theta \min_{j\leq 1}[S^A_{j,n} - S^B_{j-1,0}]}] \leq \Lambda_A(\theta), \tag{2.38}$$

for $\theta < 0$, satisfying $\Lambda_A(\theta) + \Lambda_B(-\theta) < 0$.

Using Markov's inequality we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[\min_{j\leq 1}[S^A_{j,n} - S^B_{j-1,0}] \leq na] \leq \Lambda_A(\theta) - \theta a.$$

Optimizing over θ to obtain the tightest bound we finally find (note that for $\theta < 0$ we have $\Lambda^-_A(\theta) = \Lambda_A(\theta)$)

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[\min_{j\leq 1}[S^A_{j,n} - S^B_{j-1,0}] \leq na] \leq - \sup_{\{\theta|\Lambda_A(\theta)+\Lambda_B(-\theta)<0\}} [\theta a - \Lambda^-_A(\theta)]. \tag{2.39}$$

Moreover from Assumption F we can assert that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S^B_{1,n} \leq na] \leq -\Lambda^{*-}_B(a). \tag{2.40}$$

Combining Eqs. (2.40) and (2.39) along with Eq. (2.37) we obtain (2.30).

■

Obtaining a lower bound on the tail probability of $S^D_{1,n}$ is much more involved. Assumption F which provides a LDP for the partial sums $S^A_{1,n}, S^B_{1,n}$ of the interarrival and service times is not sufficient. Assumption G which provides a LDP for the partial sum processes $\{S^A_{1,j},\ j = 1,\ldots,n\}$ and $\{S^B_{i,j},\ 1 \leq i \leq j \leq n\}$, is required. In the next proposition we derive a lower bound on the tail probability of $S^D_{1,n}$ and we prove that the departure process $\{S^D_{1,i},\ i = 1,\ldots,n\}$ satisfies Assumption G.

Proposition 2.3.2 (Lower Bound) *Under Assumptions F and G, the partial sum $S^D_{1,n}$ of the departure process of a G/GI/1 queue under FCFS satisfies*

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S^D_{1,n} \leq na] \geq -\Lambda^{*-}_D(a). \tag{2.41}$$

Moreover, the departure process $\{S^D_{1,i},\ i = 1,\ldots,n\}$ satisfies Assumption G.

Proof : Fix $\epsilon_1, \epsilon_2 > 0$, $\zeta \geq 0$ and $y_1, y_2 \geq 0$ such that $y_1 - y_2 = a$ and $\frac{y_1}{1+\zeta} \geq a$.

Consider the set of all sample paths that satisfy

$$S^B_{j,k} \leq (k+1-j)a + \epsilon_1 n, \qquad 1 \leq j \leq k \leq n, \tag{2.42}$$

$$S^A_{-\zeta n,k} \leq (\zeta n + k - 1)\frac{y_1}{1+\zeta} + \epsilon_1 n, \qquad k = 1, \ldots, n, \tag{2.43}$$

and

$$S^B_{-\zeta n-1,0} \geq n y_2 - \epsilon_1 n. \tag{2.44}$$

We state the following lemma the proof of which is deferred until the end of the current proof.

Lemma 2.3.3 *For any sample path that satisfies (2.42), (2.43) and (2.44) we have*

$$S^D_{1,k} \leq ka + 4\epsilon_1 n, \qquad k = 1, \ldots, n. \tag{2.45}$$

Therefore,

$$\begin{aligned}
&\mathbf{P}[S^D_{1,k} \leq ka + 4\epsilon_1 n,\ k = 1, \ldots, n] \geq \\
&\quad \geq \mathbf{P}[S^B_{j,k} \leq (k+1-j)a + \epsilon_1 n,\ 1 \leq j \leq k \leq n] \times \\
&\qquad \sup_{\{\zeta \geq 0 | \frac{y_1}{1+\zeta} \geq a\}} \sup_{y_1 - y_2 = a} \mathbf{P}[S^A_{-\zeta n,k} \leq (\zeta n + k - 1)\tfrac{y_1}{1+\zeta} + \epsilon_1 n,\ k = 1, \ldots, n] \times \\
&\qquad\qquad \mathbf{P}[S^B_{-\zeta n-1,0} \geq n y_2 - \epsilon_1 n] \\
&\geq \sup_{\{\zeta \geq 0 | \frac{y_1}{1+\zeta} \geq a\}} \sup_{y_1 - y_2 = a} \exp\Big\{ -n(\Lambda^{*-}_B(a) + \epsilon') - n\left[\Lambda^{*-}_A\left(\tfrac{y_1}{1+\zeta}\right)(1+\zeta) + \epsilon''\right] \\
&\qquad\qquad - n\left[\Lambda^{*+}_B\left(\tfrac{y_2 - \epsilon_1}{\zeta}\right)\zeta + \epsilon'''\right]\Big\},
\end{aligned} \tag{2.46}$$

where the last inequality holds for large n and is obtained by applying Assumption

G to the arrival and service processes. We can now choose appropriate ϵ', ϵ'' and ϵ''' such that for sufficiently large n and given ϵ_2 we have

$$\mathbf{P}[S^D_{1,k} \leq ka + 4\epsilon_1 n,\ k = 1, \ldots, n] \geq \sup_{\{\zeta \geq 0 | \frac{y_1}{1+\zeta} \geq a\}} \sup_{y_1 - y_2 = a} \exp\Big\{-n \Big[\Lambda^{*-}_B(a) + \Lambda^{*-}_A\left(\tfrac{y_1}{1+\zeta}\right)(1+\zeta) + + \Lambda^{*+}_B\left(\tfrac{y_2}{\zeta}\right)\zeta + \epsilon_2\Big]\Big\}. \quad (2.47)$$

We now argue that the constraint $\frac{y_1}{1+\zeta} \geq a$ can be removed from the optimization in (2.47). Consider a choice of $y_1 = \tilde{y}_1$, $y_2 = \tilde{y}_2$ and $\zeta = \tilde{\zeta}$ such that $\tilde{y}_1 - \tilde{y}_2 = a$ and $\frac{\tilde{y}_1}{1+\tilde{\zeta}} < a$. Let us now consider the subset of sample paths with $\zeta = 0$, $y_1 = a$ and $y_2 = 0$ from those satisfying (2.42), (2.43) and (2.44). It is easy to see that the probability of this subset is $e^{-n[\Lambda^{*-}_B(a) + \Lambda^{*-}_A(a)]}$. Now note that since $\frac{\tilde{y}_1}{1+\tilde{\zeta}} < a$ we have

$$\exp\{-n[\Lambda^{*-}_B(a) + \Lambda^{*-}_A(a)]\} > \exp\Big\{-n\Big[\Lambda^{*-}_B(a) + \Lambda^{*-}_A\left(\tfrac{\tilde{y}_1}{1+\tilde{\zeta}}\right)(1+\tilde{\zeta}) + \Lambda^{*+}_B\left(\tfrac{\tilde{y}_2}{\tilde{\zeta}}\right)\tilde{\zeta}\Big]\Big\}.$$

This shows that there exist choices of y_1, y_2 and ζ satisfying $\frac{y_1}{1+\zeta} \geq a$ that have a better exponent. Hence, the constraint $\frac{y_1}{1+\zeta} \geq a$ can indeed be removed.

We now use convex analysis to prove that $\Lambda^{*-}_\Gamma(a)$ as defined in Eq. (2.32) is equal to

$$-\sup_{\zeta \geq 0} \sup_{y_1 - y_2 = a} \left\{-(1+\zeta)\Lambda^{*-}_A\left(\tfrac{y_1}{1+\zeta}\right) - \zeta\Lambda^{*+}_B\left(\tfrac{y_2}{\zeta}\right)\right\}$$

thus, proving that the lower bound in (2.47) (taking $\epsilon_2 \to 0$) matches the upper bound obtained in Proposition 2.3.1. Dualizing the constraint $\Lambda_A(\theta) + \Lambda_B(-\theta) < 0$ we obtain (note that $\Lambda_A(\theta) + \Lambda_B(-\theta) < 0$ if and only if $\Lambda^-_A(\theta) + \Lambda^+_B(-\theta) < 0$)

$$\begin{aligned} -\Lambda^{*-}_\Gamma(a) &= -\sup_{\{\theta | \Lambda_A(\theta) + \Lambda_B(-\theta) < 0\}} [\theta a - \Lambda^-_A(\theta)] \\ &= \inf_{\{\theta | \Lambda_A(\theta) + \Lambda_B(-\theta) < 0\}} [-\theta a + \Lambda^-_A(\theta)] \\ &= \sup_{\zeta \geq 0} \left\{-\sup_\theta [\theta a - (1+\zeta)\Lambda^-_A(\theta) - \zeta\Lambda^+_B(-\theta)]\right\} \end{aligned}$$

$$= \sup_{\zeta \geq 0} \left\{ - \inf_{y_1 - y_2 = a} \left[(1+\zeta) \Lambda_A^{*-} \left(\tfrac{y_1}{1+\zeta} \right) + \zeta \Lambda_B^{*+} \left(\tfrac{y_2}{\zeta} \right) \right] \right\}$$

$$= \sup_{\zeta \geq 0} \sup_{y_1 - y_2 = a} \left[-(1+\zeta) \Lambda_A^{*-} \left(\tfrac{y_1}{1+\zeta} \right) - \zeta \Lambda_B^{*+} \left(\tfrac{y_2}{\zeta} \right) \right]. \tag{2.48}$$

To see that, note that for convex functions f, f_1, f_2 and for a scalar $c \geq 0$, it holds $(cf)^*(x^*) = cf^*(x^*/c)$, and $(f_1 + f_2)^*(x^*) = \inf_{x_1^* + x_2^* = x^*} [f_1^*(x_1^*) + f_2^*(x_2^*)]$ (see [Roc70, Thm. 16.1, Thm 16.4]).

In summary, we have verified that Assumption G holds for the departure process, i.e.,

$$\mathbf{P}[S_{1,i}^D \leq ia + 4\epsilon_1 n, \ i = 1, \ldots, n] \geq e^{-n(\Lambda_D^{*-}(a) + \epsilon_2)}. \tag{2.49}$$

By taking $\epsilon_1, \epsilon_2 \to 0$ and since $\mathbf{P}[S_{1,n}^D \leq na]$ is clearly larger than the probability in (2.49), (2.41) is verified for the same region.

■

Proof of Lemma 2.3.3: Note that for $k = 1, \ldots, n$ from (2.43) and (2.44) we obtain

$$\begin{aligned} S_{-\zeta n,k}^A \leq & (\zeta n + k - 1) \frac{y_1}{1+\zeta} + \epsilon_1 n \\ \leq & (n y_2 - \epsilon_1 n) + ((k-1)a + 2\epsilon_1 n) \\ \leq & (k-1)a + 2\epsilon_1 n + S_{-\zeta n - 1, 0}^B, \end{aligned} \tag{2.50}$$

where the second inequality holds because the two sides are equal at $k = n+1$ and because $\frac{y_1}{1+\zeta} \geq a$. The third inequality is justified by (2.42) and (2.44).

Let t be the arrival time of customer $-\zeta n - 1$. Then customer k arrives at time $t + S_{-\zeta n,k}^A$. We distinguish two cases. In case 1, customer k finds an empty system upon arrival. Then it departs at time t' where

$$t' = t + S_{-\zeta n,k}^A + B_k \leq ka + 3\epsilon_1 n + t + S_{-\zeta n - 1, 0}^B, \tag{2.51}$$

by using (2.42) and (2.50). Let t'' the departure time of the 0th customer. Clearly, $t'' \geq t + S^B_{-\zeta n-1,0}$, which along with (2.51) implies that $t' - t'' \leq ka + 3\epsilon_1 n \leq ka + 4\epsilon_1 n$. But, according to their definition $t' - t'' = S^D_{1,k}$.

In case 2, customer k finds a busy system upon arrival in which case $D_k = B_k$. Then, if this is also true for all $i = 1, \ldots, k-1$, we have $S^D_{1,k} = S^B_{1,k} \leq ka + \epsilon_1 n \leq ka + 4\epsilon_1 n$. If not, let $i \in [1, \ldots, k-1]$ the latest customer that finds the system empty (i.e., the one with maximum index). To bound $S^D_{1,i}$ we use the argument of Case 1. Thus,

$$S^D_{1,k} = S^D_{1,i} + S^D_{i+1,k} = S^D_{1,i} + S^B_{i+1,k} \leq ia + 3\epsilon_1 n + (k-i)a + \epsilon_1 n = ka + 4\epsilon_1 n,$$

where we have used (2.42) in the last inequality.

■

The proof of the above theorem indicates a most likely path along which the large deviation of $S^D_{1,n}$ occurs (in the sense that its probability equals $\mathbf{P}[S^D_{1,n} \leq na]$). Let ζ^*, y_1^* and y_2^* be a solution of the optimization problem in (2.47). The large deviation in $S^D_{1,n}$ occurs by

- Maintaining an empirical arrival rate of at least $\frac{1+\zeta^*}{y_1^*}$ from the arrival of customer $-\zeta^* n - 1$, until the departure of the nth customer, and an empirical service rate of at most $\frac{\zeta^*}{y_2^*}$ from the arrival of customer $-\zeta^* n - 1$, until the departure of the 0th customer, and by

- Maintaining an empirical service rate of at least $1/a$ from the departure of the 0th customer until the departure of the nth customer.

Figure 2-6 illustrates the situation.

Combining Propositions 2.3.1 and 2.3.2 we obtain the following theorem.

Theorem 2.3.4 *Under Assumptions F and G, the partial sum $S^D_{1,n}$ of the departure*

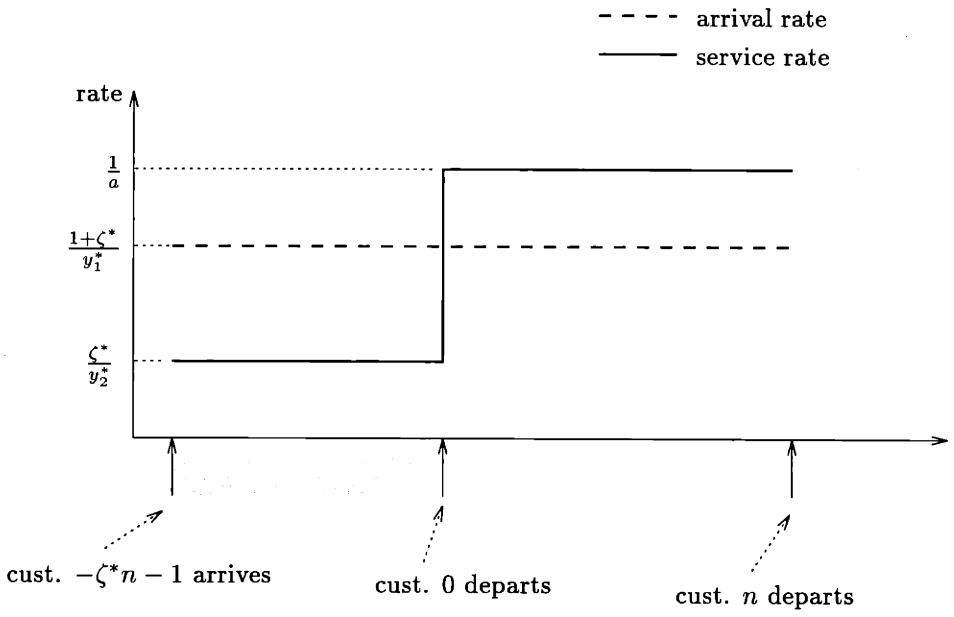

Figure 2-6: The most likely path for large deviations of $S_{1,n}^D$.

process of a G/GI/1 queue under FCFS satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_{1,n}^D \leq na] = -\Lambda_D^{*-}(a), \tag{2.52}$$

where

$$\Lambda_D^{*-}(a) = \Lambda_B^{*-}(a) + \Lambda_\Gamma^{*-}(a)$$

and

$$\Lambda_\Gamma^{*-}(a) = \sup_{\{\theta|\Lambda_A(\theta)+\Lambda_B(-\theta)<0\}} [\theta a - \Lambda_A^-(\theta)].$$

We now argue that the passing-through-a-queue operation preserves Assumption F. Proposition 2.3.2 establishes that it preserves Assumption G. To see that the departure process satisfies Assumption F, notice that we have proven an one-sided LDP for the departure process with large deviations rate function expressed as a

function of the large deviations rate function of the arrival and service processes.

Throughout this section we have assumed that the service times B_i are iid. A close examination of the proofs of Propositions 2.3.1 and 2.3.2, suggests that a weaker condition is sufficient for our purposes. Namely, we only need the random variables $S^B_{j,0}$ and $S^B_{1,n}$ to be approximately independent for every $j \leq 0$, as $n \to \infty$. A mixing condition of the type $\mathbf{E}[e^{\theta S^B_{j,0}} e^{\theta S^B_{1,n}}] = \mathbf{E}[e^{\theta S^B_{j,0}}]\mathbf{E}[e^{\theta S^B_{1,n}}]e^{n\epsilon(n)}$ for every $j \leq 0$ and θ, where $\lim_{n\to\infty} \epsilon(n) = 0$, is sufficient.

An alternative expression for $\Lambda^{*-}_D(\cdot)$ which is a consequence of the defining Eq. (2.31) is

$$\Lambda^{*-}_D(a) = \Lambda^{*-}_B(a) + \Lambda^{*-}_\Gamma(a) = \begin{cases} \Lambda^{*-}_B(a) + \Lambda^{*-}_A(a) & \text{if } a \geq \Lambda'_A(\theta^*) \\ \Lambda^{*-}_B(a) + \theta^* a - \Lambda_A(\theta^*) & \text{if } a < \Lambda'_A(\theta^*) \end{cases} \tag{2.53}$$

where θ^* is defined in the statement of Thm. 2.2.1 and $\Lambda'_A(x)$ denotes the derivative of $\Lambda_A(\cdot)$ evaluated at x. To see that consult Figure 2-2 and notice that the first branch of Eq. (2.53) corresponds to the region of a where the constraint $\Lambda_A(\theta) + \Lambda_B(-\theta) < 0$ is not tight, and the second branch to the region of a where this constraint is tight.

To obtain the limiting log-moment generating function for the partial sum of the departure process, we take the convex dual of $\Lambda^{*-}_D(\cdot)$ in (2.53). Using the duality correspondences proven in [Roc70, Sec. 16] we obtain the following corollary.

Corollary 2.3.5 *Under Assumptions F and G we have*

$$\Lambda^-_D(\theta) = \begin{cases} \inf_{\theta_1+\theta_2=\theta}\{\Lambda^-_B(\theta_1) + \Lambda^-_A(\theta_2)\} & \text{if } \theta \geq \hat{\theta} \\ \Lambda^-_B(\theta - \theta^*) + \Lambda_A(\theta^*) & \text{if } \theta < \hat{\theta} \end{cases} \tag{2.54}$$

where

$$\hat{\theta} \triangleq \frac{d}{da}\left[\Lambda^{*-}_B(a) + \Lambda^{*-}_A(a)\right]_{a=\Lambda'_A(\theta^*)}. \tag{2.55}$$

It is instructive to determine the fluctuations of the queue length that lead to a large deviation in the departure process. Let ζ^* solve the optimization problem in

(2.47). Let t be the arrival time of customer $-\zeta^* n - 1$. The 0th customer arrives at $t + S^A_{-\zeta^* n,0}$ and departs no earlier than $t + S^B_{-\zeta^* n-1,0}$. Thus, for the waiting time of customer 0 holds

$$W_0 \geq t + S^B_{-\zeta^* n-1,0} - t - S^A_{-\zeta^* n,0} = S^B_{-\zeta^* n-1,0} - S^A_{-\zeta^* n,0} \stackrel{\Delta}{=} \tilde{W}_0. \tag{2.56}$$

A close examination of the proofs of Propositions 2.3.1 and 2.3.2 suggests that $\Lambda^{*-}_\Gamma(\cdot)$ is the large deviations rate function of the process

$$\{S^A_{-\zeta^* n,k} - S^B_{-\zeta^* n-1,0},\ k = 1, \ldots, n\} \equiv \{S^A_{1,k} - \tilde{W}_0,\ k = 1, \ldots, n\}. \tag{2.57}$$

From the above discussion and Eq. (2.53) we conclude that depending on the value of a, we can distinguish two cases for the large deviation in the departure process to occur.

$a \geq \Lambda'_A(\theta^*)$: In this region, $\Lambda^{*-}_\Gamma(a) = \Lambda^{*-}_A(a)$ and from Eq. (2.57) it is clear that the most likely way for the large deviation in the departure process to occur is the 0th customer to incur $O(1)$ waiting time, which implies that it finds a queue length of $O(1)$ upon arrival.

$a < \Lambda'_A(\theta^*)$: In this region, $\Lambda^{*-}_\Gamma(a) = \theta^* a - \Lambda_A(\theta^*)$ and from Eq. (2.57) it is clear that the most likely way for the large deviation in the departure process to occur is the 0th customer to incur a large waiting time (recall from Thm. 2.2.1 that the large deviations rate function for the waiting time is linear with slope θ^*).

Hence, taking also into account Fig. 2-6 we can infer for the queue length the cases depicted in Figure 2-7. In Region 2 and in contrast with Region 1, the queue builds up to lead to a large deviation in the departure process.

2.3.1 Special Cases

In this section we apply Theorem 2.3.4 to two special cases. Namely, we study the departure process, in the large deviations regime, of an M/M/1 queue and a G/D/1

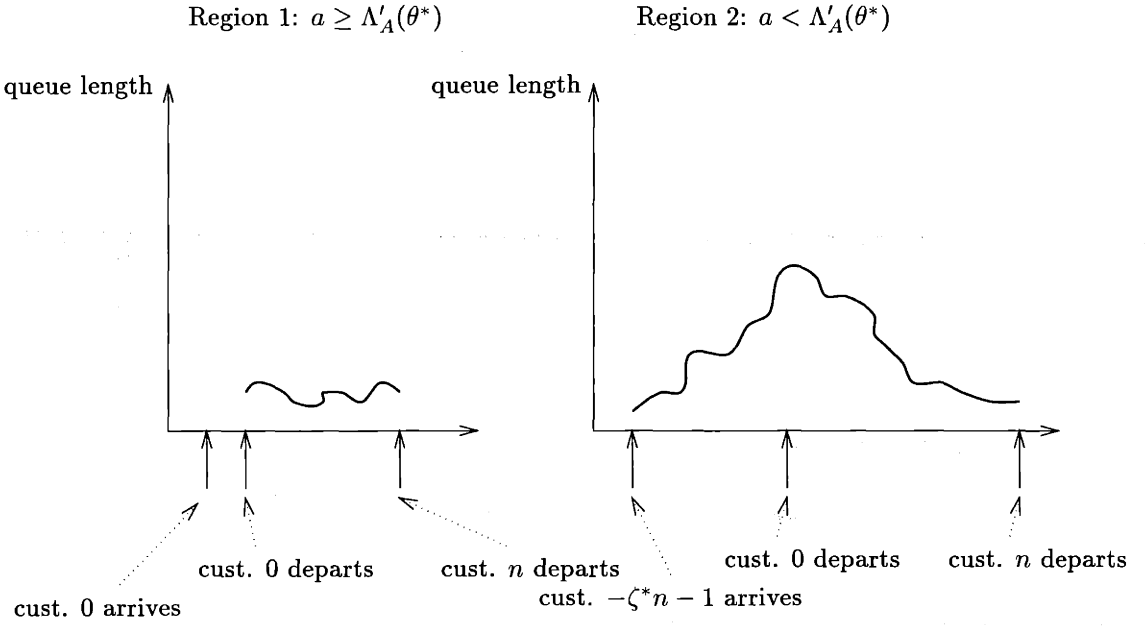

Figure 2-7: Two cases for the queue length: In Region 1, the 0th customer finds an O(1) queue upon arrival and until the nth customer departs the queue stays at an O(1) level. In Region 2, the queue first builds up (see also the arrival and service rates in Figure 2-6) and then it is depleted resulting in the large deviation in the departure process.

queue.

The departure process of a G/D/1 queue

We assume, as in Section 2.3, that the interarrival times process $\{A_i, i \in \mathbb{Z}\}$ is stationary and A_i are possibly dependent random variables. The service times B_i are iid random variables and equal to c w.p.1. Interarrival and service times are assumed independent.

It is straightforward that $\Lambda_B(\theta) = c\theta$. Therefore a simple calculation yields

$$\Lambda_B^{*-}(a) = \begin{cases} +\infty & \text{if } a < c \\ 0 & \text{if } a \geq c \end{cases} \tag{2.58}$$

Moreover,

$$\Lambda_\Gamma^{*-}(a) = \sup_{\{\theta | \Lambda_A(\theta) - c\theta < 0\}} [\theta a - \Lambda_A^-(\theta)] = \hat{\theta} a - \Lambda_A^-(\hat{\theta}), \tag{2.59}$$

where $\hat{\theta}$ is the optimizing θ. Note that by taking $a > c$ we have $\Lambda_A(\hat{\theta}) - c\hat{\theta} < 0$, which implies that for such a we have $\Lambda_\Gamma^{*-}(a) = \Lambda_A^{*-}(a)$. Therefore, using Eq. (2.31),

$$\Lambda_D^{*-}(a) = \begin{cases} +\infty & \text{if } a < c \\ \Lambda_A^{*-}(a) & \text{if } a \geq c \end{cases} \tag{2.60}$$

This is exactly the result obtained in [dVCW93] for a discrete time model. Taking the convex dual of the above we obtain

$$\Lambda_D^-(\theta) = \begin{cases} \Lambda_A^-(\theta) & \text{if } \theta \geq \frac{d}{da}[\Lambda_A^{*-}(a)]_{a=c} = \hat{\theta} \\ \theta c - \Lambda_A^{*-}(c) = \theta c - c\hat{\theta} + \Lambda_A^-(\hat{\theta}) & \text{if } \theta < \frac{d}{da}[\Lambda_A^{*-}(a)]_{a=c} = \hat{\theta}, \end{cases} \tag{2.61}$$

where $\hat{\theta}$ is the solution of the equation $\Lambda_A'(\theta) = c$.

The departure process of an M/M/1 queue

We assume that the arrival process is Poisson with rate λ and the service times are iid, distributed according to an exponential distribution with parameter μ.

It is straightforward to calculate

$$\Lambda_A(\theta) = \log\left(\frac{\lambda}{\lambda - \theta}\right), \qquad \Lambda_B(\theta) = \log\left(\frac{\mu}{\mu - \theta}\right). \tag{2.62}$$

Now, notice that

$$\Lambda_A(\theta) + \Lambda_B(-\theta) = 0 \Leftrightarrow \frac{\lambda}{\lambda - \theta}\frac{\mu}{\mu + \theta} = 1 \Leftrightarrow \theta = 0,\ \theta = \lambda - \mu, \tag{2.63}$$

which implies that $\theta^* = \lambda - \mu$, where θ^* is defined in the statement of Thm. 2.2.1. Moreover, notice that

$$\Lambda'_A(\theta^*) = \frac{\lambda - \theta^*}{\lambda}\frac{\lambda}{(\lambda - \theta^*)^2} = \frac{1}{\mu}.$$

Thus, using Eq. (2.53), we obtain for $a \geq 1/\mu$,

$$\Lambda_D^{*-}(a) = \Lambda_B^{*-}(a) + \Lambda_A^{*-}(a) = \Lambda_A^{*-}(a), \tag{2.64}$$

since by definition $\Lambda_B^{*-}(a) = 0$ for $a \geq 1/\mu$. Using the second branch of Eq. (2.53), we obtain for $a < 1/\mu$,

$$\Lambda_D^{*-}(a) = \Lambda_B^{*-}(a) + a(\lambda - \mu) - \log(\lambda/\mu). \tag{2.65}$$

But

$$\Lambda_B^{*-}(a) = \sup_{\theta}[\theta a - \Lambda_B^{-}(\theta)] = a\mu - 1 - \log(a\mu),$$

since, by differentiating, the optimal θ is found equal to $(a\mu - 1)/a$. Thus, from Eq. (2.65), for $a < 1/\mu$,

$$\Lambda_D^{*-}(a) = a\lambda - 1 - \log(a\lambda) = \Lambda_A^{*-}(a). \tag{2.66}$$

Summarizing Eq. (2.64) and (2.66) we finally obtain

$$\Lambda_D^{*-}(a) = \Lambda_A^{*-}(a). \tag{2.67}$$

This result is in accordance with Burke's output Theorem which states that the departure process of an M/M/1 queue is Poisson with rate λ (see [Kel79]).

2.4 Superposition of independent streams

In this section we treat the superposition operation of our network model. In particular, we derive a LDP for the process resulting from the superposition of independent arrival streams and we show that the superposition preserves Assumptions F and G. However, as it will become clear in the sequel, in order to derive this LDP we need a result that connects, in the large deviations regime, the Palm distribution of the arrival process (i.e., as it is seen by a random customer) with its stationary distribution as seen at a random time. This result is presented in Subsection 2.4.1 and could be of independent interest.

Consider two independent arrival streams. By A_i^1(resp. A_i^2), $i \in \mathbb{Z}$, we denote the interarrival time of the ith customer in stream 1 (resp. 2). We assume that the processes $\{A_i^1, A_i^2,\ i \in \mathbb{Z}\}$ are stationary, and mutually independent. However the interarrival times in each stream may be dependent. We impose Assumptions F and G on the arrival process of each stream. We denote by $A_i^{1,2}$, $i \in \mathbb{Z}$, the interarrival times of the process resulting from the superposition. It should be noted that in order to derive the LDP for the superposition, Assumption G is not used.

The next theorem establishes a LDP for the partial sum $S_{1,n}^{A^{1,2}}$ of the aggregate process, resulting from the superposition of streams 1 and 2.

Theorem 2.4.1 *Under Assumption F, the partial sum $S_{1,n}^{A^{1,2}}$ of the aggregate process, resulting from the superposition of the independent processes $A_i^1, A_i^2,\ i \in \mathbb{Z}$, satisfies*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_{1,n}^{A^{1,2}} \leq na] = - \inf_{\substack{\delta_1+\delta_2=1 \\ \delta_1,\delta_2\geq 0}} [\delta_1 \Lambda_{A^1}^{*-}(a/\delta_1) + \delta_2 \Lambda_{A^2}^{*-}(a/\delta_2)] \triangleq -\Lambda_{A^{1,2}}^{*-}(a). \tag{2.68}$$

Proof : Consult Figure 2-8. Consider the partial sum $S_{1,n}^{A^{1,2}}$ and let H_1 (resp. H_2) denote the event that the first customer of the aggregate process originates from stream 1 (resp. 2). We first obtain an upper bound on $\mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1]$. Notice

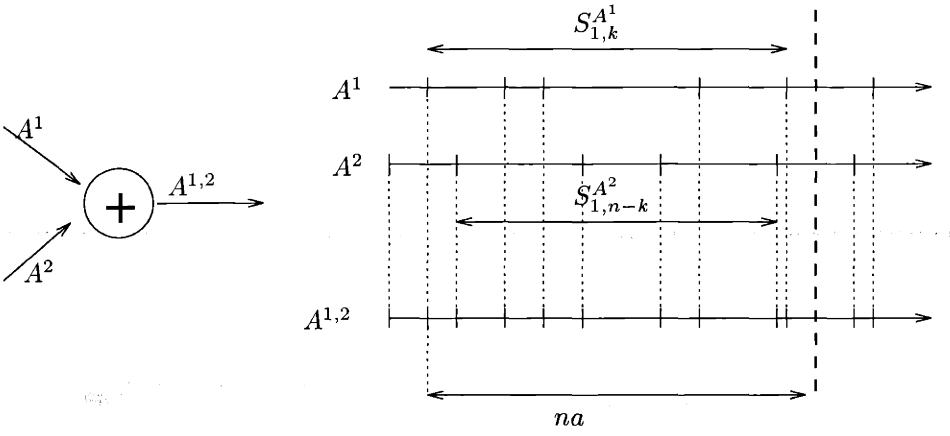

Figure 2-8: Superposition of two independent streams.

that

$$\mathbf{P}[S^{A^{1,2}}_{1,n} \leq na \mid H_1] \leq \sum_{k=1}^{n} \mathbf{P}[S^{A^1}_{1,k} \leq na]\mathbf{P}_R[S^{A^2}_{1,n-k} \leq na]. \tag{2.69}$$

Here, $\mathbf{P}[\cdot]$ denotes the probability distribution seen by a random customer (Palm distribution) and $\mathbf{P}_R[\cdot]$ denotes the probability distribution seen at a random time. Due to the independence of the two arrival streams, an arrival originating from stream 1 constitutes a random incidence in the arrival process of stream 2 and therefore we are interested in the probability distribution seen at a random time for events concerning stream 2.

In Subsection 2.4.1 it is shown that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}_R[S^{A^2}_{1,n} \leq na] = \lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S^{A^2}_{1,n} \leq na] = -\Lambda^{*-}_{A^2}(a). \tag{2.70}$$

Therefore, from (2.69), letting $k = n\delta$, $\delta \in [0,1]$ ($n\delta$ is assumed integer), and taking

large n we obtain

$$\begin{aligned}
\mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1] &\leq \sum_{\delta\in[0,1]} \mathbf{P}[S_{1,n\delta}^{A^1} \leq na]\mathbf{P}_R[S_{1,n(1-\delta)}^{A^2} \leq na] \\
&\leq n \sup_{\delta\in[0,1]} \mathbf{P}[S_{1,n\delta}^{A^1} \leq na]\mathbf{P}_R[S_{1,n(1-\delta)}^{A^2} \leq na] \Rightarrow \\
\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1] &\leq - \inf_{\delta\in[0,1]} [\delta\Lambda_{A^1}^{*-}(a/\delta) + (1-\delta)\Lambda_{A^2}^{*-}(a/(1-\delta))] \\
&= - \inf_{\substack{\delta_1+\delta_2=1 \\ \delta_1,\delta_2\geq 0}} [\delta_1\Lambda_{A^1}^{*-}(a/\delta_1) + \delta_2\Lambda_{A^2}^{*-}(a/\delta_2)].
\end{aligned} \tag{2.71}$$

To obtain a lower bound notice that

$$\begin{aligned}
\mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1] &\geq \sup_{\delta\in[0,1]} \mathbf{P}[S_{1,n\delta}^{A^1} \leq na]\mathbf{P}_R[S_{1,n(1-\delta)}^{A^2} \leq na] \Rightarrow \\
\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1] &\geq - \inf_{\delta\in[0,1]} [\delta\Lambda_{A^1}^{*-}(a/\delta) + (1-\delta)\Lambda_{A^2}^{*-}(a/(1-\delta))] \\
&= - \inf_{\substack{\delta_1+\delta_2=1 \\ \delta_1,\delta_2\geq 0}} [\delta_1\Lambda_{A^1}^{*-}(a/\delta_1) + \delta_2\Lambda_{A^2}^{*-}(a/\delta_2)].
\end{aligned} \tag{2.72}$$

Finally, observe that because of symmetry, Eqs. (2.71) and (2.72) also hold for $\mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_2]$. This along with the fact

$$\mathbf{P}[S_{1,n}^{A^{1,2}} \leq na] = \mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_1]\mathbf{P}[H_1] + \mathbf{P}[S_{1,n}^{A^{1,2}} \leq na \mid H_2]\mathbf{P}[H_2],$$

proves the theorem.

■

Remark : Let δ_1^*, δ_2^* be a solution to the optimization problem in (2.68). It can be seen that a most likely path to have a large deviation in the aggregate process is to maintain an empirical arrival rate of $\frac{\delta_1^*}{a}$ in stream 1 and a rate of $\frac{\delta_2^*}{a}$ in stream 2. Then, since $\delta_1^* + \delta_2^* = 1$ the empirical rate of the aggregate process is $\frac{1}{a}$.

Using induction on the number of streams superimposed we generalize Theorem 2.4.1 to obtain the following corollary.

Corollary 2.4.2 *Under Assumption F, the partial sum $S_{1,n}^{A^{1,\dots,m}}$ of the aggregate process, resulting from the superposition of the m independent processes $A_i^1, \dots, A_i^m$ $i \in \mathbb{Z}$, satisfies*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S_{1,n}^{A^{1,\dots,m}} \leq na] = - \inf_{\substack{\delta_1+\cdots+\delta_m=1 \\ \delta_1,\dots,\delta_m\geq 0}} \sum_{k=1}^{m} \delta_k \Lambda_{A^k}^{*-}(a/\delta_k) \triangleq -\Lambda_{A^{1,\dots,m}}^{*-}(a). \tag{2.73}$$

Using convex duality one can obtain the limiting log-moment generating function $\Lambda_{A^{1,\dots,m}}^{-}(\cdot)$ of $S_{1,n}^{A^{1,\dots,m}}$ as the convex dual of its large deviations rate function $\Lambda_{A^{1,\dots,m}}^{*-}(\cdot)$. The latter function is convex by [Roc70, Thm. 5.8].

We now proceed into proving that the aggregate process, resulting from the superposition of independent streams which satisfy Assumptions F and G also satisfies the same assumptions. It is clear that the process resulting from the superposition satisfies Assumption F, since we have proven an one sided LDP for this process with large deviations rate function expressed as a function of the large deviations rate function of the superimposed processes. The next Theorem establishes that the process resulting from the superposition satisfies Assumption G.

Theorem 2.4.3 *Assume that the m independent processes $A_i^1, \dots, A_i^m$ $i \in \mathbb{Z}$ satisfy Assumption G. The aggregate process resulting from their superposition also satisfies Assumption G.*

Proof : It suffices to prove the result for $m = 2$ since by using induction we can prove it for any m. We need to prove that for every $\epsilon_1, \epsilon_2, a > 0$, there exists M_S such that for all $n \geq M_S$

$$e^{-n(\Lambda_{A^{1,2}}^{*-}(a)+\epsilon_2)} \leq \mathbf{P}[S_{1,j}^{A^{1,2}} - ja \leq \epsilon_1 n,\ j = 1, \dots, n]. \tag{2.74}$$

Following the steps of the proof of Theorem 2.4.1 we consider the scenario that a fraction δ of customers of the aggregate process originates from the A^1 process. Again, H_1 denotes the event that customer 1 of the aggregate process originates from the A^1

process. We have

$$\begin{aligned}\mathbf{P}[S_{1,j}^{A^{1,2}} - ja \le \epsilon_1 n,\ j = 1, \dots, n \mid H_1] \ge \\ \ge \sup_{\delta \in [0,1]} \mathbf{P}[S_{1,j\delta}^{A^1} - ja \le \epsilon_1 n,\ j = 1, \dots, n] \times \\ \mathbf{P}_R[S_{1,j(1-\delta)}^{A^2} - ja \le \epsilon_1 n,\ j = 1, \dots, n]. \end{aligned} \tag{2.75}$$

Using Assumption G for the A^1 stream we obtain for large enough n

$$\mathbf{P}[S_{1,j\delta}^{A^1} - ja \le \epsilon_1 n,\ j = 1, \dots, n] \ge e^{-n\delta(\Lambda_{A^1}^{*-}(a/\delta)+\epsilon')}. \tag{2.76}$$

In Subsection 2.4.1 (Lemma 2.4.6) it is shown that for large enough n

$$\mathbf{P}_R[S_{1,j(1-\delta)}^{A^2} - ja \le \epsilon_1 n,\ j = 1, \dots, n] \ge e^{-n(1-\delta)(\Lambda_{A^2}^{*-}(a/(1-\delta))+\epsilon'')}. \tag{2.77}$$

To obtain (2.74) it suffices to choose appropriate ϵ' and ϵ'' such that for large enough n and given ϵ_2

$$e^{-n \inf_{\delta \in [0,1]}[\delta(\Lambda_{A^1}^{*-}(a/\delta)+\epsilon')+(1-\delta)(\Lambda_{A^2}^{*-}(a/(1-\delta))+\epsilon'')]} \ge e^{-n(\Lambda_{A^{1,2}}^{*-}(a)+\epsilon_2)}.$$

■

2.4.1 Connection between Palm and stationary distributions in the large deviations regime

In this subsection we show that the stationary and the Palm distribution of the same point process have the same large deviations behaviour.

Consider a stationary arrival process satisfying Assumptions F with interarrivals

A_i, $i \in \mathbb{Z}$. We have

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[S^A_{1,n} \leq na] = -\Lambda^{*-}_A(a). \tag{2.78}$$

As explained in the proof of Thm. 2.4.1, $\mathbf{P}[\cdot]$ denotes the probability distribution seen by a random customer (customer 1 in the case of Eq. (2.78)). Consider now a random time (say $t = 0$) and assume that customer 0 is the first customer to arrive after $t = 0$. Let U, V denote the duration and the age, respectively, of A_0. The situation is depicted in Figure 2-9. By $\mathbf{P}_R[\cdot]$ we denote the probability distribution

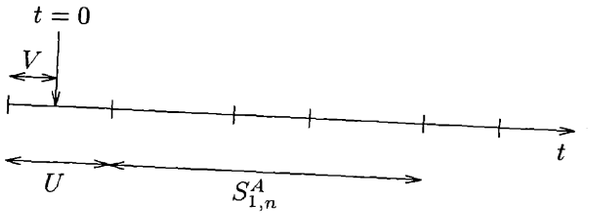

Figure 2-9: The arrival process seen at a random time.

seen at the random time $t = 0$ and we are interested in obtaining a LDP for $S^A_{1,n}$ under $\mathbf{P}_R[\cdot]$. The next theorem establishes the result. Moreover, we are also interested in obtaining a LDP result for the partial sum process $\{S^A_{1,j},\ j = 1, \ldots, n\}$ under $\mathbf{P}_R[\cdot]$ when Assumption G is satisfied. The latter result is obtained in Lemma 2.4.6.

Theorem 2.4.4 *Under Assumption F we have*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}_R[S^A_{1,n} \leq na] = -\Lambda^{*-}_A(a). \tag{2.79}$$

Proof : Let $\mathbf{E}_R[\cdot]$ denote the expectation with respect to $\mathbf{P}_R[\cdot]$. We use a standard procedure to relate $\mathbf{E}_R[\cdot]$ to $\mathbf{E}[\cdot]$ (see [Wal88]). Consider an arbitrary function $f(\cdot)$ of

$S^A_{1,n}$. It can be shown ([Wal88, ch. 7]) that

$$\mathbf{E}_R[f(S^A_{1,n}) \mid V = v, U = u] = \mathbf{E}[f(S^A_{1,n}) \mid A_0 = u].$$

Thus following the steps in [Wal88, ch. 7],

$$\begin{aligned}
\mathbf{E}_R[f(S^A_{1,n})] &= \frac{1}{\mathbf{E}[A_1]} \int_{u=0}^{\infty} \int_{v=0}^{u} \mathbf{E}[f(S^A_{1,n}) \mid A_0 = u] \, dv \, dF_{A_0}(u) \\
&= \mathbf{E}[\int_{v=0}^{A_0} f(S^A_{1,n}) \, dv] \\
&= \mathbf{E}[A_0 f(S^A_{1,n})], \qquad (2.80)
\end{aligned}$$

where we have assumed without loss of generality that $\mathbf{E}[A_1] = 1$, and we have used the notation $F_{A_0}(\cdot)$ for the distribution function of A_0.

To obtain an upper bound on $\mathbf{E}_R[e^{\theta S^A_{1,n}}]$ we set $f(\cdot) = e^{\theta \cdot}$ and use Hölder's inequality. Namely,

$$\begin{aligned}
&\mathbf{E}_R[e^{\theta S^A_{1,n}}] = \mathbf{E}[A_0 e^{\theta S^A_{1,n}}] \\
(p+q=1) \qquad &= \mathbf{E}[(A_0^{1/p})^p \left(e^{(\theta/q) S^A_{1,n}}\right)^q] \leq \mathbf{E}[A_0^{1/p}]^p \mathbf{E}[e^{(\theta/q) S^A_{1,n}}]^q, \qquad (2.81)
\end{aligned}$$

which implies

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}_R[e^{\theta S^A_{1,n}}] \leq \limsup_{n\to\infty} \frac{p \log \mathbf{E}[A_0^{1/p}]}{n} + q\Lambda_A(\theta/q) = q\Lambda_A(\theta/q), \quad (2.82)$$

since the first term of the right hand side vanishes. Taking the now limit as $q \to 1$ in the above equation we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{E}_R[e^{\theta S^A_{1,n}}] \leq \Lambda_A(\theta). \qquad (2.83)$$

Therefore using Eq. (2.83) and the Markov inequality we obtain

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbf{P}_R[S^A_{1,n} \leq na] \leq -\Lambda^{*-}_A(a). \tag{2.84}$$

To obtain now a lower bound on $\mathbf{P}_R[S^A_{1,n} \leq na]$ set $f(S^A_{1,n}) = \mathbf{1}\{S^A_{1,n} \leq na\}$ in Eq. (2.80), where $\mathbf{1}\{\cdot\}$ denotes the indicator function. We have

$$\begin{aligned} \mathbf{P}_R[S^A_{1,n} \leq na] &= \int_0^\infty u\mathbf{P}[S^A_{1,n} \leq na \mid A_0 = u]\, dF_{A_0}(u) \\ &\geq \frac{1}{n^2} \int_{1/n^2}^\infty \mathbf{P}[S^A_{1,n} \leq na \mid A_0 = u]\, dF_{A_0}(u) \\ &= \frac{1}{n^2} \mathbf{P}[S^A_{1,n} \leq na, A_0 \geq \tfrac{1}{n^2}]. \end{aligned} \tag{2.85}$$

We need the following lemma the proof of which is deferred until the end of the current proof.

