Asymptotic Buffer Overflow Probabilities in Multiclass Multiplexers, Part I: The GPS Policy [1]

Dimitris Bertsimas
dbertsim@aris.mit.edu

Ioannis Ch. Paschalidis
yannis@mit.edu

John N. Tsitsiklis
jnt@mit.edu

LABORATORY FOR INFORMATION AND DECISION SYSTEMS
AND
OPERATIONS RESEARCH CENTER
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MA 02139

June 1996

LIDS Report: LIDS-P-2341

[1] A preliminary version of these results was reported in [BPT95]. The results in this paper are included in [Pas96].
Research 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.

Abstract

In this paper and its sequel [BPT96] we consider a multiclass multiplexer, with segregated buffers for each type of traffic. Under specific scheduling policies for sharing bandwidth, we seek the asymptotic (as the buffer size goes to infinity) tail of the buffer overflow probability for each buffer. We assume dependent arrival and service processes as it is usually the case in models of bursty traffic. 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 probabilities. We relate the lower bound derivation to a *deterministic optimal control problem*, which we explicitly solve. Optimal state trajectories of the control problem correspond to typical congestion scenarios. We explicitly and in detail characterize the *most likely* modes of overflow. Here we consider the *generalized processor sharing policy (GPS)* and in [BPT96] the *generalized longest queue first policy (GLQF)*. The performance of strict priority policies is obtained as a corollary. A comparison of the loss probability characteristics of the GPS and GLQF policy is made in [BPT96]. Our results have important implications for traffic management of high-speed networks. They extend the deterministic, worst-case analysis of [PG93] to the case where a detailed statistical model of the input traffic is available and can be used as a basis for an admission control mechanism which guarantees a different loss probability for each type of traffic.

Keywords: Communication networks, ATM-based B-ISDN, Large Deviations.

1 Introduction

Future high speed, packet-switched communication networks, for example ATM-based B-ISDN networks, will accommodate various types of traffic, namely, digitized voice, encoded video, and data. One of the central and most challenging current problems in computer networking is the design and the operation of these networks.

Congestion causes packet losses, due to buffer overflows, and excessive delays, phenomena that greatly contribute to the degradation of the *quality of service (qos)* that the network delivers to its users. Since voice and video are very sensitive to such phenomena the network should have the ability to guarantee certain qos parameters to the user. We quantify qos by the probabilities of excessive delay and buffer overflow. It is desirable to operate the network in a regime where packet loss probabilities are very small, e.g., in the order of 10^{-9}. Moreover, large delays should also have a correspondingly small probability. An essential step for preventing congestion through a variety of control mechanisms (buffer dimensioning, admission control, resource allocation) is to determine how it occurs and to estimate the probabilities of congestion phenomena, i.e., buffer overflows and large delays. The problem is particularly difficult 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 the *large deviations regime* and obtain asymptotic expressions for the tails of congestion probabilities.

In this paper and its sequel [BPT96] we focus at a simplified version of the problem which retains the most salient features, that is, it is multiclass and has correlated arrival and service processes. In particular, we consider a multiclass multiplexer (one node), with segregated buffers for each type of traffic. Under specific scheduling policies for sharing bandwidth we seek the asymptotic (as the buffer size goes to infinity) tail of the buffer overflow probability for each buffer. In other words, we estimate the loss probability for each type of traffic. In this paper we consider the *generalized processor sharing policy (GPS)* (introduced in [DKS90] and further explored in [PG93, PG94]), and in its sequel [BPT96] the *generalized longest queue first policy (GLQF)*. Both of these 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 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 probabilities. We prove that overflows occur 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 the deterministic, worst-case analysis of [PG93] to the case where a detailed statistical model of the input traffic is available. They can be used as a basis for an admission control mechanism which guarantees desirable loss probability, and allows for different requirements for each type of traffic.

We wish to note at this point that although our principal motivation for studying this problem is computer networking, our results have applications in other queueing situations, e.g. service industry and manufacturing systems.

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]. The extension of these ideas to single class networks, although much harder, has been treated in various versions and degrees of rigor in [BPT94, GA94, Cha95, O'C95a, dVCW93].

Closer to the subject of this paper, 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 this paper, we provide an *optimal control formulation* of the problem. 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 overflow probability is the optimal value of the control problem, which we explicitly solve. Optimal state-trajectories of the control problem correspond to the most likely modes of overflow; 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 the two overflow 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. Moreover, our work extends the GPS results in the literature to the case of stochastic service capacity. This extension 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 overflows occur. To see this recall 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 [BPT94, CZ95].