Lemma 2.4.5 *Under Assumption F and for every positive ϵ and a, there exists $N_{a,\epsilon}$ such that for every $n \geq N_{a,\epsilon}$ it holds*

$$\mathbf{P}[S^A_{1,n} \leq na, A_0 \geq \tfrac{1}{n^2}] \geq e^{-n(\Lambda^{*-}_A(a)+\epsilon)}. \tag{2.86}$$

We now use Lemma 2.4.5 in Eq. (2.85) and take $\epsilon \to 0$ to obtain

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}_R[S^A_{1,n} \leq na] \geq -\Lambda^{*-}_A(a).$$

■

Proof of Lemma 2.4.5: Eq. (2.78) implies that for every positive ϵ' and a there exists $N'_{a,\epsilon'}$ such that for every $n \geq N'_{a,\epsilon'}$ it holds

$$e^{-n(\Lambda^{*-}_A(a)+\epsilon')} \leq \mathbf{P}[S^A_{1,n} \leq na] \leq e^{-n(\Lambda^{*-}_A(a)-\epsilon')}. \tag{2.87}$$

Fix now $a, \epsilon' > 0$, and let $\delta = \epsilon'$. We have

$$\begin{aligned}
& \mathbf{P}[S^A_{1,n} \leq na, A_0 \geq \tfrac{1}{n^2}] = \\
\text{(by stationarity)} \quad & = \frac{1}{n\delta}\sum_{i=1}^{n\delta} \mathbf{P}[S^A_{i+1,i+n} \leq na, A_i \geq \tfrac{1}{n^2}] \\
\text{(union bound)} \quad & \geq \frac{1}{n\delta}\mathbf{P}[\exists i \in [1, n\delta] \text{ s.t. } S^A_{i+1,i+n} \leq na, A_i \geq \tfrac{1}{n^2}] \\
& \geq \frac{1}{n\delta}\mathbf{P}[S^A_{1,n(1+\delta)} \leq na, \exists i \in [1, n\delta] \text{ s.t. } A_i \geq \tfrac{1}{n^2}] \\
& \geq \frac{1}{n\delta}\mathbf{P}[S^A_{1,n(1+\delta)} \leq na, \sum_{i=1}^{n\delta} A_i \geq \tfrac{n\delta}{n^2}] \\
(\mathbf{P}[A\cap B] \geq \mathbf{P}[A] - \mathbf{P}[B^C]) \quad & \geq \frac{1}{n\delta}\mathbf{P}[S^A_{1,n(1+\delta)} \leq na] - \frac{1}{n\delta}\mathbf{P}[S^A_{1,n\delta} \leq \tfrac{\delta}{n}] \\
& \geq \frac{1}{n\delta}e^{-n(1+\delta)(\Lambda^{*-}_A(\frac{a}{1+\delta})+\epsilon')} - \frac{1}{n\delta}\mathbf{P}[S^A_{1,n\delta} \leq \tfrac{\delta}{n}] \quad (2.88)
\end{aligned}$$

where the last inequality holds for all $n \geq N'_{\frac{a}{1+\delta},\epsilon'}$. Note that we have used the notation B^C to denote the complement of B. We next show that for $n \to \infty$ (keeping a, δ, ϵ' fixed) we can neglect the second term in the right hand side of (2.88). To see that note that for all β positive there exists $N_{\beta,\epsilon'}$ such that for all $n \geq N_{\beta,\epsilon'}$ it holds

$$\mathbf{P}[S^A_{1,n\delta} \leq \tfrac{\delta}{n}] \leq \mathbf{P}[S^A_{1,n\delta} \leq n\delta\beta] \leq e^{-n(\Lambda^{*-}_A(\beta)-\epsilon')}. \quad (2.89)$$

By taking β, δ and ϵ' small enough and $n \geq N_{\beta,\epsilon'}$ we can achieve

$$\Lambda^{*-}_A(\beta) - \epsilon' > (1+\delta)(\Lambda^{*-}_A(\tfrac{a}{1+\delta}) + \epsilon'). \quad (2.90)$$

Here we are using the fact that for sufficiently small β, $\Lambda^{*-}_A(\beta) > \Lambda^{*-}_A(\frac{a}{1+\delta})$ since $\Lambda^{*-}_A(\beta)$ is monotonically increasing as $\beta \downarrow 0$.

Observe now that the value of β which satisfies Eq. (2.90) is a function of a, δ and ϵ'. Therefore, using Eq. (2.89), there exists $N_{a,\delta,\epsilon'}$ such that for all $n \geq N_{a,\delta,\epsilon'}$ it

holds

$$-\frac{1}{n\delta}\mathbf{P}[S_{1,n\delta} \leq \tfrac{\delta}{n}] \geq -\frac{1}{2n\delta}e^{-n(1+\delta)(\Lambda_A^{*-}(\frac{a}{1+\delta})+\epsilon')}. \tag{2.91}$$

Combining Eqs. (2.91) and (2.88) we conclude that there exists $\hat{N}_{a,\delta,\epsilon'}$ such that for all $n \geq \hat{N}_{a,\delta,\epsilon'}$ it holds

$$\mathbf{P}[S_{1,n}^A \leq na, A_0 \geq \tfrac{1}{n^2}] \geq \frac{1}{2n\delta}e^{-n(1+\delta)(\Lambda_A^{*-}(\frac{a}{1+\delta})+\epsilon')}. \tag{2.92}$$

We now choose ϵ' such that (recall $\delta = \epsilon'$)

$$\frac{1}{2n\epsilon'}e^{-n(1+\epsilon')(\Lambda_A^{*-}(\frac{a}{1+\epsilon'})+\epsilon')} \geq e^{-n(\Lambda_A^{*-}(a)+\epsilon)},$$

for all $n \geq N_{a,\epsilon}$. This can be done due to the lower semicontinuity of $\Lambda_A^{*-}(\cdot)$ (see the argument in the proof of Lemma 2.2.5 in [DZ93b]).

■

Lemma 2.4.6 *Under Assumptions F and G we have that for every $\epsilon_1, \epsilon_2, a > 0$ there exists $N_{a,\epsilon_1,\epsilon_2}$ such that for all $n \geq N_{a,\epsilon_1,\epsilon_2}$ it holds*

$$\mathbf{P}_R[S_{1,j}^A \leq ja + \epsilon_1 n,\ j = 1, \ldots, n] \geq e^{-n(\Lambda_A^{*-}(a)+\epsilon_2)}. \tag{2.93}$$

Proof : Following the proof of the lower bound in Thm. 2.4.4 (using the argument used to derive Eq. (2.85) but applied to the sample path $S_{1,j}^A \leq ja+\epsilon_1 n,\ j = 1, \ldots, n$) we have

$$\mathbf{P}_R[S_{1,j}^A \leq ja + \epsilon_1 n,\ j = 1, \ldots, n] \geq \frac{1}{n^2}\mathbf{P}[S_{1,j}^A \leq ja + \epsilon_1 n,\ j = 1, \ldots, n,\ A_0 \geq \tfrac{1}{n^2}]. \tag{2.94}$$

Now, as in the proof of Lemma 2.4.5, fixing $a, \epsilon_1, \epsilon_2 > 0$ we obtain

$$\mathbf{P}[S_{1,j}^A \leq ja + \epsilon_1 n,\ j = 1, \ldots, n,\ A_0 \geq \tfrac{1}{n^2}] =$$

$$
\begin{aligned}
&= \frac{1}{n\delta}\sum_{k=1}^{n\delta}\mathbf{P}[S^A_{1+k,j+k} \le ja+\epsilon_1 n,\ j=1,\dots,n,\ A_k \ge \tfrac{1}{n^2}] \\
&\ge \frac{1}{n\delta}\mathbf{P}[\exists k\in[1,n\delta] \text{ s.t. } S^A_{1+k,j+k} \le ja+\epsilon_1 n,\ j=1,\dots,n,\ A_k \ge \tfrac{1}{n^2}] \\
&\ge \frac{1}{n\delta}\mathbf{P}[\forall k\in[1,n\delta]\ S^A_{1+k,j+k} \le ja+\epsilon_1 n,\ j=1,\dots,n, \exists k\in[1,n\delta] \text{ s.t. } A_k \ge \tfrac{1}{n^2}] \\
&\ge \frac{1}{n\delta}\mathbf{P}[\forall k\in[1,n\delta]\ S^A_{1+k,j+k} \le ja+\epsilon_1 n,\ j=1,\dots,n] - \frac{1}{n\delta}\mathbf{P}[S^A_{1,n\delta} \le \tfrac{\delta}{n}].
\end{aligned}
\tag{2.95}
$$

Now notice that

$$
\begin{aligned}
&\mathbf{P}[\forall k\in[1,n\delta]\ S^A_{1+k,j+k} \le ja+\epsilon_1 n,\ j=1,\dots,n] \\
&=\mathbf{P}[\forall k\in[1,n\delta]\ S^A_{1,j+k} - S^A_{1,k} \le (j+k)a - ka + \epsilon_1 n,\ j=1,\dots,n] \\
&\ge\mathbf{P}[S^A_{1,j+k} \le (j+k)a + \tfrac{\epsilon_1 n}{2},\ \forall k\in[1,n\delta],\ j=1,\dots,n, \\
&\qquad\qquad S^A_{1,k} \ge ka - \tfrac{\epsilon_1 n}{2},\ \forall k\in[1,n\delta]] \\
&=\mathbf{P}[S^A_{1,j+k} \le (j+k)a + \tfrac{\epsilon_1 n}{2},\ \forall k\in[1,n\delta],\ j=1,\dots,n] \\
&\ge e^{-n(1+\delta)(\Lambda_A^{*-}(a)+\epsilon')},
\end{aligned}
\tag{2.96}
$$

where the last equality is obtained by choosing sufficiently small δ such that $n\delta a - \frac{\epsilon_1 n}{2} < 0$ which implies that $\mathbf{P}[S^A_{1,k} \ge ka - \frac{\epsilon_1 n}{2},\ \forall k\in[1,n\delta]] = 1$. The last inequality holds, due to Assumption G, for all $n \ge N'_{a,\epsilon_1,\epsilon'}$. Now, as in Lemma 2.4.5 it can be shown that there exists $N''_{a,\delta,\epsilon'}$ such that for all $n \ge N''_{a,\delta,\epsilon'}$ it holds

$$
-\frac{1}{n\delta}\mathbf{P}[S_{1,n\delta} \le \tfrac{\delta}{n}] \ge -\frac{1}{2n\delta}e^{-n(1+\delta)(\Lambda_A^{*-}(a)+\epsilon')}. \tag{2.97}
$$

Combining Eqs. (2.94), (2.95), (2.96) and (2.97) we conclude that there exists $\hat{N}_{a,\epsilon_1,\delta,\epsilon'}$ such that for all $n \ge \hat{N}_{a,\epsilon_1,\delta,\epsilon'}$ it holds

$$
\mathbf{P}_R[S^A_{1,j} \le ja+\epsilon_1 n,\ j=1,\dots,n] \ge \frac{1}{2n^3\delta}e^{-n(1+\delta)(\Lambda_A^{*-}(a)+\epsilon')}. \tag{2.98}
$$

We now choose ϵ' and if necessary δ smaller than the one chosen above for the purposes

of (2.96), such that

$$\frac{1}{2n^3\delta}e^{-n(1+\delta)(\Lambda_A^{*-}(a)+\epsilon')} \geq e^{-n(\Lambda_A^{*-}(a)+\epsilon_2)},$$

for $n \geq N_{a,\epsilon_1,\epsilon_2}$.

■

2.5 Deterministic splitting of a stream

In this section we treat the splitting operation of our network model. In particular, we derive a LDP for the process resulting from the splitting of a stream to a number of streams and we show that splitting preserves Assumptions F and G.

Consider a stream with stationary interarrival times A_i, $i \in \mathbb{Z}$, which is split to 2 substreams. In particular, a fraction p of arrivals of the *"master"* stream is directed to substream 1 and a fraction $1-p$ to substream 2. The next theorem provides a LDP for stream 1. Since stream 1 is chosen arbitrarily, by relabelling the streams one can obtain a LDP for stream 2. The more general case in which the master stream is split to more than two substreams can be handled by successive splitting to two substreams. Let us denote by A_i^1, A_i^2, $i \in \mathbb{Z}$, the interarrival times of substreams $1, 2$, respectively. $\Lambda_A^*(\cdot)$ and $\Lambda_A(\cdot)$ denote the large deviations rate function and the limiting log-moment generating function of the master stream.

Theorem 2.5.1 *Under Assumption F, the partial sum $S_{1,n}^{A^1}$ of substream 1 satisfies*

$$\lim_{n\to\infty}\frac{1}{n}\log \mathbf{P}[S_{1,n}^{A^1} \leq na] = -\frac{1}{p}\Lambda_A^{*-}(ap). \tag{2.99}$$

Proof : To have n arrivals in substream 1 we need n/p arrivals of the master stream. Since we are interested in large values of n we will ignore integrality issues (i.e., it

holds $\lfloor n/p \rfloor / n \to 1/p$, as $n \to \infty$). Thus,

$$\mathbf{P}[S^{A^1}_{1,n} \leq na] = \mathbf{P}[S^{A}_{1,n/p} \leq na] \leq e^{-(n/p)(\Lambda^{*-}_A(ap)-\epsilon)}.$$

Similarly for the lower bound.

■

We now argue that splitting preserves Assumptions F, and G. It is clear that the process resulting from splitting satisfies Assumption F, since we have proven an one sided LDP for this process with large deviations rate function expressed as a function of the large deviations rate function of the master process. The next theorem establishes that the process resulting from splitting satisfies Assumption G.

Theorem 2.5.2 *Assume that the process $\{A_i,\ i \in \mathbb{Z}\}$, satisfies Assumptions F and G. Then the A^1 process satisfies Assumption G.*

Proof : The proof is very similar to the proof of Theorem 2.5.1.

$$\begin{aligned}\mathbf{P}[S^{A^1}_{1,j} - ja \leq \epsilon_1 n,\ j = 1, \ldots, n] \geq & \mathbf{P}[S^{A}_{1,j/p} - ja \leq \epsilon_1 n,\ j = 1, \ldots, n] \\ \geq & e^{-(n/p)(\Lambda^{*-}_A(ap)+\epsilon)},\end{aligned}$$

for n large enough and all $\epsilon_1, \epsilon > 0$ by using Assumption G for the master process.

■

2.6 An Example: Queues in Tandem

In this section we apply the results derived so far to obtain LDP's for two G/GI/1 queues in tandem. Moreover we work out a numerical example in order to get a qualitative understanding of the results. Large deviations results for tandem queues with renewal arrivals and exponential servers have been reported in [GA94].

Consider two G/GI/1 queues in tandem. Let $A_i,\ i \in \mathbb{Z}$, denote the interarrival

times in the first queue and $B_i^1, B_i^2,\ i \in \mathbb{Z}$, the service times in the first and second queue respectively. These processes are mutually independent, stationary and satisfy Assumptions F and G.

According to Corollary 2.3.5 the limiting log-moment generating function of the departure process from the first queue is given by

$$\Lambda_D^-(\theta) = \begin{cases} \inf_{x+y=\theta}\{\Lambda_{B^1}^-(x) + \Lambda_A^-(y)\} & \text{if } \theta \geq \hat{\theta} \\ \Lambda_{B^1}^-(\theta - \theta^*) + \Lambda_A(\theta^*) & \text{if } \theta < \hat{\theta} \end{cases} \tag{2.100}$$

where

$$\hat{\theta} \triangleq \frac{d}{da}\left[\Lambda_{B^1}^{*-}(a) + \Lambda_A^{*-}(a)\right]_{a=\Lambda_A'(\theta_1^*)}.$$

Applying Theorem 2.2.1 we obtain that the tail probability of the stationary waiting time, W_2, seen by a customer in the second queue, is characterized by

$$\mathbf{P}[W_2 \geq U] \sim e^{\theta_2^* U}, \tag{2.101}$$

where U is large enough and $\theta_2^* < 0$ is the smallest root of the equation $\Lambda_D(\theta) + \Lambda_{B^2}(-\theta) = 0$. Since for $\theta \leq 0$ the equation $\Lambda_D(\theta) + \Lambda_{B^2}(-\theta) = 0$ has exactly the same roots as the equation $\Lambda_D^-(\theta) + \Lambda_{B^2}^+(-\theta) = 0$, it turns out that θ_2^* is the smallest root of the equation

$$\begin{aligned} &\inf_{x+y=\theta}\{\Lambda_{B^1}^-(x) + \Lambda_A^-(y)\} + \Lambda_{B^2}^+(-\theta) = 0 \quad \text{if } \theta \geq \hat{\theta} \\ &\Lambda_{B^1}^-(\theta - \theta^*) + \Lambda_A(\theta^*) + \Lambda_{B^2}^+(-\theta) = 0 \qquad \text{if } \theta < \hat{\theta} \end{aligned}$$

It is instructive to characterize the most likely path along which the LDP for the waiting time occurs in the second queue. The remarks after the proof of Theorem 2.2.1, suggest that the most likely path for the waiting time in the second queue is characterized by

$$\mathbf{P}[W_0^2 \geq (i+1)a] \sim \sup_{x_2 - x_1 = a} \mathbf{P}[S_{-i,0}^D \leq (i+1)x_1]\mathbf{P}[S_{-i-1,-1}^{B^2} \geq (i+1)x_2], \tag{2.102}$$

where W_0^2 denotes the waiting time of the 0th customer in the second queue and i is large enough. Setting $U = (i+1)a$, we obtain for large enough U

$$\mathbf{P}[W_0^2 \geq U] \sim \exp\left\{-U \inf_{a>0} \frac{1}{a} \inf_{x_2 - x_1 = a} [\Lambda_D^{*-}(x_1) + \Lambda_{B^2}^{*+}(x_2)]\right\}. \tag{2.103}$$

Let (a^*, x_1^*, x_2^*) be an optimal solution of the optimization problem appearing in (2.103). Eq. (2.103) suggests that the waiting time in the second queue builds up by maintaining an empirical rate of $1/x_1^*$ for the process D (departure from first queue) and an empirical service rate (process B^2) of $1/x_2^*$.

We use the remarks after Theorem 2.3.4 to characterize the most likely path for the process D to maintain an empirical rate of $1/x_1^*$. Let i^* be defined by the equation $i^* + 1 = U/a^*$. From (2.102), it can be seen that it suffices to characterize the most likely path along which the event $\{S_{-i^*,0}^D \leq (i^*+1)x_1^*\}$ occurs. As shown in Theorem 2.3.4, this most likely path is characterized by

$$\mathbf{P}[S_{-i^*,0}^D \leq (i^*+1)x_1^*] \sim \exp\left\{(i^*+1) \sup_{\zeta \geq 0} \sup_{y_1 - y_2 = a} \left[-(1+\zeta)\Lambda_A^{*-}\left(\frac{y_1}{1+\zeta}\right) - \right.\right.$$
$$\left.\left. -\zeta \Lambda_{B^1}^{*+}\left(\frac{y_2}{\zeta}\right)\right] - (i^*+1)\Lambda_{B^1}^{*-}(a)\right\} \tag{2.104}$$

Let (y_1^*, y_2^*, ζ^*) be the solution of the optimization problem appearing in Eq. (2.104). We depict the most likely path in Figure 2-10.

We now proceed with a numerical example. We chose the arrival process A to be a two-state *Markov modulated* deterministic process. More precisely, we consider a two-state Markov chain with transition probability matrix

$$P = \begin{bmatrix} 0.8 & 0.2 \\ 0.3 & 0.7 \end{bmatrix}, \tag{2.105}$$

and we let the interarrival times be equal to $\frac{1}{\lambda_1} = \frac{1}{5}$ w.p.1 when the chain is at state 1, and equal to $\frac{1}{\lambda_2} = \frac{1}{10}$ w.p.1 when the chain is at state 2. The steady-state probability vector for this Markov chain is $[\pi_1 \ \pi_2] = [0.6 \ 0.4]$ and thus the mean inter-arrival is

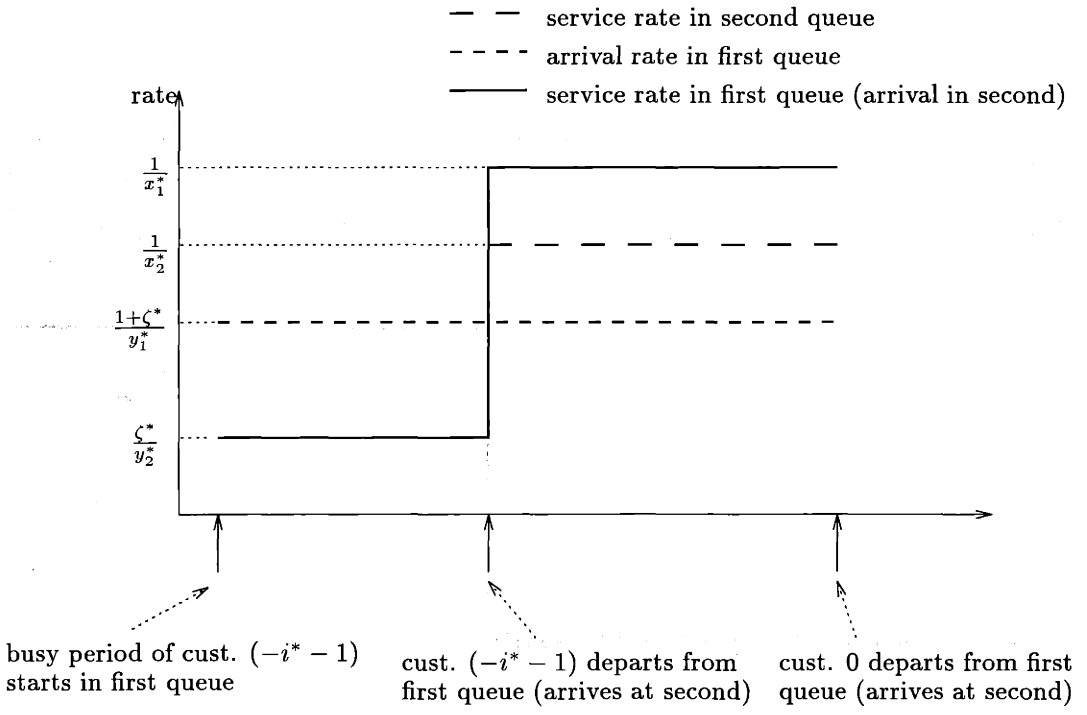

Figure 2-10: The most likely path for the waiting time in the second queue.

$\frac{1}{\lambda_1}\pi_1 + \frac{1}{\lambda_2}\pi_2 = 0.16$. We chose a deterministic server for both queues 1 and 2 with service times $c = 0.13$.

Theorem 3.1.2 in [DZ93b] calculates the limiting log-moment generating function for the arrival process as the largest eigenvalue of the matrix $P_\theta \stackrel{\triangle}{=} [p_{ij}e^{\theta/\lambda_j}]$ which in our case is

$$P = \begin{bmatrix} 0.8e^{\theta/5} & 0.2e^{\theta/10} \\ 0.3e^{\theta/5} & 0.7e^{\theta/10} \end{bmatrix}. \tag{2.106}$$

We performed several calculations using the software package *Matlab*. For the tail probability of the waiting time in the first queue we found that $\theta_1^* = -9.47$. We calculated the large deviations rate functions $\Lambda_A^{*-}(a)$ and $\Lambda_D^{*-}(a)$ for the arrival process

and the departure process from the first queue, respectively. The results appear in Figure 2-11. To calculate $\Lambda_D^{*-}(a)$ we used Eq. (2.60). It can be seen that the first

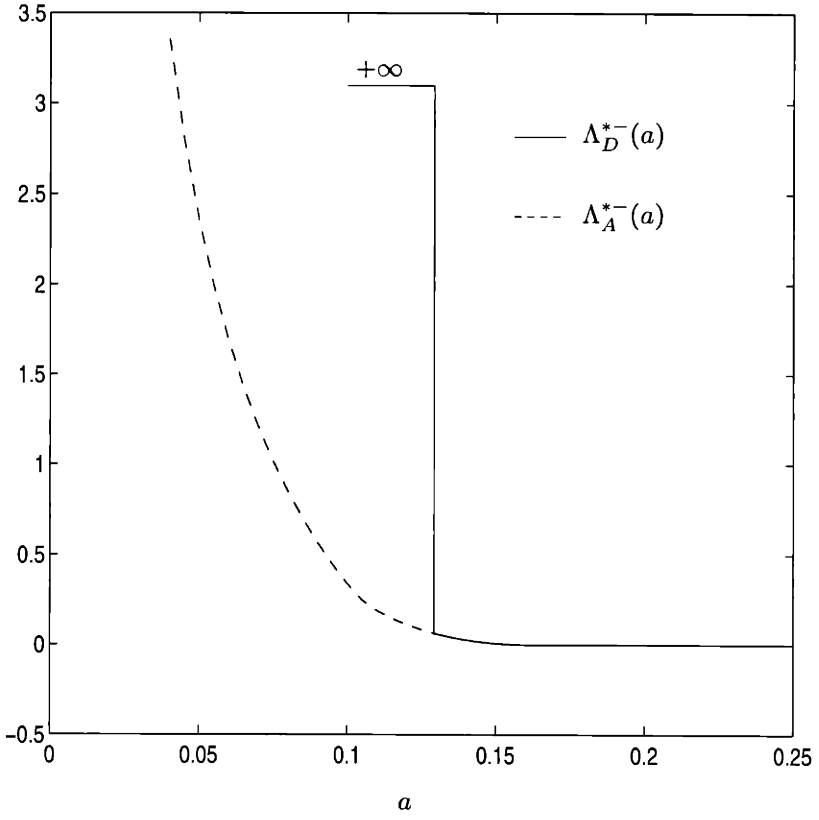

Figure 2-11: $\Lambda_A^{*-}(a)$ and $\Lambda_D^{*-}(a)$ for the numerical example.

queue has a smoothing effect on the arrival process. In other words, the departure process deviates from its mean with smaller probability than the arrival process does. We also found that $\Lambda_D(\theta) + \Lambda_{B^2}(-\theta)$ is strictly negative for all $\theta < 0$, so as it can be seen from the proof of Theorem 2.2.1 that we have $\theta_2^* = -\infty$, which means that w.p.1 a large queue does not built up in the second queue. Finally, we found that the departure process D_2 from the second queue has large deviations rate function $\Lambda_{D_2}^{*-}(a)$ equal to $\Lambda_D^{*-}(a)$. This, can also be seen analytically. Namely, observe that in

Eq. (2.59) we have $\Lambda_{\Gamma}^{*-}(a) = \Lambda_{D}^{*-}(a)$ which implies $\Lambda_{D_2}^{*-}(a) = \Lambda_{D}^{*-}(a)$.

Chapter 3

Overflow Probabilities with GPS

In this chapter we switch gears and consider a multiclass model. The motivation is to capture the interaction between distinct types (classes) of traffic in future high-speed networks and to understand how they can share common network resources in a way that congestion stays within desired limits. We therefore consider a multiclass multiplexer (switch) with two types of traffic and a distinct buffer assigned to each traffic class. We estimate the buffer overflow probability in each of the buffers, when the multiplexer is operated under the *generalized processor sharing policy (GPS)*.

Regarding the structure of this chapter, we begin in Section 3.1 by formally defining the multiclass model that we consider and by stating a set of assumptions that arrival and service processes need to conform to. In Section 3.2 we formally define the GPS policy and we provide an outline of the methodology that we follow in proving our results. In Section 3.3 we prove a lower bound on the overflow probability and in Section 3.4 we introduce the optimal control formulation and solve the control problem. In Section 3.5 we summarize the most likely modes of overflow obtained from the solution of the control problem and in Section 3.6 we prove the matching upper bound. We gather our main results in Section 3.7, where we also treat the special case of strict priority policies.

3.1 A Multiclass Model

In this section we introduce a multiclass multiplexer model that we plan to analyze, in the large deviations regime.

Consider the system depicted in Figure 3-1. We assume a slotted time model (i.e.,

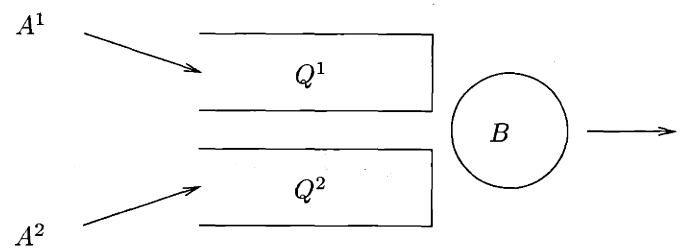

Figure 3-1: A multiclass model.

discrete time) and we let A_i^1 (resp. A_i^2), $i \in \mathbb{Z}$, denote the number of type 1 (resp. 2) customers that enter queue Q^1 (resp. Q^2) at time i. Both queues have infinite buffers and share the same server which can process B_i customers during the time interval $[i, i+1]$. We assume that the processes $\{A_i^1;\ i \in \mathbb{Z}\}$, $\{A_i^2;\ i \in \mathbb{Z}\}$ and $\{B_i;\ i \in \mathbb{Z}\}$ are stationary and mutually independent. However, we allow dependencies between the number of customers at different slots in each process.

We denote by L_i^1 and L_i^2, the queue lengths at time i (without counting arrivals at time i) in queues Q^1 and Q^2, respectively. We assume that the server allocates its capacity between queues Q^1 and Q^2 according to a work-conserving policy (i.e., the server never stays idle when there is work in the system). We also assume that the queue length processes $\{L_i^j, j = 1, 2, i \in \mathbb{Z}\}$ are stationary (under a work-conserving policy, the system reaches steady-state due to the stability condition (3.1) by assuming ergodicity for the arrival and service processes).

To simplify the analysis and avoid integrality issues we assume a "fluid" model, meaning that we will be treating A_i^1, A_i^2 and B_i as real numbers (the amount of fluid

entering or being served). This will not change the results in the large deviations regime.

For stability purposes we assume that for all i

$$\mathbf{E}[B_i] > \mathbf{E}[A_i^1] + \mathbf{E}[A_i^2]. \tag{3.1}$$

We further assume that the arrival and service processes satisfy a LDP (Assumption A), as well as Assumptions D and E. As we have noted in Section 1.3, these assumptions are satisfied by processes that are commonly used to model bursty traffic in communication networks, e.g., renewal processes, Markov-modulated processes and more generally stationary processes with mild mixing conditions.

3.2 The GPS policy

In this section we introduce the *generalized processor sharing* (GPS) policy that was proposed in [DKS90] and further explored in [PG93, PG94]. According to this policy the server allocates a fraction $\phi_1 \in [0,1]$ of its capacity to queue Q^1, and the remaining fraction $\phi_2 = 1 - \phi_1$ to queue Q^2. The policy is defined to be work-conserving, which implies that one of the queues, say queue Q^1, may get more than a fraction ϕ_1 of the server's capacity during times that the other queue, Q^2, is empty. More formally, we can define the GPS to be the policy that satisfies (work-conservation)

$$L_{i+1}^1 + L_{i+1}^2 = [L_i^1 + L_i^2 + A_i^1 + A_i^2 - B_i]^+,$$

and

$$L_{i+1}^j \leq [L_i^j + A_i^j - \phi_j B_i]^+, \qquad j = 1, 2,$$

where $[x]^+ \triangleq \max\{x, 0\}$.

We are interested in estimating the overflow probability $\mathbf{P}[L_i^1 > U]$ for large values of U, at an arbitrary time slot i, in steady-state. Having determined this, the overflow probability of the second queue can be obtained by a symmetrical argument.

We will prove that the overflow probability satisfies

$$\mathbf{P}[L_i^1 > U] \sim e^{-U\theta^*_{GPS}}, \tag{3.2}$$

asymptotically, as $U \to \infty$. To this end, we will develop a lower bound on the overflow probability, along with a matching upper bound. Consider all scenarios (paths) that lead to an overflow. We will show that the probability of each such scenario ω asymptotically behaves as $e^{-U\theta(\omega)}$, for some function $\theta(\omega)$. This probability is a lower bound on $\mathbf{P}[L_i^1 > U]$ for all ω. We select the tightest lower bound by performing the minimization $\theta^*_{GPS} = \min_\omega \theta(\omega)$, by solving a deterministic optimal control problem. Optimal trajectories (paths) of the control problem correspond to *most likely* overflow scenarios. We show that these must be of one out of two possible types. In other words, with high probability, overflow occurs in one out of two possible modes. We will obtain an upper bound on $\mathbf{P}[L_i^1 > U]$ by first obtaining a sample path upper bound, i.e., $L_i^1 \leq \tilde{L}_i^1$ (which implies $\mathbf{P}[L_i^1 > U] \leq \mathbf{P}[\tilde{L}_i^1 > U]$) and establishing that $\mathbf{P}[\tilde{L}_i^1 > U]$ is at most $e^{-U\theta^*_{GPS}}$.

3.3 A Lower Bound

In this section we establish a lower bound on the overflow probability $\mathbf{P}[L_i^1 > U]$.

Proposition 3.3.1 ***(GPS Lower Bound)*** *Assuming that the arrival and service processes satisfy Assumptions A and D, and under the GPS policy, the steady-state queue length L^1 of queue Q^1 satisfies*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \geq -\theta^*_{GPS}, \tag{3.3}$$

where θ^*_{GPS} is given by

$$\theta^*_{GPS} = \min\left[\inf_{a>0} \frac{1}{a}\Lambda^{I*}_{GPS}(a), \inf_{a>0} \frac{1}{a}\Lambda^{II*}_{GPS}(a)\right], \tag{3.4}$$

and the functions $\Lambda^{I*}_{GPS}(\cdot)$ *and* $\Lambda^{II*}_{GPS}(\cdot)$ *are defined as follows*

$$\Lambda^{I*}_{GPS}(a) \stackrel{\Delta}{=} \inf_{\substack{x_1+x_2-x_3=a \\ x_2 \le \phi_2 x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)], \tag{3.5}$$

and

$$\Lambda^{II*}_{GPS}(a) \stackrel{\Delta}{=} \inf_{\substack{x_1-\phi_1 x_3=a \\ x_2 \ge \phi_2 x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \tag{3.6}$$

Proof : Let $-n \le 0$ and $a > 0$. Fix $x_1, x_2, x_3 \ge 0$ and $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ and consider the event

$$\{ |S^{A^1}_{-n,-i-1} - (n-i)x_1| \le \epsilon_1 n,\ |S^{A^2}_{-n,-i-1} - (n-i)x_2| \le \epsilon_2 n,$$
$$|S^{B}_{-n,-i-1} - (n-i)x_3| \le \epsilon_3 n,\ i = 0, 1, \ldots, n-1\}.$$

Notice that x_1, x_2 (resp. x_3) have the interpretation of empirical arrival (resp. service) rates during the interval $[-n, -1]$. We focus on two particular scenarios

$$\underline{\text{Scenario 1}}\text{:}\quad \begin{array}{l} x_1 + x_2 - x_3 = a \\ x_2 \le \phi_2 x_3 \end{array} \qquad \underline{\text{Scenario 2}}\text{:}\quad \begin{array}{l} x_1 - \phi_1 x_3 = a \\ x_2 \ge \phi_2 x_3. \end{array} \tag{3.7}$$

Under Scenario 1, the first queue receives the maximum capacity (at a rate of $x_3 - x_2$) when the second queue stays always empty during the interval $[-n, 0]$. Thus, $L^1_0 \ge na - n\epsilon'_1$, where $\epsilon'_1 \to 0$ as $\epsilon_1, \epsilon_2, \epsilon_3 \to 0$. Similarly, under Scenario 2, the second queue is almost always backlogged during the interval $[-n, 0]$, and the first queue gets capacity roughly $\phi_1 x_3$, implying also $L^1_0 \ge na - n\epsilon'_2$, where $\epsilon'_2 \to 0$ as $\epsilon_1, \epsilon_2, \epsilon_3 \to 0$.

Now, the probability of Scenario 1 is a lower bound on $\mathbf{P}[L^1_0 \ge n(a - \epsilon'_1)]$. Calculating the probability of Scenario 1, maximizing over x_1, x_2 and x_3, to obtain the tightest bound, and using Assumption D we have

$$\mathbf{P}[L^1_0 \ge n(a-\epsilon'_1)] \ge \sup_{\substack{x_1+x_2-x_3=a \\ x_2 \le \phi_2 x_3}} \mathbf{P}[\ |S^{A^1}_{-n,-i-1} - (n-i)x_1| \le \epsilon_1 n,\ i = 0, 1, \ldots, n-1]$$

$$\times \mathbf{P}[\ |S^{A^2}_{-n,-i-1} - (n-i)x_2| \leq \epsilon_2 n,\ i = 0, 1, \ldots, n-1]$$
$$\times \mathbf{P}[\ |S^{B}_{-n,-i-1} - (n-i)x_3| \leq \epsilon_3 n,\ i = 0, 1, \ldots, n-1]$$
$$\geq \exp\left\{-n\left(\inf_{\substack{x_1+x_2-x_3=a \\ x_2 \leq \phi_2 x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)] + \epsilon\right)\right\}$$
$$= \exp\{-n(\Lambda^{I*}_{GPS}(a) + \epsilon)\}, \tag{3.8}$$

where n is large enough, and $\epsilon, \epsilon'_1 \to 0$ as $\epsilon_1, \epsilon_2, \epsilon_3 \to 0$.

Similarly, calculating the probability of Scenario 2, we obtain

$$\mathbf{P}[L^1_0 \geq n(a - \epsilon'_2)] \geq \exp\{-n(\Lambda^{II*}_{GPS}(a) + \epsilon')\}, \tag{3.9}$$

for n large enough, and with $\epsilon', \epsilon'_2 \to 0$ as $\epsilon_1, \epsilon_2, \epsilon_3 \to 0$.

Combining Eqs. (3.8) and (3.9) (taking the limit of all ϵ's going to zero) we obtain

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[L^1_0 \geq na] \geq -\min(\Lambda^{I*}_{GPS}(a), \Lambda^{II*}_{GPS}(a)). \tag{3.10}$$

As a final step to this proof, by letting $U = na$, we obtain

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] = \lim_{n\to\infty} \frac{1}{na} \log \mathbf{P}[L^1_0 \geq na] \geq -\frac{1}{a} \min(\Lambda^{I*}_{GPS}(a), \Lambda^{II*}_{GPS}(a)).$$

Since a, in the above, is arbitrary we can select it properly to make the bound tighter. Namely,

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \geq -\min\left[\inf_{a>0} \frac{1}{a} \Lambda^{I*}_{GPS}(a), \inf_{a>0} \frac{1}{a} \Lambda^{II*}_{GPS}(a)\right].$$

■

3.4 The optimal control problem

In this section we introduce an optimal control problem and show that θ^*_{GPS} is its optimal value.

To motivate the control problem, we relate it, heuristically, with the problem of obtaining an asymptotically tight estimate of the overflow probability [1]. For every overflow sample path, leading to $L^1_0 > U$, there exists some time $-n \leq 0$ that both queues are empty. Since we are interested in the asymptotics as $U \to \infty$, we scale time and the levels of the processes A^1, A^2 and B by U. We then let $T = \frac{n}{U}$ and define the following continuous-time functions in $D[-T, 0]$ (these are right-continuous functions with left-limits):

$$L^j(t) = \frac{1}{U} L^j_{\lfloor Ut \rfloor}, \; j = 1, 2, \quad S^X(t) = \frac{1}{U} S^X_{-UT, \lfloor Ut \rfloor}, \; X \in \{A^1, A^2, B\}, \quad \text{for } t \in [-T, 0].$$

Notice that the empirical rate of a process X is roughly equal to the rate of growth of $S^X(t)$. More formally, we will say that a process X has empirical rate $x(t)$ in the interval $[-T, 0]$ if for large U and small $\epsilon > 0$ it is true

$$|S^X(t) - \int_{-T}^{t} x(\tau)\, d\tau| < \epsilon, \qquad \forall t \in [-T, 0],$$

where $x(t)$ are arbitrary non-negative functions. We let, $x_1(t), x_2(t)$ and $x_3(t)$ denote the empirical rates of the processes A^1, A^2 and B, respectively. The probability of sustaining rates $x_1(t), x_2(t)$ and $x_3(t)$, in the interval $[-UT, 0]$ for large values of U is given (up to first degree in the exponent) by

$$\exp\left\{-U \int_{-T}^{0} [\Lambda^*_{A^1}(x_1(t)) + \Lambda^*_{A^2}(x_2(t)) + \Lambda^*_B(x_3(t))]\, dt\right\}.$$

This cost functional is a consequence of Assumption D. With the scaling introduced here as $U \to \infty$ the sequence of slopes $a_0, a_1, \ldots, a_{m-1}$ appearing there converges

[1] Such a relation can be rigorously established using the sample path LDP for the arrival and service processes, as it is defined in [DZ93a] and [Cha94a].

to the empirical rate $x(\cdot)$ and the sum of rate functions appearing in the exponent converges to an integral.

We seek a path with maximum probability, i.e., a minimum cost path where the cost functional is given by the integral in the above expression. This optimization is subject to the constraints $L^1(-T) = L^2(-T) = 0$ and $L^1(0) = 1$. The fluid levels in the two queues $L^1(t)$ and $L^2(t)$ are the state variables and the empirical rates $x_1(t), x_2(t)$ and $x_3(t)$ are the control variables. The dynamics of the system depend on the state. We distinguish three regions:

Region A: $L^1(t), L^2(t) > 0$, where according to the GPS policy

$$\dot{L}^1 = x_1(t) - \phi_1 x_3(t) \quad \text{and} \quad \dot{L}^2 = x_2(t) - \phi_2 x_3(t),$$

Region B: $L^1(t) = 0, L^2(t) > 0$, where according to the GPS policy

$$\dot{L}^2 = x_1(t) + x_2(t) - x_3(t),$$

Region C: $L^1(t) > 0, L^2(t) = 0$, where according to the GPS policy

$$\dot{L}^1 = x_1(t) + x_2(t) - x_3(t).$$

Dotted variables in the above expressions denote derivatives [2]. Let (GPS-DYNAMICS) denote the set of state trajectories $L^j(t)$, $j = 1, 2$, $t \in [-T, 0]$, that obey the dynamics given above.

Motivated by this discussion we now formally define the following optimal control problem (GPS-OVERFLOW). The control variables are $x_j(t)$, $j = 1, 2, 3$, and the state variables are $L^j(t)$, $j = 1, 2$, for $t \in [-T, 0]$, which obey the dynamics given in

[2]Here we use the notion of derivative for simplicity of the exposition. Note that these derivatives may not exist everywhere. Thus, in Region B for example, the rigorous version of the statement $\dot{L}^2 = x_1(t) + x_2(t) - x_3(t)$ is $L^2(t_2) = L^2(t_1) + \int_{t_1}^{t_2} (x_1(t) + x_2(t) - x_3(t))\, dt$, for all intervals (t_1, t_2) that the system remains in Region B.

the previous paragraph.

$$\text{(GPS-OVERFLOW) minimize } \int_{-T}^{0} [\Lambda^*_{A^1}(x_1(t)) + \Lambda^*_{A^2}(x_2(t)) + \Lambda^*_B(x_3(t))]\, dt \tag{3.11}$$

$$\begin{aligned} \text{subject to: } & L^1(-T) = L^2(-T) = 0 \\ & L^1(0) = 1 \\ & L^2(0) : \text{ free} \\ & T : \text{ free} \\ & \{L^j(t) : \ t \in [-T, 0],\ j = 1, 2\} \in \text{(GPS-DYNAMICS)}. \end{aligned}$$

The first property of (GPS-OVERFLOW) that we show is that *optimal control trajectories can be taken to be constant* within each of the three regions. The result is established in the next lemma, where only Region A is considered in the proof. The other regions can be treated similarly.

Lemma 3.4.1 *Fix a time interval* $[-T_1, -T_2]$. *Consider a segment of a control trajectory* $\{x_1(t), x_2(t), x_3(t);\ t \in [-T_1, -T_2]\}$, *achieving cost* V, *such that the corresponding state trajectory* $\{L^1(t), L^2(t);\ t \in (-T_1, -T_2)\}$ *stays in one of the regions A, B, or C. Then there exist scalars* $\bar{x}_1$, $\bar{x}_2$ *and* $\bar{x}_3$ *such that the segment of the control trajectory* $\{x_1(t) = \bar{x}_1, x_2(t) = \bar{x}_2, x_3(t) = \bar{x}_3;\ t \in [-T_1, -T_2]\}$ *achieves cost at most* V, *with the same corresponding state trajectory* $\{L^1(t), L^2(t);\ t \in (-T_1, -T_2)\}$.

Proof : Consider a segment of any arbitrary control trajectory $\{x_1(t), x_2(t), x_3(t);\ t \in [-T_1, -T_2]\}$, that satisfies

$$\begin{aligned} L^1(-T_1) = a_1 > 0, \qquad & L^1(-T_2) = b_1 > 0, \\ L^2(-T_1) = a_2 > 0, \qquad & L^2(-T_2) = b_2 > 0, \end{aligned} \tag{3.12}$$

and stays in Region A, i.e., $L^1(t), L^2(t) > 0$ for all $t \in (-T_1, -T_2)$. We will prove that

the time-average control trajectory

$$\bar{x}_i(\tau) = \frac{1}{T_1 - T_2} \int_{-T_1}^{-T_2} x_i(t)\, dt, \qquad i = 1, 2, 3,\ \forall \tau \in [-T_1, -T_2], \tag{3.13}$$

is no more costly. To this end, notice that to stay in Region A, the state variables have to be positive, which by the system dynamics implies

$$L^j(t) = a_j + \int_{-T_1}^{t} [x_j(\tau) - \phi_j x_3(\tau)] > 0, \qquad j = 1, 2,\ t \in (-T_1, -T_2). \tag{3.14}$$

Moreover, we also have

$$L^j(-T_2) = a_j + \int_{-T_1}^{-T_2} [x_j(\tau) - \phi_j x_3(\tau)] = b_j, \qquad j = 1, 2. \tag{3.15}$$

Notice now that the time-average trajectory, has the same end points (i.e., satisfies (3.12)), moves along a straight line and thus stays in Region A for $t \in (-T_1, -T_2)$. Moreover, by convexity of the rate functions we have

$$\int_{-T_1}^{-T_2} [\Lambda^*_{A^1}(x_1(t)) + \Lambda^*_{A^2}(x_2(t)) + \Lambda^*_B(x_3(t))]\, dt \geq (T_1 - T_2)[\Lambda^*_{A^1}(\bar{x}_1) + \Lambda^*_{A^2}(\bar{x}_2) + \Lambda^*_B(\bar{x}_3)].$$

■

Given this property, to solve (GPS-OVERFLOW) it suffices to restrict ourselves to state trajectories with constant control variables in each of the regions A, B and C. A trajectory is called optimal if it achieves the lowest cost among all trajectories with the same initial and final state. Since we have a free time problem, any segment of an optimal trajectory is also optimal.

Consider now a control trajectory $\{x_i^L(t);\ t \in [-T, 0]\}$ with corresponding state trajectory $\{L^1(t), L^2(t);\ t \in [-T, 0]\}$, which leads to a final state $(L^1(0), L^2(0))$. Define a scaled trajectory as

$$x_i^Q(t) = x_i^L(t/\alpha), \qquad i = 1, 2, 3,\ t \in [-\alpha T, 0],$$

$$Q^j(t) = \alpha L^j(t/\alpha), \qquad j = 1, 2,\ t \in [-\alpha T, 0],$$

and note that it leads to the final state $(\alpha L^1(0), \alpha L^2(0))$. Then, the cost of the Q trajectory is given by

$$\int_{-\alpha T}^{0} [\Lambda^*_{A^1}(x_1^Q(t)) + \Lambda^*_{A^2}(x_2^Q(t)) + \Lambda^*_B(x_3^Q(t))]\ dt = \alpha \int_{-T}^{0} [\Lambda^*_{A^1}(x_1^L(t)) + \Lambda^*_{A^2}(x_2^L(t)) + \Lambda^*_B(x_3^L(t))]\ dt.$$

Using this observation, it follows easily that every scaled version of an optimal trajectory is optimal for the corresponding terminal state. Given this *homogeneity* property we can compare the state trajectories in Figure 3-2(a), (b) and (c). If the trajectory in Figure 3-2(a) is optimal then so does the scaled version (by $\alpha = a_2/a_1$) in Figure 3-2(b) and as consequence its segment which appears in Figure 3-2(c) is also optimal (since we have a free time problem).

Using the homogeneity property we can make the reduction in Figure 3-2(e), starting from any arbitrary trajectory with constant controls as the one appearing in Figure 3-2(d) (by appropriately scaling the dashed segment). Therefore, we conclude that optimal state trajectories which have $L^1(t) = 0$ for some initial segment can be restricted to have one of the forms depicted in Figure 3-3(c) and (d). Similarly, optimal state trajectories which have $L^1(t) > 0$ for some initial segment can be restricted to have one of the forms depicted in Figure 3-3(a) and (b). Consider now the trajectories in Figure 3-3(c) and (c'). The segment of (c) and (c') that is in Region A has the same slope, thus the same controls, which implies that the trajectory in (c') is at least as cheap since it spends less time on the L^2 axis. Hence, we have reduced the candidates for optimal trajectories to the ones in Figure 3-3 (a), (b) and (d).

Finally, consider the state trajectory in Figure 3-3(d). Assume, without loss of generality that it spends a ζ fraction of its total time T on the L^2 axis (Region B) and the remaining $1 - \zeta$ fraction in Region A. Let also, $\{x_j;\ j = 1, 2, 3, \}$ be the controls in Region B and $\{y_j;\ j = 1, 2, 3, \}$ the controls in Region A. The feasibility

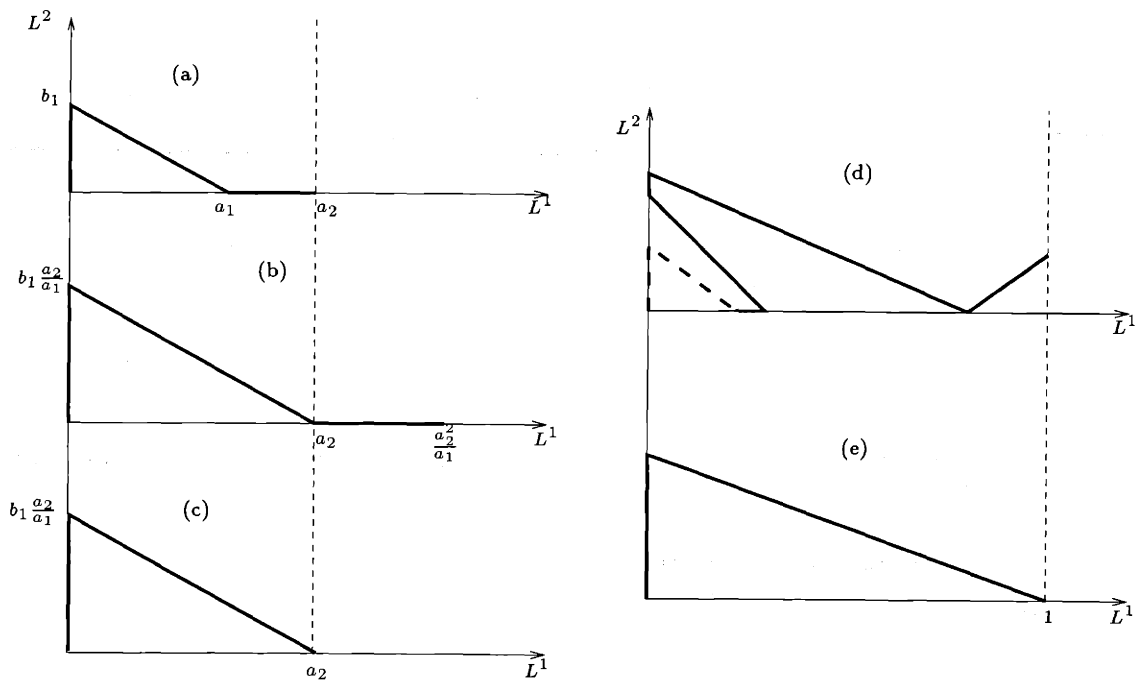

Figure 3-2: By the homogeneity property, optimality of the trajectory in (a) implies optimality of the trajectory in (b) which by its turn implies optimality of the trajectory in (c). Using the homogeneity property the trajectory in (d) reduces to the one in (e).

constraints are

$$x_1 \leq \phi_1 x_3,$$
$$\zeta T(x_1 + x_2 - x_3) + (1-\zeta)T(y_2 - \phi_2 y_3) = 0,$$
$$(1-\zeta)T(y_1 - \phi_1 y_3) = 1.$$

Note that the time average control over x_2, y_2, i.e., $\bar{x}_2 = \zeta x_2 + (1-\zeta)y_2$, satisfies the same feasibility constraints and therefore by convexity (using the argument in the proof of Lemma 3.4.1) it is at least as profitable to have $x_2 = y_2 = \bar{x}_2$. Now, to have

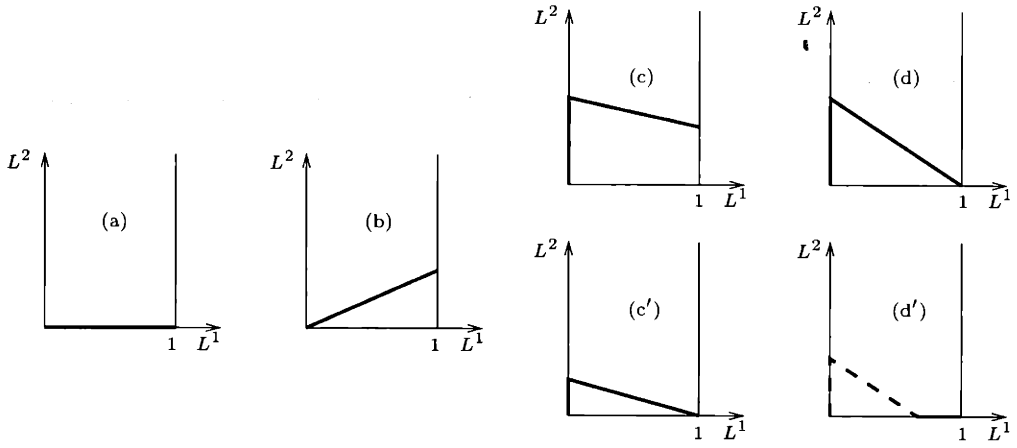

Figure 3-3: Candidates for optimal state trajectories are depicted in (a), (b), (c) and (d). The trajectory in (c) is reduced to the one in (c') which has the same form as the one in (d). The trajectory in (d) is reduced to the one in (d') which is contradicted by the time-homogeneity property. Hence, optimal state trajectories have only the form in (a) and (b).

the trajectory in Figure 3-3(d), it has to be the case (since $x_2 > \phi_2 x_3$ and $x_2 < \phi_2 y_3$)

$$\bar{x}_2 > \phi_2 x_3,$$
$$\bar{x}_2 < \phi_2 y_3.$$

Consider the trajectory with $x_3' = x_3 + \frac{\epsilon}{\zeta}$ and $y_3' = y_3 - \frac{\epsilon}{1-\zeta}$ for some small $\epsilon > 0$. This latter trajectory serves the same total number of customers as the former in the interval $[-T, 0]$ (equal to $\zeta T x_3 + (1-\zeta) T y_3$) and it is at least as cheap by convexity of the rate functions. It is depicted in Figure 3-3(d'). We can now apply the same argument to its dashed segment. If we keep doing that we conclude that the trajectory in Figure 3-3(a) is at least as cheap.

Therefore, optimal state trajectories of (GPS-OVERFLOW) can be restricted to have one of the forms depicted in Figure 3-3(a) and (b). We next calculate the optimal

value of (GPS-OVERFLOW). The best trajectory of the form shown in Figure 3-3(a) has value

$$\inf_T \inf_{\substack{x_1+x_2-x_3=\frac{1}{T} \\ x_2 \le \phi_2 x_3}} T[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)], \tag{3.16}$$

which is equal to $\inf_T[T\Lambda^{I*}_{GPS}(1/T)]$ by the definition in (3.5). The best trajectory of the form shown in Figure 3-3(b) has value

$$\inf_T \inf_{\substack{x_1-\phi_1 x_3=\frac{1}{T} \\ x_2 \ge \phi_2 x_3}} T[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)], \tag{3.17}$$

which is equal to $\inf_T[T\Lambda^{II*}_{GPS}(1/T)]$ by the definition in (3.6). Thus, the optimal value of (GPS-OVERFLOW) is equal to the minimum of the two expressions above which is identical to θ^*_{GPS} as it is defined in (3.4). In summary we have established the following:

Theorem 3.4.2 *The optimal value of the problem (GPS-OVERFLOW) is given by θ^*_{GPS}, as it is defined in (3.4).*

It is of interest to investigate under what conditions on the parameters of the arrival and service processes the trajectory in Figure 3-3(a) dominates the one in (b) and vice versa. We will distinguish two cases: $\mathbf{E}[A^2] \ge \phi_2\mathbf{E}[B]$ and $\mathbf{E}[A^2] < \phi_2\mathbf{E}[B]$, where for $j = 1, 2$, $\mathbf{E}[A^j]$ (resp. $\mathbf{E}[B]$) denote the expected number of customers arriving from stream j (resp. expected potential number of departures). In the first case we will establish that the trajectory in Figure 3-3(b) dominates the one in (a). In the second case, however, the relationship between expectations is not sufficient to discard one of the two trajectories and which one dominates depends on the distribution of the arrival and service processes. The following theorem describes the result.

Theorem 3.4.3 *If $\mathbf{E}[A^2] \ge \phi_2\mathbf{E}[B]$ then optimal state trajectories of (GPS-OVERFLOW) can be restricted to have the form in Figure 3-3(b) with optimal*

value

$$\inf_{T} \inf_{x_1 - \phi_1 x_3 = \frac{1}{T}} T[\Lambda^*_{A^1}(x_1) + \Lambda^*_B(x_3)].$$

Proof : Assume $\mathbf{E}[A^2] \geq \phi_2 \mathbf{E}[B]$ and consider the state trajectory in Figure 3-3(a) which has optimal value given by the expression in (3.16). Since $x_2 \leq \phi_2 x_3$, either $x_2 \leq \mathbf{E}[A^2]$ or $x_3 \geq \mathbf{E}[B]$. Then we can increase x_2 and decrease x_3 until $x_2 = \phi_2 x_3$, making $x_1 + x_2 - x_3 \geq \frac{1}{T}$. The segment of this trajectory with terminal point at $L^1 = \frac{1}{T}$ has the form of the state trajectory in Figure 3-3(b). Thus we have reduced optimal state trajectories to Figure 3-3(b). To determine the optimal value, notice that if $x_3 > \mathbf{E}[B]$ we can decrease x_3 to $\mathbf{E}[B]$, without violating the constraint $x_2 \geq \phi_2 x_3$, making $x_1 - \phi_1 x_3 \geq \frac{1}{T}$, and keeping the segment of the resulting trajectory with terminal point at $L^1 = \frac{1}{T}$. Thus, it has to be the case $x_3 \leq \mathbf{E}[B]$. Then we can actually fix x_2 to $\mathbf{E}[A^2]$, without violating the constraint $x_2 \geq \phi_2 x_3$ (since $x_2 = \mathbf{E}[A^2] \geq \phi_2 \mathbf{E}[B] \geq \phi_2 x_3$). This proves that the optimal value is given by the expression appearing in the statement of this theorem.

■

3.5 The most likely paths

As we have explained in the Section 3.2 we will prove an upper bound that matches the lower bound in Proposition 3.3.1. This is sufficient to guarantee that the two scenarios identified in the proof of Proposition 3.3.1 (or equivalently the two optimal state trajectories of (GPS-OVERFLOW)) are two generic ways that queue Q^1 overflows. We summarize here these two modes of overflow.

In particular, we distinguish two cases:

Case 1: Suppose $\theta^*_{GPS} = \inf_a \Lambda^{I*}_{GPS}(a)/a$ holds. Let $a^* > 0$ the optimal solution of this optimization problem. In this case, the first queue is building up to an $O(U)$ level while the second queue stays at an $o(U)$ level. The first queue builds up linearly with rate a^*, during a period with duration U/a^*. During

this period the empirical rates of the processes A^1, A^2 and B, are roughly equal to the optimal solution (x_1^*, x_2^*, x_3^*), respectively, of the optimization problem appearing in the definition of $\Lambda_{GPS}^{I*}(a^*)$ (Eq. (3.5)). The trajectory in L^1-L^2 space is depicted in Figure 3-3(a).

Case 2: Suppose $\theta_{GPS}^* = \inf_a \Lambda_{GPS}^{II*}(a)/a$ holds. Let $a^* > 0$ the optimal solution of this optimization problem. In this case, both queues are building up to an $O(U)$ level. The first queue builds up linearly with rate a^*, during a period with duration U/a^*. During this period the empirical rates of the processes A^1, A^2 and B, are roughly equal to the optimal solution (x_1^*, x_2^*, x_3^*), respectively, of the optimization problem appearing in the definition of $\Lambda_{GPS}^{II*}(a^*)$ (Eq. (3.6)). The trajectory in L^1-L^2 space is depicted in Figure 3-3(b).

It is interesting to reflect at this point on the implications of this result on admission control for ATM multiplexers operating under the GPS policy. Consider the admission control mechanism for queue Q^1 and suppose that the objective of this mechanism is to keep the overflow probability below a given desirable threshold. A worst-case analysis as in [PG93] would conclude that the admission control mechanism has to be designed with the assumption that the second queue always uses a fraction ϕ_2 of the service capacity. If instead the results of this chapter are used (assuming that a detailed statistical model of the input traffic streams is available) a statistical multiplexing gain can be realized. In the overflow mode described in Case 1 above, the second queue consumes less than the fraction ϕ_2 of the total service capacity, implying that more Type 1 connections can be allowed without compromising the quality of service. Even if the overflow mode described in Case 2 above prevails, the overflow probability is explicitly calculated (in an exponential scale) and can be taken into account in the design of the admission control mechanism.

3.6 An Upper Bound

In this section we develop an upper bound on the probability $\mathbf{P}[L_0^1 > U]$. In particular, we will prove that as $U \to \infty$ we have $\mathbf{P}[L_0^1 > U] \leq e^{-\theta^*_{GPS}U+o(U)}$, where $o(U)$ denotes functions with the property $\lim_{U\to\infty} \frac{o(U)}{U} = 0$.

In proving the upper bound we will distinguish two cases:

Case 1. $\mathbf{E}[A^2] < \phi_2\mathbf{E}[B]$.

Case 2. $\mathbf{E}[A^2] \geq \phi_2\mathbf{E}[B]$.

3.6.1 Upper Bound: Case 2

We will first establish the proof for Case 2, which is easier.

We consider a busy period of the first queue, Q^1, that starts at some time $-n^* \leq 0$ ($L^1_{-n^*} = 0$) and has not ended until time 0. Notice that due to the stability condition (3.1) and the fact $\mathbf{E}[A^2] \geq \phi_2\mathbf{E}[B]$, it is true that $\mathbf{E}[A^1] < \phi_1\mathbf{E}[B]$, which implies that such a time $-n^*$ always exists. We will focus on sample paths of the system in $[-n^*, 0]$ that lead to $L_0^1 > U$. Note that

$$L_0^1 \leq S^{A^1}_{-n^*,-1} - \phi_1 S^B_{-n^*,-1}. \tag{3.18}$$

Thus,

$$\begin{aligned}\mathbf{P}[L_0^1 > U] \leq& \mathbf{P}[\exists n \geq 0 \text{ s.t. } S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1} > U] \\ \leq& \mathbf{P}[\max_{n\geq 0}(S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1}) > U].\end{aligned} \tag{3.19}$$

We next upper bound the moment generating function of $\max_{n\geq 0}(S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1})$. Applying the LDP for the arrival and service processes for $\theta \geq 0$ we can obtain

$$\mathbf{E}[e^{\theta \max_{n\geq 0}(S^{A^1}_{-n,-1}-\phi_1 S^B_{-n,-1})}] \leq \sum_{n\geq 0} \mathbf{E}[e^{\theta(S^{A^1}_{-n,-1}-\phi_1 S^B_{-n,-1})}]$$

$$\begin{aligned}
&\leq \sum_{n\geq 0} e^{n(\Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)+\epsilon)} \\
&\leq K(\theta,\epsilon) \qquad \text{if } \Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)<0,
\end{aligned} \tag{3.20}$$

since when the exponent is negative (for sufficiently small ϵ), the infinite geometric series converges to a constant, with respect to n, $K(\theta,\epsilon)$. We can now apply the Markov inequality in (3.19) to obtain

$$\begin{aligned}
\mathbf{P}[L_0^1 > U] \leq & \mathbf{E}[e^{\theta \max_{n\geq 0}(S^{A^1}_{-n,-1}-\phi_1 S^B_{-n,-1})}]e^{-\theta U} \\
\leq & K(\theta,\epsilon) \qquad \text{if } \Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)<0.
\end{aligned} \tag{3.21}$$

Taking the limit as $U\to\infty$ and minimizing over θ to obtain the tightest bound we establish the following proposition.

Proposition 3.6.1 *If* $\mathbf{E}[A^2] \geq \phi_2\mathbf{E}[B]$ *and assuming an LDP for the arrival and service processes (Assumption A)*

$$\lim_{U\to\infty}\frac{1}{U}\log\mathbf{P}[L_0^1>U] \leq -\sup_{\{\theta\geq 0:\ \Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)<0\}}\theta.$$

We are now left with proving that this upper bound matches the lower bound, θ^*_{GPS}, which in Case 2 is given by the expression in Thm. 3.4.3.

In preparation for this result, consider a convex function $f(u)$ with the property $f(0)=0$. We define the *largest root* of $f(u)$ to be the solution of the optimization problem $\sup_{u:f(u)<0} u$. If $f(\cdot)$ has negative derivative at $u=0$, there are two cases: either $f(\cdot)$ has a single positive root or it stays below the horizontal axis $u=0$, for all $u>0$. In the latter, case we will say that $f(\cdot)$ has a root at $u=\infty$.

Lemma 3.6.2 *For* $\Lambda^*(\cdot)$ *and* $\Lambda(\cdot)$ *being convex duals it holds*

$$\inf_{a>0}\frac{1}{a}\Lambda^*(a)=\theta^*,$$

where θ^* is the largest root of the equation $\Lambda(\theta) = 0$.

Proof :

$$\begin{aligned}\inf_{a>0} \frac{1}{a}\Lambda^*(a) &= \inf_{a>0} \sup_{\theta} \frac{1}{a}[\theta a - \Lambda(\theta)] \\ &= \inf_{a'>0} \sup_{\theta}[\theta - a'\Lambda(\theta)] \\ &= \sup_{\theta:\ \Lambda(\theta)<0} \theta.\end{aligned}$$

In the second equality above, we have made the substitution $a' := \frac{1}{a}$ and in the last one we have used duality.

■

Based on this lemma and Proposition 3.6.1 we establish the following proposition.

Proposition 3.6.3 ***(GPS Upper bound, Case 2)*** *If $\mathbf{E}[A^2] \geq \phi_2 \mathbf{E}[B]$ and assuming that the arrival and service processes satisfy Assumption A, the steady-state queue length, L^1, of queue Q^1, at an arbitrary time slot satisfies*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \leq -\theta^*_{GPS}.$$

Proof : It suffices to prove that $\theta^*_{GPS} = \sup_{\{\theta\geq 0:\ \Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)<0\}} \theta$. Since we are in Case 2, θ^*_{GPS} is given by the expression in Thm. 3.4.3. Due to Lemma 3.6.2 it suffices to prove that $\Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)$ is the convex dual of $\Lambda^*(a) \stackrel{\Delta}{=} \inf_{x_1-\phi_1 x_3=a}[\Lambda^*_{A^1}(x_1) + \Lambda^*_B(x_3)]$. Notice that the latter is a convex function of a as the value function of a convex optimization problem with a appearing only in the right hand side of the constraints. Indeed

$$\begin{aligned}\sup_{a} \sup_{x_1-\phi_1 x_3=a} &[\theta a - \Lambda^*_{A^1}(x_1) - \Lambda^*_B(x_3)] = \\ &= \sup_{x_1,x_3} [\theta(x_1 - \phi_1 x_3) - \Lambda^*_{A^1}(x_1) - \Lambda^*_B(x_3)]\end{aligned}$$

$$=\Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta).$$

■

3.6.2 Upper bound: Case 1

We now proceed to establish the upper bound in Case 1.

Consider all sample paths that lead to $L_0^1 > U$. Looking backwards in time from time 0, let $-k^* \leq 0$ be the first time that $L^1 = 0$. Since the system is busy during the interval $[-k^*, 0]$, the server operates at capacity and

$$L_0^1 \leq L_0^1 + L_0^2 = L_{-k^*}^2 + S_{-k^*,-1}^{A^1} + S_{-k^*,1}^{A^2} - S_{-k^*,1}^{B}. \tag{3.22}$$

Since according to the GPS policy Q^2 gets at least a fraction ϕ_2 of the capacity, we can upper bound $L_{-k^*}^2$ by the queue length at a *virtual system* which gives to Q^2 exactly a ϕ_2 fraction of the capacity (wasting some capacity at times that Q^1 is empty). This trick of using the virtual system to upper bound the queue length in the second queue has been introduced in [dVK95] and used in [Zha95], although the upper bound proofs there do not extend to the general services case. To establish the upper bound we will use the fact that θ_{GPS}^* is the optimal value of (GPS-OVERFLOW). Let $-n^* \leq -k^*$ be the first time (looking backwards in time from $-k^*$) that the queue length of Q^2 becomes zero in the virtual system. Notice that such a time $-n_i^*$ always exists since we are in Case 1, and Q^2 is stable when it gets exactly a fraction ϕ_2 of the capacity. Then

$$L_{-k^*}^2 \leq S_{-n^*,-k^*-1}^{A^2} - \phi_2 S_{-n^*,-k^*-1}^{B}, \tag{3.23}$$

which when combined with (3.22) yields

$$L_0^1 \leq S_{-k^*,-1}^{A^1} + S_{-n^*,-1}^{A^2} - S_{-k^*,-1}^{B} - \phi_2 S_{-n^*,-k^*-1}^{B}. \tag{3.24}$$

Now, since Q^1 is non-empty during the interval $[-k^*, 0]$

$$L_0^1 \leq S^{A^1}_{-k^*,-1} - \phi_1 S^B_{-k^*,-1}. \tag{3.25}$$

We will use the bound in (3.24) when $S^{A^2}_{-n^*,-1} \leq \phi_2 S^B_{-n^*,-1}$ and the bound in (3.25) otherwise. Namely we will use

$$L_0^1 \leq \begin{cases} S^{A^1}_{-k^*,-1} + S^{A^2}_{-n^*,-1} - S^B_{-k^*,-1} - \phi_2 S^B_{-n^*,-k^*-1} & \text{if } S^{A^2}_{-n^*,-1} \leq \phi_2 S^B_{-n^*,-1} \\ S^{A^1}_{-k^*,-1} - \phi_1 S^B_{-k^*,-1} & \text{if } S^{A^2}_{-n^*,-1} \geq \phi_2 S^B_{-n^*,-1}. \end{cases} \tag{3.26}$$

Let Ω_1 the set of sample paths that satisfy $S^{A^2}_{-n^*,-1} \leq \phi_2 S^B_{-n^*,-1}$ and Ω_2 its complement. We have

$$\begin{aligned} &\mathbf{P}[L_0^1 > U \text{ and } \Omega_1] \leq \\ &\leq \mathbf{P}[\exists n \geq k \geq 0 \text{ s.t. } S^{A^2}_{-n,-1} \leq \phi_2 S^B_{-n,-1} \text{ and} \\ &\qquad S^{A^1}_{-k,-1} + S^{A^2}_{-n,-1} - S^B_{-k,-1} - \phi_2 S^B_{-n,-k-1} > U] \\ &\leq \mathbf{P}[\max_{\{n \geq k \geq 0:\ S^{A^2}_{-n,-1} \leq \phi_2 S^B_{-n,-1}\}} (S^{A^1}_{-k,-1} + S^{A^2}_{-n,-1} - S^B_{-k,-1} - \phi_2 S^B_{-n,-k-1}) > U]. \end{aligned} \tag{3.27}$$

For sample paths in Ω_2 we have

$$\begin{aligned} &\mathbf{P}[L_0^1 > U \text{ and } \Omega_2] \leq \\ &\leq \mathbf{P}[\exists n \geq k \geq 0 \text{ s.t. } S^{A^2}_{-n,-1} \geq \phi_2 S^B_{-n,-1} \text{ and } S^{A^1}_{-k,-1} - \phi_1 S^B_{-k,-1} > U] \\ &\leq \mathbf{P}[\max_{\{n \geq k \geq 0:\ S^{A^2}_{-n,-1} \geq \phi_2 S^B_{-n,-1}\}} (S^{A^1}_{-k,-1} - \phi_1 S^B_{-k,-1}) > U]. \end{aligned} \tag{3.28}$$

Let us now define

$$L^I_{GPS,1} \triangleq \max_{\{n \geq k \geq 0:\ S^{A^2}_{-n,-1} \leq \phi_2 S^B_{-n,-1}\}} (S^{A^1}_{-k,-1} + S^{A^2}_{-n,-1} - S^B_{-k,-1} - \phi_2 S^B_{-n,-k-1}),$$

and

$$L^{II}_{GPS,1} \stackrel{\triangle}{=} \max_{\{n\geq k\geq 0:\ S^{A^2}_{-n,-1}\geq \phi_2 S^B_{-n,-1}\}} (S^{A^1}_{-k,-1} - \phi_1 S^B_{-k,-1}),$$

which after bringing the constraints in the objective function become

$$L^{I}_{GPS,1} = \max_{n\geq k\geq 0} \inf_{u\geq 0} [S^{A^1}_{-k,-1} + (1-u)S^{A^2}_{-n,-1} - (1-u\phi_2)S^B_{-k,-1} - \phi_2(1-u)S^B_{-n,-k-1}], \tag{3.29}$$

and

$$L^{II}_{GPS,1} = \max_{n\geq k\geq 0} \inf_{u\geq 0} [S^{A^1}_{-k,-1} + uS^{A^2}_{-n,-1} + (-u\phi_2 - \phi_1)S^B_{-k,-1} - u\phi_2 S^B_{-n,-k-1}]. \tag{3.30}$$

Next we will upper bound the moment generating functions of $L^I_{GPS,1}$ and $L^{II}_{GPS,1}$ by using Assumption E. For the moment generating function of $L^I_{GPS,1}$ and $\theta \geq 0$ we have

$$\begin{aligned}
&\mathbf{E}[e^{\theta L^I_{GPS,1}}] \leq \\
&\leq \sum_{n\geq 0} \sum_{0\leq k\leq n} \inf_{u\geq 0} \mathbf{E}[\exp\{\theta[S^{A^1}_{-k,-1} + (1-u)S^{A^2}_{-n,-1} - (1-u\phi_2)S^B_{-k,-1} \\
&\qquad\qquad - \phi_2(1-u)S^B_{-n,-k-1}]\}] \\
&\leq \sum_{n\geq 0} \sum_{0\leq k\leq n} \inf_{u\geq 0} \exp\{(n-k)[\Lambda_{A^2}(\theta-\theta u) + \Lambda_B(-\theta\phi_2(1-u))] \\
&\qquad\qquad + k[\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta-\theta u) + \Lambda_B(-\theta(1-u\phi_2))] + \Gamma(\theta,u)\} \\
&\leq \sum_{n\geq 0} n \sup_{\zeta\in[0,1]} \inf_{u\geq 0} \exp\{n[\zeta(\Lambda_{A^2}(\theta-\theta u) + \Lambda_B(-\theta\phi_2(1-u))) \\
&\quad + (1-\zeta)(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta-\theta u) + \Lambda_B(-\theta(1-u\phi_2))) + \tfrac{\Gamma(\theta,u)}{n}]\},
\end{aligned} \tag{3.31}$$

where we let $\zeta = \frac{n-k}{n}$. In the second inequality above we have used Assumption E. Let us now define

$$\Lambda^I_{GPS,1}(\theta) \stackrel{\triangle}{=} \sup_{\zeta\in[0,1]} \inf_{u\geq 0} [\zeta(\Lambda_{A^2}(\theta-\theta u) + \Lambda_B(-\theta\phi_2(1-u))) +$$

$$+ (1-\zeta)(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta - \theta u) + \Lambda_B(-\theta(1 - u\phi_2)))].$$

Let $u^*(\theta)$ be the optimal u in the above optimization problem. From (3.31) we have

$$\begin{aligned}
&\mathbf{E}[e^{\theta L^I_{GPS,1}}] \leq \\
&\leq \sum_{n\geq 0} n \sup_{\zeta\in[0,1]} \exp\{n[\zeta(\Lambda_{A^2}(\theta - \theta u^*) + \Lambda_B(-\theta\phi_2(1-u^*))) \\
&\quad + (1-\zeta)(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta - \theta u^*) + \Lambda_B(-\theta(1-u^*\phi_2))) + \tfrac{\Gamma(\theta,u^*)}{n}]\}.
\end{aligned} \tag{3.33}$$

Now for every $\epsilon > 0$ and $\theta \geq 0$ we can take n large enough such that $\frac{\Gamma(\theta,u^*)}{n} < \epsilon$. For sufficiently small ϵ and if $\Lambda^I_{GPS,1}(\theta) < 0$ then the infinite geometric series in the right hand side of (3.33) converges to a constant, with respect to n, $K_1(\theta,\epsilon)$. That is,

$$\mathbf{E}[e^{\theta L^I_{GPS,1}}] \leq K_1(\theta,\epsilon), \qquad \text{if } \Lambda^I_{GPS,1}(\theta) < 0. \tag{3.34}$$

Similarly, for the moment generating function of $L^{II}_{GPS,1}$ and $\theta \geq 0$ we have

$$\begin{aligned}
&\mathbf{E}[e^{\theta L^{II}_{GPS,1}}] \leq \\
&\leq \sum_{n\geq 0}\sum_{0\leq k\leq n} \inf_{u\geq 0} \mathbf{E}[\exp\{\theta[S^{A^1}_{-k,-1} + uS^{A^2}_{-n,-1} + (-u\phi_2 - \phi_1)S^B_{-k,-1} \\
&\qquad\qquad - u\phi_2 S^B_{-n,-k-1}]\}] \\
&\leq \sum_{n\geq 0}\sum_{0\leq k\leq n} \inf_{u\geq 0} \exp\{(n-k)[\Lambda_{A^2}(\theta u) + \Lambda_B(-\theta\phi_2 u))] \\
&\qquad\qquad + k[\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta u) + \Lambda_B(-\theta(\phi_1 + u\phi_2))] + \Gamma'(\theta,u)\} \\
&\leq \sum_{n\geq 0} n \sup_{\zeta\in[0,1]} \inf_{u\geq 0} \exp\{n[\zeta(\Lambda_{A^2}(\theta u) + \Lambda_B(-\theta\phi_2 u))) \\
&\quad + (1-\zeta)(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta u) + \Lambda_B(-\theta(\phi_1 + u\phi_2))) + \tfrac{\Gamma'(\theta,u)}{n}]\}.
\end{aligned} \tag{3.35}$$

In the second inequality above we have used Assumption E. Let us now define

$$\Lambda^{II}_{GPS,1}(\theta) \stackrel{\triangle}{=} \sup_{\zeta\in[0,1]} \inf_{u\geq 0}[\zeta(\Lambda_{A^2}(\theta u) + \Lambda_B(-\theta\phi_2 u))) + (1-\zeta)(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta u) + \Lambda_B(-\theta(\phi_1 + u\phi_2)))].$$

Let $\hat{u}^*(\theta)$ be the optimal u in the above optimization problem. From (3.35) we have

$$\begin{aligned}&\mathbf{E}[e^{\theta L^{II}_{GPS,1}}] \leq \\ &\leq \sum_{n\geq 0} n \sup_{\zeta\in[0,1]} \exp\{n[\zeta(\Lambda_{A^2}(\theta\hat{u}^*) + \Lambda_B(-\theta\phi_2\hat{u}^*))) \\ &\quad + (1-\zeta)(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta\hat{u}^*) + \Lambda_B(-\theta(\phi_1 + \hat{u}^*\phi_2))) + \tfrac{\Gamma'(\theta,\hat{u}^*)}{n}]\}.\end{aligned} \tag{3.37}$$

Now for every $\epsilon' > 0$ and $\theta \geq 0$ we can take n large enough such that $\frac{\Gamma'(\theta,\hat{u}^*)}{n} < \epsilon'$. For sufficiently small ϵ' and if $\Lambda^{II}_{GPS,1}(\theta) < 0$ then the infinite geometric series in the right hand side of (3.37) converges to a constant, with respect to n, $K_2(\theta,\epsilon')$. That is,

$$\mathbf{E}[e^{\theta L^{II}_{GPS,1}}] \leq K_2(\theta,\epsilon'), \qquad \text{if } \Lambda^{II}_{GPS,1}(\theta) < 0. \tag{3.38}$$

We can now invoke the Markov inequality and by using the bounds (3.31) and (3.35) on (3.27) and (3.28) obtain

$$\begin{aligned}\mathbf{P}[L^1_0 > U] \leq& \mathbf{P}[L^1_0 > U \text{ and } \Omega_1] + \mathbf{P}[L^1_0 > U \text{ and } \Omega_2] \\ \leq& (\mathbf{E}[e^{\theta L^I_{GPS,1}}] + \mathbf{E}[e^{\theta L^{II}_{GPS,1}}])e^{-\theta U} \\ \leq& (K_1(\theta,\epsilon) + K_2(\theta,\epsilon'))e^{-\theta U}, \qquad \text{if } \max(\Lambda^I_{GPS,1}(\theta), \Lambda^{II}_{GPS,1}(\theta)) < 0.\end{aligned} \tag{3.39}$$

Optimizing over θ to get the tightest bound we establish the following proposition.

Proposition 3.6.4 *If* $\mathbf{E}[A^1] < \phi_2\mathbf{E}[B]$ *and assuming that the arrival and service*

processes satisfy Assumptions A and E

$$\lim_{U\to\infty}\frac{1}{U}\log\mathbf{P}[L_0^1>U]\leq-\sup_{\{\theta\geq 0:\ \max(\Lambda^I_{GPS,1}(\theta),\Lambda^{II}_{GPS,1}(\theta))<0\}}\theta.$$

We are now left with proving that this upper bound matches the lower bound, θ^*_{GPS}. The result which is based on Lemma 3.6.2 and convex duality is established in the next proposition.

Proposition 3.6.5 ***(GPS Upper bound, Case 1)*** *If* $\mathbf{E}[A^2]<\phi_2\mathbf{E}[B]$ *and assuming that the arrival and service processes satisfy Assumption A and E, the steady-state queue length,* L^1*, of queue* Q^1*, at an arbitrary time slot satisfies*

$$\lim_{U\to\infty}\frac{1}{U}\log\mathbf{P}[L^1>U]\leq-\theta^*_{GPS}.$$

Proof : It suffices to prove that $\theta^*_{GPS}=\sup_{\{\theta\geq 0:\ \max(\Lambda^I_{GPS,1}(\theta),\Lambda^{II}_{GPS,1}(\theta))<0\}}\theta$. Consider the following expressions

$$\Lambda^{I*}_{GPS,1}(a)\triangleq\inf_{\substack{\zeta(x_2-\phi_2x_3)+(1-\zeta)(y_1+y_2-y_3)=a\\ \zeta(x_2-\phi_2x_3)+(1-\zeta)(y_2-\phi_2y_3)\leq 0\\ 0\leq\zeta\leq 1}}[\zeta(\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3))+(1-\zeta)(\Lambda^*_{A^1}(y_1)+\Lambda^*_{A^2}(y_2)+\Lambda^*_B(y_3))],\quad(3.40)$$

and

$$\Lambda^{II*}_{GPS,1}(a)\triangleq\inf_{\substack{(1-\zeta)(y_1-\phi_1y_3)=a\\ \zeta(x_2-\phi_2x_3)+(1-\zeta)(y_2-\phi_2y_3)\geq 0\\ 0\leq\zeta\leq 1}}[\zeta(\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3))+(1-\zeta)(\Lambda^*_{A^1}(y_1)+\Lambda^*_{A^2}(y_2)+\Lambda^*_B(y_3))],\quad(3.41)$$

which by a change of variables can be written as

$$\Lambda^{I*}_{GPS,1}(a)=\inf_{\substack{(x_2-\phi_2x_3)+(y_1+y_2-y_3)=a\\ (x_2-\phi_2x_3)+(y_2-\phi_2y_3)\leq 0}}\inf_{\zeta\in[0,1]}[\zeta(\Lambda^*_{A^2}(x_2/\zeta)+\Lambda^*_B(x_3/\zeta))+$$

$$+ (1-\zeta)(\Lambda^*_{A^1}(y_1/(1-\zeta)) + \Lambda^*_{A^2}(y_2/(1-\zeta)) + \Lambda^*_B(y_3/(1-\zeta)))], \quad (3.42)$$

and

$$\Lambda^{II*}_{GPS,1}(a) = \inf_{\substack{(y_1-\phi_1 y_3)=a \\ (x_2-\phi_2 x_3)+(y_2-\phi_2 y_3)\geq 0}} \inf_{\zeta\in[0,1]} [\zeta(\Lambda^*_{A^2}(x_2/\zeta) + \Lambda^*_B(x_3/\zeta))$$
$$+ (1-\zeta)(\Lambda^*_{A^1}(y_1/(1-\zeta)) + \Lambda^*_{A^2}(y_2/(1-\zeta)) + \Lambda^*_B(y_3/(1-\zeta)))]. \quad (3.43)$$

By [Roc70, Thm. 5.8] the function

$$\inf_{\zeta\in[0,1]} [\zeta(\Lambda^*_{A^2}(x_2/\zeta) + \Lambda^*_B(x_3/\zeta))$$
$$+ (1-\zeta)(\Lambda^*_{A^1}(y_1/(1-\zeta)) + \Lambda^*_{A^2}(y_2/(1-\zeta)) + \Lambda^*_B(y_3/(1-\zeta)))]$$

is convex in $(x_2, x_3, y_1, y_2, y_3)$ and therefore the functions $\Lambda^{I*}_{GPS,1}(a)$ and $\Lambda^{II*}_{GPS,1}(a)$ are convex in a as optimal value functions of a convex optimization problem with a appearing only in the right hand side of the constraints. We will next show that the convex duals of these functions are $\Lambda^I_{GPS,1}(\theta)$ and $\Lambda^{II}_{GPS,1}(\theta)$, respectively. Indeed, by using convex duality, we have

$$\sup_a [\theta a - \Lambda^{I*}_{GPS,1}(a)] =$$
$$= \sup_{\zeta\in[0,1]} \sup_a \sup_{\substack{\zeta(x_2-\phi_2 x_3)+(1-\zeta)(y_1+y_2-y_3)=a \\ \zeta(x_2-\phi_2 x_3)+(1-\zeta)(y_2-\phi_2 y_3)\leq 0 \\ 0\leq\zeta\leq 1}} [\theta a - \zeta(\Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3))$$
$$- (1-\zeta)(\Lambda^*_{A^1}(y_1) + \Lambda^*_{A^2}(y_2) + \Lambda^*_B(y_3))]$$
$$= \sup_{\zeta\in[0,1]} \inf_{u\geq 0} \sup_{\substack{x_2,x_3 \\ y_1,y_2,y_3}} [\theta\zeta(x_2-\phi_2 x_3) + \theta(1-\zeta)(y_1+y_2-y_3) - u\zeta(x_2-\phi_2 x_3)$$
$$- u(1-\zeta)(y_2-\phi_2 y_3) - \zeta(\Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3))$$
$$- (1-\zeta)(\Lambda^*_{A^1}(y_1) + \Lambda^*_{A^2}(y_2) + \Lambda^*_B(y_3))]$$
$$= \sup_{\zeta\in[0,1]} \inf_{u\geq 0} [\zeta(\Lambda_{A^2}(\theta-u) + \Lambda_B(-\theta\phi_2 + u\phi_2))$$
$$+ (1-\zeta)(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta-u) + \Lambda_B(-\theta+u\phi_2)))]$$

$$=\Lambda^{I}_{GPS,1}(\theta).$$

Similarly it can be shown that $\Lambda^{II}_{GPS,1}(\theta)$ is the convex dual of $\Lambda^{I*}_{GPS,1}(a)$. Let now

$$\theta_I \stackrel{\Delta}{=} \inf_{a>0} \frac{1}{a}\Lambda^{I*}_{GPS,1}(a), \tag{3.44}$$

and

$$\theta_{II} \stackrel{\Delta}{=} \inf_{a>0} \frac{1}{a}\Lambda^{II*}_{GPS,1}(a). \tag{3.45}$$

Using the result of Lemma 3.6.2, θ_I (resp. θ_{II}) is the largest positive root of $\Lambda^{I}_{GPS,1}(\theta) = 0$ (resp. $\Lambda^{II}_{GPS,1}(\theta) = 0$). As Figure 3-4 indicates, due to convexity, $\theta^*_{GPS,1} \stackrel{\Delta}{=} \min(\theta_I, \theta_{II})$ is the largest positive root of the equation $\Lambda_{GPS,1}(\theta) \stackrel{\Delta}{=} \max[\Lambda^{I}_{GPS,1}(\theta), \Lambda^{II}_{GPS,1}(\theta)] = 0$, that is $-\theta^*_{GPS,1}$ is equal to the upper bound established in Prop. 3.6.4. The last thing we have to show is that $\theta^*_{GPS,1} = \theta^*_{GPS}$. This is

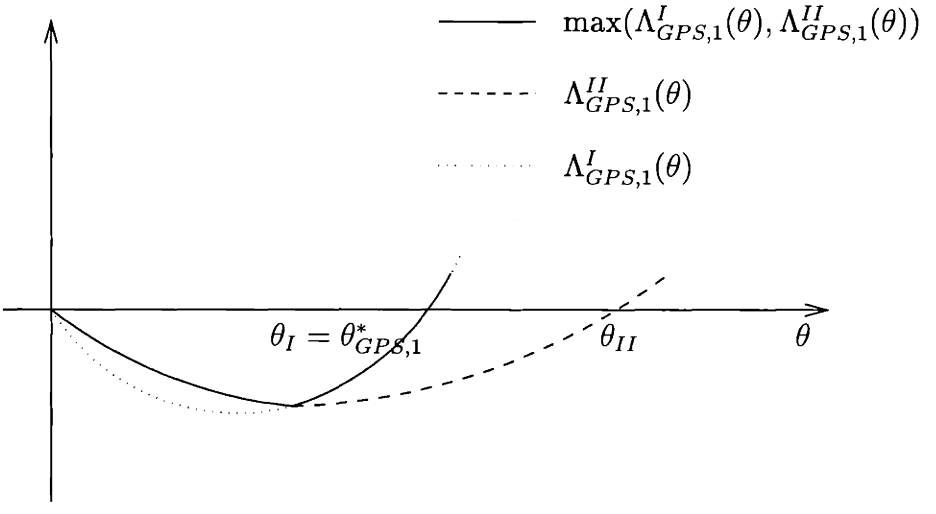

Figure 3-4: $\theta^*_{GPS,1}$ as the largest positive root of the equation $\Lambda_{GPS,1}(\theta) = 0$.

based on $\theta^*_{GPS,1}$ being equal to $\min(\theta_I, \theta_{II})$. Note, from (3.44), that θ_I corresponds to

the optimal solution of a control problem very similar to (GPS-OVERFLOW) with a trajectory of the form appearing in Figure 3-5(a). Also, from (3.45), θ_{II} corresponds to the optimal solution of a control problem with a trajectory of the form appearing in Figure 3-5(b) [3]. The only difference from (GPS-OVERFLOW) is that on the L^2-axis the cost functional is $\Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)$ instead of $\Lambda^*_{A^2}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)$. Using exactly the same techniques as in Section 3.4, that is convexity and the homogeneity property, it can be established that optimal state trajectories do not spend any time on the L^2 axis. Thus, Figure 3-5(a) and (b) can be reduced to the ones in Figure 3-3(a) and (b), respectively. This establishes the desired result $\theta^*_{GPS,1} = \theta^*_{GPS}$ and concludes the proof of the theorem.