Regarding the structure of this paper, we begin in Section 2 with a brief review of the large deviations results that we use in this paper. We also state a set of assumptions that arrival and service processes need to conform to. In Section 3 we formally define the multiclass model that we consider and in Section 4 we formally define the GPS policy. Moreover, in the latter section, we provide an outline of the methodology that we follow in proving our results. In Section 5 we prove a lower bound on the overflow probability and in Section 6 we introduce the optimal control formulation and solve the control problem. In Section 7 we summarize the most likely modes of overflow obtained from the solution of the control problem and in Section 8 we prove the matching upper bound. We gather our main results in Section 9, where we also treat the special case of strict priority policies. Conclusions are in Section 10.

2 Preliminaries

In this section we review some basic results on the Large Deviations Theory [DZ93b, SW95, Buc90] that will be used in the sequel.

We first state the Gärtner-Ellis Theorem (see Bucklew [Buc90], and Dembo and Zeitouni [DZ93b]) which establishes a *Large Deviations Principle (LDP)* for dependent random variables in $\mathbb{R}$. It is a generalization of Cramer's theorem which applies to independent and identically distributed (iid) random variables.

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

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

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*

$$\Lambda(\theta) \triangleq \lim_{n \to \infty} \Lambda_n(\theta) = \lim_{n \to \infty} \frac{1}{n} \log \mathbf{E}[e^{\theta S_n}] \tag{2}$$

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

Upper Bound: *For every closed set F*

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbf{P}\left[\frac{S_n}{n} \in F\right] \leq - \inf_{a \in F} \Lambda^*(a). \tag{3}$$

Lower Bound: *For every open set G*

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbf{P}\left[\frac{S_n}{n} \in G\right] \geq - \inf_{a \in G} \Lambda^*(a), \tag{4}$$

where

$$\Lambda^*(a) \triangleq \sup_\theta (\theta a - \Lambda(\theta)). \tag{5}$$

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 (5), it also holds

$$\Lambda(\theta) = \sup_a(\theta a - \Lambda^*(a)). \tag{6}$$

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

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

A stronger concept than the LDP for the partial sum random *variable* $S_n \in \mathbb{R}$, is the LDP for the partial sum *process* (*Sample path LDP*)

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

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], under certain mild mixing conditions on the stationary sequence $\{X_i;\ i \geq 1\}$, the authors establish an LDP for the process $S_n(\cdot)$ in $D[0,1]$ (the space of right continuous functions with left limits).

Their result is a starting point for our analysis in this paper. In particular, we will be assuming the following version of the sample path LDP.

Assumption B
For all $m \in \mathbb{N}$, for every $\epsilon_1, \epsilon_2 > 0$ and for every scalars $a_0, \dots, a_{m-1}$, there exists $M > 0$ such that for all $n \geq M$ and all $k_0, \dots, k_m$ with $1 = k_0 \leq k_1 \leq \cdots \leq k_m = n$,

$$\begin{aligned} e^{-\left(n\epsilon_2 + \sum_{i=0}^{m-1}(k_{i+1}-k_i)\Lambda^*(a_i)\right)} &\leq \mathbf{P}[|S_{k_{i+1}} - S_{k_i} - (k_{i+1}-k_i)a_i| \leq \epsilon_1 n,\ i = 0, \dots, m-1] \\ &\leq e^{\left(n\epsilon_2 - \sum_{i=0}^{m-1}(k_{i+1}-k_i)\Lambda^*(a_i)\right)}. \end{aligned} \tag{7}$$

A detailed discussion of this Assumption, and the technical conditions under which it is satisfied is given by Dembo and Zajic in [DZ93a]. 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), Assumption B is a consequence of Mogulskii's theorem (see [DZ93b]). Intuitively, Assumption B deals with the probability of sample paths that are constrained to be within a tube around a "polygonal" path made up with linear segments of slopes $a_0, \dots, a_{m-1}$. In [DZ93a] it is proved that this assumption is satisfied by processes that are commonly used in modeling the input traffic to communication networks, that is, renewal processes, Markov modulated processes and correlated stationary processes with mild mixing conditions.