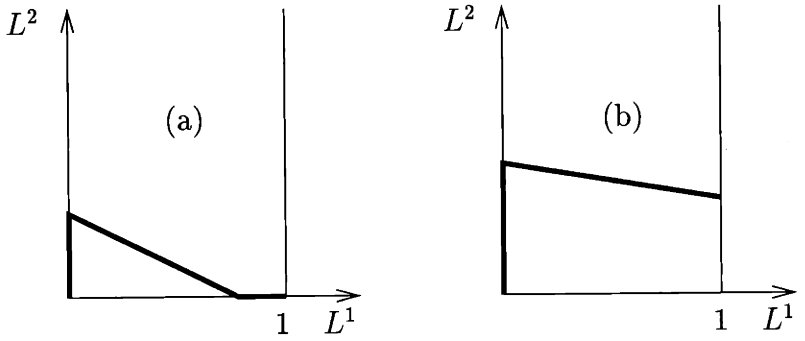

Figure 3-5: Trajectories for the control problems corresponding to θ_I and θ_{II}.

■

We summarize Propositions 3.6.5 and 3.6.3 in the following proposition.

Proposition 3.6.6 ***(GPS Upper Bound)*** *Assuming that the arrival and service processes satisfy Assumptions A and E, and under the GPS policy, the steady-state*

[3] For both trajectories we let ζ be the fraction of time that they spend on the L^2 axis and x_2, x_2 (resp. y_1, y_2, y_3) the controls for the initial ζ (resp. last $1-\zeta$) fraction of the time.

queue length, L^1, of queue Q^1, at an arbitrary time slot satisfies

$$\lim_{U\to\infty}\frac{1}{U}\log \mathbf{P}[L^1 > U] \leq -\theta^*_{GPS}. \tag{3.46}$$

3.7 Main Results

In this section we combine Propositions 3.3.1 and 3.6.6 and summarize our main results for the GPS policy. As a corollary we obtain results for priority policies.

Theorem 3.7.1 ***(GPS Main)*** *Under the GPS policy, assuming that the arrival and service processes satisfy Assumptions A, D, and E the steady-state queue length, L^1, of queue Q^1, at an arbitrary time slot satisfies*

$$\lim_{U\to\infty}\frac{1}{U}\log \mathbf{P}[L^1 > U] = -\theta^*_{GPS}, \tag{3.47}$$

where θ^*_{GPS} is given by

$$\theta^*_{GPS} = \min\left[\inf_{a>0}\frac{1}{a}\Lambda^{I*}_{GPS}(a), \inf_{a>0}\frac{1}{a}\Lambda^{II*}_{GPS}(a)\right], \tag{3.48}$$

and the functions $\Lambda^{I*}_{GPS}(\cdot)$ and $\Lambda^{II*}_{GPS}(\cdot)$ are defined as follows

$$\Lambda^{I*}_{GPS}(a) \stackrel{\triangle}{=} \inf_{\substack{x_1+x_2-x_3=a \\ x_2\leq\phi_2 x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)], \tag{3.49}$$

and

$$\Lambda^{II*}_{GPS}(a) \stackrel{\triangle}{=} \inf_{\substack{x_1-\phi_1 x_3=a \\ x_2\geq\phi_2 x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \tag{3.50}$$

An interesting observation is that strict priority policies are a special case of the GPS policy. Type 1 customers have higher priority when $\phi_1 = 1$ and lower priority when $\phi_1 = 0$. We can therefore obtain the performance of these two priority policies as a by-product of our analysis. Note that the result for the policy that

assigns higher priority to Type 1 customers, matches the FCFS single class result (see [Kel91, GW94, BPT94]) since under this policy, Type 1 customers are oblivious of Type 2 customers. We summarize the performance of priority policies in the next corollary. The discussion of Section 3.5 can be easily adapted to the cases $\phi_1 = 1$ and $\phi_1 = 0$ to characterize the *most likely ways* that lead to overflow under priority policies.

Corollary 3.7.2 ***(Priority policies)*** *Under strict priority policy for Type 1 customers (P_1), assuming that the arrival and service processes satisfy Assumptions A, D, and E the steady-state queue length, L^1, of queue Q^1, at an arbitrary time slot satisfies*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] = -\theta^*_{P_1}, \tag{3.51}$$

where $\theta^*_{P_1}$ is given by

$$\theta^*_{P_1} = \inf_{a>0} \frac{1}{a} \Lambda^*_{P_1}(a), \tag{3.52}$$

and where

$$\Lambda^*_{P_1}(a) \stackrel{\triangle}{=} \inf_{x_1 - x_3 = a} [\Lambda^*_{A^1}(x_1) + \Lambda^*_B(x_3)]. \tag{3.53}$$

Under strict priority policy for Type 2 customers (P_2), the steady-state queue length, L^1, of queue Q^1, at an arbitrary time slot satisfies

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] = -\theta^*_{P_2}, \tag{3.54}$$

where $\theta^*_{P_2}$ is given by

$$\theta^*_{P_2} = \inf_{a>0} \frac{1}{a} \Lambda^*_{P_2}(a), \tag{3.55}$$

and where

$$\Lambda^*_{P_2}(a) \stackrel{\triangle}{=} \inf_{\substack{x_1+x_2-x_3=a \\ x_2 \le x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)], \tag{3.56}$$

Proof : For policy P_1 apply Theorem 3.7.1 with $\phi_1 = 1$. For such ϕ_1, it is easy to verify that $\Lambda^{I*}_{GPS}(a) \ge \Lambda^{II*}_{GPS}(a)$, for all a. Thus, we define $\Lambda^*_{P_1}(a)$ to be equal to $\Lambda^{II*}_{GPS}(a)$ with ϕ_1 set to 1.

For policy P_2 apply Theorem 3.7.1 with $\phi_1 = 0$. Application of $\phi_1 = 0$ to $\Lambda^{I*}_{GPS}(a)$ yields

$$\Lambda^{I*}_{GPS}(a) = \inf_{\substack{x_1+x_2-x_3=a \\ x_2 \le x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \tag{3.57}$$

Also, application of $\phi_1 = 0$ to $\Lambda^{II*}_{GPS}(a)$ yields

$$\Lambda^{II*}_{GPS}(a) = \inf_{\substack{x_1=a \\ x_2 \ge x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \tag{3.58}$$

The functions $\Lambda^*_{A^2}(x_2)$ and $\Lambda^*_B(x_3)$ are non-negative, convex, and achieve their minimum value, which is equal to 0, at $x_2 = \mathbf{E}[A^2_0]$ and $x_3 = \mathbf{E}[B_0]$, respectively. Since $\mathbf{E}[B_0] > \mathbf{E}[A^2_0]$, the inequality $x_2 \ge x_3$ implies that either $x_2 > \mathbf{E}[A^2_0]$ or $x_3 < \mathbf{E}[B_0]$. If the former is the case, we can decrease x_2 and reduce the cost, as long $x_2 \ge x_3$ holds. Also, if $x_3 < \mathbf{E}[B_0]$ is the case, we can increase x_3 and reduce the cost, as long $x_2 \ge x_3$ holds. Thus, at optimality $x_2 = x_3$ in (3.58). But, the region characterized by $x_1 = a$ and $x_2 = x_3$ is included in the region defined by the constraints in the optimization problem in (3.57). Hence, for all a, and when $\phi_1 = 0$, $\Lambda^{I*}_{GPS}(a) \le \Lambda^{II*}_{GPS}(a)$. Therefore, we define $\Lambda^*_{P_2}(a)$ to be equal to the expression in (3.57).

■

As the results of Theorem 3.7.1 and Corollary 3.7.2 indicate, the calculation of the overflow probabilities involves the solution of an optimization problem. We will next show that because of the special structure that these problems exhibit, this is

equivalent to finding the maximum root of a convex function. Such a task might be easier to perform in some cases, analytically or computationally. This equivalence relies mainly on Lemma 3.6.2. Hence, using duality, we express θ^*_{GPS} as the largest root of a convex function. On a notational remark, we will be denoting by $\Lambda^I_{GPS}(\cdot)$ and $\Lambda^{II}_{GPS}(\cdot)$, the convex duals of $\Lambda^{I*}_{GPS}(\cdot)$ and $\Lambda^{II*}_{GPS}(\cdot)$, respectively. Notice, that $\Lambda^{I*}_{GPS}(a)$ and $\Lambda^{II*}_{GPS}(a)$ are convex functions of a as the value functions of a convex optimization problem with a appearing only in the right hand side of the constraints.

Theorem 3.7.3 *θ^*_{GPS} is the largest positive root of the equation*

$$\Lambda_{GPS}(\theta) \stackrel{\triangle}{=} \Lambda_{A^1}(\theta) + \inf_{0\le u\le\theta}[\Lambda_{A^2}(\theta-u) + \Lambda_B(-\theta+\phi_2 u)] = 0. \tag{3.59}$$

Proof : The first thing to note is that $\Lambda_{GPS}(\theta)$ is a convex function of θ. This can be seen when we write it as the value function of a convex optimization problem with θ appearing only in the right hand side of the constraints, i.e.,

$$\Lambda_{GPS}(\theta) = \Lambda_{A^1}(\theta) + \inf_{\substack{z=\theta\\ 0\le u\le\theta}} [\Lambda_{A^2}(z-u) + \Lambda_B(-z+\phi_2 u)].$$

Next we show that Equation (3.59) has a positive, possibly infinite, root. To this end, observe that

$$\Lambda_{GPS}(\theta) \le \Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta) + \Lambda_B(-\theta),$$

and that both sides of the above inequality are 0 at $\theta = 0$. This implies that their derivatives at $\theta = 0$ satisfy

$$\Lambda'_{GPS}(0) \le \Lambda'_{A^1}(0) + \Lambda'_{A^2}(0) - \Lambda'_B(0) < 0,$$

where the last inequality follows from the stability condition (3.1). The convexity of $\Lambda_{GPS}(\cdot)$ is sufficient to guarantee the existence of a positive, possible infinite, root.

We now calculate the functions $\Lambda^I_{GPS}(\theta)$ and $\Lambda^{II}_{GPS}(\theta)$, using convex duality. We have

$$\begin{aligned}
\Lambda^I_{GPS}(\theta) &= \sup_a[\theta a - \Lambda^{I*}_{GPS}(a)] \\
&= \sup_a \sup_{\substack{x_1+x_2-x_3=a \\ x_2\le\phi_2 x_3}} [\theta a - \Lambda^*_{A^1}(x_1) - \Lambda^*_{A^2}(x_2) - \Lambda^*_B(x_3)] \\
&= \sup_a \sup_{\substack{x_1+x_2-x_3=a \\ x_2\le\phi_2 x_3}} [\theta(x_1+x_2-x_3) - \Lambda^*_{A^1}(x_1) - \Lambda^*_{A^2}(x_2) - \Lambda^*_B(x_3)] \\
&= \sup_{x_2\le\phi_2 x_3} [\theta(x_1+x_2-x_3) - \Lambda^*_{A^1}(x_1) - \Lambda^*_{A^2}(x_2) - \Lambda^*_B(x_3)] \\
&= \Lambda_{A^1}(\theta) + \inf_{u\ge 0} \sup_{x_2,x_3} [\theta(x_2-x_3) - \Lambda^*_{A^2}(x_2) - \Lambda^*_B(x_3) + u(\phi_2 x_3 - x_2)] \\
&= \Lambda_{A^1}(\theta) + \inf_{u\ge 0}[\Lambda_{A^2}(\theta-u) + \Lambda_B(-\theta+u\phi_2)].
\end{aligned}$$

In the fifth equality above we have dualized the constraint $x_2 \le \phi_2 x_3$ and used the definition of $\Lambda_{A^1}(\theta)$. Similarly, the convex dual of $\Lambda^{II*}_{GPS}(\cdot)$ is

$$\begin{aligned}
\Lambda^{II}_{GPS}(\theta) &= \sup_a[\theta a - \Lambda^{II*}_{GPS}(a)] \\
&= \sup_a \sup_{\substack{x_1-\phi_1 x_3=a \\ x_2\ge\phi_2 x_3}} [\theta a - \Lambda^*_{A^1}(x_1) - \Lambda^*_{A^2}(x_2) - \Lambda^*_B(x_3)] \\
&= \Lambda_{A^1}(\theta) + \inf_{u\ge 0} \sup_{x_2,x_3} [\theta(-\phi_1 x_3) - \Lambda^*_{A^2}(x_2) - \Lambda^*_B(x_3) + u(-\phi_2 x_3 + x_2)] \\
&= \Lambda_{A^1}(\theta) + \inf_{u\ge 0}[\Lambda_{A^2}(u) + \Lambda_B(-\theta\phi_1 - u\phi_2)] \\
&= \Lambda_{A^1}(\theta) + \inf_{u\le\theta}[\Lambda_{A^2}(\theta-u) + \Lambda_B(-\theta+u\phi_2)].
\end{aligned}$$

In the fifth equality above we have made the substitution $u := \theta - u$.

Using the result of Lemma 3.6.2, $\theta_1 \stackrel{\triangle}{=} \inf_{a>0} \frac{1}{a}\Lambda^{I*}_{GPS}(a)$ is the largest positive root of $\Lambda^I_{GPS}(\theta) = 0$ (this equation has a positive, possibly, infinite root by the argument used to establish that $\Lambda_{GPS}(\theta) = 0$ does). Similarly, $\theta_2 \stackrel{\triangle}{=} \inf_{a>0} \frac{1}{a}\Lambda^{II*}_{GPS}(a)$ is the largest positive root of $\Lambda^{II}_{GPS}(\theta) = 0$. By Equation (3.48), $\theta^*_{GPS} = \min(\theta_1, \theta_2)$. The situation is exactly the same as in Figure 3-4, that is θ^*_{GPS} is the largest positive root

of the equation $\max[\Lambda^I_{GPS}(\theta), \Lambda^{II}_{GPS}(\theta)] = 0$.

The last thing we have to show to conclude the proof is that $\Lambda_{GPS}(\theta) = \max[\Lambda^I_{GPS}(\theta), \Lambda^{II}_{GPS}(\theta)]$. Indeed, we have

$$\begin{aligned}\max(\Lambda^I_{GPS}(\theta), \Lambda^{II}_{GPS}(\theta)) =& \max(\Lambda_{A^1}(\theta) + \inf_{u\geq 0}[\Lambda_{A^2}(\theta-u) + \Lambda_B(-\theta+u\phi_2)],\\ & \Lambda_{A^1}(\theta) + \inf_{u\leq\theta}[\Lambda_{A^2}(\theta-u) + \Lambda_B(-\theta+u\phi_2)])\\ =& \Lambda_{A^1}(\theta) + \inf_{0\leq u\leq\theta}[\Lambda_{A^2}(\theta-u) + \Lambda_B(-\theta+u\phi_2)]\\ \overset{(3.59)}{=}& \Lambda_{GPS}(\theta).\end{aligned}$$

■

Again, as it was the case with Theorem 3.7.1, the result of Theorem 3.7.3 can be specialized to the case of priority policies.

Corollary 3.7.4 *θ^*_{P1} is the largest positive root of the equation*

$$\Lambda_{P1}(\theta) \triangleq \Lambda_{A^1}(\theta) + \Lambda_B(-\theta) = 0. \tag{3.60}$$

Also, θ^*_{P2} is the largest positive root of the equation

$$\Lambda_{P2}(\theta) \triangleq \Lambda_{A^1}(\theta) + \inf_{0\leq u\leq\theta}[\Lambda_{A^2}(\theta-u) + \Lambda_B(-\theta+u)] = 0. \tag{3.61}$$

We conclude this section noting that, by symmetry, all the results obtained here can be easily adapted (it suffices to substitute everywhere $1 := 2$ and $2 := 1$) to estimate the overflow probability of the second queue and characterize the most likely ways that it builds up.

Chapter 4

Overflow Probabilities with GLQF

In this chapter we continue the line of development of Chapter 3. We consider the multiclass multiplexer model that we introduced there and we estimate the overflow probabilities for each buffer under another scheduling policy, the *generalized longest queue first (GLQF)* policy which we formally define in the sequel. The notation of Section 3.1 is in place as well as the assumptions on the arrival and service processes that are reported there.

Regarding the structure of this chapter, we begin in Section 4.1 by formally defining the GLQF policy and the probabilities of which we seek the asymptotic tails. Moreover, in the latter section, we provide an orientation of the methodology that we follow in proving our results. In Section 4.2 we prove a lower bound on the overflow probability and in Section 4.3 we introduce the optimal control formulation and solve the control problem. In Section 4.4 we summarize the most likely modes of overflow obtained from the solution of the control problem and in Section 4.5 we prove the matching upper bound. We gather our main results in Section 4.6. Finally, in Section 4.7 we compare the performance of the GPS and GLQF policy.

4.1 The GLQF policy

In this section we introduce the *generalized longest queue first policy (GLQF)*. We will obtain in this chapter, the tail distributions of the queue lengths in the multiclass system, defined in Section 3.1, operated under this policy. Moreover, in the course of the analysis, we identify the most likely ways leading to large queue length values (overflows).

Figure 4-1 depicts the operation of the GLQF policy in the L^1-L^2 space. Fix the parameter of the policy $\beta \geq 0$. There is a threshold line, of slope β, which divides the positive orthant of the $L^1 - L^2$ space in two regions. The GLQF policy serves Type 2 customers above the threshold line and Type 1 below it. The value $\beta = 1$ corresponds to the longest queue first (LQF) policy. More formally, we define the GLQF policy to be the work-conserving policy that at each time slot i serves Type 1 customers when

$$L_i^2 < \beta L_i^1 \quad \text{and} \quad L_i^2 + A_i^2 \leq \beta(L_i^1 + A_i^1 - B_i).$$

It serves Type 2 customers when

$$L_i^2 > \beta L_i^1 \quad \text{and} \quad L_i^2 + A_i^2 - B_i \geq \beta(L_i^1 + A_i^1).$$

When

$$L_i^2 < \beta L_i^1 \quad \text{and} \quad L_i^2 + A_i^2 > \beta(L_i^1 + A_i^1 - B_i),$$

or when

$$L_i^2 > \beta L_i^1 \quad \text{and} \quad L_i^2 + A_i^2 - B_i < \beta(L_i^1 + A_i^1),$$

then the GLQF policy allocates appropriate capacity to both types of customers such that $L_{i+1}^2 = \beta L_{i+1}^1$. Whenever $L_i^2 = \beta L_i^1$, the GLQF policy arbitrarily allocates its capacity to Type 1 and 2 customers.

As in Section 3.1, we assume that the queue length processes $\{L_i^j, j = 1, 2, i \in \mathbb{Z}\}$ are stationary. We are interested in estimating the overflow probability $\mathbf{P}[L_i^1 > U]$

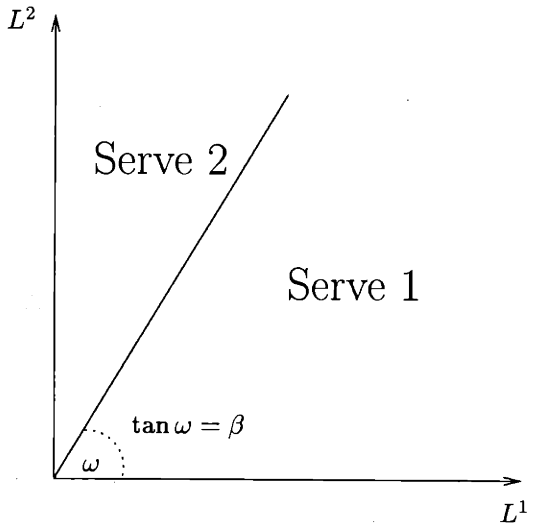

Figure 4-1: The operation of the GLQF policy.

for large values of U, at an arbitrary time slot i in steady-state. Having determined this, the overflow probability of the second queue can be obtained by a symmetrical argument.

We will prove that the overflow probability satisfies

$$\mathbf{P}[L_i^1 > U] \sim e^{-U\theta^*_{GLQF}}, \tag{4.1}$$

asymptotically, as $U \to \infty$. Our methodology is similar to the one we used in analyzing the GPS policy in Chapter 3. To this end, we will develop a lower bound on the overflow probability, along with a matching upper bound. Consider all scenarios (paths) that lead to an overflow. We will show that the probability of each such scenario ω asymptotically behaves as $e^{-U\theta(\omega)}$, for some function $\theta(\omega)$. This probability is a lower bound on $\mathbf{P}[L_i^1 > U]$ for all ω. We select the tightest lower bound by performing the minimization $\theta^*_{GLQF} = \min_\omega \theta(\omega)$. This is a deterministic optimal control problem, which we will solve. Optimal trajectories (paths) of the control problem

correspond to *most likely* overflow scenarios. We show that these must be of one out of two possible types. In other words, with high probability, overflow occurs in one out of two possible modes. For the upper bound, we will consider the probability of all sample paths that lead to overflow and show that it is, asymptotically, no more that $e^{-U\theta^*_{GLQF}}$.

4.2 A Lower Bound

In this section we derive a lower bound on the overflow probability $\mathbf{P}[L_i^1 > U]$.

Proposition 4.2.1 ***(GLQF Lower Bound)*** *Assuming that the arrival and service processes satisfy Assumptions A and D, and under the GLQF policy, the steady-state queue length, L^1, of queue Q^1, at an arbitrary time slot satisfies*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \geq -\theta^*_{GLQF}, \tag{4.2}$$

where θ^*_{GLQF} is given by

$$\theta^*_{GLQF} = \min\left[\inf_{a>0} \frac{1}{a}\Lambda^{I*}_{GLQF}(a), \inf_{a>0} \frac{1}{a}\Lambda^{II*}_{GLQF}(a)\right], \tag{4.3}$$

and the functions $\Lambda^{I*}_{GLQF}(\cdot)$ and $\Lambda^{II*}_{GLQF}(\cdot)$ are defined as follows

$$\Lambda^{I*}_{GLQF}(a) \triangleq \inf_{\substack{x_1-x_3=a \\ x_2\leq\beta(x_1-x_3)}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)], \tag{4.4}$$

and

$$\Lambda^{II*}_{GLQF}(a) \triangleq \inf_{\substack{x_1-\phi x_3=a \\ x_2-(1-\phi)x_3=\beta a \\ 0\leq\phi<1}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \tag{4.5}$$

Proof : Let $-n \leq 0$ and $a > 0$. Fix $x_1, x_2, x_3 \geq 0$ and $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ and consider the event

$$\mathcal{A} \triangleq \{ |S^{A^1}_{-n,-i-1} - (n-i)x_1| \leq \epsilon_1 n,\ |S^{A^2}_{-n,-i-1} - (n-i)x_2| \leq \epsilon_2 n,$$
$$|S^{B}_{-n,-i-1} - (n-i)x_3| \leq \epsilon_3 n,\ i = 0, 1, \ldots, n-1\}.$$

Notice that x_1, x_2 (resp. x_3) have the interpretation of empirical arrival (resp. service) rates during the interval $[-n, -1]$. We focus on two particular scenarios

$$\begin{array}{lll} \underline{\text{Scenario 1}}: & x_1 - x_3 = a & \underline{\text{Scenario 2}}: \quad x_1 - \phi x_3 = a \\ & x_2 \leq \beta(x_1 - x_3) & x_2 - (1-\phi)x_3 = \beta a \\ & & 0 \leq \phi < 1 \end{array} \tag{4.6}$$

Under Scenario 1, even if the server always serves Type 1 customers [1] in $[-n, 0]$ we have that $L^1_0 \geq na - n\epsilon'_1$, where $\epsilon'_1 \to 0$ as $\epsilon_1, \epsilon_2, \epsilon_3 \to 0$.

Consider now Scenario 2. Let $(L^1_{-n}, L^2_{-n}) = (x, y)$ and let for the moment ignore ϵ's (i.e., $\epsilon_1 = \epsilon_2 = \epsilon_3 = 0$). If $y = \beta x$ and for given x_1, x_2, x_3 we can find ϕ such that both queues build up together with the relation $L^2 = \beta L^1$ holding in the interval $[-n, 0]$. According to the GLQF policy the server arbitrarily allocates its capacity to the two queues, giving fraction ϕ to Q^1 and the remaining $1 - \phi$ to Q^2 (here ϕ is subject to optimization), yielding $L^1_0 = na + x \geq na$. If $y > \beta x$ then the first queue receives less capacity in $[-n, 0]$ than $n\phi x_3$, resulting also in $L^1_0 \geq na$. Finally, consider the case $y < \beta x$. Then at time $-t \in [-n, 0]$ we have $L^1_{-t} = x + (n-t)(x_1 - x_3)$ and $L^2_{-t} = y + (n-t)x_2$. Notice that $x_2 > \beta(x_1 - x_3)$. Otherwise, we have a contradiction, i.e.,

$$\beta a \leq x_2 \leq \beta(x_1 - x_3) < \beta a.$$

[1] which is the case if we start from an empty system at $-n$ and the arrival and service rates are exactly x_1, x_2, x_3, respectively. Then the second queue, since it receives zero capacity, builds up with rate x_2, and its level always stays below βL^1, a necessary condition for the first queue to be receiving all the capacity.

Thus, for large enough n, there exists some t such that $L^2_{-t} = \beta L^1_{-t}$. From that time on, both queues build up together with the relation $L^2 = \beta L^1$ holding. Therefore and since $L^2_0 + L^1_0 \geq (1+\beta)a$, we have $L^1_0 \geq na$.

With $\epsilon_1, \epsilon_2, \epsilon_3 > 0$, and with the same ϕ there exists $\epsilon_2' > 0$ such that queue lengths are within an ϵ_2' band of their values in the previous paragraph, resulting in $L^1_0 \geq na - n\epsilon_2'$, where $\epsilon_2' \to 0$ as $\epsilon_1, \epsilon_2, \epsilon_3 \to 0$.

The probability of Scenario 1 is a lower bound on $\mathbf{P}[L^1_0 \geq na]$. Calculating the probability of Scenario 1, maximizing over x_1, x_2 and x_3, to obtain the tightest bound, and using Assumption D we have

$$\begin{aligned}
\mathbf{P}[L^1_0 \geq n(a-\epsilon_1')] \geq & \sup_{\substack{x_1 - x_3 = a \\ x_2 \leq \beta(x_1 - x_3)}} \mathbf{P}[\, |S^{A^1}_{-n,-i-1} - (n-i)x_1| \leq \epsilon_1 n,\ i = 0, 1, \ldots, n-1] \\
& \times \mathbf{P}[\, |S^{A^2}_{-n,-i-1} - (n-i)x_2| \leq \epsilon_2 n,\ i = 0, 1, \ldots, n-1] \\
& \times \mathbf{P}[\, |S^{B}_{-n,-i-1} - (n-i)x_3| \leq \epsilon_3 n,\ i = 0, 1, \ldots, n-1] \\
\geq & \exp\left\{-n\left(\inf_{\substack{x_1 - x_3 = a \\ x_2 \leq \beta(x_1 - x_3)}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)] + \epsilon\right)\right\} \\
= & \exp\{-n(\Lambda^{I*}_{GLQF}(a) + \epsilon)\}, \qquad (4.7)
\end{aligned}$$

where n is large enough, and the $\epsilon_1', \epsilon \to 0$ as $\epsilon_1, \epsilon_2, \epsilon_3 \to 0$.

Similarly, calculating the probability of Scenario 2, we have

$$\begin{aligned}
\mathbf{P}[L^1_0 \geq n(a-\epsilon_2')] \geq & \sup_{\substack{x_1 - \phi x_3 = a \\ x_2 - (1-\phi)x_3 = \beta a \\ 0 \leq \phi < 1}} \mathbf{P}[\, |S^{A^1}_{-n,-i-1} - (n-i)x_1| \leq \epsilon_1 n,\ i = 0, 1, \ldots, n-1] \\
& \times \mathbf{P}[\, |S^{A^2}_{-n,-i-1} - (n-i)x_2| \leq \epsilon_2 n,\ i = 0, 1, \ldots, n-1] \\
& \times \mathbf{P}[\, |S^{B}_{-n,-i-1} - (n-i)x_3| \leq \epsilon_3 n,\ i = 0, 1, \ldots, n-1] \\
\geq & \exp\left\{-n\left(\inf_{\substack{x_1 - \phi x_3 = a \\ x_2 - (1-\phi)x_3 = \beta a \\ 0 \leq \phi < 1}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)] + \epsilon'\right)\right\} \\
= & \exp\{-n(\Lambda^{II*}_{GLQF}(a) + \epsilon')\}, \qquad (4.8)
\end{aligned}$$

where n is large enough, and the $\epsilon_2', \epsilon' \to 0$ as $\epsilon_1, \epsilon_2, \epsilon_3 \to 0$.

Combining Eqs. (4.7) and (4.8) (taking the limit of all ϵ's going to zero) we obtain

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[L_0^1 \geq na] \geq -\min(\Lambda^{I*}_{GLQF}(a), \Lambda^{II*}_{GLQF}(a)). \tag{4.9}$$

As a final step to this proof, letting $U = na$, we obtain

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] = \lim_{n\to\infty} \frac{1}{na} \log \mathbf{P}[L_0^1 \geq na] \geq -\frac{1}{a} \min(\Lambda^{I*}_{GLQF}(a), \Lambda^{II*}_{GLQF}(a)).$$

Since a, in the above, is arbitrary we can select it in order to make the bound tighter. Namely,

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \geq -\min\left[\inf_{a>0} \frac{1}{a}\Lambda^{I*}_{GLQF}(a), \inf_{a>0} \frac{1}{a}\Lambda^{II*}_{GLQF}(a)\right].$$

■

4.3 The optimal control problem

In this section we introduce an optimal control problem and show that θ^*_{GLQF} is its optimal value. The ideas are similar to the case of the GPS policy, we will therefore keep the discussion brief.

The scaling of time and fluid levels is done in exactly the same manner, as in Section 3.4, therefore the resulting control problem is identical to (GPS-OVERFLOW) with the exception of the system dynamics that are different in the case of the GLQF policy. In particular, we distinguish three regions depending on the state as follows

Region A: $L^2(t) > \beta L^1(t)$, where according to the GLQF policy

$$\dot{L}^1 = x_1(t) \quad \text{and} \quad \dot{L}^2 = x_2(t) - x_3(t),$$

Region B: $L^2(t) < \beta L^1(t)$, where according to the GLQF policy

$$\dot{L}^1 = x_1(t) - x_3(t) \quad \text{and} \quad \dot{L}^2 = x_2(t),$$

Region C: $L^2(t) = \beta L^1(t)$, where according to the GLQF policy

$$\dot{L}^1 + \dot{L}^2 = x_1(t) + x_2(t) - x_3(t)$$

Let (GLQF-DYNAMICS) denote the set of state trajectories $L^j(t),\ j = 1, 2,\ t \in [-T, 0]$, that obey the dynamics given above.

We now formally define the following optimal control problem (GLQF-OVERFLOW). The control variables are $x_j(t),\ j = 1, 2, 3$, and the state variables are $L^j(t),\ j = 1, 2$, for $t \in [-T, 0]$, which obey the dynamics given in the previous paragraph.

$$\text{(GLQF-OVERFLOW)} \quad \inf \int_{-T}^{0} [\Lambda^*_{A^1}(x_1(t)) + \Lambda^*_{A^2}(x_2(t)) + \Lambda^*_B(x_3(t))]\, dt \tag{4.10}$$

$$\begin{aligned} \text{subject to: } & L^1(-T) = L^2(-T) = 0 \\ & L^1(0) = 1 \\ & L^2(0) : \text{ free} \\ & T : \text{ free} \\ & \{L^j(t) :\ t \in [-T, 0],\ j = 1, 2\} \in \text{(GLQF-DYNAMICS)}. \end{aligned}$$

This problem exhibits both the properties of constant control trajectories within each region of system dynamics, and time-homogeneity. We omit the proofs since they are similar to the GPS case. Using these properties we can make the reductions appearing in Figure 4-2(a), (b) and (c), starting from an arbitrary trajectory with piecewise constant controls. We conclude that optimal state trajectories can be reduced to having one of the forms depicted in Figure 4-2(d), (e) and (f).

The trajectory in Figure 4-2(d) has value equal to $\inf_T[T\Lambda^{I*}_{GLQF}(\frac{1}{T})]$ and the trajectory in Figure 4-2(e) has value equal to $\inf_T[T\Lambda^{II*}_{GLQF}(\frac{1}{T})]$, where $\Lambda^{I*}_{GLQF}(\cdot)$ and $\Lambda^{II*}_{GLQF}(\cdot)$ are defined in Equations (4.4) and (4.5), respectively. Consider now the

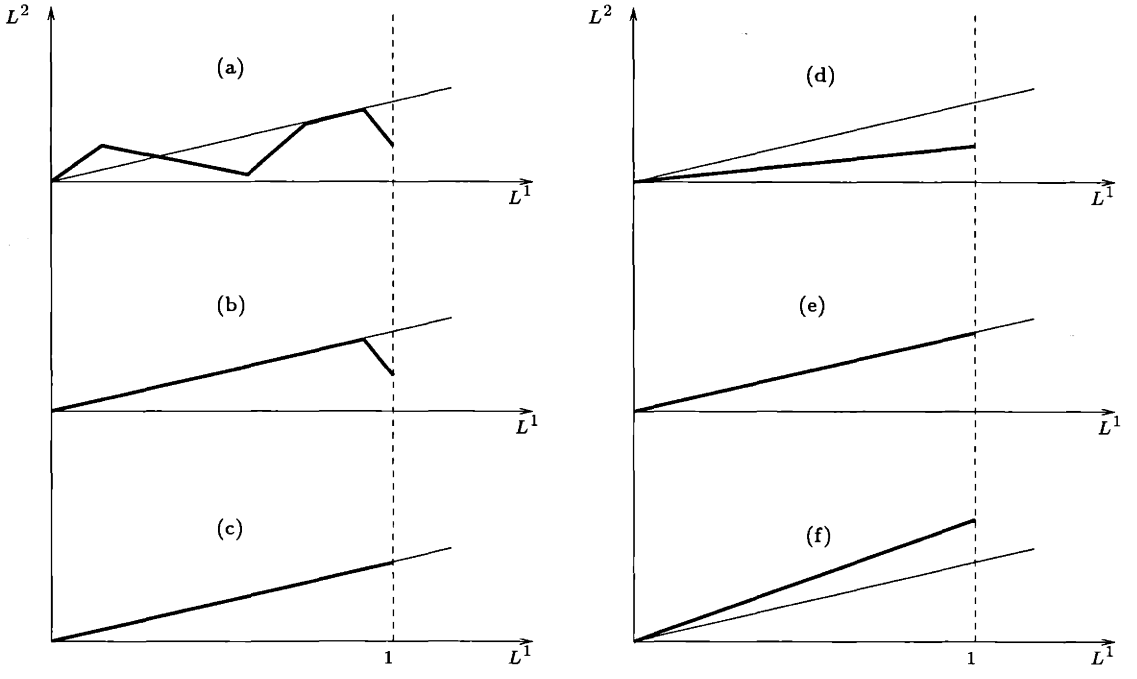

Figure 4-2: By the property of constant controls within each region of system dynamics the state trajectory in (b) is no more costly than the trajectory in (a). Also, by the time-homogeneity property, optimality of the state trajectory in (b) implies optimality of the trajectory in (c). Candidates for optimal state trajectories are depicted in (d), (e) and (f). The trajectory in (f) is eliminated as less profitable to the one in (e). Hence, without loss of optimality we can restrict attention to trajectories of the form in (d) and (e).

trajectory in Figure 4-2(f) which has value

$$\inf_{T} \inf_{\substack{x_1=\frac{1}{T} \\ x_2-x_3\geq\beta\frac{1}{T}}} [\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)]. \tag{4.11}$$

The functions $\Lambda^*_{A^2}(x_2)$ and $\Lambda^*_B(x_3)$ are non-negative, convex, and achieve their minimum value which is equal to 0 at $x_2 = \mathbf{E}[A_0^2]$ and $x_3 = \mathbf{E}[B_0]$, respectively. Since $\frac{1}{T} \geq 0$, and due to the stability condition (3.1), for $x_2 - x_3 \geq \beta\frac{1}{T}$, it has to be the case that either $x_2 > \mathbf{E}[A_0^2]$ or $x_3 < \mathbf{E}[B_0]$. If the former is the case, we can decrease x_2 and reduce the cost, as long $x_2 - x_3 \geq \beta\frac{1}{T}$ holds. Also, if $x_3 < \mathbf{E}[B_0]$ is the case, we can increase x_3 and reduce the cost, as long $x_2 - x_3 \geq \beta\frac{1}{T}$ holds. Thus, at optimality it is true that $x_2 - x_3 = \beta\frac{1}{T}$. Then, the expression in (4.11) is equal to $\inf_T[T\Lambda^{II*}_{GLQF}(\frac{1}{T})]$ with $\phi = 0$ in the definition of $\Lambda^{II*}_{GLQF}(\frac{1}{T})$. Thus, since the calculation of $\Lambda^{II*}_{GLQF}(\frac{1}{T})$ involves optimization over ϕ, we conclude that the state trajectory Figure 4-2(f) is less profitable than the one in Figure 4-2(e), leaving us with only the trajectories in Figure 4-2(d) and (e) as possible candidates for optimality. We summarize the discussion of this section in the following theorem.

Theorem 4.3.1 *The optimal value of the problem (GLQF-OVERFLOW) is given by* θ^*_{GLQF}.

4.4 The most likely path

As we have explained in the Sec. 4.1 we will prove a matching upper bound to the one in Proposition 4.2.1. This is sufficient to guarantee that the two scenarios identified in the proof of Proposition 4.2.1 (or equivalently the two optimal state trajectories of (GLQF-OVERFLOW)) are the most likely ways that queue Q^1 overflows. We summarize here these two most likely modes of overflow. We distinguish two cases:

Case 1: Suppose $\theta^*_{GLQF} = \inf_a \Lambda^{I*}_{GLQF}(a)/a$ holds. Let $a^* > 0$ the optimal solution of this optimization problem. The first queue builds up linearly with rate a^*, dur-

ing a period with duration U/a^*. During this period the empirical rates of the processes A^1, A^2 and B, are roughly equal to the optimal solution (x_1^*, x_2^*, x_3^*), respectively, of the optimization problem appearing in the definition of $\Lambda^{I*}_{GLQF}(a^*)$ (Eq. (4.4)). In this case the first queue is building up to an $O(U)$ level while the second queue builds up at a rate of x_2^*, in such a way that the level of the second queue is always below the level of the first, which results in the server allocating its entire capacity to the first queue. The trajectory in L^1-L^2 space is depicted in Figure 4-2(d).

Case 2: Suppose $\theta^*_{GLQF} = \inf_a \Lambda^{II*}_{GLQF}(a)/a$ holds. Let $a^* > 0$ the optimal solution of this optimization problem. Again, the first queue builds up linearly with rate a^*, during a period of duration U/a^*, and with the empirical rates of the processes A^1, A^2 and B being roughly equal to the optimal solution (x_1^*, x_2^*, x_3^*), respectively, of the optimization problem appearing in the definition of $\Lambda^{II*}_{GLQF}(a^*)$ (Eq. (4.5)). In this case both queues are building up, the first to an $O(U)$ level and the second to an $O(\beta U)$ level. The trajectory in L^1-L^2 space is depicted in Figure 4-2(e).

4.5 An Upper Bound

In this section we develop an upper bound on the probability $\mathbf{P}[L_0^1 > U]$. In particular, we will prove that as $U \to \infty$ we have $\mathbf{P}[L_0^1 > U] \leq e^{-\theta^*_{GLQF}U + o(U)}$, where $o(U)$ denotes functions with the property $\lim_{U\to\infty} \frac{o(U)}{U} = 0$.

Before we proceed into the proof of the upper bound, we derive an alternative expression for θ^*_{GLQF} which will be essential in the proof. In the next theorem, we will show that the calculation of θ^*_{GLQF} is equivalent to finding the maximum root of a convex function. The equivalence relies mainly on Lemma 3.6.2.

In the derivation of such an equivalence we will be using the same convention for the term *infinite root* that we introduced in Section 3.6. Namely, consider a convex function $f(u)$ with the property $f(0) = 0$. We define the *largest root* of $f(u)$ to be the

solution of the optimization problem $\sup_{u:f(u)<0} u$. If $f(\cdot)$ has negative derivative at $u = 0$, there are two cases: either $f(\cdot)$ has a single positive root or it stays below the horizontal axis $u = 0$, for all $u > 0$. In the latter, case we will say that $f(\cdot)$ has a root at $u = \infty$. On a notational remark, we will be denoting by $\Lambda^{I}_{GLQF}(\cdot)$ and $\Lambda^{II}_{GLQF}(\cdot)$, the convex duals of $\Lambda^{I*}_{GLQF}(\cdot)$ and $\Lambda^{II*}_{GLQF}(\cdot)$, respectively. Notice, that the latter are convex functions. For $\Lambda^{I*}_{GLQF}(a)$, convexity is implied by the fact that it is the value function of a convex optimization problem with a appearing only in the right hand side of the constraints. For $\Lambda^{II*}_{GLQF}(a)$, the same argument applies when we note the following reformulation

$$\begin{aligned}\Lambda^{II*}_{GLQF}(a) &= \inf_{\substack{x_1-\phi x_3=a \\ x_2-(1-\phi)x_3=\beta a \\ 0\le\phi<1}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)] \\ &= \inf_{\substack{x_1-x'_3=a \\ x_2-(x_3-x'_3)=\beta a \\ 0\le x'_3\le x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)].\end{aligned}$$

In preparation for the following theorem we prove the next monotonicity lemma.

Lemma 4.5.1 ***(Monotonicity)*** *Consider a random process $\{X_i;\ i \in \mathbb{Z}\}$ that satisfies Assumption A. Assume $X_i \ge 0,\ i \in \mathbb{Z}$. Then for all $\theta \le \theta'$ we have $\Lambda_X(\theta) \le \Lambda_X(\theta')$.*

Proof : $X_i \ge 0,\ i \in \mathbb{Z}$, implies $S^X_{1,n} \ge 0$ which in turn implies

$$\mathbf{E}[e^{\theta S^X_{1,n}}] \le \mathbf{E}[e^{\theta' S^X_{1,n}}],$$

for all $\theta \le \theta'$.

■

This Lemma, clearly applies to the arrival and service processes.

Theorem 4.5.2 *θ^*_{GLQF} is the largest positive root of the equation*

$$\Lambda_{GLQF}(\theta) \stackrel{\triangle}{=} \max[\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)] = 0, \tag{4.12}$$

where $\Lambda^I_{GLQF}(\cdot)$ is the convex dual of $\Lambda^{I*}_{GLQF}(\cdot)$ and is given by

$$\Lambda^I_{GLQF}(\theta) = \inf_{u\leq 0}[\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta + u\beta)], \tag{4.13}$$

and $\Lambda^{II}_{GLQF}(\cdot)$ is the convex dual of $\Lambda^{II*}_{GLQF}(\cdot)$ and for $\theta \geq 0$ satisfies

$$\Lambda^{II}_{GLQF}(\theta) = \inf_{u\geq 0}[\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \max(\Lambda_B(-u), \Lambda_B(-\theta + u\beta))]. \tag{4.14}$$

Proof : Let us first calculate $\Lambda^I_{GLQF}(\cdot)$ and $\Lambda^{II}_{GLQF}(\cdot)$ by using convex duality. We have

$$\begin{aligned}
\Lambda^I_{GLQF}(\theta) &= \sup_a[\theta a - \Lambda^{I*}_{GLQF}(a)] \\
&= \sup_a \sup_{\substack{x_1 - x_3 = a \\ x_2 \leq \beta(x_1 - x_3)}} [\theta a - \Lambda^*_{A^1}(x_1) - \Lambda^*_{A^2}(x_2) - \Lambda^*_B(x_3)] \\
&= \sup_a \sup_{\substack{x_1 - x_3 = a \\ x_2 \leq \beta(x_1 - x_3)}} [\theta(x_1 - x_3) - \Lambda^*_{A^1}(x_1) - \Lambda^*_{A^2}(x_2) - \Lambda^*_B(x_3)] \\
&= \sup_{x_2 \leq \beta(x_1 - x_3)} [\theta(x_1 - x_3) - \Lambda^*_{A^1}(x_1) - \Lambda^*_{A^2}(x_2) - \Lambda^*_B(x_3)] \\
&= \inf_{u\leq 0} \sup_{x_1, x_2, x_3} [\theta(x_1 - x_3) - \Lambda^*_{A^1}(x_1) - \Lambda^*_{A^2}(x_2) - \Lambda^*_B(x_3) \\
&\qquad\qquad - u(\beta x_1 - \beta x_3 - x_2)] \\
&= \inf_{u\leq 0}[\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta + u\beta)].
\end{aligned}$$

Similarly,

$$\Lambda^{II}_{GLQF}(\theta) = \sup_a[\theta a - \Lambda^{II*}_{GLQF}(a)]$$

$$
\begin{aligned}
&= \sup_{a} \sup_{\substack{x_1-\phi x_3=a \\ x_2-(1-\phi)x_3=\beta(x_1-\phi x_3) \\ 0\le\phi<1}} [\theta a - \Lambda^*_{A^1}(x_1) - \Lambda^*_{A^2}(x_2) - \Lambda^*_B(x_3)] \\
&= \inf_u \sup_{\substack{x_1,x_2,x_3 \\ 0\le\phi<1}} [\theta(x_1-\phi x_3) - \Lambda^*_{A^1}(x_1) - \Lambda^*_{A^2}(x_2) - \Lambda^*_B(x_3) \\
&\qquad\qquad + u(x_2 - \beta x_1 + (\beta\phi+\phi-1)x_3)] \\
&= \inf_u [\Lambda_{A^1}(\theta-u\beta) + \Lambda_{A^2}(u) + \sup_{0\le\phi<1} \Lambda_B(-\theta\phi + (\beta\phi+\phi-1)u)] \\
&= \inf_u [\Lambda_{A^1}(\theta-u\beta) + \Lambda_{A^2}(u) + \max(\Lambda_B(-u), \Lambda_B(-\theta+u\beta))] \\
&= \inf_{u\ge 0} [\Lambda_{A^1}(\theta-u\beta) + \Lambda_{A^2}(u) + \max(\Lambda_B(-u), \Lambda_B(-\theta+u\beta))].
\end{aligned}
$$

In the fifth equality above, we have used the monotonicity of $\Lambda_B(\cdot)$ (see Lemma 4.5.1), and the fact that the argument $-\theta\phi + (\beta\phi+\phi-1)u$ is linear in ϕ, thus, taking its maximum value at either $\phi = 0$ or $\phi = 1$. For the sixth equality above, notice that because $\Lambda_B(\cdot)$ is non-decreasing it holds

$$
\begin{aligned}
&\Lambda_{A^1}(\theta-u\beta) + \Lambda_{A^2}(u) + \max(\Lambda_B(-u), \Lambda_B(-\theta+u\beta)) = \\
&\qquad = \begin{cases} \Lambda_{A^1}(\theta-u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u) & \text{if } u < \frac{\theta}{1+\beta} \\ \Lambda_{A^1}(\theta-u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta+u\beta) & \text{if } u \ge \frac{\theta}{1+\beta}, \end{cases}
\end{aligned} \tag{4.15}
$$

since at the upper branch $-u > -\theta + u\beta$ and at the lower branch $-u \le -\theta + u\beta$. Differentiating the above at $u = 0$, and for $\theta \ge 0$, we obtain

$$
\underbrace{-\beta\dot{\Lambda}_{A^1}(\theta)}_{\le 0} + \underbrace{\dot{\Lambda}_{A^2}(0) - \dot{\Lambda}_B(0)}_{\overset{(3.1)}{\le} 0} \le 0,
$$

which implies (by convexity) that the infimum is achieved at some $u \ge 0$. Thus, the infimum over unrestricted u has to be the same with the infimum over $u \ge 0$.

Using the result of Lemma 3.6.2, $\rho_1 \overset{\triangle}{=} \inf_a \frac{1}{a}\Lambda^{I*}_{GLQF}(a)$ is the largest positive root of $\Lambda^I_{GLQF}(\theta) = 0$ (it is not hard to verify that this equation has a positive, possibly, infinite root). Similarly, $\rho_2 \overset{\triangle}{=} \inf_a \frac{1}{a}\Lambda^{II*}_{GLQF}(a)$ is the largest positive root of

$\Lambda^{II}_{GLQF}(\theta) = 0$. By Equation (4.3), $\theta^*_{GLQF} = \min(\rho_1, \rho_2)$. This implies that θ^*_{GLQF} is the largest positive root of the equation $\max[\Lambda^{I}_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)] = 0$.

■

We next prove the upper bound for the overflow probability.

Proposition 4.5.3 ***(GLQF Upper Bound)*** *Under the GLQF policy, assuming that the arrival and service processes satisfy Assumptions A and E, the steady-state queue length, L^1, of queue Q^1, at an arbitrary time slot satisfies*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \leq -\theta^*_{GLQF}. \tag{4.16}$$

Proof : Without loss of generality we derive an upper bound for $\mathbf{P}[L^1_0 > U]$. We will restrict ourselves to sample paths with $L^1_0 > 0$ since the remaining sample paths, with $L^1_0 = 0$, do not contribute to the probability $\mathbf{P}[L^1_0 > U]$.

Consider a busy period for the system that starts at some time $-n < 0$ ($L^1_{-n} = L^2_{-n} = 0$), and has not ended until time 0. Such a time $-n$ exists due to the stability condition (3.1). Note that since the system is busy in the interval $[-n, 0]$, the server works at capacity and therefore serves B_i customers at slot i, for $i \in [-n, 0]$. We will partition the set of sample paths, with $L^1_0 > 0$, in three subsets Ω_1, Ω_2 and Ω_3. The first subset, Ω_1, contains all sample paths at which only Type 1 customers get serviced in the interval $[-n, 0]$. As a consequence,

$$L^1_{-k} = S^{A^1}_{-n,-k-1} - S^{B}_{-n,-k-1},\ L^2_{-k} = S^{A^2}_{-n,-k-1}, \text{ and } \beta L^1_{-k} \geq L^2_{-k}, \qquad \forall k \in [0, n],$$

which implies

$$L^1_0 = S^{A^1}_{-n,-1} - S^{B}_{-n,-1}, \quad \text{and} \quad \beta(S^{A^1}_{-n,-1} - S^{B}_{-n,-1}) \geq S^{A^2}_{-n,-1}.$$

Thus

$$\mathbf{P}[L^1_0 > U \text{ and } \Omega_1] \leq$$

$$\leq \mathbf{P}[\exists n \geq 0 \text{ s.t. } S^{A^1}_{-n,-1} - S^B_{-n,-1} > U \text{ and } \beta(S^{A^1}_{-n,-1} - S^B_{-n,-1}) \geq S^{A^2}_{-n,-1}]$$
$$= \mathbf{P}[\max_{\{n\geq 0:\ \beta(S^{A^1}_{-n,-1} - S^B_{-n,-1}) \geq S^{A^2}_{-n,-1}\}} (S^{A^1}_{-n,-1} - S^B_{-n,-1}) > U]. \tag{4.17}$$

The second subset, Ω_2, contains sample paths at which Type 1 customers do not receive the entire capacity, and $\beta L^1_0 \leq L^2_0$. That is, there exists a $\phi \in [0,1]$ such that Type 1 customers receive only a ϕ fraction of the total capacity ($\phi S^B_{-n,-1}$). Then we have

$$\begin{aligned}
&\mathbf{P}[L^1_0 > U \text{ and } \Omega_2] \leq \\
&\quad \leq \mathbf{P}[\exists n \geq 0, 0 \leq \phi < 1, \text{ s.t. } S^{A^1}_{-n,-1} - \phi S^B_{-n,-1} > U \text{ and} \\
&\qquad\qquad \beta(S^{A^1}_{-n,-1} - \phi S^B_{-n,-1}) \leq S^{A^2}_{-n,-1} - (1-\phi) S^B_{-n,-1}] \\
&\quad = \mathbf{P}[\max_{\{n\geq 0,\ 0\leq\phi<1:\ \beta(S^{A^1}_{-n,-1} - \phi S^B_{-n,-1}) \leq S^{A^2}_{-n,-1} - (1-\phi) S^B_{-n,-1}\}} (S^{A^1}_{-n,-1} - \phi S^B_{-n,-1}) > U].
\end{aligned} \tag{4.18}$$

Finally, the third subset, Ω_3 contains sample paths at which Type 1 customers do not receive the entire capacity, and $\beta L^1_0 \geq L^2_0$. Then there exists $k \in [0,n]$, such that the interval $[-k,0]$ is the maximal interval that only Type 1 customers get serviced. That is, $\beta L^1_{-i} \geq L^2_{-i}$, $i \in [0,k-1]$, and $\beta L^1_{-k} \leq L^2_{-k}$. Since Type 1 customers do not receive the entire capacity, there exists $0 \leq \phi < 1$ such that $L^1_{-k} = S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1}$. Since $\beta L^1_{-k} \leq L^2_{-k}$, we have

$$\beta(S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1}) \leq S^{A^2}_{-n,-k-1} - (1-\phi) S^B_{-n,-k-1}. \tag{4.19}$$

Now, due to the way we defined k we have $L^1_{-i} = L^1_{-k} + S^{A^1}_{-k,-i-1} - S^B_{-k,-i-1}$, $i \in [0,k-1]$, and the inequality $\beta L^1_{-i} \geq L^2_{-i}$ becomes

$$\beta(S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1} + S^{A^1}_{-k,-i-1} - S^B_{-k,-i-1}) \geq S^{A^2}_{-n,-k-1} - (1-\phi) S^B_{-n,-k-1} + S^{A^2}_{-k,-i-1},$$

which by (4.19) implies

$$\beta(S^{A^1}_{-k,-i-1} - S^{B}_{-k,-i-1}) \geq S^{A^2}_{-k,-i-1}, \qquad i \in [0, k-1].$$

Thus,

$$\begin{aligned}
&\mathbf{P}[L_0^1 > U \text{ and } \Omega_3] \leq \\
&\leq \mathbf{P}[\exists n \geq 0, 0 \leq k \leq n, 0 \leq \phi < 1, \text{ s.t. } S^{A^1}_{-n,-k-1} - \phi S^{B}_{-n,-k-1} + S^{A^1}_{-k,-1} - S^{B}_{-k,-1} > U \\
&\qquad \text{and } \beta(S^{A^1}_{-n,-k-1} - \phi S^{B}_{-n,-k-1}) \leq S^{A^2}_{-n,-k-1} - (1-\phi) S^{B}_{-n,-k-1} \\
&\qquad \text{and } \beta(S^{A^1}_{-k,-1} - S^{B}_{-k,-1}) \geq S^{A^2}_{-k,-1}] \\
&\leq \mathbf{P}[\max_{\substack{n \geq 0, 0 \leq k \leq n, 0 \leq \phi < 1 \\ \beta(S^{A^1}_{-n,-k-1} - \phi S^{B}_{-n,-k-1}) \leq S^{A^2}_{-n,-k-1} - (1-\phi) S^{B}_{-n,-k-1} \\ \beta(S^{A^1}_{-k,-1} - S^{B}_{-k,-1}) \geq S^{A^2}_{-k,-1}}} (S^{A^1}_{-n,-k-1} - \phi S^{B}_{-n,-k-1} + S^{A^1}_{-k,-1} - S^{B}_{-k,-1}) > U].
\end{aligned} \tag{4.20}$$

Let us now define

$$L^{I}_{GLQF} \triangleq \max_{\{n \geq 0:\ \beta(S^{A^1}_{-n,-1} - S^{B}_{-n,-1}) \geq S^{A^2}_{-n,-1}\}} (S^{A^1}_{-n,-1} - S^{B}_{-n,-1}),$$

$$L^{II}_{GLQF} \triangleq \max_{\{n \geq 0,\ 0 \leq \phi < 1:\ \beta(S^{A^1}_{-n,-1} - \phi S^{B}_{-n,-1}) \leq S^{A^2}_{-n,-1} - (1-\phi) S^{B}_{-n,-1}\}} (S^{A^1}_{-n,-1} - \phi S^{B}_{-n,-1}),$$

and

$$L^{III}_{GLQF} \triangleq \max_{\substack{n \geq 0, 0 \leq k \leq n, 0 \leq \phi < 1 \\ \beta(S^{A^1}_{-n,-k-1} - \phi S^{B}_{-n,-k-1}) \leq S^{A^2}_{-n,-k-1} - (1-\phi) S^{B}_{-n,-k-1} \\ \beta(S^{A^1}_{-k,-1} - S^{B}_{-k,-1}) \geq S^{A^2}_{-k,-1}}} (S^{A^1}_{-n,-k-1} - \phi S^{B}_{-n,-k-1} + S^{A^1}_{-k,-1} - S^{B}_{-k,-1}),$$

which after bringing the constraints in the objective function become

$$L^{I}_{GLQF} \triangleq \max_{n \geq 0} \inf_{u \geq 0} [(1 + u\beta) S^{A^1}_{-n,-1} - u S^{A^2}_{-n,-1} + (-1 - \beta u) S^{B}_{-n,-1}], \tag{4.21}$$

$$L^{II}_{GLQF} \stackrel{\triangle}{=} \max_{\substack{n\geq 0\\ 0\leq\phi<1}} \inf_{u\geq 0}[(1-u\beta)S^{A^1}_{-n,-1} + uS^{A^2}_{-n,-1} + (-\phi+u\beta\phi-u+u\phi)S^{B}_{-n,-1}], \tag{4.22}$$

and

$$L^{III}_{GLQF} \stackrel{\triangle}{=} \max_{\substack{n\geq 0\\ 0\leq k\leq n\\ 0\leq\phi<1}} \left\{ \inf_{u_1\geq 0}[(1-u_1\beta)S^{A^1}_{-n,-k-1} + u_1 S^{A^2}_{-n,-k-1} + (-\phi+u_1\beta\phi-u_1+ \right.$$

$$\left. u_1\phi)S^{B}_{-n,-k-1}] + \inf_{u_2\geq 0}[(1+u_2\beta)S^{A^1}_{-k,-1} - u_2 S^{A^2}_{-k,-1} + (-1-u_2\beta)S^{B}_{-k,-1}] \right\}. \tag{4.23}$$

Next, we will first upper bound the moment generating functions of L^{I}_{GLQF}, L^{II}_{GLQF} and L^{III}_{GLQF}. For L^{I}_{GLQF} and for $\theta \geq 0$ we have

$$\begin{aligned}
&\mathbf{E}[e^{\theta L^{I}_{GLQF}}] \\
&\leq \sum_{n\geq 0} \mathbf{E}[\exp\{\theta \inf_{u\geq 0}[(1+u\beta)S^{A^1}_{-n,-1} - uS^{A^2}_{-n,-1} + (-1-\beta u)S^{B}_{-n,-1}]\}\,] \\
&\leq \sum_{n\geq 0} \inf_{u\geq 0} \mathbf{E}[\exp\{\theta[(1+u\beta)S^{A^1}_{-n,-1} - uS^{A^2}_{-n,-1} + (-1-\beta u)S^{B}_{-n,-1}]\}\,] \\
&\leq \sum_{n\geq 0} e^{n(\inf_{u\geq 0}[\Lambda_{A^1}(\theta+\theta u\beta)+\Lambda_{A^2}(-u\theta)+\Lambda_B(-\theta-u\beta\theta)]+\epsilon_1)} \\
&\leq K^I(\theta,\epsilon_1) \qquad \text{if } \Lambda^{I}_{GLQF}(\theta) < 0.
\end{aligned} \tag{4.24}$$

In the third inequality above we have used the LDP for the arrival and service processes. In the last inequality above, when the exponent is negative (for sufficiently small ϵ_1), the infinite geometric series converges to a constant, with respect to n, $K^I(\theta,\epsilon_1)$. Also, in the last inequality, we have made the substitution $u := -\theta u$ in the expression in the exponent and used the definition of $\Lambda^{I}_{GLQF}(\theta)$ (Eq. (4.13)).

Similarly, for L^{II}_{GLQF} and for $\theta \geq 0$ we have

$$\begin{aligned}
&\mathbf{E}[e^{\theta L^{II}_{GLQF}}] \\
&\leq \sum_{n\geq 0} \mathbf{E}[\exp\{\theta \max_{0\leq\phi<1} \inf_{u\geq 0}[(1-u\beta)S^{A^1}_{-n,-1} + uS^{A^2}_{-n,-1} + (-\phi+u\beta\phi-u+u\phi)S^{B}_{-n,-1}]\}\,]
\end{aligned}$$

$$\leq \sum_{n\geq 0} \inf_{u\geq 0} \mathbf{E}[\exp\{\theta \max_{0\leq\phi<1}[(1-u\beta)S^{A^1}_{-n,-1} + uS^{A^2}_{-n,-1} + (-\phi+u\beta\phi-u+u\phi)S^{B}_{-n,-1}]\}\,]$$

$$\leq \sum_{n\geq 0} \inf_{u\geq 0}\left(e^{n([\Lambda_{A^1}(\theta-\theta u\beta)+\Lambda_{A^2}(u\theta)+\Lambda_B(-\theta u)]+\epsilon_2')} + e^{n([\Lambda_{A^1}(\theta-\theta u\beta)+\Lambda_{A^2}(u\theta)+\Lambda_B(-\theta+\theta u\beta)]+\epsilon_2'')}\right)$$

$$\leq 2\sum_{n\geq 0} e^{n(\inf_{u\geq 0}[\Lambda_{A^1}(\theta-\theta u\beta)+\Lambda_{A^2}(u\theta)+\max(\Lambda_B(-\theta u),\Lambda_B(-\theta+\theta u\beta))]+\epsilon_2)}$$

$$\leq K^{II}(\theta,\epsilon_2), \qquad \text{if } \Lambda^{II}_{GLQF}(\theta) < 0. \tag{4.25}$$

In the third inequality above, the expression to be maximized over ϕ is linear, thus, the maximum is achieved at either $\phi = 0$ or $\phi = 1$, which implies that we can upper bound it by the sum of the terms for $\phi = 0$ and $\phi = 1$.

Also, for L^{III}_{GLQF} and for $\theta \geq 0$ we have

$$\mathbf{E}[e^{\theta L^{III}_{GLQF}}]$$

$$\leq \sum_{n\geq 0}\sum_{0\leq k\leq n} \mathbf{E}\Bigg[\exp\Bigg\{\theta \max_{0\leq\phi<1}\inf_{u_1\geq 0}[(1-u_1\beta)S^{A^1}_{-n,-k-1} + u_1 S^{A^2}_{-n,-k-1} + (-\phi+u_1\beta\phi -$$

$$u_1+u_1\phi)S^{B}_{-n,-k-1}] + \inf_{u_2\geq 0}[(1+u_2\beta)S^{A^1}_{-k,-1} - u_2 S^{A^2}_{-k,-1} + (-1-u_2\beta)S^{B}_{-k,-1}]\Bigg\}\Bigg]$$

$$\leq \sum_{n\geq 0}\sum_{0\leq k\leq n} \inf_{u_1,u_2\geq 0} \mathbf{E}\Bigg[\exp\Bigg\{\theta \max_{0\leq\phi<1}[(1-u_1\beta)S^{A^1}_{-n,-k-1} + u_1 S^{A^2}_{-n,-k-1} + (-\phi+u_1\beta\phi -$$

$$u_1+u_1\phi)S^{B}_{-n,-k-1}] + [(1+u_2\beta)S^{A^1}_{-k,-1} - u_2 S^{A^2}_{-k,-1} + (-1-u_2\beta)S^{B}_{-k,-1}]\Bigg\}\Bigg]$$

$$\leq \sum_{n\geq 0}\sum_{0\leq k\leq n} \inf_{u_1,u_2\geq 0}\Bigg[e^{(n-k)(\Lambda_{A^1}(\theta-\theta u_1\beta)+\Lambda_{A^2}(u_1\theta)+\Lambda_B(-\theta u_1)+\epsilon_3')} +$$

$$e^{(n-k)(\Lambda_{A^1}(\theta-\theta u_1\beta)+\Lambda_{A^2}(u_1\theta)+\Lambda_B(-\theta+\theta u_1\beta)+\epsilon_3'')}\Bigg]e^{k(\Lambda_{A^1}(\theta+\theta u_2\beta)+\Lambda_{A^2}(-u_2\theta)+\Lambda_B(-\theta-\theta u_2\beta)+\epsilon_3''')}$$

$$\leq 2\sum_{n\geq 0}\sum_{0\leq k\leq n} e^{(n-k)(\Lambda^{II}(\theta)+\hat{\epsilon}_3)}e^{k(\Lambda^{I}(\theta)+\hat{\epsilon}_3')}$$

$$\leq 2\sum_{n\geq 0} n e^{n(\Lambda^{II}(\theta)+\hat{\epsilon}_3)} + 2\sum_{n\geq 0} n e^{n(\Lambda^{I}(\theta)+\hat{\epsilon}_3')}$$

$$\leq K^{III}(\theta,\epsilon_3), \qquad \text{if } \max(\Lambda^{I}_{GLQF}(\theta),\Lambda^{II}_{GLQF}(\theta)) < 0. \tag{4.26}$$

In the third inequality above we have used the LDP for arrival and service processes, as well as Assumption E. Concerning the maximization over ϕ, we have used the same argument as in Eq. (4.25). In the fifth inequality above, since the exponent is linear in k, the maximum over k is either at $k = 0$ or at $k = n$. Thus, we bound the term by the sum of the terms for $k = 0$ and $k = n$. Finally, for the last inequality, both series converge to a constant if both their exponents are negative, which requires $\max(\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) < 0$.

To summarize (4.24), (4.25) and (4.26), the moment generating functions of L^I_{GLQF}, L^{II}_{GLQF} and L^{III}_{GLQF} are upper bounded by some constant $K(\theta, \epsilon_1, \epsilon_2, \epsilon_3)$ if $\max(\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) < 0$, where $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ are sufficiently small. We can now apply the Markov inequality to obtain (using Eqs. (4.17), (4.18) and (4.20))

$$\begin{aligned}
&\mathbf{P}[L^1_0 > U] \\
&\leq \mathbf{P}[L^1_0 > U \text{ and Case 1}] + \mathbf{P}[L^1_0 > U \text{ and Case 2}] + \mathbf{P}[L^1_0 > U \text{ and Case 3}] \\
&\leq \left(\mathbf{E}[e^{\theta \Lambda^I(\theta)}] + \mathbf{E}[e^{\theta \Lambda^{II}(\theta)}] + \mathbf{E}[e^{\theta \Lambda^{III}(\theta)}] \right) e^{-\theta U} \\
&\leq 3K(\theta, \epsilon_1, \epsilon_2, \epsilon_3) e^{-\theta U} \qquad \text{if } \max(\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) < 0.
\end{aligned}$$

Taking the limit as $U \to \infty$ and minimizing the upper bound with respect to $\theta \geq 0$, in order to obtain the tightest bound, we have

$$\lim_{U \to \infty} \frac{1}{U} \log \mathbf{P}[L^1_0 > U] \leq - \sup_{\{\theta \geq 0:\ \max(\Lambda^I(\theta), \Lambda^{II}(\theta)) < 0\}} \theta.$$

The right hand side of the above is equal to $-\theta^*_{GLQF}$ by Theorem 4.5.2.

■

4.6 Main Results

In this section we summarize our main results for the GLQF policy.

Combining Propositions 4.2.1 and 4.5.3 we obtain the following main theorem.

An exact characterization of the *most likely ways* that lead to overflow were discussed in Section 4.4.

Theorem 4.6.1 ***(GLQF Main)*** *Under the GLQF policy, assuming that the arrival and service processes satisfy Assumptions A, D, and E, the steady-state queue length, L^1, of queue Q^1, at an arbitrary time slot satisfies*

$$\lim_{U\to\infty}\frac{1}{U}\log\mathbf{P}[L^1>U]=-\theta^*_{GLQF}, \tag{4.27}$$

where θ^*_{GLQF} is given by

$$\theta^*_{GLQF}=\min\left[\inf_{a>0}\frac{1}{a}\Lambda^{I*}_{GLQF}(a),\inf_{a>0}\frac{1}{a}\Lambda^{II*}_{GLQF}(a)\right], \tag{4.28}$$

and the functions $\Lambda^{I*}_{GLQF}(\cdot)$ and $\Lambda^{II*}_{GLQF}(\cdot)$ are defined as follows

$$\Lambda^{I*}_{GLQF}(a)\stackrel{\triangle}{=}\inf_{\substack{x_1-x_3=a\\ x_2\le\beta(x_1-x_3)}}[\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)], \tag{4.29}$$

and

$$\Lambda^{II*}_{GLQF}(a)\stackrel{\triangle}{=}\inf_{\substack{x_1-\phi x_3=a\\ x_2-(1-\phi)x_3=\beta a\\ 0\le\phi<1}}[\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)]. \tag{4.30}$$

It should be noted that the performance of strict priority policies, which is characterized by Corollary 3.7.2, can be also obtained as a corollary of the above theorem. We obtain the performance of strict priority to Type 2 (P_2) when $\beta=0$, and the performance of strict priority to Type 1 (P_1) when $\beta=\infty$. It is not hard to verify that the result is identical to Corollary 3.7.2. The above Theorem indicates that the calculation of the overflow probabilities involves the solution of a convex optimization problem. In Section 4.5, and for the purposes of proving Proposition 4.5.3, we proved in Theorem 4.5.2 that the exponent of the overflow probability can also be obtained as the maximum root of a convex function. This may be easier to do in some cases.

Here, we restate this latter result, simplifying the expression for $\Lambda_{GLQF}(\cdot)$.

Theorem 4.6.2 *θ^*_{GLQF} is the largest positive root of the equation*

$$\Lambda_{GLQF}(\theta) = \max\{\Lambda_{A^1}(\theta) + \Lambda_B(-\theta), \inf_{0\le u\le \frac{\theta}{1+\beta}} [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)]\} = 0. \tag{4.31}$$

Proof : Due to Theorem 4.5.2 it suffices to prove that the expression in (4.31) is equal to $\max[\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)]$. Recall the definitions of $\Lambda^I_{GLQF}(\theta)$ in (4.13) and of $\Lambda^{II}_{GLQF}(\theta)$ in (4.14). Recall also the expression in (4.15) for the objective function of the optimization problem corresponding to $\Lambda^{II}_{GLQF}(\theta)$. Let now u^* be the optimal solution of the optimization problem in the definition of $\Lambda^{II}_{GLQF}(\theta)$. We distinguish two cases:

<u>Case 1:</u> where $u^* \ge \frac{\theta}{1+\beta}$. Then, notice that u^* is also the minimizer of the objective function in the definition of $\Lambda^I_{GLQF}(\theta)$. Thus, due to convexity, the constraint $u \le 0$ is tight for the problem corresponding to $\Lambda^I_{GLQF}(\theta)$, and

$$\max(\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) = \Lambda_{A^1}(\theta) + \Lambda_B(-\theta), \qquad \text{if } u^* \ge \tfrac{\theta}{1+\beta}. \tag{4.32}$$

But,

$$\begin{aligned}
&\inf_{0\le u\le \frac{\theta}{1+\beta}} [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)] \\
&\qquad \le [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)]_{u=\frac{\theta}{1+\beta}} \\
&\qquad = [\Lambda_{A^1}(\tfrac{\theta}{1+\beta}) + \Lambda_{A^2}(\tfrac{\theta}{1+\beta}) + \Lambda_B(-\tfrac{\theta}{1+\beta})] \\
&\qquad = [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta + u\beta)]_{u=\frac{\theta}{1+\beta}} \\
&\qquad \le [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta + u\beta)]_{u=0} \\
&\qquad = \Lambda_{A^1}(\theta) + \Lambda_B(-\theta).
\end{aligned}$$

In the second inequality above we have used the assumption $u^* \ge \frac{\theta}{1+\beta}$ and

convexity. Therefore, combining it with (4.32) we obtain

$$\max(\Lambda^{I}_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) = \max\{\Lambda_{A^1}(\theta) + \Lambda_B(-\theta), \inf_{0\le u\le \frac{\theta}{1+\beta}} [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)] \} = \Lambda_{GLQF}(\theta) \qquad \text{if } u^* \ge \tfrac{\theta}{1+\beta}. \tag{4.33}$$

Case 2: where $0 \le u^* < \frac{\theta}{1+\beta}$. To conclude the proof we need to show that $\max(\Lambda^{I}_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta))$ is not $\Lambda^{I}_{GLQF}(\theta)$ when the optimal solution, of the optimization problem appearing in the definition of $\Lambda^{I}_{GLQF}(\theta)$, is some $\hat{u} < 0$. Let us, indeed, assume that this optimal solution is some $\hat{u} < 0$. Then, for all $u \in [0, \frac{\theta}{1+\beta})$ (hence for u^*) we have

$$\begin{aligned}\Lambda^{I}_{GLQF}(\theta) &= [\Lambda_{A^1}(\theta - \hat{u}\beta) + \Lambda_{A^2}(\hat{u}) + \Lambda_B(-\theta + \hat{u}\beta)] \\ &\le [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta + u\beta)] \\ &\le [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)],\end{aligned}$$

where in the last inequality we have used the fact that $u < \frac{\theta}{1+\beta}$ which implies (see also (4.15)) $\Lambda_B(-u) \ge \Lambda_B(-\theta + u\beta)$. Therefore, for $0 \le u^* \le \frac{\theta}{1+\beta}$ also, we have

$$\max(\Lambda^{I}_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) = \max\{\Lambda_{A^1}(\theta) + \Lambda_B(-\theta), \inf_{0\le u\le \frac{\theta}{1+\beta}} [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)] \} = \Lambda_{GLQF}(\theta).$$

■

The results of this Theorem can be also specialized to the case of priority policies, to obtain the characterization of Corollary 9.2 3.7.4.

We conclude this section, noting that, by symmetry, all the results obtained here can be easily adapted (it suffices to substitute everywhere $1 := 2$, $2 := 1$, and $\beta = \frac{1}{\beta}$)

to estimate the overflow probability of the second queue and characterize the most likely ways that it builds up.

4.7 A Comparison

In this section we compare the overflow probabilities achieved by the GPS and the GLQF policy.

Let π be an arbitrary work-conserving policy to allocate the capacity of the server to the two queues Q^1 and Q^2, and Π the set of all work-conserving policies π. Let L^1 and L^2 denote the queue lengths of Q^1 and Q^2, respectively, at an arbitrary time slot, when the system operates under π. Let us now define θ^π the vector $(\theta_1^\pi, \theta_2^\pi)$ where

$$\theta_1^\pi = \lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \qquad \text{and} \qquad \theta_2^\pi = \lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^2 > U]. \quad (4.34)$$

The GPS policy analyzed in Chapter 3 is a parametric policy with performance depending on the parameter ϕ_1. To make this dependence explicit we will be using the notation GPS(ϕ_1). Also, the GLQF policy analyzed in Section 4.1 is a parametric policy with performance depending on the parameter β. For the same reason we will be using the notation GLQF(β). Special cases of a work-conserving policy π are the GPS(ϕ_1) policy, the GLQF(β) policy, the strict priority to Type 1 policy (P_1 policy), and the strict priority to Type 2 policy (P_2 policy). Using Theorems 3.7.1, 4.6.1 and Corollary 3.7.2 one can readily obtain the corresponding θ^π for the policies GPS(ϕ_1), GLQF(β), P_1 and P_2.

It is intuitively obvious that

$$\theta^{P_1} = (\max_{\pi\in\Pi} \theta_1^\pi, \min_{\pi\in\Pi} \theta_2^\pi) \qquad \text{and} \qquad \theta^{P_2} = (\min_{\pi\in\Pi} \theta_1^\pi, \max_{\pi\in\Pi} \theta_2^\pi).$$

In Figure 4-3 we plot $\theta^{GPS(\phi_1)}$ as ϕ_1 varies in $[0, 1]$, and $\theta^{GLQF(\beta)}$ as β varies in $[0, \infty)$. For simplicity the calculations were performed with the arrival and service processes being Bernoulli (we say that a process $\{X_i;\ i \in \mathbb{Z}\}$ is Bernoulli with parameter p,

denoted by $X \sim$Ber(p), when X_i are i.i.d. and $X_i = 1$ with probability p and $X_i = 0$ with probability $1 - p$). Also, for the calculations we used the expressions for θ^*_{GPS}

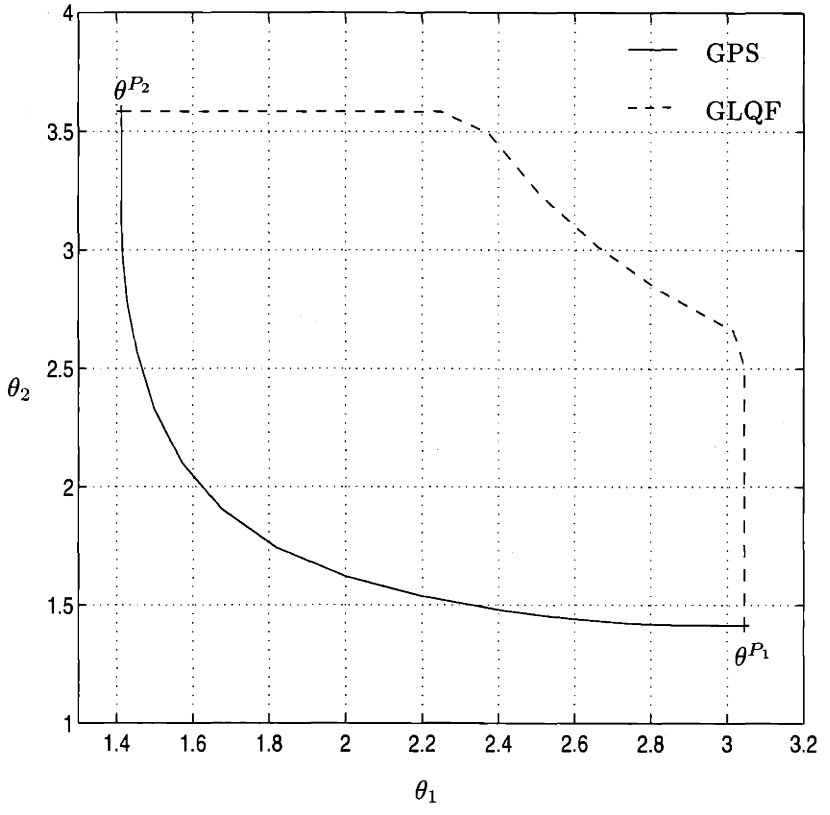

Figure 4-3: The performance $\theta^{GPS(\phi_1)}$ of the GPS(ϕ_1) policy as ϕ_1 varies in $[0, 1]$, and the performance $\theta^{GLQF(\beta)}$ of the GLQF(β) policy as β varies in $[0, \infty)$, when $A^1 \sim$Ber(0.3), $A^2 \sim$Ber(0.2) and $B \sim$Ber(0.9).

and θ^*_{GLQF} given in Theorems 3.7.3 and 4.6.2, respectively, because they were more efficient to perform numerically than the equivalent expressions in Theorems 3.7.1 and 4.6.1. Note that $\theta^{P_1} = \theta^{GPS(1)} = \theta^{GLQF(\infty)}$ and that $\theta^{P_2} = \theta^{GPS(0)} = \theta^{GLQF(0)}$.

Figure 4-3 indicates that the GLQF curve dominates the GPS curve, i.e., the GLQF policy achieves smaller overflow probabilities than the GPS policy. The ques-

tion that arises is whether this depends on the particular distributions and parameters chosen in the figure or is a general property. In the sequel we show that the latter is the case, that is, for all arrival and service processes that our analysis holds (processes satisfying Assumptions A, D, and E) the GLQF curve dominates the GPS curve. The intuition behind this result is that the GLQF policy, which adaptively depends on the current queue lengths, allocates capacity to the queue that builds up, thus, achieving smaller overflow probabilities than the GPS policy which is static. This suggests than when one has to deal with delay insensitive traffic (i.e., when there are no delay constraints) GLQF is more suitable than GPS. On the other hand, GLQF does not have the fairness property of GPS, that is it may allow a bursty class of traffic to be using all the available capacity until the backlog of the other class reaches the level of the bursty one.

Let us first formally define the term *the GLQF curve dominates the GPS curve.*

Definition 4.7.1
We say that the GLQF curve dominates the GPS curve *when there does not exist a pair of* $\phi_1 \in [0,1]$ *and* $\beta \in [0,\infty)$ *satisfying* $\theta_1^{GPS(\phi_1)} > \theta_1^{GLQF(\beta)}$ *and* $\theta_2^{GPS(\phi_1)} > \theta_2^{GLQF(\beta)}$.

In order to establish that the GLQF curve dominates the GPS curve, we need to prove the three lemmata that follow.

Lemma 4.7.2 *If* $\phi_1 \leq \phi_1'$ *we have*

$$\theta_1^{GPS(\phi_1)} \leq \theta_1^{GPS(\phi_1')} \qquad \text{and} \qquad \theta_2^{GPS(\phi_1)} \geq \theta_2^{GPS(\phi_1')}.$$

Proof : We only prove the first relation. The second can be obtained by a symmetrical argument. We use the result of Theorem 3.7.3. Note that $\phi_1 \leq \phi_1'$, implies $\phi_2' = (1-\phi_1') \leq \phi_2 = (1-\phi_1)$. Thus, by Lemma 4.5.1, for all $u, \theta \geq 0$ we have that $\Lambda_B(-\theta+\phi_2 u) \geq \Lambda_B(-\theta+\phi_2' u)$, which by Thm. 3.7.3 implies $\Lambda_{GPS(\phi_1)}(\theta) \geq \Lambda_{GPS(\phi_1')}(\theta)$ for all θ. Therefore, by convexity, for θ^*_{GPS}, as it is defined in Thm. 3.7.3, we have

$\theta^*_{GPS(\phi_1)} \leq \theta^*_{GPS(\phi'_1)}$.

■

A similar property is proven for the GLQF policy.

Lemma 4.7.3 *If $\beta \leq \beta'$ we have*

$$\theta_1^{GLQF(\beta)} \leq \theta_1^{GLQF(\beta')} \qquad \text{and} \qquad \theta_2^{GLQF(\beta)} \geq \theta_2^{GLQF(\beta')}.$$

Proof : Again we only prove the first relation. The second can be obtained by a symmetrical argument. We use the optimal control formulation of Section 4.3. We argued there that optimal trajectories have the form of Figure 4-2(d) and (e), with cost $\inf_a \frac{1}{a}\Lambda^{I*}_{GLQF}(a)$ and $\inf_a \frac{1}{a}\Lambda^{II*}_{GLQF}(a)$, respectively. Let us fix β and consider how the cost is affected by using the policy with $\beta' = \beta + \epsilon$, for small $\epsilon > 0$.

Consider first trajectories of the form in Figure 4-2(e). Note that we can rewrite $\Lambda^{II*}_{GLQF(\beta)}(a)$ as

$$\Lambda^{II*}_{GLQF(\beta)}(a) = \inf_{\substack{x_1-\phi x_3=a \\ x_1+x_2-x_3=\beta(1+a) \\ 0\leq\phi<1}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)].$$

We shall show $\Lambda^{II*}_{GLQF(\beta')}(a) \geq \Lambda^{II*}_{GLQF(\beta)}(a)$ for all $a \geq 0$. Assume the contrary. Consider the optimal solution of the problem corresponding to β' which satisfies the feasibility constraints

$$x'_1 - \phi' x'_3 = a$$
$$x'_1 + x'_2 - x'_3 = \beta'(1+a)$$
$$0 \leq \phi' < 1$$

We distinguish two cases: $\phi' > 0$ and $\phi' = 0$. We provide an argument only for the first case. The second case can be handled similarly. Since $\beta, a \geq 0$, at least one of the following holds: $x'_1 > \mathbf{E}[A^1_0]$ or $x'_2 > \mathbf{E}[A^2_0]$ or $x'_3 < \mathbf{E}[B_0]$. Depending on which

one is the case we can decrease x_1', or x_2', or increase x_3', respectively, reducing the cost, until $x_1' + x_2' - x_3' = \beta(1+a)$. Thus, we have constructed a feasible solution of the problem corresponding to β with smaller cost than $\Lambda^{II*}_{GLQF(\beta')}(a)$. This contradicts our initial assumption. We conclude that by increasing β to β' we also increase the optimal cost of trajectories having the form in Figure 4-2(e).

If now, an optimal trajectory has the form in Figure 4-2(d), then it will still be the optimal, by convexity, when β is increased to β'. Thus, in this case, the optimal cost does not change.

We summarize by considering how the cost is affected as β is increased from 0 to ∞. At $\beta = 0$, possible optimal trajectories have the form of Figure 4-2(e). There is a threshold value $\bar{\beta}$ such that for all $\beta \leq \bar{\beta}$ optimal trajectories have the form of Figure 4-2(e) with values increasing as β increases from 0 to $\bar{\beta}$. For all $\beta > \bar{\beta}$ optimal trajectories have the form of Figure 4-2(d) with slope $\bar{\beta}$ and do not change as β increases from $\bar{\beta}$ to ∞.

■

We next prove a sufficient condition for the GLQF curve dominating the GPS curve.

Lemma 4.7.4 *If for all $\beta \in [0, \infty)$ there exists $\phi_1 \in [0, 1)$ such that*

$$\theta_1^{GPS(\phi_1)} \leq \theta_1^{GLQF(\beta)} \qquad \text{and} \qquad \theta_2^{GPS(\phi_1)} \leq \theta_2^{GLQF(\beta)},$$

then the GLQF curve dominates the GPS curve.

Proof : We use contradiction. Assume that the condition given in the statement holds but the GLQF curve does not dominate the GPS curve. Then, by definition, there exist β' and ϕ_1' such that

$$\theta_1^{GPS(\phi_1')} > \theta_1^{GLQF(\beta')} \qquad \text{and} \qquad \theta_2^{GPS(\phi_1')} > \theta_2^{GLQF(\beta')}.$$

By Lemma 4.7.2 all points with $\phi_1 < \phi_1'$ have $\theta_2^{GPS(\phi_1)} \geq \theta_2^{GPS(\phi_1')} > \theta_2^{GLQF(\beta')}$. Also,

by the same lemma, all points with $\phi_1 \geq \phi_1'$ have $\theta_1^{GPS(\phi_1)} \geq \theta_1^{GPS(\phi_1')} > \theta_1^{GLQF(\beta')}$. This contradicts our initial assumption.

■

We now have all the necessary tools to prove that the GLQF curve dominates the GPS curve.

Theorem 4.7.5 *Assuming that the arrival and service processes satisfy Assumptions A, E, and D, the GLQF curve dominates the GPS curve.*

Proof : Fix an arbitrary β. We will prove that there exists ϕ_1 satisfying the condition of Lemma 4.7.4. It suffices to prove that for both queues and such ϕ_1, overflow with the GLQF(β) policy implies overflow with the GPS(ϕ_1) policy. Then, the overflow probability of GLQF(β) is a lower bound on the corresponding probability of GPS(ϕ_1), i.e., it holds

$$\mathbf{P}[L^j_{GLQF(\beta)} > U] \leq \mathbf{P}[L^j_{GPS(\phi_1)} > U], \qquad j = 1, 2,$$

which implies

$$\theta_1^{GPS(\phi_1)} \leq \theta_1^{GLQF(\beta)} \qquad \text{and} \qquad \theta_2^{GPS(\phi_1)} \leq \theta_2^{GLQF(\beta)}.$$

Since we have established that both in the GPS and the GLQF case, the overflow probability is equal to the probability of overflowing according to one out of two scenarios, it suffices to establish the above only for these scenarios. In particular, we distinguish the following cases depending on the possible modes of overflow for GLQF(β), which are described in Section 4.4.

Case 1: Mode 1 for overflow of Q^1 and mode 1 for overflow of Q^2.

Case 2: Mode 1 for overflow of Q^1 and mode 2 for overflow of Q^2.

Case 3: Mode 2 for overflow of Q^1 and mode 1 for overflow of Q^2.

Case 4: Mode 2 for overflow of Q^1 and mode 2 for overflow of Q^2.

In Case 1 and 2, we have

$$x_1 - x_3 = a,$$
$$x_2 \leq \beta a,$$

where $x_j,\ j = 1, 2, 3,\ a$, solve the optimization problem corresponding to the overflow of Q^1 in mode 1. Then, since $x_1 - \phi_1 x_3 \geq x_1 - x_3 = a\ \forall \phi_1$, it is clear that for all ϕ_1 the GPS policy will overflow Q^1. If we are in Case 1, then also for all ϕ_1 the GPS policy will overflow Q^2. If we are in Case 2, we have

$$y_2 - \phi y_3 = a,$$
$$y_1 - (1 - \phi) y_3 = a/\beta,$$
$$0 \leq \phi < 1,$$

where $y_j,\ j = 1, 2, 3,\ a, \phi$, solve the optimization problem corresponding to the overflow of Q^2 in mode 2. Then, the GPS policy with $\phi_1 \geq 1 - \phi$ will overflow Q^2.