In [Cha95] a uniform bounding condition is given under which the above Assumption is true, and is verified that the condition is satisfied by renewal, Markov-modulated and stationary processes with mild mixing conditions. Using this uniform bounding condition it is not hard to verify (see [Cha95] for a proof) that the following assumption is satisfied. This assumption can be viewed as the "convex dual analog" of Assumption B.

Assumption C

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 $1 = k_0 \leq k_1 \leq \cdots \leq k_m = n$,

$$\mathbf{E}[e^{\theta \cdot Z}] \leq \exp\{\sum_{j=1}^{m}[(k_j - k_{j-1})\Lambda(\theta_j) + \Gamma(\theta_j)]\}, \tag{8}$$

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

On a notational remark, in the rest of the paper we will be denoting by $S^X_{i,j} \triangleq \sum_{k=i}^{j} X_k$, $i \leq j$, 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. (2) and (5) for definitions), respectively, of the process X.

3 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 1. We assume a slotted time model (i.e., discrete time) and we let A^1_i (resp. A^2_i), $i \in \mathbb{Z}$, 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 $\{A^1_i;\ i \in \mathbb{Z}\}$, $\{A^2_i;\ i \in \mathbb{Z}\}$ 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^1_i and L^2_i, 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 $\{L^j_i, j = 1, 2, i \in \mathbb{Z}\}$ are stationary (under a work-conserving policy, the system reaches

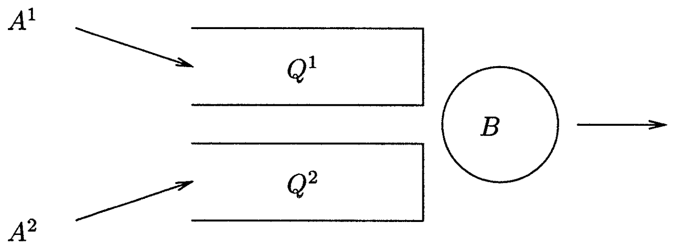

Figure 1: A multiclass model.

steady-state due to the stability condition (9) 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

$$\mathbf{E}[B_i] > \mathbf{E}[A_i^1] + \mathbf{E}[A_i^2]. \tag{9}$$

We further assume that the arrival and service processes satisfy a LDP (Assumption A), as well as Assumptions B and C. As we have noted in Section 2, 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.

4 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)

$$L^1_{i+1} + L^2_{i+1} = [L^1_i + L^2_i + A^1_i + A^2_i - B_i]^+,$$

and

$$L^j_{i+1} \leq [L^j_i + A^j_i - \phi_j B_i]^+, \qquad j = 1, 2,$$

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

We are interested in estimating the overflow probability $\mathbf{P}[L^1_i > U]$ 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

$$\mathbf{P}[L^1_i > U] \sim e^{-U\theta^*_{GPS}}, \tag{10}$$

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 $\theta(\omega)$. For every ω, this probability is a lower bound on $\mathbf{P}[L^1_i > U]$. We select the tightest lower bound by performing the minimization $\theta^*_{GPS} = \min_\omega \theta(\omega)$, which amounts to 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 $\mathbf{P}[L^1_i > U]$ by first obtaining a sample path upper bound, i.e., $L^1_i \leq \tilde{L}^1_i$ (which implies $\mathbf{P}[L^1_i > U] \leq \mathbf{P}[\tilde{L}^1_i > U]$) and establishing that $\mathbf{P}[\tilde{L}^1_i > U]$ is at most $e^{-U\theta^*_{GPS}}$.

5 A Lower Bound

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

Proposition 5.1 ***(GPS Lower Bound)*** *Assuming that the arrival and service processes satisfy Assumptions A and B, and under the GPS policy, the steady-state queue length L^1 of queue Q^1 satisfies*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \geq -\theta^*_{GPS}, \tag{11}$$

where θ^*_{GPS} is given by

$$\theta^*_{GPS} = \min\left[\inf_{a>0} \frac{1}{a}\Lambda^{I*}_{GPS}(a), \inf_{a>0} \frac{1}{a}\Lambda^{II*}_{GPS}(a)\right], \tag{12}$$

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

$$\Lambda^{I*}_{GPS}(a) \triangleq \inf_{\substack{x_1+x_2-x_3=a \\ x_2\le\phi_2 x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)], \tag{13}$$

and

$$\Lambda^{II*}_{GPS}(a) \triangleq \inf_{\substack{x_1-\phi_1 x_3=a \\ x_2\ge\phi_2 x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \tag{14}$$