Consider now Cases 3 and 4. We have

$$x_1 - \phi x_3 = a,$$
$$x_2 - (1 - \phi) x_3 = a\beta,$$
$$0 \leq \phi < 1,$$

where $x_j,\ j = 1, 2, 3,\ a, \phi$, solve the optimization problem corresponding to the overflow of Q^1 in mode 2. Then the GPS policy with $\phi_1 \leq \phi$ will overflow Q^2. In Case 3, for reasons explained in the previous paragraph, the GPS policy will overflow Q^2 for all ϕ_1. If, finally, we are in Case 4, we have

$$y_2 - (1 - \phi') y_3 = a',$$
$$y_1 - \phi' y_3 = a'/\beta,$$
$$0 \leq \phi' < 1,$$

where y_j, $j = 1,2,3$, a', ϕ', solve the optimization problem corresponding to the overflow of Q^2 in mode 2. Then the GPS policy with $\phi_1 \geq \phi'$ will overflow Q^2. To show that there is at least one ϕ_1 that overflows both queues we need to show $\phi = \phi'$. To see that notice that (by making the substitution $a' := \beta a'$)

$$\inf_{a'} \frac{1}{a'} \inf_{\substack{y_2-(1-\phi')y_3=a' \\ y_1-\phi' y_3=a'/\beta \\ 0\leq\phi'<1}} [\Lambda^*_{A^1}(y_1) + \Lambda^*_{A^2}(y_2) + \Lambda^*_B(y_3)] =$$

$$\frac{1}{\beta} \inf_{a} \frac{1}{a} \inf_{\substack{y_1-\phi' y_3=a' \\ y_2-(1-\phi')y_3=\beta a' \\ 0\leq\phi'<1}} [\Lambda^*_{A^1}(y_1) + \Lambda^*_{A^2}(y_2) + \Lambda^*_B(y_3)].$$

The right hand side is exactly the problem corresponding to the overflow of Q^1 in mode 2.

■

Chapter 5

Delay in GPS

In this chapter we analyze the probability of large delay for each queue Q^1 and Q^2, in the multiclass model of Chapter 3, operated under the GPS policy. We assume that the FCFS policy is implemented for customers of the same class. The same notation and assumptions (i.e., Assumptions A, D, and E) are in effect.

We first establish in Section 5.1 a general result for the delay that the customers in each of the queues Q^1 and Q^2 are facing (see Figure 3-1). Next, in Section 5.2 we establish a lower bound on the probability of large delays using the optimal control approach that we introduced in Chapter 3 and finally in Section 5.3 we establish a matching (up to first degree in the exponent) upper bound on the same probability.

5.1 Delay in the Multiclass Model

Consider the multiclass model of Section 3.1. We denote by D_i^1 and D_i^2 the sojourn time in the system of a virtual customer arriving at time i (we assume that the virtual customer arrives at the beginning of time slot i before any other customer arrives or departs at the same slot). In this section we establish a relationship between the distribution of the delay and the corresponding queue length. These relationships are typically termed as *distributional laws* in the literature [BN95, BM92]. We will do

that only for the delay in the first queue Q^1, the delay in the second queue can be obtained by a symmetrical argument.

Theorem 5.1.1 *Assuming that customers in queue Q^1 are served in the order they arrive (FCFS policy), for each $m \in \mathbb{N}_+$ we have that*

$$\mathbf{P}[D_0^1 > m] = \mathbf{P}[L_m^1 > S_{0,m-1}^{A^1}]$$

Proof : Consider a virtual customer arriving right before time 0 in Q^1. If $D_0^1 > m$ then the customer should be in the system at time $m-1$, and because Q^1 operates with the FCFS policy, the queue length at time m, L_m^1 (recall that this does not includes arrivals and departures at time m) should include all the arrivals after the virtual customer. Thus, $D_0^1 > m$ implies $L_m^1 > S_{0,m-1}^{A^1}$. Hence $\mathbf{P}[D_0^1 > m] \leq \mathbf{P}[L_m^1 > S_{0,m-1}^{A^1}]$.

Similarly, $L_m^1 > S_{0,m-1}^{A^1}$ implies that the customer arriving right before time 0 is still in the system at time $m-1$.

■

We are interested in obtaining the probability $\mathbf{P}[D_0^1 > m]$, up to first degree in the exponent, for large values of m. Using stationarity, the above theorem implies

$$\mathbf{P}[D_{-m}^1 > m] = \mathbf{P}[L_0^1 > S_{-m,-1}^{A^1}],$$

and we will be using the latter expression to calculate the probability that the delay gets large.

5.2 Lower Bound: Optimal Control

In this section we establish a lower bound on the probability of large delay $\mathbf{P}[D_0^1 > m]$, for large values of m. We use the same optimal control approach that we used in Section 3.4 to obtain a lower bound on the overflow probability.

Consider a virtual customer arriving at time $-m$. To calculate the probability

that this customer suffers large delay, as we argued in the previous section, we will be using the expression

$$\mathbf{P}[D^1_{-m} > m] = \mathbf{P}[L^1_0 > S^{A^1}_{-m,-1}]. \tag{5.1}$$

For every sample path that leads to large delay, there exists some time $-n \leq -m$ at which both queues are empty. Since we are interesting in the asymptotics as $m \to \infty$ we scale both time and the levels of the processes A^1, A^2 and B by m. In particular, we let $T = \frac{n-m}{m}$ and define the following continuous-time functions in $D[-1-T, 0]$

$$L^j(t) = \frac{1}{m} L^j_{\lfloor mt \rfloor}, \ j = 1, 2, \quad S^X(t) = \frac{1}{m} S^X_{-m(1+T), \lfloor mt \rfloor}, \ X \in \{A^1, A^2, B\}.$$

In formulating the control problem, we use exactly the same notation as in Section 3.4. The empirical rates $x_1(t), x_2(t)$ and $x_3(t)$, of the processes A^1, A^2 and B, respectively, are the control variables and the levels of the two queues $L^1(t)$ and $L^2(t)$, the state variables. The cost functional is

$$\exp\left\{-U \int_{-T-1}^{0} [\Lambda^*_{A^1}(x_1(t)) + \Lambda^*_{A^2}(x_2(t)) + \Lambda^*_B(x_3(t))] \, dt\right\}.$$

The state trajectories are in the set (GPS-DYNAMICS), as it is defined in Section 3.4. We next formally define the deterministic optimal control problem (GPS-DELAY).

$$\begin{aligned} \text{(GPS-DELAY)} \quad \text{minimize} & \int_{-T-1}^{0} [\Lambda^*_{A^1}(x_1(t)) + \Lambda^*_{A^2}(x_2(t)) + \Lambda^*_B(x_3(t))] \, dt \\ \text{subject to: } & L^1(-T-1) = L^2(-T-1) = 0 \\ & L^1(0) > \int_{-1}^{0} x_1(t) \, dt \\ & L^2(0) : \text{ free} \\ & T : \text{ free} \\ & \{L^j(t) : \ t \in [-T-1, 0], \ j = 1, 2\} \in \text{(GPS-DYNAMICS)}. \end{aligned} \tag{5.2}$$

To solve the above problem we decompose it into the two time intervals $[-1-T, -1]$

and $[-1,0]$. First note that for all $t \in [-1,0]$ we have

$$\int_{-1}^{0} x_1(\tau)\, d\tau < L^1(0) \leq L^1(t) + \int_{t}^{0} x_1(\tau)\, d\tau \leq L^1(t) + \int_{-1}^{0} x_1(\tau)\, d\tau,$$

which implies

$$L^1(t) > 0, \qquad \forall t \in [-1,0] \tag{5.3}$$

Thus, the state trajectory in the interval $[-1,0]$ does not touch the L^2 axis in the $L^1 - L^2$ space. Observe, now, that the problem in the time interval $[-1-T,-1]$ is the same as (GPS-OVERFLOW) with the exception of the final value being $L^1(-1) > 0$ instead of 1 as was the case in (GPS-OVERFLOW). The development in Section 3.4 suggests that the difference in the final value affects only the optimal value and not the optimal trajectories. Thus the segment of the optimal trajectory of (GPS-DELAY) in $[-1-T,-1]$ has exactly the same form as in Figure 3-3(a) and (b).

We next focus in the time interval $[-1,0]$. Using exactly the same argument (taking time averages) as in the proof of Lemma 3.4.1, we can restrict the search for optimal trajectories to trajectories with constant controls in each of the Regions A, B and C. Thus, depending on the form of the segment of the state trajectory in $[-1-T,-1]$ we distinguish two different sets of candidates for optimality. These are depicted in Figure 5-1. For candidates belonging to Set I (Set II, respectively), the segment of the state trajectory in $[-1-T,-1]$ has the form of Figure 3-3(a) (Figure 3-3(b), respectively).

Let us first examine the state trajectories in Set I. Consider the trajectory in Figure 5-1(b). Let y_j and x_j, $j = 1,2,3$, be the controls in the time intervals $[-1-T,-1]$ and $[-1,0]$, respectively. We have

$$y_2 \leq \phi_2 y_3,$$
$$x_2 \geq \phi_2 x_3,$$
$$T(y_1 + y_2 - y_3) + (x_1 - \phi_1 x_3) > x_1,$$

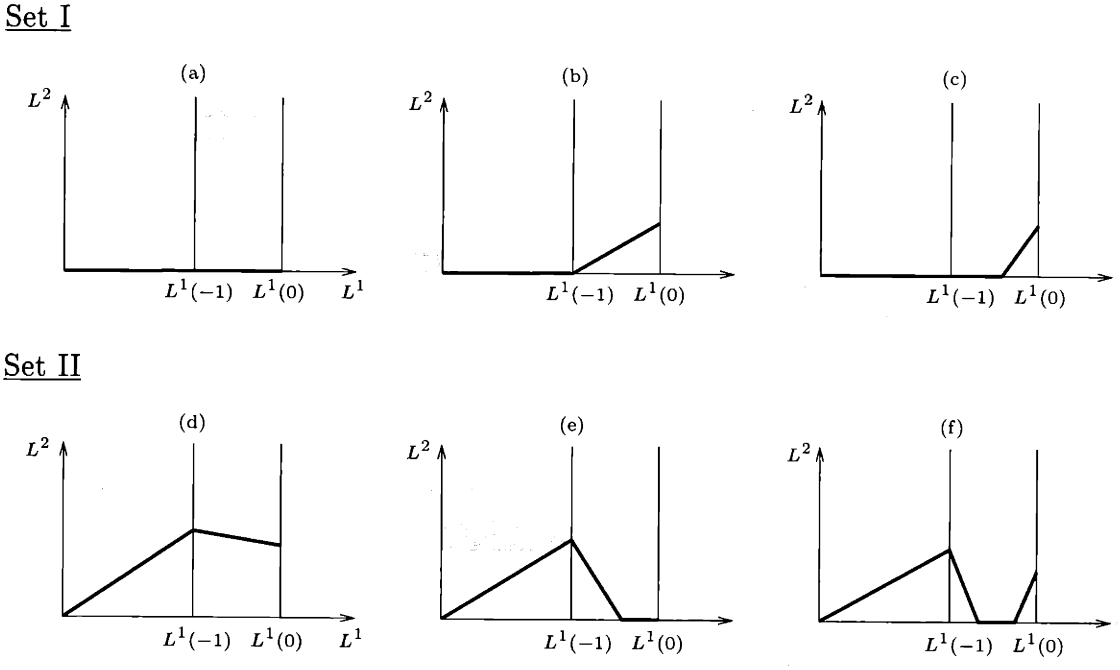

Figure 5-1: Candidates for optimal state trajectories of (GPS-DELAY). From Set I, candidates for optimal trajectories are reduced to case (a). From Set II, candidates for optimal trajectories are reduced to case (d).

which implies

$$y_2 \leq \phi_2 y_3, \tag{5.4}$$

$$x_2 \geq \phi_2 x_3, \tag{5.5}$$

$$T(y_1 + y_2 - y_3) > \phi_1 x_3. \tag{5.6}$$

We now claim that $x_3 \geq y_3$. To show this we assume that $x_3 < y_3$ and we will arrive at a contradiction. With $x_3 < y_3$, and for small $\epsilon > 0$, we increase x_3 to $x_3 + \epsilon$ and decrease y_3 to $y_3 - \frac{\epsilon}{T}$, such that the total number of services in $[-1-T, 0]$ stays constant. Note that the constraint (5.6) is not violated since $T(y_1 + y_2 - y_3) + \epsilon > \phi_1 x_3 + \phi_1 \epsilon$. Also, due to convexity the cost is decreased. We can keep doing this until

some of the constraints (5.4) or (5.5) is violated. This however contradicts the initial assumption that the trajectory has the form of Figure 5-1(b). Thus, we conclude that $x_3 \geq y_3$. This implies that $y_2 \leq x_2$ since

$$y_2 \leq \phi_2 y_3 \leq \phi_2 x_3 \leq x_2.$$

For small $\epsilon > 0$, we can now keep increasing y_2 to $y_2 + \frac{\epsilon}{T}$, and decreasing x_2 to $x_2 - \epsilon$, without violating (5.6), until one of the constraints (5.4) or (5.5) is violated. This also contradicts the initial assumption that the trajectory has the form of Figure 5-1(b). We finally conclude that we can exclude the trajectory in Figure 5-1(b) from our search for optimality. The same argument also excludes the trajectory in Figure 5-1(c) from this search. Hence, from trajectories in Set I, candidates for optimality are restricted to trajectories of the form of Figure 5-1(a).

We next examine trajectories in Set II. Consider the trajectory in Figure 5-1(e). Let $-(1-\zeta)$ the time that this trajectory hits the L^1 axis in the interval $[-1, 0]$. Let y_i, i=1,2,3, the rates during $[-1-T, -1]$ and x_i, i=1,2,3, the rates during $[-1, -(1-\zeta)]$. By taking the time average over the controls in the interval $[-1-T, -(1-\zeta)]$ we have constant controls during this interval. Let $\bar{y}_2$ and $\bar{y}_3$ the arrival rate in the second buffer and the service rate, respectively, during the same interval. From the form of the trajectory we should have $(T+\zeta)(\bar{y}_2 - \phi_2 \bar{y}_3) = 0$, which implies $\bar{y}_2 = \phi_2 \bar{y}_3$. Thus, the trajectory reduces to the one in Figure 5-1(a). The same argument applies in the trajectory in Figure 5-1(f) which reduces to the one in Figure 5-1(c). Hence, from trajectories in Set II, candidates for optimality are restricted to trajectories of the form of Figure 5-1(d). We summarize this discussion in the following proposition.

Proposition 5.2.1 *The state trajectories in Figure 5-1(a) and (d) are optimal.*

5.2.1 Optimal Value of (GPS-DELAY)

Next we calculate the optimal value of the control problem (GPS-DELAY). The result of the above proposition allows us to consider only trajectories of the form of

Figure 5-1(a) and (d).

Consider first the former. Let y_i, and x_i, i=1,2,3, the rates during the time intervals $[-1-T,-1]$ and $[-1,0]$, respectively. The feasibility constraints are

$$y_2 \leq \phi_2 y_3,$$
$$x_2 \leq \phi_2 x_3,$$
$$T(y_1+y_2-y_3)+(x_2-x_3)>0.$$

Taking the time average for x_2, y_2 (i.e., $(1+T)\bar{x}_2 = Ty_2 + x_2$) and for x_3, y_3 (i.e., $(1+T)\bar{x}_3 = Ty_3 + x_3$), we improve the cost and we obtain

$$\bar{x}_2 \leq \phi_2 \bar{x}_3, \tag{5.7}$$
$$Ty_1 + (1+T)(\bar{x}_2 - \bar{x}_3) > 0. \tag{5.8}$$

Therefore for trajectories of the form of Figure 5-1(a) the optimal cost is

$$\theta^*_{D,1} = \inf_T \inf_{\substack{\bar{x}_2 \leq \phi_2 \bar{x}_3 \\ Ty_1+(1+T)(\bar{x}_2-\bar{x}_3)>0}} [T\Lambda^*_{A^1}(y_1) + \Lambda^*_{A^1}(x_1) + (1+T)(\Lambda^*_{A^2}(\bar{x}_2) + \Lambda^*_B(\bar{x}_3))]$$

Notice in the optimization above we can take $x_1 = \mathbf{E}[A^1]$, making $\Lambda^*_{A^1}(x_1) = 0$. We next manipulate the above expression, using convex duality, to arrive at a more compact formula. We have

$$\begin{aligned}
\theta^*_{D,1} &= \inf_T \left[- \sup_{\substack{(1+T)\bar{x}_2 \leq (1+T)\phi_2 \bar{x}_3 \\ Ty_1+(1+T)(\bar{x}_2-\bar{x}_3)>0}} [-T\Lambda^*_{A^1}(y_1) - (1+T)(\Lambda^*_{A^2}(\bar{x}_2) + \Lambda^*_B(\bar{x}_3))] \right] \\
&= \inf_T \Big[- \inf_{u_1,u_2 \geq 0} \sup [u_1(1+T)\phi_2\bar{x}_3 - u_1(1+T)\bar{x}_2 + u_2(1+T)(\bar{x}_2 - \bar{x}_3) \\
&\qquad + u_2 T y_1 - T\Lambda^*_{A^1}(y_1) - (1+T)(\Lambda^*_{A^2}(\bar{x}_2) + \Lambda^*_B(\bar{x}_3))] \Big] \\
&= \inf_T \Big[- \inf_{u_1,u_2 \geq 0} [T\Lambda_{A^1}(u_2) + (1+T)(\Lambda_{A^2}(u_2-u_1) + \Lambda_B(-u_2+u_1\phi_2))] \Big]
\end{aligned}$$

$$= \inf_T \left[- \inf_{u_2 \geq 0} [T\Lambda_{A^1}(u_2) + (1+T)(\Lambda^I_{GPS}(u_2) - \Lambda_{A^1}(u_2))] \right]$$
$$= \inf_T \sup_{u_2 \geq 0} [\Lambda_{A^1}(u_2) - (1+T)\Lambda^I_{GPS}(u_2)]. \tag{5.9}$$

In the fourth equality above we have used the expression of $\Lambda^I_{GPS}(\cdot)$ as it appears in the proof of Thm. 3.7.3.

We next consider the trajectory of Figure 5-1(d). We again let y_i, and x_i, i=1,2,3, the rates during the time intervals $[-1-T, -1]$ and $[-1, 0]$, respectively. The feasibility constraints are

$$y_2 \geq \phi_2 y_3,$$
$$x_2 \geq \phi_2 x_3,$$
$$T(y_1 - \phi_1 y_3) + (x_1 - \phi_1 x_3) > x_1.$$

Taking the time average for x_2, y_2 (i.e., $(1+T)\bar{x}_2 = Ty_2 + x_2$) and for x_3, y_3 (i.e., $(1+T)\bar{x}_3 = Ty_3 + x_3$), we improve the cost and we obtain

$$\bar{x}_2 \geq \phi_2 \bar{x}_3, \tag{5.10}$$
$$Ty_1 > (1+T)\phi_1 \bar{x}_3. \tag{5.11}$$

Therefore for trajectories of the form of Figure 5-1(d) the optimal cost is

$$\theta^*_{D,2} = \inf_T \inf_{\substack{\bar{x}_2 \geq \phi_2 \bar{x}_3 \\ Ty_1 > (1+T)\phi_1 \bar{x}_3}} [T\Lambda^*_{A^1}(y_1) + \Lambda^*_{A^1}(x_1) + (1+T)(\Lambda^*_{A^2}(\bar{x}_2) + \Lambda^*_B(\bar{x}_3))]$$

Notice again in the optimization above we can take $x_1 = \mathbf{E}[A^1]$, making $\Lambda^*_{A^1}(x_1) = 0$. We next manipulate the above expression, using convex duality, to arrive at a more compact formula. We have

$$\theta^*_{D,2} = \inf_T \left[- \sup_{\substack{(1+T)\bar{x}_2 \geq (1+T)\phi_2 \bar{x}_3 \\ Ty_1 > (1+T)\phi_1 \bar{x}_3}} [-T\Lambda^*_{A^1}(y_1) - (1+T)(\Lambda^*_{A^2}(\bar{x}_2) + \Lambda^*_B(\bar{x}_3))] \right]$$

$$
\begin{aligned}
&= \inf_T \Big[- \inf_{u_1,u_2\geq 0} \sup[-u_1(1+T)\phi_2\bar{x}_3 + u_1(1+T)\bar{x}_2 + u_2 T y_1 - u_2(1+T)\phi_1\bar{x}_3 \\
&\qquad - T\Lambda^*_{A^1}(y_1) - (1+T)(\Lambda^*_{A^2}(\bar{x}_2) + \Lambda^*_B(\bar{x}_3))] \Big] \\
&= \inf_T \Big[- \inf_{u_1,u_2\geq 0} [T\Lambda_{A^1}(u_2) + (1+T)(\Lambda_{A^2}(u_1) + \Lambda_B(-u_2\phi_1 - u_1\phi_2))] \Big] \\
&= \inf_T \Big[- \inf_{u_2\geq 0} [T\Lambda_{A^1}(u_2) + (1+T) \inf_{u\leq u_2} (\Lambda_{A^2}(u_2 - u) + \Lambda_B(-u_2 + u\phi_2))] \Big] \\
&= \inf_T \Big[- \inf_{u_2\geq 0} [T\Lambda_{A^1}(u_2) + (1+T)(\Lambda^{II}_{GPS}(u_2) - \Lambda_{A^1}(u_2))] \Big] \\
&= \inf_T \sup_{u_2\geq 0} [\Lambda_{A^1}(u_2) - (1+T)\Lambda^{II}_{GPS}(u_2)]. \qquad (5.12)
\end{aligned}
$$

In the fifth equality above we have used the expression of $\Lambda^{II}_{GPS}(\cdot)$ as it appears in the proof of Thm. 3.7.3.

Hence the optimal value of (GPS-DELAY) is $\theta^*_D = \min(\theta^*_{D,1}, \theta^*_{D,1})$ which yields

$$
\begin{aligned}
\theta^*_D &= \min(\theta^*_{D,1}, \theta^*_{D,2}) \\
&= \inf_T \sup_{u_2\geq 0} [\Lambda_{A^1}(u_2) - (1+T)\Lambda_{GPS}(u_2)] \\
&= \sup_{u_2\geq 0:\ \Lambda_{GPS}(u_2)<0} [\Lambda_{A^1}(u_2) - \Lambda_{GPS}(u_2)], \qquad (5.13)
\end{aligned}
$$

by recalling by the proof of Thm. 3.7.3 that $\Lambda_{GPS}(\theta) = \max[\Lambda^I_{GPS}(\theta), \Lambda^{II}_{GPS}(\theta)]$.

We have proved the following theorem.

Theorem 5.2.2 *The optimal value, θ^*_D, of the control problem (GPS-DELAY) is given by the following expression*

$$
\theta^*_D = \sup_{u\geq 0:\ \Lambda_{GPS}(u)<0} [\Lambda_{A^1}(u) - \Lambda_{GPS}(u)]
$$

Moreover, following exactly the same argument as in Chapter 3, the solution to the control problem provides a lower bound on the probability of large delay and the optimal trajectories identify the most likely ways that large delays occur.

Proposition 5.2.3 ***(GPS Delay Lower Bound)*** *Assuming that the arrival and service processes satisfy Assumptions A and D, and under the GPS policy, the steady-state delay D^1 of queue Q^1 satisfies*

$$\lim_{m\to\infty} \frac{1}{m} \log \mathbf{P}[D^1 > m] \geq -\theta_D^*. \tag{5.14}$$

5.3 A Matching Upper Bound

In this section we establish an upper bound on the probability of large delay, establishing that the lower bound of the previous section is asymptotically tight (up to first degree in the exponent). Namely we will show that for large m the steady-state delay in Q^1 satisfies $\mathbf{P}[D^1 > m] \leq e^{-m\theta_D^* + o(m)}$. A symmetrical argument provides the bound for the delay in Q^2.

As we argued in Section 5.1 we will work with the quantity $\mathbf{P}[L_0^1 > S_{-m,-1}^{A^1}]$ since

$$\mathbf{P}[D_{-m}^1 > m] = \mathbf{P}[L_0^1 > S_{-m,-1}^{A^1}].$$

We first argue that Q^1 never empties in $[-m, 0]$.

Lemma 5.3.1 *If $L_0^1 > S_{-m,-1}^{A^1}$ then $L_{-k}^1 > 0$ for all $k \in [0, m]$.*

Proof : We have

$$L_{-k}^1 + S_{-m,-1}^{A^1} \geq L_{-k}^1 + S_{-k,-1}^{A^1} \geq L_0^1 > S_{-m,-1}^{A^1}.$$

■

Following the methodology of Section 3.6 we distinguish the following two cases:

Case 1. $\mathbf{E}[A^2] < \phi_2 \mathbf{E}[B]$.

Case 2. $\mathbf{E}[A^2] \geq \phi_2 \mathbf{E}[B]$.

5.3.1 Upper Bound: Case 2

We will first establish the upper bound for Case 2. We follow the line of development of Section 3.6.1.

We consider a busy period of the first queue, Q^1, that starts at some time $-n^* \leq -m$ ($L^1_{-n^*} = 0$) and has not ended until time 0. Notice that due to the stability condition (3.1) and the fact $\mathbf{E}[A^2] \geq \phi_2 \mathbf{E}[B]$, it is true that $\mathbf{E}[A^1] < \phi_1 \mathbf{E}[B]$, which implies that such a time $-n^*$ always exists. We will focus on sample paths of the system in $[-n^*, 0]$ that lead to $L^1_0 > S^{A^1}_{-m,-1}$. Note that

$$L^1_0 \leq S^{A^1}_{-n^*,-1} - \phi_1 S^B_{-n^*,-1}. \tag{5.15}$$

Thus,

$$\begin{aligned}\mathbf{P}[L^1_0 > S^{A^1}_{-m,-1}] \leq & \mathbf{P}[\exists n \geq m \text{ s.t. } S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1} > S^{A^1}_{-m,-1}] \\ \leq & \mathbf{P}[\max_{n\geq m}(S^{A^1}_{-n,-m-1} - \phi_1 S^B_{-n,-1}) > 0].\end{aligned} \tag{5.16}$$

We next upper bound the moment generating function of $\max_{n\geq m}(S^{A^1}_{-n,-m-1} - \phi_1 S^B_{-n,-1})$. Applying the LDP for the arrival and service processes for $\theta \geq 0$ we can obtain

$$\begin{aligned}&\mathbf{E}[e^{\theta \max_{n\geq m}(S^{A^1}_{-n,-m-1} - \phi_1 S^B_{-n,-1})}] \leq \sum_{n\geq m} \mathbf{E}[e^{\theta(S^{A^1}_{-n,-m-1} - \phi_1 S^B_{-n,-1})}] \\ &\leq \sum_{n\geq m} e^{(n-m)\Lambda_{A^1}(\theta) + n\Lambda_B(-\phi_1\theta) + n\epsilon} \\ &\leq e^{-m\Lambda_{A^1}(\theta)} e^{m(\Lambda_{A^1}(\theta) + \Lambda_B(-\phi_1\theta))} K(\theta,\epsilon) \qquad \text{if } \Lambda_{A^1}(\theta) + \Lambda_B(-\phi_1\theta) < 0,\end{aligned} \tag{5.17}$$

since when the exponent is negative (for sufficiently small ϵ), the infinite geometric series converges. We can now apply the Markov inequality in (5.16) to obtain

$$\mathbf{P}[L^1_0 > S^{A^1}_{-m,-1}] \leq \mathbf{E}[e^{\theta \max_{n\geq m}(S^{A^1}_{-n,-m-1} - \phi_1 S^B_{-n,-1})}]$$

$$\leq K(\theta,\epsilon)e^{m\Lambda_B(-\phi_1\theta)} \qquad \text{if } \Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)<0. \tag{5.18}$$

Taking the limit as $m\to\infty$ and minimizing over θ to obtain the tightest bound we establish the following proposition.

Proposition 5.3.2 *If* $\mathbf{E}[A^2]\geq\phi_2\mathbf{E}[B]$ *and assuming an LDP for the arrival and service processes (Assumption A)*

$$\lim_{m\to\infty}\frac{1}{m}\log\mathbf{P}[L_0^1>S^{A^1}_{-m,-1}]\leq-\sup_{\{\theta\geq 0:\ \Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)<0\}}[-\Lambda_B(-\phi_1\theta)].$$

We are now left with proving that the above upper bound is tight. This is done in the following proposition.

Proposition 5.3.3 ***(GPS Delay Upper bound, Case 2)*** *If* $\mathbf{E}[A^2]\geq\phi_2\mathbf{E}[B]$ *and assuming that the arrival and service processes satisfy Assumption A, the steady-state delay,* D^1*, of queue* Q^1*, at an arbitrary time slot satisfies*

$$\lim_{m\to\infty}\frac{1}{m}\log\mathbf{P}[D^1>m]\leq-\theta^*_D.$$

Proof : Given the result of Proposition 5.3.2 it suffices to prove that $\theta^*_D=\sup_{\{\theta\geq 0:\ \Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)<0\}}[-\Lambda_B(-\phi_1\theta)]$. Consider the analysis of Section 5.2 that yields θ^*_D as the solution to (GPS-DELAY). The argument of Theorem 3.4.3 applies, implying that optimal state trajectories can be restricted to having the form depicted in Figure 5-1(d). Consider the feasibility constraints of such a trajectory given in Eqs. (5.10) and (5.11). Notice that $\bar{x}_3>\mathbf{E}[B]$ (if otherwise, we can decrease $\bar{x}_3$ to $\mathbf{E}[B]$ improving the cost without violating the constraints). Then we can actually fix $\bar{x}_2$ to $\mathbf{E}[A^2]$ without violating the constraint (5.10) since

$$\bar{x}_2=\mathbf{E}[A^2]\geq\phi_2\mathbf{E}[B]\geq\phi_2\bar{x}_3.$$

Finally, fixing also x_1 to $\mathbf{E}[A^1]$ the expression for $\theta^*_{D,2}$ given in Section 5.2.1, which

as we argued equals θ_D^* in Case 2, becomes

$$\theta_D^* = \inf_T \inf_{Ty_1 > (1+T)\phi_1 \bar{x}_3} [T\Lambda_{A^1}^*(y_1) + (1+T)\Lambda_B^*(\bar{x}_3)].$$

Manipulating the above expression as in (5.12), using convex duality, it can be verified that it is equal to $\sup_{\{\theta \geq 0:\ \Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)<0\}}[-\Lambda_B(-\phi_1\theta)]$.

■

5.3.2 Upper Bound: Case 1

We now proceed with establishing an upper bound in Case 1. We follow the line of development of Section 3.6.2.

Consider all sample paths that lead to $L_0^1 > S_{-m,-1}^{A^1}$. Looking backwards in time from time 0, let $-k^* \leq -m$ be the first time that $L^1 = 0$. We use again the virtual system idea of Section 3.6.2. Let $-n^* \leq -k^*$ be the first time (looking backwards in time from $-k^*$) that the queue length of Q^2 becomes zero in the virtual system. Notice that such a time $-n^*$ always exists since we are in Case 1, and Q^2 is stable when it gets exactly a fraction ϕ_2 of the capacity. Eq. (3.24) and (3.25), which we repeat for convenience, hold, i.e., we have

$$L_0^1 \leq S_{-k^*,-1}^{A^1} + S_{-n^*,-1}^{A^2} - S_{-k^*,-1}^{B} - \phi_2 S_{-n^*,-k^*-1}^{B}, \tag{5.19}$$

and

$$L_0^1 \leq S_{-k^*,-1}^{A^1} - \phi_1 S_{-k^*,-1}^{B}. \tag{5.20}$$

We will use the bound in (5.19) when $S_{-n^*,-1}^{A^2} \leq \phi_2 S_{-n^*,-1}^{B}$ and the bound in (5.20)

otherwise. Namely, we repeat (3.26)

$$L_0^1 \leq \begin{cases} S^{A^1}_{-k^*,-1} + S^{A^2}_{-n^*,-1} - S^B_{-k^*,-1} - \phi_2 S^B_{-n^*,-k^*-1} & \text{if } S^{A^2}_{-n^*,-1} \leq \phi_2 S^B_{-n^*,-1} \\ S^{A^1}_{-k^*,-1} - \phi_1 S^B_{-k^*,-1} & \text{if } S^{A^2}_{-n^*,-1} \geq \phi_2 S^B_{-n^*,-1}. \end{cases} \tag{5.21}$$

Let Ω_1 the set of sample paths that satisfy $S^{A^2}_{-n^*,-1} \leq \phi_2 S^B_{-n^*,-1}$ and Ω_2 its complement. We have

$$\begin{aligned} &\mathbf{P}[L_0^1 > S^{A^1}_{-m,-1} \text{ and } \Omega_1] \leq \\ &\leq \mathbf{P}[\exists n \geq k \geq m \text{ s.t. } S^{A^2}_{-n,-1} \leq \phi_2 S^B_{-n,-1} \text{ and} \\ &\qquad S^{A^1}_{-k,-1} + S^{A^2}_{-n,-1} - S^B_{-k,-1} - \phi_2 S^B_{-n,-k-1} > S^{A^1}_{-m,-1}] \\ &\leq \mathbf{P}[\max_{\{n \geq k \geq m:\ S^{A^2}_{-n,-1} \leq \phi_2 S^B_{-n,-1}\}} (S^{A^1}_{-k,-m-1} + S^{A^2}_{-n,-1} - S^B_{-k,-1} - \phi_2 S^B_{-n,-k-1}) > 0]. \end{aligned} \tag{5.22}$$

To bound the above we will make use of the following lemma, which establishes that $n^* = O(m)$ with high probability, meaning that the probability of $n^* > lm$, for some constant l, is smaller that $e^{-m\theta^*_D + o(m)}$.

Lemma 5.3.4 *Assuming that the arrival and service processes satisfy Assumptions A and E, there exists $l > 0$ such that*

$$\mathbf{P}[L_0^1 > S^{A^1}_{-m,-1} \text{ and } n^* > lm] \leq e^{-mq + o(m)},$$

where $q > \theta^*_D$.

Proof : Using (5.19) and the Markov inequality we obtain

$$\begin{aligned} &\mathbf{P}[L_0^1 > S^{A^1}_{-m,-1} \text{ and } n^* > lm] \leq \\ &\leq \mathbf{P}[\exists n \geq k \geq m,\ n > lm, \text{ s.t. } S^{A^1}_{-k,-1} + S^{A^2}_{-n,-1} - S^B_{-k,-1} - \phi_2 S^B_{-n,-k-1} > 0] \end{aligned}$$

$$\leq \mathbf{P}\left[\max_{\substack{n\geq k\geq m \\ n>lm}} [S^{A^1}_{-k,-1} + S^{A^2}_{-n,-1} - S^{B}_{-k,-1} - \phi_2 S^{B}_{-n,-k-1}] > 0\right]$$

$$\leq \mathbf{E}\left[e^{\theta \max_{\substack{n\geq k\geq m \\ n>lm}} [S^{A^1}_{-k,-1}+S^{A^2}_{-n,-1}-S^{B}_{-k,-1}-\phi_2 S^{B}_{-n,-k-1}]}\right], \tag{5.23}$$

for some $\theta \geq 0$ to be specified in the sequel. The above moment generating function can be bounded as follows:

$$\begin{aligned}
&\mathbf{E}\left[e^{\theta \max_{\substack{n\geq k\geq m \\ n>lm}} [S^{A^1}_{-k,-1}+S^{A^2}_{-n,-1}-S^{B}_{-k,-1}-\phi_2 S^{B}_{-n,-k-1}]}\right] \leq \\
&\leq \sum_{n>lm} \sum_{k=0}^{n} e^{(n-k)[\Lambda_{A^2}(\theta)+\Lambda_B(-\phi_2\theta)]+k[\Lambda_{A^1}(\theta)+\Lambda_{A^2}(\theta)+\Lambda_B(-\theta)]+\Gamma(\theta)} \\
&\leq \sum_{n>lm} \left(e^{n[\Lambda_{A^2}(\theta)+\Lambda_B(-\phi_2\theta)]+\Gamma(\theta)} + e^{n[\Lambda_{A^1}(\theta)+\Lambda_{A^2}(\theta)+\Lambda_B(-\theta)]+\Gamma(\theta)}\right) \\
&\leq K(\theta)\left(e^{lm[\Lambda_{A^2}(\theta)+\Lambda_B(-\phi_2\theta)]} + e^{lm[\Lambda_{A^1}(\theta)+\Lambda_{A^2}(\theta)+\Lambda_B(-\theta)]}\right)
\end{aligned} \tag{5.24}$$

In the first inequality above we have used Assumption E and in the second the fact that the exponent is linear in k and hence the expression is upper bounded by the sum of the terms at $k = 0$ and $k = n$. In the third inequality above, for sufficiently large m, the infinite geometric series converge if the exponents are negative for some θ. Indeed this is the case, that is, there exists a θ at which both exponents are negative, since both $[\Lambda_{A^2}(\theta)+\Lambda_B(-\phi_2\theta)]$ and $[\Lambda_{A^1}(\theta)+\Lambda_{A^2}(\theta)+\Lambda_B(-\theta)]$ are zero at $\theta = 0$ and have negative derivatives at $\theta = 0$, i.e, $\mathbf{E}[A^2] - \phi_2\mathbf{E}[B] < 0$ and $\mathbf{E}[A^1] + \mathbf{E}[A^2] - \mathbf{E}[B] < 0$, respectively. We can now choose large enough l to make the exponents in the right hand side of (5.24) sufficiently small (equal to $-q$).

■

We now return to Eq. (5.22) and note that due to the above Lemma we can constrain $n \leq lm$. Let

$$L^I_{D,1} \triangleq \max_{\{lm\geq n\geq k\geq m:\ S^{A^2}_{-n,-1}\leq \phi_2 S^{B}_{-n,-1}\}} (S^{A^1}_{-k,-m-1} + S^{A^2}_{-n,-1} - S^{B}_{-k,-1} - \phi_2 S^{B}_{-n,-k-1}),$$

which after bringing the constraints in the objective function becomes

$$L^I_{D,1} = \max_{lm\geq n\geq k\geq m} \inf_{u\geq 0}[S^{A^1}_{-k,-m-1} + (1-u)S^{A^2}_{-n,-1} - (1-u\phi_2)S^B_{-k,-1} - \phi_2(1-u)S^B_{-n,-k-1}]. \tag{5.25}$$

Next we will upper bound the moment generating functions of $L^I_{D,1}$ using Assumption E. For $\theta \geq 0$ we have

$$\begin{aligned}
\mathbf{E}[e^{\theta L^I_{D,1}}] \leq & \\
\leq & \sum_{lm\geq n\geq m} \sum_{m\leq k\leq n} \inf_{u\geq 0} \mathbf{E}[\exp\{\theta[S^{A^1}_{-k,-m-1} + (1-u)S^{A^2}_{-n,-1} - (1-u\phi_2)S^B_{-k,-1} \\
& \qquad - \phi_2(1-u)S^B_{-n,-k-1}]\}] \\
\leq & \sum_{lm\geq n\geq m} \sum_{m\leq k\leq n} \inf_{u\geq 0} \exp\{(n-k)[\Lambda_{A^2}(\theta-\theta u) + \Lambda_B(-\theta\phi_2(1-u))] \\
& \qquad + (k-m)[\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta-\theta u) + \Lambda_B(-\theta(1-u\phi_2))] \\
& \qquad + m[\Lambda_{A^2}(\theta-\theta u) + \Lambda_B(-\theta(1-u\phi_2))] + \Gamma(\theta,u)\}
\end{aligned} \tag{5.26}$$

Let now $\frac{n-m}{m} = \tau$ and $\frac{n-k}{m} = \tau\zeta$ for $\zeta \in [0,1]$. For large enough m we have

$$\mathbf{E}[e^{\theta L^I_{D,1}}] \leq lm^2 e^{m\Lambda^I_{D,1}(\theta)}, \tag{5.27}$$

where

$$\begin{aligned}
\Lambda^I_{D,1}(\theta) \triangleq \sup_{\tau} \sup_{\zeta\in[0,1]} \inf_{u\geq 0} \Big(& \tau\zeta[\Lambda_{A^2}(\theta-\theta u) + \Lambda_B(-\theta\phi_2(1-u))] \\
& + \tau(1-\zeta)[\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta-\theta u) + \Lambda_B(-\theta(1-u\phi_2))] \\
& + [\Lambda_{A^2}(\theta-\theta u) + \Lambda_B(-\theta(1-u\phi_2))]\Big).
\end{aligned} \tag{5.28}$$

Thus, invoking the Markov inequality from (5.22) we establish

$$\lim_{m\to\infty} \frac{1}{m} \log \mathbf{P}[L_0^1 > S_{-m,-1}^{A^1} \text{ and } \Omega_1] \leq \inf_\theta \Lambda_{D,1}^I(\theta). \tag{5.29}$$

To show that the above upper bound matches the lower bound of Proposition 5.2.3 we have to establish that $-\theta_D^* \geq \inf_\theta \Lambda_{D,1}^I(\theta)$. Let $\theta_D^{I*} \triangleq \inf_\theta \Lambda_{D,1}^I(\theta)$. This is done using the ideas in the proof of Proposition 3.6.5. That is, we consider the trajectory of Figure 5-2 with associated cost

$$\inf_\tau \inf_{\substack{\zeta\tau(x_2-\phi_2 x_3)+(1-\zeta)\tau(y_2-\phi_2 y_3)+(w_2-\phi_2 w_3)\leq 0 \\ \zeta\tau(x_2-\phi_2 x_3)+(1-\zeta)\tau(y_1+y_2-y_3)+(w_2-w_3)>0 \\ 0\leq\zeta\leq 1}} [\zeta\tau(\Lambda_{A^2}^*(x_2)+\Lambda_B^*(x_3))$$
$$+ (1-\zeta)\tau(\Lambda_{A^1}^*(y_1)+\Lambda_{A^2}^*(y_2)+\Lambda_B^*(y_3)) + (\Lambda_{A^2}^*(w_2)+\Lambda_B^*(w_3))].$$

Manipulating the above expression as in (5.9) and using convex duality it can be

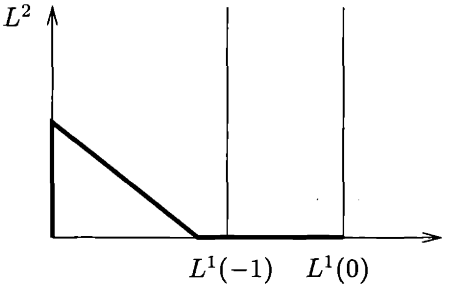

Figure 5-2: Trajectory for the control problems corresponding to θ_D^{I*}.

verified that it is equal to θ_D^{I*}. Now, using exactly the same techniques as in Section 5.2, that is convexity and the homogeneity property, it can be established that optimal state trajectories do not spend any time on the L^2 axis. Thus, the trajectory in Figure 5-2 can be reduced to the one in Figure 5-1(a). This establishes the desired

result $\theta_D^{I*} = \theta_{D,1}^* \geq \theta_D^*$, which implies

$$\lim_{m\to\infty} \frac{1}{m} \log \mathbf{P}[L_0^1 > S_{-m,-1}^{A^1} \text{ and } \Omega_1] \leq -\theta_D^*.$$

Similarly, it can be shown that

$$\lim_{m\to\infty} \frac{1}{m} \log \mathbf{P}[L_0^1 > S_{-m,-1}^{A^1} \text{ and } \Omega_2] \leq -\theta_D^*.$$

Hence we have established the following Proposition.

Proposition 5.3.5 ***(GPS Delay Upper bound, Case 1)*** *If* $\mathbf{E}[A^2] < \phi_2 \mathbf{E}[B]$ *and assuming that the arrival and service processes satisfy Assumption A and E, the steady-state delay,* D^1*, of queue* Q^1*, at an arbitrary time slot satisfies*

$$\lim_{m\to\infty} \frac{1}{m} \log \mathbf{P}[D^1 > m] \leq -\theta_D^*.$$

We summarize Propositions 5.3.3 and 5.3.5 in the following:

Proposition 5.3.6 ***(GPS Delay Upper bound)*** *Assuming that the arrival and service processes satisfy Assumption A and E, the steady-state delay,* D^1*, of queue* Q^1*, at an arbitrary time slot satisfies*

$$\lim_{m\to\infty} \frac{1}{m} \log \mathbf{P}[D^1 > m] \leq -\theta_D^*.$$

The main result of this chapter is the following theorem.

Theorem 5.3.7 *Under the GPS policy, assuming that the arrival and service processes satisfy Assumption A, D and E, the steady-state delay,* D^1*, of queue* Q^1*, at an arbitrary time slot satisfies*

$$\lim_{m\to\infty} \frac{1}{m} \log \mathbf{P}[D^1 > m] = -\theta_D^*,$$

where

$$\theta_D^* = \sup_{u_2 \geq 0:\ \Lambda_{GPS}(u_2) < 0} [\Lambda_{A^1}(u_2) - \Lambda_{GPS}(u_2)],$$

and where $\Lambda_{GPS}(\cdot)$ *is as defined in Thm. 3.7.3.*

Chapter 6

Admission Control

In this chapter we discuss the application of the performance analysis results that we have obtained in the previous chapters to the problem of admission control in high speed networks. As we discussed in the introduction, admission control in high speed networks (e.g., ATM-based networks) is necessary since real-time services as interactive-TV, video, video-conferencing and voice are very sensitive to packet losses (due to buffer overflows) and to large delays. In our treatment we take only two *Quality of Service (QoS)* measures of interest: the loss probability and the probability of large delay. It is desirable that both these probabilities are upper bounded by some given constant which depends on the particular service. In practice there are additional QoS measures of interest, for example, the delay jitter (i.e., delay variation) and the short term fraction of packets that are lost or delayed. Notice, that the latter QoS measure is concerned with losses and delays that occur almost consecutively, in bursts. In contrast, the probability of loss or delay captures the long term fraction of packets that are lost or delayed (assuming stationarity and ergodicity).

We propose in this chapter an admission control scheme on a call by call basis. This means that when a call is admitted, the network does not take any control action to regulate the traffic produced by the call. One reason for doing that, besides the analytical tractability of the problem, is that regulation (with a leaky bucket for example) will also introduce delays or losses that will degrade the quality of the call.

Therefore, if we were to use regulation we would also need to evaluate the quality degradation that it produces. To implement such a call-based admission control mechanism two are the critical decisions to be made. First to assign buffer sizes and second to restrict the number of connections (calls) that the network services in order to ensure that the QoS measures are within specifications.

A prerequisite for using the performance analysis results obtained in earlier chapters to do admission control is that we have a detailed statistical model for the traffic (knowledge of moment generating functions). For voice traffic Markov-modulated processes with an underlying Markov chain of very few states (usually two states) are satisfactory models (see [AMS82, MAS88, EM93]). For video traffic, Markov modulated processes with higher dimensionality are often required. When such a model is not available off-line, the admission control mechanism has to be coupled with an on-line estimator. In practice, there are types of traffic (especially data) for which appropriate statistical models are not available, or are too complicated and do not satisfy Assumptions A and D and E, which basically require short range dependencies. In such cases the results of this chapter are not applicable and worst case analysis might be used instead.

In this chapter we will focus on an isolated node (switch) and devise an admission control mechanism. Although our network analysis of Chapter 2 provides the basis for admission control in the single class case, for a network, we choose to focus on a single node for two primary reasons: a) we consider the multiclass effect of primary interest and there are not yet developed analytical results for multiclass networks, and b) even if such results were available the admission scheme would be too complicated and too computationally burdensome to be implementable in real-time. For modeling traffic we will use the discrete-time model that we introduced in Section 3.1 in page 86.

Regarding the structure of this chapter we begin in Section 6.1 with the simpler case of single class admission control. We also present there several examples (based both on some simple traffic models and on actual MPEG video traffic) that indicate the relevance of the large deviations asymptotics and the statistical multiplexing gains that can be obtained by such an admission control scheme. In Section 6.2 we consider

the much more involved multiclass case and propose an admission control algorithm under the GPS scheduling policy.

6.1 Single Class: Effective Bandwidth

In this section we review a single class admission control scheme. This scheme is based on the notion of the effective bandwidth. Numerous papers deal in various ways with this concept [Hui88, GH91, Kel93, KWC93, dVW93, CW93, EM93, EHL+94], and is probably this concept that generated so much interest in large deviations techniques for communication networks.

Consider the architecture of Figure 6-1. We assume a deterministic service capac-

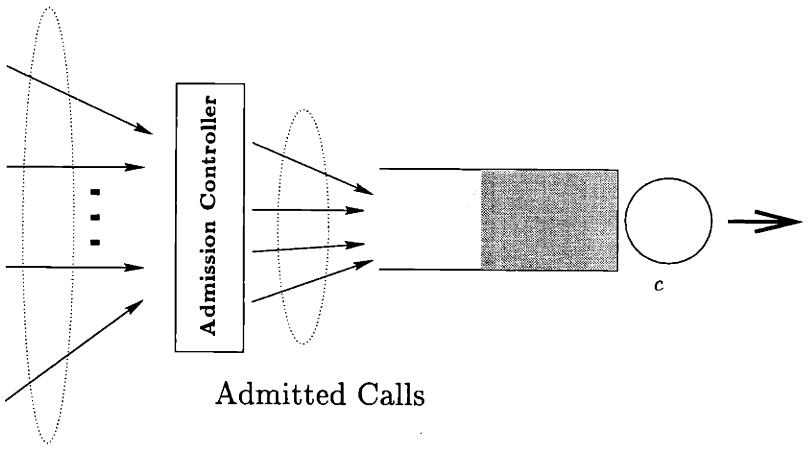

Figure 6-1: An architecture for single-class admission control.

ity of c b/s (bits/sec). Calls request to be connected to the system. We assume that all calls are of the same type and generate traffic according to a stochastic process $\{A_i \; i \in \mathbb{N}\}$, where A_i is the traffic (in b/s) generated in time slot i. We also assume that the process A satisfies Assumptions A and D. Let $\Lambda_A(\theta)$ denote the limiting

log-moment generating function of the process A. If the admission controller admits N calls then it is easy to verify that the aggregate arrival process in the buffer, say $\tilde{A}$, has limiting log-moment generating function

$$\Lambda_{\tilde{A}}(\theta) = N\Lambda_A(\theta), \qquad \forall \theta \in \mathbb{R}.$$

Since the service process is deterministic it is characterized by the limiting log-moment generating function $\Lambda_B(\theta) = c\theta$. Let U (in bits) denote the buffer size of the system, and L_U, D_U, the steady-state queue length and delay, respectively. The subscript U denotes the fact that these steady-state distributions depend on the buffer size.

The QoS parameters are given in terms of a constant $D_{\max}$, which denotes the desirable maximum allowed delay in the buffer, and a scalar δ such that

$$\mathbf{P}[L_U > U] < \delta, \tag{6.1}$$

$$\mathbf{P}[D_U > D_{\max}] < \delta. \tag{6.2}$$

The admission controller has the freedom to select the appropriate buffer size U and restrict the number of admitted calls in order to guarantee (6.1) and (6.2). We next argue that the appropriate buffer size is $U = cD_{\max}$. Consider first setting $U > cD_{\max}$. Then the system will be admitting packets that will need more than $D_{\max}$ time to clear the buffer. We can view $D_{\max}$ as a threshold value set appropriately, depending on the application, such that packets which are not transmitted within $D_{\max}$ will severely degrade the performance and cannot therefore be considered of use in the receiving end. In practice, such packets are often discarded before transmitted. Constraints (6.1) and (6.2) guarantee that the fraction of such packets as well as lost packets remains sufficiently small. Thus, for engineering reasons the system should not be admitting such packets and hence the buffer size should be set to $U \leq cD_{\max}$. Consider now setting the buffer size to $U < cD_{\max}$. Then the system is discarding packets that can be transmitted within $D_{\max}$. Moreover, the loss probability for a fixed number of connections is increased as we decrease U. Therefore we set the buffer size to $U = cD_{\max}$. Note that this immediately guarantees (6.2) for all $\delta > 0$. We are

now left with guaranteeing (6.1).

We will use the level crossing probability of an infinite buffer system to approximate the loss probability of the system with finite buffer. In our discrete time model the queue length at time 0 in an infinite buffer G/G/1 queue with aggregate arrival process $\tilde{A}$ and service process B, respectively, is given by the Lindley equation

$$L_0 = [L_{-1} + \tilde{A}_{-1} - B_{-1}]^+,$$

which is very similar to Eq. (2.9). The following theorem is the discrete-time analog of Theorem 2.2.1 in page 46; the proof is identical and is omitted.

Theorem 6.1.1 *The steady-state queue length L in a G/G/1 queue with arrival and service processes satisfying Assumption A is characterized by*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L > U] = -\theta^*, \tag{6.3}$$

where $\theta^* > 0$ is the maximum root of the equation

$$\Lambda_{\tilde{A}}(\theta) + \Lambda_B(-\theta) = 0. \tag{6.4}$$

In the discrete-time model the stability condition asserts $\mathbf{E}[\tilde{A}] < \mathbf{E}[B]$, thus the function $\Lambda_{\tilde{A}}(\theta) + \Lambda_B(-\theta)$ has negative derivative at 0. The situation of Figure 2-2 holds when we change the direction of the θ-axis ($\theta^* > 0$). By convexity we have that $\Lambda_{\tilde{A}}(\theta) + \Lambda_B(-\theta) < 0$ for all $0 < \theta < \theta^*$. Let now $\delta_L = -\frac{\log \delta}{U}$. We can ensure

$$\mathbf{P}[L > U] \sim e^{-\theta^* U} < \delta = e^{-U\delta_L},$$

if and only if $\delta_L < \theta^*$ which holds if and only if $\Lambda_{\tilde{A}}(\delta_L) + \Lambda_B(-\delta_L) < 0$. For the system of Figure 6-1, when N calls are admitted, this implies

$$N\frac{\Lambda_A(\delta_L)}{\delta_L} < c. \tag{6.5}$$

The quantity $v(\delta_L) \stackrel{\triangle}{=} \frac{\Lambda_A(\delta_L)}{\delta_L}$ depends only on the QoS parameters and the statistics of a call and is referred to as *effective bandwidth* of the call due to the similarity of (6.5) with the admission criterion for loss networks (circuit switched). That is, the number of calls times the capacity that each call consumes has to be at most the total capacity of the system. Here, that a stochastic process characterizes the traffic generated by a call, the effective bandwidth captures, in a single number, the capacity that the call requires for the quality of service to remain within the given specifications. The next theorem establishes that the effective bandwidth is constrained within the mean and the peak arrival rate.

Theorem 6.1.2 *Assume that $A_i \leq A_{\max} < \infty$ for all i (with non-zero mass at $A_{\max}$) w.p.1. Then*

$$\mathbf{E}[A] \leq v(\theta) \leq A_{\max},$$

and the lower bound becomes tight as $\theta \to 0$, while the upper bound becomes tight as $\theta \to \infty$.

Proof : For the upper bound note that for all $\theta > 0$ and n

$$\mathbf{E}[e^{\theta \sum_{i=1}^{n} A_i}] \leq e^{\theta n A_{\max}},$$

which implies that

$$v(\theta) = \frac{1}{\theta}\Lambda_A(\theta) \leq A_{\max}. \tag{6.6}$$

For the lower bound, notice that

$$v(\theta) = \frac{\sup_a[\theta a - \Lambda_A^*(a)]}{\theta} \geq \mathbf{E}[A], \tag{6.7}$$

since $\Lambda_A^*(\mathbf{E}[A]) = 0$.

When $\theta \to 0$, $\Lambda_A(\theta) = \theta \mathbf{E}[A] + o(\theta)$, which implies that (6.7) gets tight. To prove

that the upper bound is tight for $\theta \to \infty$, arguing as in (6.7), we have

$$v(\theta) \geq A_{\max} - \frac{\Lambda_A^*(A_{\max})}{\theta}.$$

This implies that as $\theta \to \infty$, $v(\theta) \geq A_{\max}$, where we have assumed that $\Lambda_A^*(A_{\max}) < \infty$ (since $A_{\max}$ has non-zero mass).

■

The above theorem implies that as $\delta \to 0$ (hence, $\delta_L \to \infty$), which means that we require no loss and no large delays, then the effective bandwidth converges to the peak arrival rate. As $\delta \to 1$ (hence, $\delta_L \to 0$), which means that we relax the QoS requirements, the constraint (6.5) degenerates to the stability condition of the system.

6.1.1 An example

We present a simple example that shows how effective bandwidth-based admission control performs. We consider two types of traffic with the parameters of Table 6.1. Both types of traffic conform to the ON-OFF model of Figure 6-2. Traffic is generated according to a continuous-time Markov process, with embedded Markov chain depicted in Figure 6-2. In the ON state, traffic is produced with a constant rate of p b/s. We refer to this as the peak rate. In the OFF state no traffic is generated. The traffic source stays in the ON state a fraction $\frac{a}{a+b}$ of the time and for an expected number of $1/b$ transitions of the embedded Markov chain. It generates traffic with an average rate of $p\frac{a}{a+b}$ b/s.

A few comments about the traffic and QoS parameters are in order. Type 1 traffic has parameters which are typical of a video-conferencing call which consists of the transmission of relatively low activity scenes (people sitting around a table). As a consequence the peak rate is close to average rate. Type 2 traffic is more typical of a bursty video call (e.g., action movie). To put $D_{\max}$ into perspective, with a packet size of 53 bytes (size of an ATM cell) and with a 64 Kb/s rate for voice, the

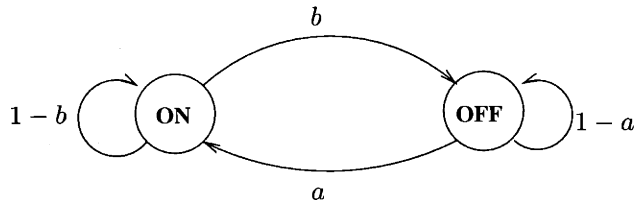

Figure 6-2: The ON-OFF source model.

	Traffic Parameters					QoS Parameters	
	Peak	Avg.	$\mathbf{E}[t_{ON}]$	a	b	$D_{\max}$	δ
Type 1	2 Mb/s	1 Mb/s	25 ms	0.04	0.04	10 ms	10^{-6}
Type 2	10 Mb/s	2 Mb/s	5 ms	0.05	0.2	30 ms	10^{-9}

Table 6.1: Traffic Parameters for the ON-OFF model. $\mathbf{E}[t_{ON}]$ denotes the expected amount of time that the traffic source stays in the ON state (expected duration of burst). For both types of traffic it can be easily verified that the embedded Markov chain makes one transition every 1 ms.

packetization delay is about 6 ms. It is usually assumed that about 300 ms end to end delay for voice and video calls is satisfactory, thus with about 10 nodes end to end path, maximum delay should be in the order of 30 ms per node. A discussion of typical traffic and QoS parameters can be found in [HW94].

Consider now the system of Figure 6-1 with a service capacity of 135 Mb/s. One naive way to admit calls is to take a worst case analysis stance and assume that traffic sources are always transmitting in their peak rate. Thus, we would be able to allow at most 67 calls of Type 1, or 13 calls of Type 2. We refer to this as *peak rate assignment.* The stability condition for the system implies that we cannot admit more than 135 Type 1 calls, or 67 Type 2 calls. If we apply the recommendations of this section, Eq. (6.5) restricts the number of calls to at most 120 for Type 1 calls, or

58 for Type 2 calls. This is about twice the peak rate assignment for Type 1 calls and more than three times for Type 2 calls. Hence, a significant statistical multiplexing gain can be realized when we take into account the statistical behaviour of traffic, the exact way that we do this being prescribed by Eq. (6.5). The following table summarizes the discussion of this paragraph.

	Max Number of Calls		
	Peak rate assignment	Stability condition	Effective bandwidth
Type 1	67	135	120
Type 2	13	67	58

Table 6.2: Comparing peak rate assignment, the stability condition and the the effective bandwidth-based assignment.

6.1.2 An example with actual MPEG video traffic

We next present an example where the input traffic is actual traffic generated by transmitting an MPEG-coded (for a description of the coding algorithm see [LeG91]) video signal of the Star Wars motion picture. The data were obtained from the Bellcore site (see [Gar93] for details).

The video trace is plotted in Figure 6-3. It depicts the number of bits used by the MPEG encoder per frame; frames are generated at a rate of 24 per second. This trace has a peak rate of 4.44 Mb/s and an average rate of 0.37 Mb/s. Consider the system of Figure 6-1 with a service capacity of 50 Mb/s. Worst case analysis demands that we admit traffic based on the peak rate. In this example, this means that we can allow at most 11 movies to be transmitted through the system. The stability condition for the system implies that we can allow at most 135 movies.

To apply the recommendations of this section we need to estimate a statistical model from the data. We will be using a Markov-modulated source model. Traffic is

generated according to a continuous-time Markov process, the state of the process being the bit rate. More specifically, consider a Markov chain with M states $s_1, \ldots, s_M$.

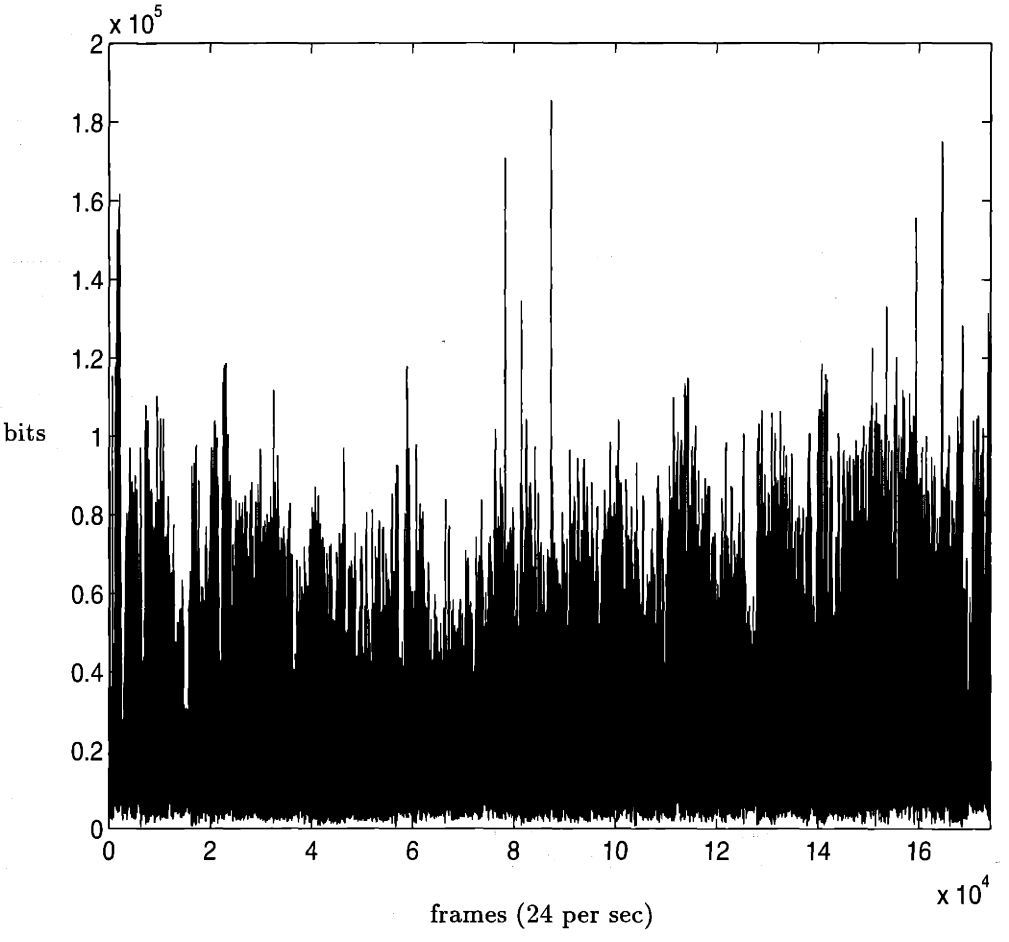

Figure 6-3: The MPEG video trace of the Star Wars movie.

In state s_i the source model generates traffic at a rate of r_i b/s. The chain makes transitions at the frame transmission rate (i.e., one transition every 1/24th of a second). Let now R and r be the peak and the mean rate, respectively, of the MPEG video trace. We build the model from the data as follows:

1. We split the bit rate range $[r, R]$ in M equal intervals of length $(R-r)/M$. Interval $i = 1, \ldots, M$ corresponds to rates in $[r + (i-1)\frac{R-r}{M}, r + i\frac{R-r}{M}]$ and is assigned to state s_i of the Markov modulated source model.

2. We define $r_i = r + i\frac{R-r}{M}$, which is the maximum rate in interval i.

3. We estimate the transition probabilities of the Markov chain using maximum likelihood estimation.

Since the source model is assumed to transmit at the maximum possible rate when in state s_i, it is, in this sense, a "worst case" model within the set of Markov modulated models with M states. We use such a model to obtain "safer" estimates of the quality of service probabilities. In other words, given the fact that traffic statistics are not known a priori and have to be estimated, it is better to use conservative modeling than to run the risk of violating QoS specifications.

We have performed a small number of experiments to assess the performance of such a scheme. We estimated, off-line, a Markov modulated model from the data, as outlined above, and we determined by Eq. (6.5) the maximum number N of movies that can be transmitted for a maximum delay of 40ms and a QoS parameter of 10^{-5}, when the available capacity is 50 Mb/s. We then compared the analytical estimate for the loss probability with an estimate obtained from simulating the system when N movies randomly [1] synchronized are transmitted. The results are outlined in Table 6.3.

As this table indicates we have used three different Markov modulated models A, B, and C, with 76, 58, and 39 states, respectively. There is an issue on how to select the dimensionality of the Markov modulated model, which is not very well understood at the moment. Very few states do not capture well the structure of the data and result in conservative estimates of the loss probability, while, on the other hand, many states result in large estimation errors since the transition probabilities

[1] Each movie starts at a frame t uniformly distributed in $[1, T]$, where T is the length of the movie in frames. Frames are transmitted in the order $t, t+1, \ldots, T, 1, \ldots, t-1$.

	No. of States (M)	No. of Movies (N)	Analytical	Simulation
Model A	76	118	$1.4\ 10^{-6}$	$5.0\ 10^{-6}$
Model B	58	116	$6.2\ 10^{-6}$	$3.5\ 10^{-7}$
Model C	39	112	$5.7\ 10^{-6}$	–

Table 6.3: Experimental results with actual MPEG video traffic.

of the Markov model are estimated from a finite trace. From Table 6.3 we see that for the particular Star Wars trace, a model with 76 states describes the data sufficiently well and yields a loss probability estimate very close to the actual one (obtained via simulation). In practice it might be desirable to use a model with relatively few states (Model C for instance) that yields a conservative loss probability estimate. Such a conservative model results in only a 5% loss in capacity (transmitting 112 movies instead of 118). Notice that transmitting 112 movies is an enormous gain over the peak rate assignment (11 movies) and not very far away from the upper bound of 135 movies implied by the stability condition.

6.2 Multiclass Admission Control

In this section we turn our attention to the multiclass case and propose an admission control mechanism which takes advantage of the performance analysis results developed in Chapters 3, 4 and 5.

The problem with the effective bandwidth-based admission control, although simple and elegant, is that all calls are treated the same. This is not desirable in truly multimedia situations in future high speed networks where different types of services are going to be operational. In such situations, different models are required to characterize distinct types of traffic (for instance voice traffic can be adequately represented with a Markov modulated process with only two states, whereas video

traffic requires Markov modulated processes of high dimensionality). Moreover, distinct types of traffic often have different quality of service requirements (for instance real-time data have a more stringent loss rate requirement than voice). The scheme that we propose in this section addresses these issues by using the multiclass model of Section 3.1 and the GPS policy. It provides for admission which guarantees different QoS requirements for each type of traffic.

Consider the architecture of Figure 6-4. The notation remains the same as in

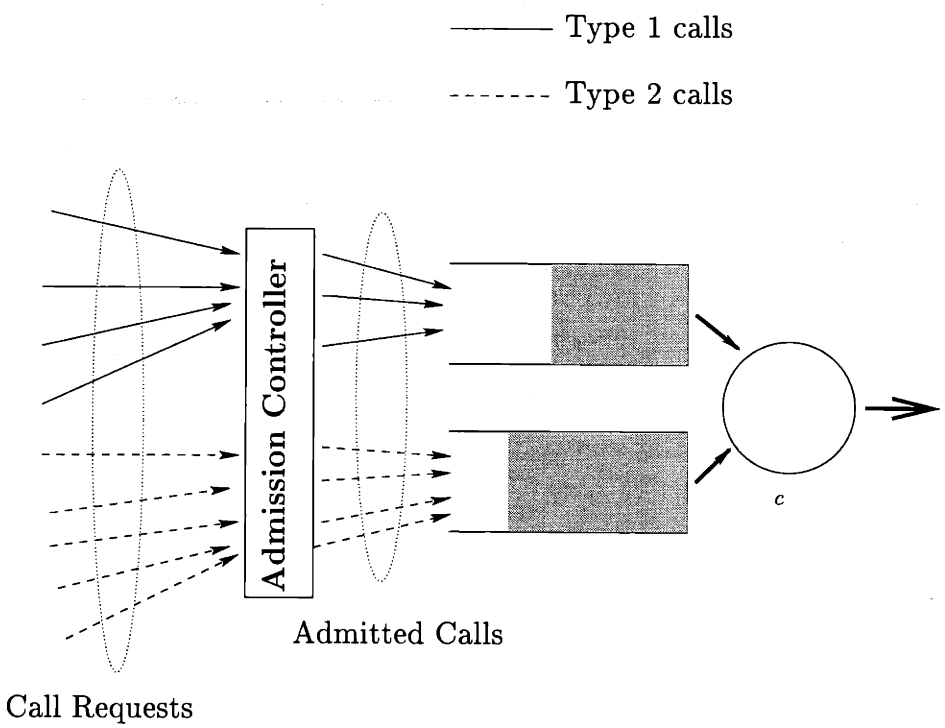

Figure 6-4: An architecture for multiclass admission control.

the previous Section and Chapter 3. Let A^1 (resp. A^2) denote the arrival process for Type 1 (resp. Type 2) calls. With N_1 (resp. N_2) Type 1 (resp. Type 2) calls admitted, the aggregate arrival process in the first (resp. second) buffer, say $\tilde{A}^1$ (resp. $\tilde{A}^2$) is characterized by $\Lambda_{\tilde{A}^1}(\theta) = N_1\Lambda_{A^1}(\theta)$ (resp. $\Lambda_{\tilde{A}^2}(\theta) = N_2\Lambda_{A^2}(\theta)$), for all θ. The

service process is deterministic with rate c b/s, hence $\Lambda_B(\theta) = c\theta$ for all θ.

We now have to satisfy four QoS constraints, namely,

$$\mathbf{P}[L_U^j > U_j] < \delta_j, \qquad j = 1, 2, , \tag{6.8}$$

$$\mathbf{P}[D_U^j > D_{\max}^j] < \delta_j, \qquad j = 1, 2. \tag{6.9}$$

The admission controller has the freedom to select the appropriate buffer sizes U_j and restrict the number of admitted calls in order to guarantee (6.8) and (6.9). We next argue that the appropriate buffer sizes are $U_j = cD_{\max}^j$, $j = 1, 2$. For the same reasons that we presented in the previous section it is not recommended to have buffer sizes $U_j > cD_{\max}^j$, since the system will be admitting packets that will not be able to clear it within $D_{\max}^j$, with certainty. Thus, it has to be the case $U_j \leq cD_{\max}^j$, $j = 1, 2$. Consider now the case $U_1 < cD_{\max}^1$. The system is discarding packets that will depart within $D_{\max}^1$, if the second buffer is and remains empty. In fact, the system is doing this "blindly" without checking the status of the second buffer. It is more efficient to be discarding packets only when there is a danger of violating the QoS constraints. We therefore, set buffer sizes to $U_j = cD_{\max}^j$, $j = 1, 2$. Notice that this no more guarantees the delay constraints, as it was the case in the single class setting of the previous section. That is, if both buffers get full, certain packets will violate their delay QoS constraint (Eq. (6.9)). We deal with this by restricting the number of admitted calls.

Hereafter, as in the previous section we approximate the loss probability with the level crossing probability in an infinite buffer system and we upper bound the delay probability with the one in the infinite buffer system. Notice, now, that with $U_j = cD_{\max}^j$, $j = 1, 2$, the event $L^j > U_j$ implies the event $D^j > D_{\max}^j$. Thus,

$$\mathbf{P}[L^j > U_j] \leq \mathbf{P}[D^j > D_{\max}^j], \qquad j = 1, 2. \tag{6.10}$$

Therefore we need to guarantee, through admission control, only (6.9). Let us denote by θ_{Dj}^* the decay rate of the delay probability that is obtained by applying Thm. 5.3.7.

Namely, for $j = 1, 2$,

$$\theta^*_{Dj} = \sup_{u \geq 0:\ \Lambda_{GPS}(u) < 0} [\Lambda_{Aj}(u) - \Lambda_{GPS}(u)],$$

and the delay probability, for large values of $D^j_{\max}$, is given by

$$\mathbf{P}[D^j > D^j_{\max}] \sim e^{-D^j_{\max}\theta^*_{Dj}}.$$

We can ensure (6.9) if and only if $\theta^*_{Dj} \geq \delta^j_D$, where $\delta^j_D \stackrel{\triangle}{=} -\frac{\log \delta_j}{D^j_{\max}}$. The following definition is of relevance.

Definition 6.2.1

We define the admission region *for the system of Figure 6-4 operated under the GPS policy the set*

$$\mathcal{A} = \{(\phi_1, N_1, N_2):\ \phi_1 \in [0,1],\ N_1, N_2 \in \mathbb{N}_+,\ \theta^*_{Dj} \geq \delta^j_D,\ j = 1, 2\}$$

If a vector $(\phi_1, N_1, N_2) \in \mathcal{A}$, we can ensure (6.9) and due to (6.10) we can also ensure (6.8), that is, all QoS requirements are satisfied.

When an adequate model for the arrival processes is available off-line then the admission region can be calculated also off-line and be used on-line by the admission controller. The calculations required for the admission region are computationally burdensome and can be done in the order of minutes, depending on the complexity of the arrival model. Thus, it may not be appropriate to repeat them for every call request. In practice, these calculations should be performed based on data from an on-line estimator of the call traffic (one estimator for each traffic type) and be updated frequently. Calls are admitted based on the current version of the admission region according to the following algorithm: (assume without loss of generality a type 1 call request)

if $\exists\ \phi_1\ :\ (\phi_1, N_1 + 1, N_2) \in \mathcal{A}$
  then *accept*;
  else *reject*;
end

Notice that the algorithm performs a look-up-table operation which can be done on-line. According to this proposed scheme the admission controller provides the input to the GPS scheduler and the scheduling parameter ϕ_1 can be adjusted to accommodate the current load. One significant advantage of this scheme, over other priority schemes proposed in the literature ([EM94]), is that each type of traffic gets the capacity required by its QoS specifications which include both a measure of loss and one of delay. It allows, for instance, Type 1 traffic to suffer less delay with a larger loss probability than Type 2 traffic, something that can not be achieved with a priority scheme.

6.2.1 The admission region: An example

Here we provide an example of the admission region for the two types of traffic that are described in Subsection 6.1.1. The traffic parameters and the QoS parameters are given by Table 6.1.

Figure 6-5 depicts the admission region for this particular example. For every fixed ϕ_1 and N_1 we plot the maximum allowed number of type 2 calls, N_2, such that $(\phi_1, N_1, N_2) \in \mathcal{A}$. That is, as long as we operate the system under the plotted surface the QoS constraints are satisfied. In Figure 6-6 we show waterfall plots of the admission region to depict better the shape of the region for N_1 constant and for ϕ_1 constant (first and second plot in Figure 6-6, respectively). Finally, for better understanding of the structure of the admission region, in Figure 6-7 we project the region in the N_1-N_2 space for various values of ϕ_1.

Some observations are in order. Recall that the admission region is defined to

satisfy both constraints

$$\theta^*_{D^1} \geq \delta^1_D, \tag{6.11}$$

$$\theta^*_{D^2} \geq \delta^2_D. \tag{6.12}$$

Consider the first plot of Figure 6-6. Notice that for small values of N_1, the maximum allowed N_2 is non-decreasing as ϕ_1 increases in $[0, 1]$. To explain this, notice that for large ϕ_1 we favor type 1 calls and since these are few the constraint (6.11) is not tight. The maximum allowed N_2 is set such that (6.12) is tight. The situation stays the same (i.e., maximum allowed N_2 is constant) as we decrease ϕ_1 until some threshold value ϕ_1^* at which (6.11) gets tight. For smaller ϕ_1 than ϕ_1^*, and since we keep N_1 fixed, to accommodate type 1 calls (i.e., satisfy (6.11)) we need to decrease N_2.

An antipodal phenomenon, in the same plot, is occurring for large values of N_1. For small values of ϕ_1 (6.11) is tight while (6.12) is not tight. Increasing ϕ_1 more than some threshold point ϕ_1^* makes (6.12) tight and thus we can guarantee the QoS constraints only by dropping the maximum allowed N_2.

Let us now turn our attention to the second plot in Figure 6-6 (or, alternatively, Figure 6-7), which depicts cross sections of the admission region for ϕ_1 fixed. Consider cross sections around $\phi_1 = 0.2$ to make the discussion clearer. We can distinguish roughly three regions: (a) small values of N_1, (b) moderate values of N_1, and (c) large values of N_1. In region (a) the maximum allowed N_2 drops almost linearly as we increase N_1. This is occurring because in this region (6.11) is not tight while (6.12) is tight and the only way to increase N_1, without compromising the quality of type 2 calls, is to decrease N_2. The decrease is roughly linear for the following reason: in this region the dominant congestion event is large delays in the second buffer, which are occurring according to the scenario depicted in Figure 5-1(a) in page 155. Recall that large delays are occurring because the second buffer builds up in the first part of that path (interval $[-1 - T, -1]$ with the notation there). Since according to that

path the first buffer stays roughly empty the second buffer gets capacity of $c - N_1 y_1$, where y_1 is the most likely arrival rate of type 1 calls during periods of congestion in the second buffer (solution of an optimization problem similar to the one appearing in Eq. (5.9)). To have (6.12) tight, this capacity should be equal to $N_2 y_2$, where y_2 is the most likely arrival rate of type 2 calls during periods of congestion in the second buffer. Thus $N_2 = \frac{c - N_1 y_1}{y_2}$ which is linear in N_1.

Now, from region (a), as we keep increasing N_1 we enter region (b). Still (6.11) is not tight, however the most likely way that the second buffer generates large delays becomes the scenario of Figure 5-1(d), that is by building up the first buffer also. This means that the first buffer requires capacity $\phi_1 c$ and to accommodate type 2 calls we have decreased their number such that they are satisfied with capacity $\phi_2 c$. We can therefore increase N_1 even more, until we make (6.11) tight, without having to decrease N_2 (notice that in region (b) N_2 is roughly constant). When (6.11) becomes tight we enter region (c) and the only way to further increase N_1 is to drop N_2. The drop is roughly linear for a similar reason to the one explained above. The discussion extends to other values of ϕ_1 away from $\phi_1 = 0.2$, with the three regions mentioned above degenerating to two (see that for ϕ_1 around 1 we can distinguish only regions (a) and (b)).

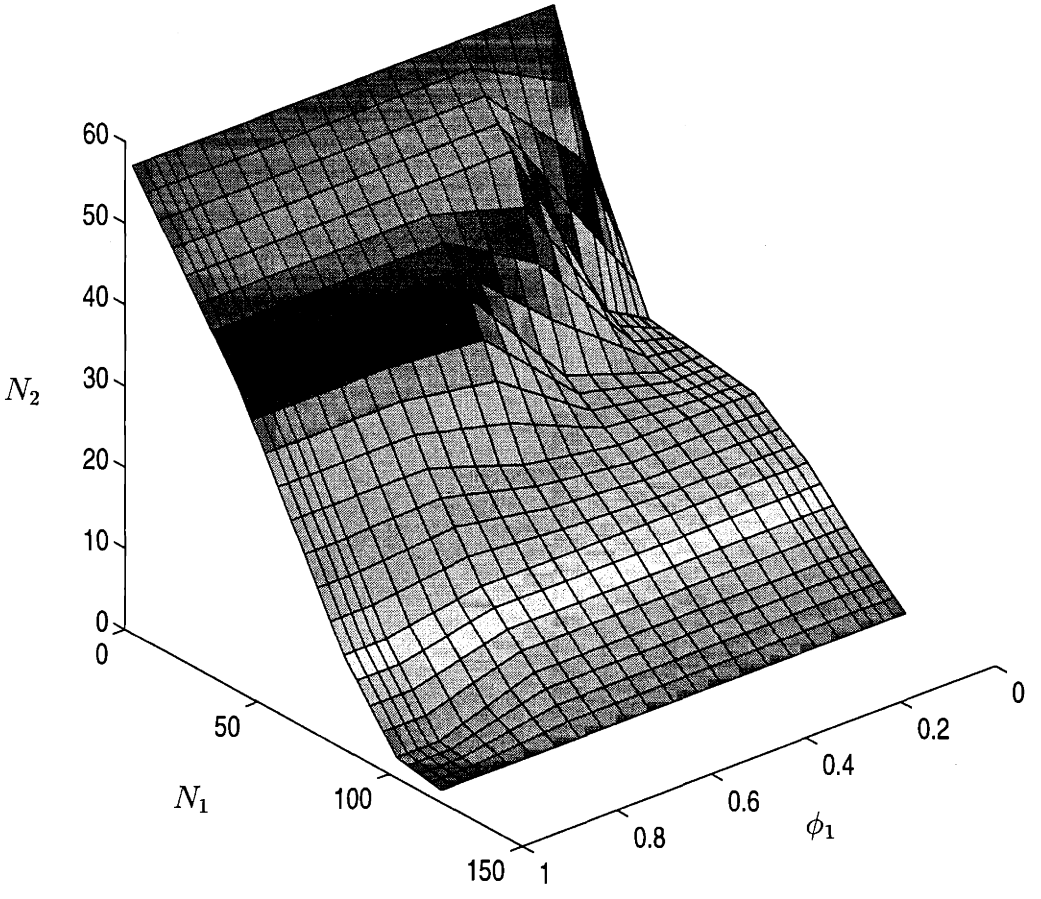

Figure 6-5: The admission region for the traffic model and parameters of Subsection 6.1.1.

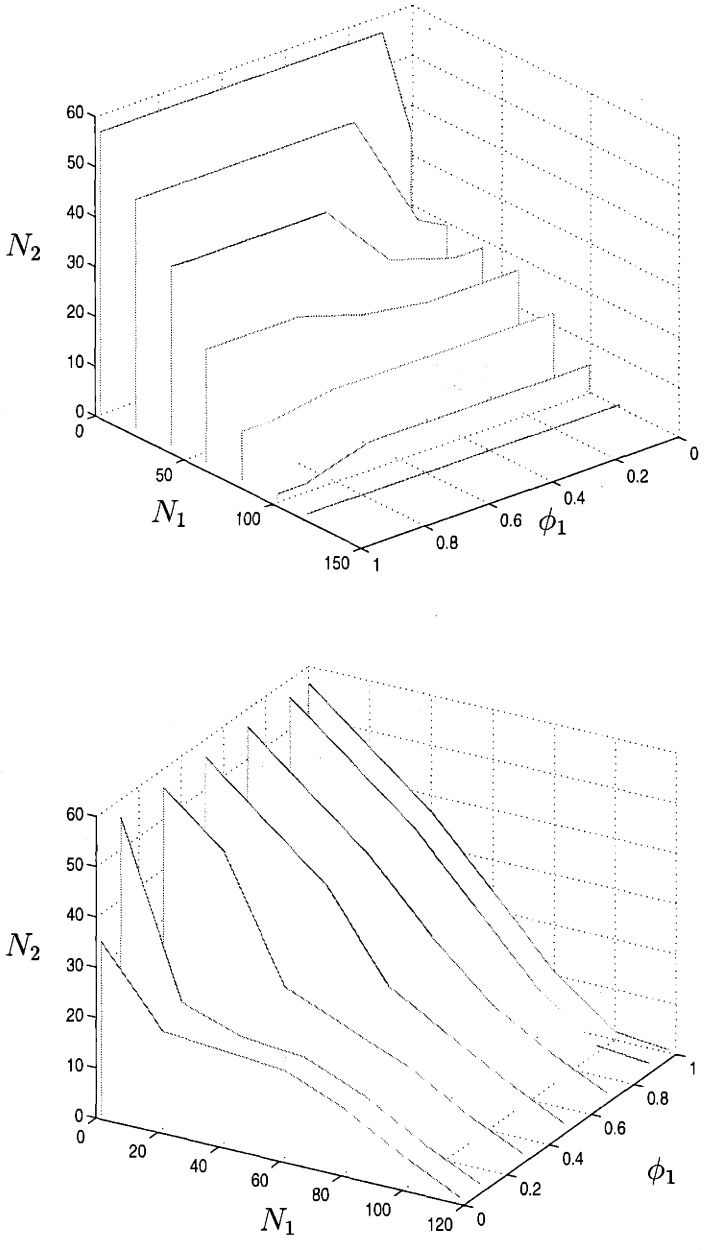

Figure 6-6: Waterfall plots of the admission region for the traffic model and parameters of Subsection 6.1.1.

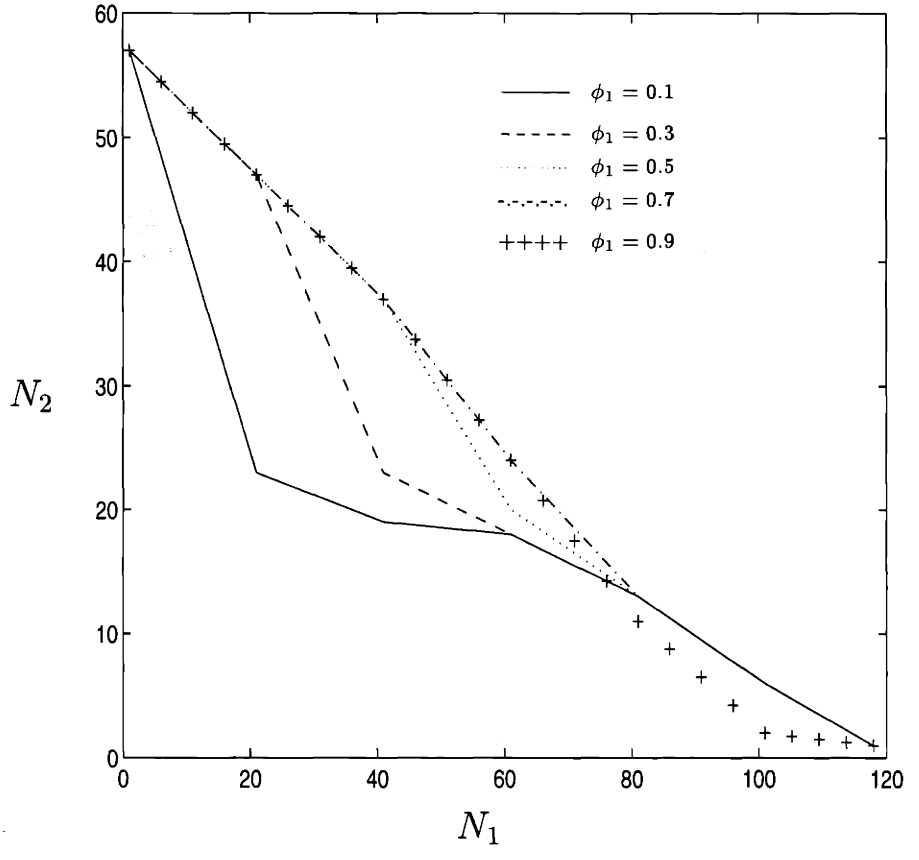

Figure 6-7: Plots of the admission region in the N_1-N_2 space for various values of ϕ_1 for the traffic model and parameters of Subsection 6.1.1.

Chapter 7

Loss Probabilities via Quick Simulation

In this chapter we switch gears once more and attempt to estimate the loss probability in a particular single class buffer through simulation. One reason that one might want to resort to simulation is that in some cases a very accurate estimate of the overflow probability is required, and the asymptotic decay rate that can be calculated analytically, via large deviations, is not sufficient.

However, since we need to estimate probabilities of rare events, direct simulation is very computationally burdensome due to the huge sample size that it requires. We will use the technique of importance sampling to speed up the simulation.

Regarding the structure of this chapter we begin in Section 7.1 with a primer on importance sampling. In Section 7.2 we define the system that we simulate. In Section 7.3 we describe a large deviation result for this system and we derive a change of measure that we use to speed up the simulation. Finally, in Section 7.4 we compare the performance of the quick simulation, the direct Monte Carlo simulation and the analytical large deviation result.

7.1 Importance Sampling Primer

In this section we briefly discuss the main idea behind the importance sampling technique. A detailed discussion of quick simulation techniques can be found in [Buc90]. The idea is rather simple. Consider a random variable X, and assume that we want to estimate $l = \mathbf{E}_P[1_B(X)]$ where $1_B(X)$ denotes the indicator function of the event B, and the expectation is with respect to the distribution P of X. Assume also that X has a density $p(\cdot)$.

If we were to calculate the expectation through direct simulation we would generate a sequence of i.i.d. samples $x_1, \ldots, x_K$ from X and obtain the estimate

$$\hat{l}_P = \frac{1}{K} \sum_{i=1}^{K} 1_B(x_i). \tag{7.1}$$

However, when the event B is very rare we need a huge sample size K to obtain a good estimate. Let now Q be some arbitrary distribution with density $q(\cdot)$ and consider a sequence of i.i.d. samples $y_1, \ldots, y_K$ drawn from Q. We can now form the estimate

$$\hat{l}_Q = \frac{1}{K} \sum_{i=1}^{K} \frac{p(y_i)}{q(y_i)} 1_B(y_i). \tag{7.2}$$

Notice, that the expected value of $\hat{l}_Q$ with respect to Q is l, since

$$\mathbf{E}_Q[\hat{l}_Q] = \frac{1}{K} \sum_{i=1}^{K} \int \frac{p(y_i)}{q(y_i)} 1_B(y_i) q(y_i)\, dy_i = l.$$

We will hereafter call the distribution Q *change of measure* and the ratio $\frac{p(x)}{q(x)}$ *likelihood ratio of P versus Q*. The problem is to find a *change of measure* which reduces the variance of the estimator. We will call *optimal change of measure* the one that minimizes the variance of the estimator.

The variance of the estimator is given by:

$$\begin{aligned}\mathrm{Var}(\hat{l}_Q) =& \frac{1}{K}\int\left(\frac{p(x)}{q(x)}1_B(x) - l\right)^2 q(x)\,dx \\ =& \frac{1}{K}\left[\int \frac{(p(x)1_B(x))^2}{q(x)}\,dx - l^2\right] \end{aligned} \tag{7.3}$$

The above is minimized when $q(x)$ is proportional to $p(x)1_B(x)$. But the normalizing constant is $1/l$, precisely what we are trying to estimate. In general it is hard to obtain the optimal change of measure since it essentially requires solving a large deviations problem. The intuition behind the change of measure idea, is that we find Q under which the event B is typical (and not rare) and we use it to obtain an estimate that has the desired mean. In this chapter we do this heuristically, and show numerical evidence that indeed the simulation speeds up drastically.

7.2 Traffic Model and Problem Definition

In this section we formally define the system of interest and the associated loss probability that we want to estimate.

Consider the system of Figure 7-1. There are N independent traffic sources that are multiplexed into one buffer. The traffic sources conform to the model of Figure 7-2. They generate traffic in a periodic fashion. Assume, without loss of generality that the period is 1. The source can be in one of two states: ON and OFF. In the ON state generates traffic at a constant rate of λ and stays in the ON state for a period of $\delta < 0.5$. The phase of the source, t_{ON}, is uniformly distributed in $[0, 1]$. In the OFF state the source is silent. We assume that the system is stable, that is, it holds $N\delta\lambda < c$.

We want to estimate the loss probability in the above system for some fixed buffer size U, as the number of sources becomes very large. We approximate the loss probability with the level crossing probability in an infinite buffer system. Assume that the sources start feeding the buffer at time $-\infty$. Let $S^A_{t_1,t_2}$ the amount of work

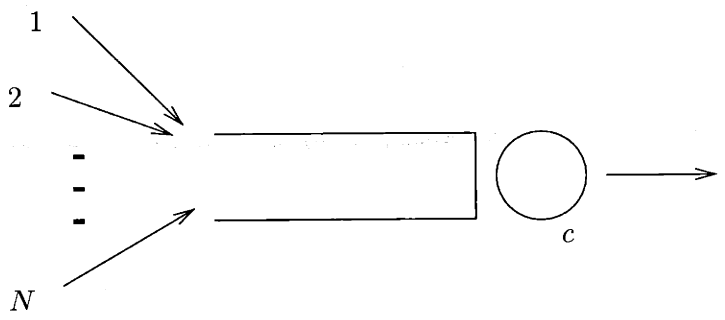

Figure 7-1: A system that multiplexes N traffic sources.

that arrives in the interval $[t_1, t_2]$ $(t_1 \leq t_2)$, per source, and $S^{A,N}_{t_1,t_2}$ the sum of N i.i.d. copies of $S^{A}_{t_1,t_2}$. Let L_t be the queue length, in the infinite buffer system, at some arbitrary time t. Since we are interested in the loss probability for large values of N we scale both the buffer size and the service capacity by N, that is we define $b = U/N$ and $s = c/N$. In particular the quantity that we want to estimate is $\mathbf{P}[L_t > Nb]$, asymptotically as $N \to \infty$.

7.3 Loss Probability and Change of Measure

In this section we discuss a large deviation result for the loss probability that was obtained in [BD94] and infer from it a change of measure that we use in simulating the system.

From the Lindley equation we obtain

$$L_0 = \max_{t \geq 0} [S^{A,N}_{-t,0} - sNt]. \tag{7.4}$$

Let us now define

$$\lambda_t(\theta) = \frac{1}{t} \log \mathbf{E}[e^{\theta(S^{A}_{-t,0} - st)}], \tag{7.5}$$

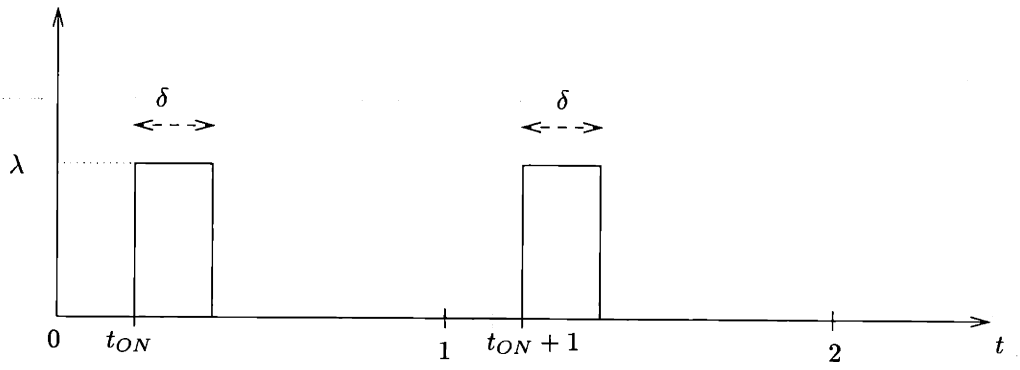

Figure 7-2: The source model.

and let us also denote by $\lambda_t^*(\cdot)$ the convex dual of $\lambda_t(\cdot)$, i.e.,

$$\lambda_t^*(a) = \sup_\theta [\theta a - \lambda_t(\theta)].$$

Since the source model is periodic with period 1 it is not hard to verify that the queue length L_t is also periodic with the same period. Thus, in Eq. (7.4) the maximum can be taken only over $0 \leq t \leq 1$, without loss of generality.

Under an assumption on $\lambda_t(\cdot)$ very similar to our Assumption A, and a local regularity condition on the sample paths of the workload process $\{S_{-t,0}^{A,N} - sNt;\ t \geq 0\}$ the following theorem is proved in [BD94].

Theorem 7.3.1 *([BD94]) For each $b > 0$*

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbf{P}[L_0 > Nb] = -I(b),$$

where

$$I(b) = \inf_{t\geq 0} t\lambda_t^*(b/t).$$

Proof : We present the proof of the lower bound since this is informative on the

change of measure that we use in order to estimate the above probability through simulation. The upper bound proof can be found in [BD94].

We have that

$$\begin{aligned}\liminf_{N\to\infty}\frac{1}{N}\log\mathbf{P}[\sup_{t\geq 0}(S^{A,N}_{-t,0}-sNt)>Nb] \\ \geq \liminf_{N\to\infty}\frac{1}{N}\sup_{t\geq 0}\log\mathbf{P}[(S^{A,N}_{-t,0}-sNt)>Nb] \\ \geq \sup_{t\geq 0}\liminf_{N\to\infty}\frac{1}{N}\log\mathbf{P}[(S^{A,N}_{-t,0}-sNt)>Nb] \qquad (7.6)\end{aligned}$$

Now note that the moment generating function for $(S^{A,N}_{-t,0}-sNt)$ is

$$\mathbf{E}[e^{\theta(S^{A,N}_{-t,0}-sNt)}]=\left(\mathbf{E}[e^{\theta(S^{A}_{-t,0}-st)}]\right)^{N},$$

which by using the definition of $\lambda_t(\cdot)$ implies that

$$t\lambda_t(\theta)=\log\mathbf{E}[e^{\theta(S^{A}_{-t,0}-st)}]=\frac{1}{N}\log\mathbf{E}[e^{\theta(S^{A,N}_{-t,0}-sNt)}].$$

Thus applying the lower bound of Cramér's theorem (the i.i.d. analog of Gärtner-Ellis theorem) to the right hand side of (7.6), we obtain

$$\liminf_{N\to\infty}\frac{1}{N}\log\mathbf{P}[(S^{A,N}_{-t,0}-sNt)>Nb]\geq-(t\lambda_t)^*(b)=-t\lambda_t^*(b/t),$$

where the superscript $*$ denotes the convex dual and the last equality above is obtained by using convex duality properties ([Roc70]). Combining the above with (7.6) we finally obtain the desired result.

■

The above theorem intuitively asserts that for large N the overflow probability behaves as

$$\mathbf{P}[L_0>Nb]\sim e^{-NI(b)}.$$

Let t^* the solution of the optimization problem associated with the large deviations rate function $I(b)$. We can interpret t^* as the most likely duration of the busy period that leads to the overflow. Due to the periodicity of the queue length process we have $t^* \in [0, 1]$.

We next present a heuristic change of measure that we use to speed up the simulation. Let us fix to $t \in [0, 1]$ the duration of the busy period that leads to overflow and ask the question how the distribution of the random source phases should look like in order to have an overflow. Then the queue length at time 0 is

$$L_0^t = S_{-t,0}^{A,N} - sNt, \tag{7.7}$$

The superscript t on L_0 denotes the fact that we have fixed the duration of the busy period (i.e., the maximizing t in Eq. (7.4)). Now, L_0^t is a sum of the N i.i.d random variables, $W_t \stackrel{\Delta}{=} S_{-t,0}^A - st$. For the sum of i.i.d. random variables an optimal change of measure is known. It is the exponential change of measure that is used to prove the lower bound in Cramér's theorem (see [Buc90, DZ93b]). In particular, if we let $\tilde{W}_t$ be the random variable with the changed measure we have

$$dF_{\tilde{W}_t}(w) = \frac{e^{\theta_t^* w} dF_{W_t}(w)}{\mathbf{E}[e^{\theta_t^* W_t}]}, \tag{7.8}$$

where θ_t^* is the optimal solution of

$$\lambda_t^*(b/t) = \sup_{\theta}[\theta b/t - \lambda_t(\theta)]. \tag{7.9}$$

Recall now that only the phase of the source is random, thus, W_t is a function of the phase, say u, which is uniformly distributed in $[0, 1]$. To explicitly denote this we will write $W_t(u) = S_{-t,0}^A(u) - st$. Thus from (7.8) we obtain the change of measure for the phase

$$q(u) = p(u)\frac{e^{\theta_t^* S_{-t,0}^A(u)}}{\mathbf{E}[e^{\theta_t^* S_{-t,0}^A(u)}]}, \tag{7.10}$$

where $p(u)$ is the original uniform density in $[0,1]$, i.e., $p(u) = 1$ for all $u \in [0,1]$.

Let now $\hat{l}_p$ the estimate of $\mathbf{P}[L_0 > bN]$ obtained from the direct simulation and $\hat{l}_q$ the one obtained from the quick simulation. The above change of measure q is optimal, for t fixed, in the sense that it minimizes the following speed factor

$$SF(\hat{l}_q) \stackrel{\triangle}{=} \lim_{N\to\infty} \frac{1}{N} \log[K\mathrm{Var}(\hat{l}_q)],$$

where K is the sample size.

Let us know explicitly calculate, for the particular source model that we are considering, the density q given by (7.10). We divide $[0,1]$ in the three subintervals [1] $[0,\delta]$, $[\delta, 1-\delta]$ and $[1-\delta, 1]$. After a fair amount of routine calculations we obtain that for $t \in [0,\delta]$

$$S^A_{-t,0} = \begin{cases} (t-u)\lambda & \text{if } u \in [0,t] \\ 0 & \text{if } u \in [t, 1-\delta] \\ (u+\delta-1)\lambda & \text{if } u \in [1-\delta, t+1-\delta] \\ t\lambda & \text{if } u \in [t+1-\delta, 1] \end{cases} \qquad t \in [0,\delta], \tag{7.11}$$

and

$$\mathbf{E}[e^{\theta S^A_{-t,0}}] = 2\frac{e^{\theta t\lambda} - 1}{\theta\lambda} + (1-\delta-t) + e^{\theta t\lambda}(\delta - t), \qquad t \in [0,\delta]. \tag{7.12}$$

For $t \in [\delta, 1-\delta]$ we obtain

$$S^A_{-t,0} = \begin{cases} \lambda\delta & \text{if } u \in [0, t-\delta] \\ (t-u)\lambda & \text{if } u \in [t-\delta, t] \\ 0 & \text{if } u \in [t, 1-\delta] \\ (u+\delta-1)\lambda & \text{if } u \in [1-\delta, 1] \end{cases} \qquad t \in [\delta, 1-\delta], \tag{7.13}$$

[1] Recall that we have assumed $\delta < 0.5$.

and

$$\mathbf{E}[e^{\theta S^A_{-t,0}}] = 2\frac{e^{\theta\delta\lambda}-1}{\theta\lambda} + (1-\delta-t) + e^{\theta\delta\lambda}(t-\delta), \qquad t \in [\delta, 1-\delta]. \tag{7.14}$$

Finally, for $t \in [1-\delta, 1]$ we obtain

$$S^A_{-t,0} = \begin{cases} \lambda\delta & \text{if } u \in [0, t-\delta] \\ (t-u)\lambda & \text{if } u \in [t-\delta, 1-\delta] \\ (\delta-1+t)\lambda & \text{if } u \in [1-\delta, t] \\ (u+\delta-1)\lambda & \text{if } u \in [t, 1] \end{cases} \qquad t \in [1-\delta, 1], \tag{7.15}$$

and

$$\mathbf{E}[e^{\theta S^A_{-t,0}}] = e^{\theta\delta\lambda}(t-\delta) + 2\frac{e^{\theta\delta\lambda} - e^{\theta(t-1+\delta)\lambda}}{\theta\lambda} + (t-1+\delta)e^{\theta(\delta-1+t)\lambda}, \qquad t \in [1-\delta, 1]. \tag{7.16}$$

In Figure 7-3 and for $\lambda = 2513.4$ packets/period, $N = 160$ sources, buffer size $U = 1000$ packets, $\delta = 0.1$ periods and $c = 75260$ cells/period, we plot the density $q(u)$ for $t := t^*$ (i.e., the optimal solution of the optimization problem associated with $I(b)$) and for $\theta^*_t := \theta^*_{t^*}$ (i.e., the optimal solution of the optimization problem associated with $\lambda^*_{t^*}(b/t^*)$). For the particular parameters we have chosen it turns out that $t^* = 0.0846$, so Eq. (7.11) and (7.12) are applicable. As it is shown in the plot, in order to have an overflow at t^*, the sources will get to their on state with much higher probability in the intervals $[1-\delta, 1]$ and $[0, t^*]$.

In all the above discussion about the change of measure we have fixed t, the duration of the busy period that leads to overflow, although this is a random variable as well. So, the change of measure in (7.10) is the optimal change of measure conditional on the overflow occurring at t. Initially we used this in the simulation and the results that we were getting were not satisfactory. In fact, instead of obtaining a reduction of the variance we were observing variance inflation of the estimator. The reason for this was that the likelihood ratio $\prod_{i=1}^N p(u_i)/\prod_{i=1}^N q(u_i) = 1/\prod_{i=1}^N q(u_i)$ for a sequence of

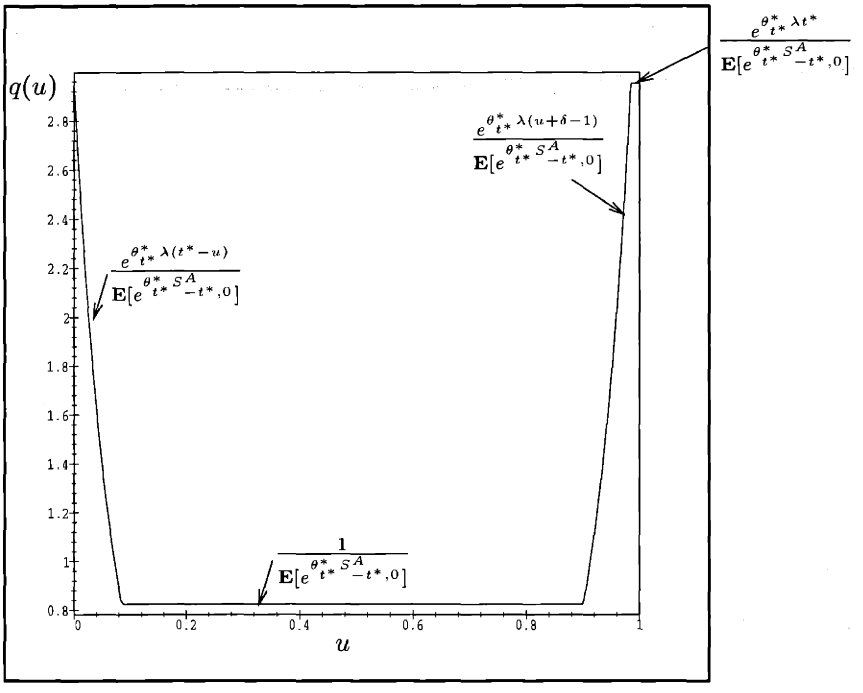

Figure 7-3: The change of measure when the duration of the busy period that leads to overflow is fixed to t^*.

N random phases $u_1, \ldots, u_N$ was behaving very erratically, experiencing huge jumps once in a while. Actually, when we tried to also calculate $\mathbf{E}_q[1/\prod_{i=1}^N q(u_i)]$, through the simulation, which should be exactly 1, we were observing a value very far away from 1. The crux of the matter is that this change of measure is in a sense singular. It forces the sources to be clustered in a way that the overflow occurs at some fixed time t, while in the original problem the overflow occurs at a random time which is uniformly distributed in $[0, 1]$.

To remedy this we employ the following idea. We first generate a sequence of random phases from $q(u)$ and sort them. Let $(U_1, \ldots, U_N)$ be the resulting random

sequence. We then let T be a uniform random variable in $[0,1]$ and shift the vector $(U_1,\ldots,U_N)$ by T to obtain $(\tilde{U}_1,\ldots,\tilde{U}_N) = (U_1+T,\ldots,U_N+T) \pmod 1$. The intuitive idea is that with the shift the overflow time will be uniformly distributed in $[0,1]$. We use samples of $(\tilde{U}_1,\ldots,\tilde{U}_N)$ to generate arrivals in the simulation.

Note now that the likelihood ratio for a sample $(\tilde{u}_1,\ldots,\tilde{u}_N)$ is $1/\mathbf{P}[(\tilde{U}_1=\tilde{u}_1,\ldots,\tilde{U}_N=\tilde{u}_N)]$, where

$$\begin{aligned}\mathbf{P}[(\tilde{U}_1=\tilde{u}_1,\ldots,\tilde{U}_N=\tilde{u}_N)] &= \int_0^1 \mathbf{P}[(U_1=\tilde{u}_1-a,\ldots,U_N=\tilde{u}_N-a) \mid T=a]\,da \\ &= \int_0^1 \mathbf{P}[(U_1=\tilde{u}_1-a,\ldots,U_N=\tilde{u}_N-a)]\,da \\ &= \int_0^1 \prod_{i=1}^{N} q(\tilde{u}_i-a)\,da \qquad (7.17)\end{aligned}$$

The second inequality above holds since T is independent of $(U_1,\ldots,U_N)$. We used the above outlined procedure to obtain samples of arrival patterns in the simulation and the results, which we report in the next section, were very satisfactory. The averaging operation in the above resulted in a much smoother behaviour of the likelihood ratio in the simulation. A similar shifting idea to smooth the likelihood ratios was also used in [Rob91].

7.4 Numerical results

Here we report numerical results that indicate the speed up in the simulation with the change of measure.

We simulated the system of Figure 7-1 for different values of N with the following parameters: $\lambda = 2513.4$ packets/period, $\delta = 0.1$ periods, $b = 6.25$ packets, $s = 470.375$ packets/period. The estimate obtained from the simulation is the loss probability (the fraction of time that buffer stays above the level $U = bN$), denoted by $\mathbf{P}_{\text{loss}}$. In Table 7.4 we compare the estimate obtained from the direct Monte Carlo simulation, the quick simulation with the change of measure described above, and the

estimate obtained from the analytical result of Thm. 7.3.1.

	Direct Simulation		Quick Simulation			
N	$\mathbf{P}_{\text{loss}}$	K	$\mathbf{P}_{\text{loss}}$	K	SU	LD
20	$(5.36 \pm .35) \cdot 10^{-2}$	$1.5 \cdot 10^3$	$(5.27 \pm .26) \cdot 10^{-2}$	$1.0 \cdot 10^3$	1.5	$1.8 \cdot 10^{-1}$
40	$(7.39 \pm .59) \cdot 10^{-3}$	$7.0 \cdot 10^3$	$(7.52 \pm .42) \cdot 10^{-3}$	$2.0 \cdot 10^3$	3.5	$3.5 \cdot 10^{-2}$
60	$(1.10 \pm .09) \cdot 10^{-3}$	$4.0 \cdot 10^4$	$(1.12 \pm .07) \cdot 10^{-3}$	$3.0 \cdot 10^3$	13.3	$6.7 \cdot 10^{-3}$
80	$(2.05 \pm .16) \cdot 10^{-4}$	$2.0 \cdot 10^5$	$(1.96 \pm .12) \cdot 10^{-4}$	$3.0 \cdot 10^3$	66.6	$1.2 \cdot 10^{-3}$
100	$(3.18 \pm .25) \cdot 10^{-5}$	$1.1 \cdot 10^6$	$(3.29 \pm .18) \cdot 10^{-5}$	$5.0 \cdot 10^3$	220.0	$2.4 \cdot 10^{-4}$
120	$(5.40 \pm .44) \cdot 10^{-6}$	$5.5 \cdot 10^6$	$(5.79 \pm .36) \cdot 10^{-6}$	$5.0 \cdot 10^3$	1100.0	$4.6 \cdot 10^{-5}$
160	$(1.75 \pm .18) \cdot 10^{-7}$	$1.0 \cdot 10^8$	$(1.77 \pm .11) \cdot 10^{-7}$	$9.0 \cdot 10^3$	11111.1	$1.6 \cdot 10^{-6}$

Table 7.1: A comparison of results from direct Monte Carlo simulation, quick simulation, and analytical large deviations (LD) results. K denotes the number of iterations (sample size) that the simulation needs to obtain a confident estimate. We define SU (Speed-up) to be the ratio of the iterations needed for the direct simulation versus the iterations needed for the quick simulation.

To obtain the last row of the table, the direct simulation was running for approximately two weeks on a Sparc 20, while the quick simulation needed only 5 min on a Sparc 5 (which is about 50% slower than a Sparc 20). An other interesting observation is that the quick simulation was very robust to changes in the pseudo-random number generator. We used various pseudo-random number generators, from a very sophisticated one with a period of 10^{422} to the standard `random()` with a period of 10^{10} and the results were consistent. In [KLE] the authors use massive parallelism to speed-up the direct simulation for the same problem and they report sensitivity of their results to the pseudo-random number generator.

Chapter 8

Conclusions

In this thesis, we have used the probabilities of packet losses (due to buffer overflows) and large delays to quantify the QoS delivered by the network to the users. Using large deviations techniques we estimated such congestion probabilities in various environments, including a networking and a multimedia environment.

Among the main methodological contributions of the thesis we consider:

1. The proof of a *rigorous induction step* in an acyclic, single class network of G/G/1 queues. This enables us to obtain the tail of congestion (loss and delay) probabilities in all the nodes of the network. This network result is based on a set of mild technical assumptions on external arrival processes [1] which are proven to hold for the internal traffic as well.

2. The introduction of a *deterministic optimal control approach* to establish tight lower bounds on congestion probabilities in multiclass multiplexers. This approach yields the leading exponent of congestion probabilities (under the GPS and the GLQF policy) and characterizes the most likely way that congestion occurs. Tightness of the lower bound is proved independently through the

[1] These assumptions as we note in Chapter 2 are satisfied by processes that are, typically, used in modeling traffic in communication networks such as renewal, Markov modulated and stationary processes with mild mixing conditions.

derivation of a matching upper bound in each case. We believe that optimal control techniques have substantial potential in attacking communication network problems, especially in the large deviations regime. The main reason is that large deviations theory is in a sense reducing the solution of a complicated stochastic problem to the solution of an associated variational problem.

8.1 Summary of Results

We gather here the main results proven in this thesis.

In Chapter 2 we considered an acyclic, single class network of G/G/1 queues. Using a continuous time model (where the arrival and service processes are described via the sequence of interarrivals and service times) we established that for a single G/G/1 queue in isolation the steady-state waiting time, W, satisfies (Thm. 2.2.1)

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[W \geq U] = \theta^*,$$

where $\theta^* < 0$ is the smallest root of the equation

$$\Lambda_A(\theta) + \Lambda_B(-\theta) = 0,$$

and where $\Lambda_A(\cdot)$ and $\Lambda_B(\cdot)$ denote the limiting log-moment generating functions of the arrival and the service process, respectively. Moreover the steady-state queue length, L, satisfies (Thm. 2.2.3)

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbf{P}[L \geq n] = \Lambda_A(\theta^*).$$

External traffic in this network is transformed by three "filtering" operations as it is transmitted within the network. These operations are passing-through-a-queue, superposition of independent processes, and deterministic splitting. An LDP for their output is also obtained in Chapter 2. For the output process, D, of the

passing-through-a-queue operation (i.e., departure of a G/G/1 queue) we established (Thm. 2.3.4)

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbf{P}[S^D_{1,n}\le na]=-\Lambda^{*-}_D(a),$$

where

$$\Lambda^{*-}_D(a)=\Lambda^{*-}_B(a)+\Lambda^{*-}_\Gamma(a)$$

and

$$\Lambda^{*-}_\Gamma(a)=\sup_{\{\theta|\Lambda_A(\theta)+\Lambda_B(-\theta)<0\}}[\theta a-\Lambda^-_A(\theta)].$$

We paid particular attention to the *most likely* way that large deviations of the departure process occur (see Figure 2-7).

For the partial sum $S^{A^1,\cdots,^m}_{1,n}$ of the aggregate process, resulting from the superposition of the m independent processes $A^1_i,\ldots,A^m_i$ $i\in\mathbb{Z}$, we established (Cor. 2.4.2)

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbf{P}[S^{A^1,\cdots,^m}_{1,n}\le na]=-\inf_{\substack{\delta_1+\cdots+\delta_m=1\\ \delta_1,\cdots,\delta_m\ge 0}}\sum_{k=1}^m\delta_k\Lambda^{*-}_{A^k}(a/\delta_k)\triangleq-\Lambda^{*-}_{A^1,\cdots,^m}(a).$$

To show the latter result we proved a result connecting the Palm and the stationary distributions in the large deviations regime.

For the output of the splitting operation where a fraction p of the arrivals from a "master" process A are routed to form the process A^1 we established (Thm. 2.5.1)

$$\lim_{n\to\infty}\frac{1}{n}\log\mathbf{P}[S^{A^1}_{1,n}\le na]=-\frac{1}{p}\Lambda^{*-}_A(ap).$$

Finally, we showed that a set of technical assumptions imposed on the external arrival processes are satisfied by the outputs of the above operations. These operations, along with the results for a single queue in isolation, establish a *calculus* of acyclic single class networks and yield the tail of waiting times and queue lengths in all the nodes of the network.

In Chapter 3, we switched gears and considered the switch of Figure 3-1. Under the GPS policy and by using a discrete time model we showed that the tail probability

of the steady-state queue length, L^1, in the first buffer is characterized by (Thms. 3.7.1 and 3.7.3)

$$\lim_{U\to\infty}\frac{1}{U}\log\mathbf{P}[L^1>U]=-\theta^*_{GPS},$$

where θ^*_{GPS} is the largest positive root of the equation

$$\Lambda_{GPS}(\theta)\stackrel{\Delta}{=}\Lambda_{A^1}(\theta)+\inf_{0\le u\le\theta}[\Lambda_{A^2}(\theta-u)+\Lambda_B(-\theta+\phi_2 u)]=0,$$

and where $\Lambda_{A^1}(\cdot)$, $\Lambda_{A^2}(\cdot)$ and $\Lambda_B(\cdot)$ denote the limiting log-moment generating functions of the two arrival processes and the service process, respectively. In Chapter 5, we characterized the tail probability of the steady-state delay, D^1, in the first buffer, namely

$$\lim_{m\to\infty}\frac{1}{m}\log\mathbf{P}[D^1>m]=-\theta^*_D,$$

where

$$\theta^*_D=\sup_{u_2\ge 0:\ \Lambda_{GPS}(u_2)<0}[\Lambda_{A^1}(u_2)-\Lambda_{GPS}(u_2)].$$

By symmetry results for the queue length in the second buffer can be also obtained. Moreover the performance of priority policies was obtained as a corollary.

For the same switch, operated under the GLQF policy, in Chapter 4 we obtained (Thms. 4.6.1 and 4.6.2)

$$\lim_{U\to\infty}\frac{1}{U}\log\mathbf{P}[L^1>U]=-\theta^*_{GLQF},$$

where θ^*_{GLQF} is the largest positive root of the equation

$$\Lambda_{GLQF}(\theta)=\max\{\Lambda_{A^1}(\theta)+\Lambda_B(-\theta),\ \inf_{0\le u\le\frac{\theta}{1+\beta}}[\Lambda_{A^1}(\theta-u\beta)+\Lambda_{A^2}(u)+\Lambda_B(-u)]\}=0.$$

Critical in our analysis of the multiclass switch (results of Chapters 3, 4 and 5) was the use of deterministic optimal control techniques. These techniques provide a tight lower bound on the various congestion probabilities, and, more importantly, completely characterize the most likely way that congestion builds up.

In Chapter 6, we utilized the GPS results for the multiclass switch (multiplexer) to dimension the buffers in the system of Figure 6-4 and devised an admission control algorithm that guarantees loss and large delay probabilities to both types of traffic. We reported experiments with traffic models and actual traffic that show the relevance of the traffic models and the asymptotics that we have used in this thesis.

Finally, in Chapter 7, we applied an importance sampling technique (deriving an appropriate change of measure) to speed up simulations of the loss probability in the system of Figure 7-1, for large values of the number N of accommodated calls.

8.2 Directions for Future Research

In this section we suggest some directions for future research.

We start by discussing some immediate research goals that follow directly from the results of this thesis. One such short term goal would be to perform a more extensive experimental investigation of the admission control scheme that we proposed in Chapter 6, especially in the multiclass case. In that chapter we reported a limited number of experiments with real MPEG video traffic and we identified the basic trade off that exists in the selection of the dimensionality of the Markov modulated model. However, more research is needed in quantifying this trade off and selecting the proper dimensionality of the Markov modulated model. Moreover, it would also be very interesting to develop approximations of the admission region that we defined in Definition 6.2.1. Calculating this region requires a fair amount of computations, since we need to solve several nonlinear optimization problems. We have proposed to do that off-line and perform admission control on-line, based on the most current evaluation of the admission region. Developing techniques to evaluate the admission region approximately and fast, would allow us to update the admission region very frequently or even to evaluate it on-line. Regarding the quick simulation result of Chapter 7, it would be interesting to extend it in the multiclass case, and incorporate the features (traffic models, scheduling policies) that we used in Chapter 6, in order to speed up the simulations reported there.

Among the long term and challenging research goals we consider the development of techniques to treat multiclass networks. It would be very interesting to bring together the techniques that we used in Chapter 2 to solve the network problem, with the techniques of Chapters 3, 4 and 5, that address the multiclass switch case, in order to establish performance analysis results for multiclass networks operating under various scheduling policies. The more challenging question is how to use admission control in a multiclass and networking setting to prevent congestion and provide end-to-end type-dependent QoS guarantees.

Appendix A

On Assumption F

Here we consider an arbitrary process $\{X_i,\ i \in \mathbb{Z}\}$ that satisfies Assumption F and the following:

For every $\epsilon_1, \epsilon_2, \delta, a > 0$, there exists M_X such that for all $n \geq M_X$

$$e^{-n(\Lambda_X^{*-}(a)+\epsilon_2)} \leq$$
$$\mathbf{P}[S_{i,j}^X - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n \text{ s.t. } (j-i+1) > \delta n]. \quad \text{(A.1)}$$

Inequality (A.1) is implied by the results in [DZ93a], under some mild mixing assumptions on the process $\{X_i,\ i \in \mathbb{Z}\}$. We prove that the process $\{X_i,\ i \in \mathbb{Z}\}$ satisfies Assumption G for the service times (Eq. (2.7)), i.e.,

For every $\epsilon_1, \epsilon_2, a > 0$, there exists M'_X such that for all $n \geq M'_X$

$$e^{-n(\Lambda_X^{*-}(a)+\epsilon_2)} \leq \mathbf{P}[S_{i,j}^X - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n]. \quad \text{(A.2)}$$

Since Assumption G for the arrivals (Eq. (2.6)) is a weaker version of the above it is also satisfied by the process $\{X_i,\ i \in \mathbb{Z}\}$.

Fix positive ϵ_1, ϵ_2 and a. We have

$$\begin{aligned}
&\mathbf{P}[S^X_{i,j} - (j-i+1)a \le \epsilon_1 n,\ 1 \le i \le j \le n] = \\
&= \mathbf{P}[S^X_{i,j} - (j-i+1)a \le \epsilon_1 n,\ 1 \le i \le j \le n \text{ s.t. } (j-i+1) > \delta n, \\
&\qquad S^X_{i,j} - (j-i+1)a \le \epsilon_1 n,\ 1 \le i \le j \le n \text{ s.t. } (j-i+1) \le \delta n] \\
&\ge \mathbf{P}[S^X_{i,j} - (j-i+1)a \le \epsilon_1 n,\ 1 \le i \le j \le n \text{ s.t. } (j-i+1) > \delta n] - \\
&\qquad \mathbf{P}[\exists\ i \le j \in [1,n] \text{ s.t. } (j-i+1) \le \delta n \text{ and } S^X_{i,j} - (j-i+1)a \ge \epsilon_1 n].
\end{aligned} \tag{A.3}$$

where we have used the inequality $\mathbf{P}[A \cap B] \ge \mathbf{P}[A] - \mathbf{P}[B^C]$. Using the union bound and the Gärtner-Ellis Thm. we obtain that for all $\epsilon_3 > 0$ there exists N_1 such that for all $n \ge N_1$

$$\begin{aligned}
&\mathbf{P}[\exists\ i \le j \in [1,n] \text{ s.t. } (j-i+1) \le \delta n \text{ and } S^X_{i,j} - (j-i+1)a \ge \epsilon_1 n] \le \\
&\qquad \le \sum_{\substack{i \le j \in [1,n] \\ (j-i+1) \le \delta n}} \mathbf{P}[S^X_{i,j} - (j-i+1)a \ge \epsilon_1 n] \\
&\qquad \le \sum_{\substack{i \le j \in [1,n] \\ (j-i+1) \le \delta n}} \mathbf{P}[S^X_{1,\delta n} \ge \epsilon_1 n] \\
&\qquad \le n^2 e^{-n\delta(\Lambda^{*+}_X(\frac{\epsilon_1}{\delta}) - \epsilon_3)}.
\end{aligned} \tag{A.4}$$

Now for given $\epsilon_2' > 0$ choose ϵ_3 and δ small enough in order for large n to have

$$n^2 e^{-n\delta(\Lambda^{*+}_X(\frac{\epsilon_1}{\delta}) - \epsilon_3)} \le \frac{1}{2} e^{-n(\Lambda^{*-}_X(a) + \epsilon_2')} \tag{A.5}$$

This can be done since $\Lambda^{*+}_X(\beta) \to \infty$ as $\beta \to \infty$.

Also, by using (A.1) we have that there exists N'' such that for all $n \ge N''$

$$\mathbf{P}[S^X_{i,j} - (j-i+1)a \le \epsilon_1 n,\ 1 \le i \le j \le n \text{ s.t. } (j-i+1) > \delta n] \ge e^{-n(\Lambda^{*-}_X(a) + \epsilon_2')} \tag{A.6}$$

Combining (A.6), (A.5) and (A.4) with (A.3) we obtain that there exists $\hat{N}$ such that for all $n \geq \hat{N}$

$$\mathbf{P}[S_{i,j}^X - (j-i+1)a \leq \epsilon_1 n,\ 1 \leq i \leq j \leq n] \geq \frac{1}{2} e^{-n(\Lambda_X^{*-}(a)+\epsilon_2')} \tag{A.7}$$

Finally, to obtain (A.2) it suffices to choose ϵ_2' such that for large enough n

$$\frac{1}{2} e^{-n(\Lambda_X^{*-}(a)+\epsilon_2')} \geq e^{-n(\Lambda_X^{*-}(a)+\epsilon_2)}$$

■

Bibliography

[AMS82] D. Anick, D. Mitra, and M. M. Sondhi, *Stochastic theory of Data-Handling system with multiple sources*, The Bell System Technical Journal **61** (1982), no. 8, 1871–1894.

[Ana88] V. Anantharam, *How large delays build up in a GI/G/1 queue*, Queueing Systems **5** (1988), 345–368.

[Asm82] S. Asmussen, *Conditioned limit theorems relating a random walk to its associate, with applications to risk reserve processes and the GI/G/1 queue*, Advances in Applied Probability **14** (1982), 143–170.

[BD94] D.D. Botvich and N.G. Duffield, *Large deviations, the shape of the loss curve, and economies of scale in large multiplexers*, Preprint, 1994.

[BG91] D.P. Bertsekas and R.G. Gallager, *Data networks*, 2nd ed., Prentice-Hall, 1991.

[BM92] D. Bertsimas and G. Mourtzinou, *A unified method to analyze overtake free queueing systems*, Working paper, Operations Research Center, MIT, 1992, To appear in *Advances of Applied Probability*, 1996.

[BN90] D. Bertsimas and D. Nakazato, *The departure process from a GI/G/1 queue and its applications to the analysis of tandem queues*, Working paper OR 245-91, Operations Research Center, MIT, 1990.

[BN95] D. Bertsimas and D. Nakazato, *The general distributional Little's law and its applications*, Operations Research **43** (1995), no. 2, 298–310.

[BPT94] D. Bertsimas, I.Ch. Paschalidis, and J.N. Tsitsiklis, *On the large deviations behaviour of acyclic networks of G/G/1 queues*, Tech. Report LIDS-P-2278, Laboratory for Information and Decision Systems, MIT, December 1994.

[Buc90] J. A. Bucklew, *Large deviation techniques in decision, simulation, and estimation*, Wiley, New York, 1990.

[Cha94a] C.S. Chang, *Sample path large deviations and intree networks*, Preprint, 1994.

[Cha94b] C.S. Chang, *Stability, queue length and delay of deterministic and stochastic queueing networks*, IEEE Transactions on Automatic Control **39** (1994), no. 5, 913–931.

[CHJS94] C.S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin, *Effective bandwidth and fast simulation of ATM intree networks*, Performance Evaluation **20** (1994), 45–65.

[Cra38] H. Cramér, *Sûr un nouveau théorème-limite de la théorie des probabilités*, In *Actualités Scientifiques et Industrielles*, no. 736 in *Colloque consacré à la théorie des probabilités*, pages 5–23, Hermann, Paris, 1938.

[Cru91a] R. L. Cruz, *A calculus for network delay, Part I: Network elements in isolation*, IEEE Transactions on Information Theory **37** (1991), no. 1, 114–131.

[Cru91b] R. L. Cruz, *A calculus for network delay, Part II: Network analysis*, IEEE Transactions on Information Theory **37** (1991), no. 1, 132–141.

[CW93] C. Courcoubetis and R. Weber, *Effective bandwidths for stationary sources*, Preprint, 1993.

[CW95] C. Courcoubetis and R. Weber, *Estimation of overflow probabilities for state-dependent service of traffic streams with dedicated buffers*, Talk at the RSS Workshop in Stochastic Networks, Edinburgh, U.K., 1995.

[CZ95] C.S. Chang and T. Zajic, *Effective bandwidths of departure process from queues with time varying capacities*, Proceedings IEEE Infocom '95 (Boston, Massachusetts), vol. 3, April 1995, pp. 1001–1009.

[Dal61] Rae Dalven, *The complete poems of Cavafy translated by Rae Dalven*, Harcourt, Brace & World, Inc., New York, 1961.

[DKS90] A. Demers, S. Keshav, and S. Shenker, *Analysis and simulation of a fair queueing algorithm*, Journal of Internetworking: Research and Experience 1 (1990), 3–26.

[DLO+94] N. G. Duffield, J.T. Lewis, N. O'Connell, R. Russell, and F. Toomey, *Statistical issues raised by the Bellcore data*, Preprint, 1994.

[dVCW93] G. de Veciana, C. Courcoubetis, and J. Walrand, *Decoupling bandwidths for networks: A decomposition approach to resource management*, Memorandum, Electronics Research Laboratory, U.C. Berkeley, 1993.

[dVK95] G. de Veciana and G. Kesidis, *Bandwidth allocation for multiple qualities of service using Generalized Processor Sharing*, Technical report SCC-94-01, Systems Communications & Control group, Department of Electrical and Computer Engineering, The University of Texas at Austin, 1994; Revised 1995.

[dVW93] G. de Veciana and J. Walrand, *Effective bandwidths: Call admission, traffic policing & filtering for ATM networks*, Memorandum, Electronics Research Laboratory, U.C. Berkeley, 1993.

[DZ93a] A. Dembo and T. Zajic, *Large deviations: From empirical mean and measure to partial sums processes*, Preprint, 1993.

[DZ93b] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, Jones and Bartlett, 1993.

[EHL+94] A. I. Elwalid, D. Heyman, T.V. Lakshman, D. Mitra, and A. Weiss, *Fundamental bounds and approximations for ATM multiplexers with applications to video teleconferencing*, Preprint, 1994, To appear *J. on Selected Areas in Communications*.

[EM93] A. I. Elwalid and D. Mitra, *Effective bandwidth of general Markovian traffic sources and admission control of high speed networks*, IEEE/ACM Transactions on Networking **1** (1993), no. 3, 329–343.

[EM94] A. I. Elwalid and D. Mitra, *Analysis, approximations and admission control of a multiple-service multiplexing system with priorities*, Preprint, 1994.

[GA94] A. Ganesh and V. Anantharam, *The stationary tail probability of an exponential server tandem fed by renewal arrivals*, Preprint, 1994.

[Gar93] M.W. Garrett, *Contributions toward real–time services on packet switched networks*, Ph.D. thesis, Columbia University, 1993.

[GGG+93] H.R. Gail, G. Grover, R. Guérin, S.L. Hantler, Z. Rosberg, and M. Sidi, *Buffer size requirements under longest queue first*, Performance Evaluation **18** (1993), 133–140.

[GH91] R.J. Gibbens and P.J. Hunt, *Effective bandwidths for the multi-type UAS channel*, Queueing Systems **9** (1991), 17–28.

[GW94] P.W. Glynn and W. Whitt, *Logarithmic asymptotics for steady-state tail probabilities in a single-server queue*, J. Appl. Prob. **31A** (1994), 131–156.

[Hui88] J. Y. Hui, *Resource allocation for broadband networks*, IEEE Journal on Selected Areas in Communications **6** (1988), no. 9, 1598–1608.

[HW94] I. Hsu and J. Walrand, *Admission control for ATM networks*, Memorandum UCB/ERL M94/44, Electronics Research Laboratory, University of California, Berkeley, 1994.

[Kel79] F.P. Kelly, *Reversibility and stochastic networks*, Wiley, New York, 1979.

[Kel91] F. P. Kelly, *Effective bandwidths at multi-class queues*, Queueing Systems **9** (1991), 5–16.

[Kel93] F. P. Kelly, *On tariffs, policing and admission control for multiservice networks*, Research report, Statistical Laboratory, University of Cambridge, England, 1993.

[Kel96] F. P. Kelly, *Notes on effective bandwidths*, Stochastic Networks: Theory and Applications (S. Zachary, I.B. Ziedins, and F.P. Kelly, eds.), vol. 9, Oxford University Press, 1996.

[Kin70] J.F.C. Kingman, *Inequalities in the theory of queues*, Journal of the Royal Statistical Society **32** (1970), 102–110.

[KLE] K. Kumaran, B. Lubachevsky, and A. Elwalid, *Massively parallel simulations of ATM systems*, Preprint.

[KWC93] G. Kesidis, J. Walrand, and C.S. Chang, *Effective bandwidths for multiclass Markov fluids and other ATM sources*, IEEE/ACM Transactions on Networking **1** (1993), no. 4, 424–428.

[LeG91] D. LeGall, *MPEG: A video compression standard for multimedia applications*, Communications of the ACM (1991), 47–58.

[LTWW94] W.E. Leland, M.S. Taqqu, W. Willinger, and D.V. Wilson, *On the self-similar nature of ethernet traffic (extended version)*, IEEE/ACM Transactions on Networking **2** (1994), 1–15.

[MAS88] B. Maglaris, D. Anastassiou, and P. Sen, *Statistical multiplexing in packet video communications*, IEEE Transactions on Communications **36** (1988), no. 7, 834–843.

[Nor94] I. Norros, *A storage model with self-similar input*, Queueing Systems **16** (1994), 387–396.

[O'C95a] N. O'Connell, *Large deviations in queueing networks*, Preprint, 1995.

[O'C95b] N. O'Connell, *Queue lengths and departures at single-server resources*, Talk at the RSS Workshop in Stochastic Networks, Edinburgh, U.K., 1995.

[PG93] 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 Transactions on Networking **1** (1993), no. 3, 344–357.

[PG94] A.K. Parekh and R.G. Gallager, *A generalized processor sharing approach to flow control in integrated services networks: The multiple node case*, IEEE/ACM Transactions on Networking **2** (1994), no. 2, 137–150.

[PW89] S. Parekh and J. Walrand, *A quick simulation method for excessive backlogs in networks of queues*, IEEE Transactions on Automatic Control **34** (1989), no. 1, 54–66.

[Rob91] J.W. Roberts (ed.), *Performance evaluation and design of multiservice networks*, Commission of the European Communities, 1991.

[Roc70] R.T. Rockafellar, *Convex analysis*, Princeton University Press, 1970.

[SW95] A. Shwartz and A. Weiss, *Large deviations for performance analysis*, Chapman and Hall, New York, 1995.

[TGT95] D. Tse, R.G. Gallager, and J.N. Tsitsiklis, *Statistical multiplexing of multiple time-scale Markov streams*, J. on Selected Areas in Communications **13** (1995), no. 6.

[Wal88] J. Walrand, *An introduction to queueing networks*, Prentice Hall, 1988.

[Wei95] A. Weiss, *An introduction to large deviations for communication networks*, J. on Selected Areas in Communications **13** (1995), no. 6, 938–952.

[Wil95] W. Willinger, *Traffic modeling for high-speed networks: Theory versus practice*, Stochastic Networks (F.P. Kelly and R.J. Williams, eds.), Springer Verlag, New York, 1995.

[YS93] O. Yaron and M. Sidi, *Performance and stability of communication networks via robust exponential bounds*, IEEE/ACM Transactions on Networking **1** (1993), no. 3, 372–385.

[Zha95] Zhi-Li Zhang, *Large deviations and the generalized processor sharing scheduling: Upper and lower bounds. Part I:Two-queue systems*, Technical report, Computer Science Department, University of Massachusetts at Amherst, 1995.