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

$$\begin{array}{llll} \underline{\text{Scenario 1}}: & x_1 + x_2 - x_3 = a & \underline{\text{Scenario 2}}: & x_1 - \phi_1 x_3 = a \\ & x_2 \le \phi_2 x_3 & & x_2 \ge \phi_2 x_3. \end{array} \tag{15}$$

Under Scenario 1, the first queue receives the maximum capacity (at a rate of $x_3 - x_2$) while 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 B we have

$$\mathbf{P}[L^1_0 \ge n(a - \epsilon'_1)] \ge \sup_{\substack{x_1+x_2-x_3=a \\ x_2\le\phi_2 x_3}} \mathbf{P}[\, |S^{A^1}_{-n,-i-1} - (n-i)x_1| \le \epsilon_1 n, \; i = 0, 1, \ldots, n-1]$$
$$\times \mathbf{P}[\, |S^{A^2}_{-n,-i-1} - (n-i)x_2| \le \epsilon_2 n, \; i = 0, 1, \ldots, n-1]$$

$$\times \mathbf{P}[\,|S^B_{-n,-i-1} - (n-i)x_3| \leq \epsilon_3 n,\ i = 0,1,\ldots,n-1]$$
$$\geq \exp\Big\{-n\Big(\inf_{\substack{x_1+x_2-x_3=a \\ x_2\leq\phi_2 x_3}}[\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)]+\epsilon\Big)\Big\}$$
$$= \exp\{-n(\Lambda^{I*}_{GPS}(a)+\epsilon)\}, \tag{16}$$

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

$$\mathbf{P}[L^1_0 \geq n(a-\epsilon'_2)] \geq \exp\{-n(\Lambda^{II*}_{GPS}(a)+\epsilon')\}, \tag{17}$$

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

Combining Eqs. (16) and (17), we obtain that for all $\epsilon, \epsilon' > 0$ there exists N such that for all $n > N$

$$\frac{1}{n}\log \mathbf{P}[L^1_0 \geq n(a-\epsilon)] \geq -(\min(\Lambda^{I*}_{GPS}(a), \Lambda^{II*}_{GPS}(a)) + \epsilon'). \tag{18}$$

As a final step to this proof, by letting $U = n(a-\epsilon)$, we obtain that for all $\epsilon, \epsilon' > 0$ there exists U_0 such that for all $U > U_0$

$$\frac{1}{U}\log \mathbf{P}[L^1 > U] = \frac{1}{n(a-\epsilon)}\log \mathbf{P}[L^1_0 \geq n(a-\epsilon)] \geq -\frac{1}{a-\epsilon}(\min(\Lambda^{I*}_{GPS}(a), \Lambda^{II*}_{GPS}(a)) + \epsilon'),$$

which implies

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

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

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

■

6 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 $L_0^1 > U$, 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 B. With the scaling introduced here as $U \to \infty$ the sequence of slopes $a_0, a_1, \ldots, a_{m-1}$ appearing there converges 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

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

[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 [Cha95].

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 the previous paragraph.

$$\begin{aligned} \text{(GPS-OVERFLOW) minimize} & \int_{-T}^{0} [\Lambda^*_{A^1}(x_1(t)) + \Lambda^*_{A^2}(x_2(t)) + \Lambda^*_B(x_3(t))]\ dt \qquad (19) \\ \text{subject to: } & L^1(-T) = L^2(-T) = 0 \\ & L^1(0) = 1 \\ & L^2(0) :\ \text{free} \\ & T :\ \text{free} \\ & \{L^j(t) :\ t \in [-T, 0],\ j = 1, 2\} \in \text{(GPS-DYNAMICS)}. \end{aligned}$$

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 6.1 *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*

[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.

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 states at $t = -T_1$ and $t = -T_2$.

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

$$\begin{aligned} L^1(-T_1) &= a_1 > 0, \qquad L^1(-T_2) = b_1 > 0, \\ L^2(-T_1) &= a_2 > 0, \qquad L^2(-T_2) = b_2 > 0, \end{aligned} \tag{20}$$

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

$$\bar{x}_i(\tau) = \frac{1}{T_1 - T_2} \int_{-T_1}^{-T_2} x_i(t)\, dt, \qquad i = 1, 2, 3,\ \forall \tau \in [-T_1, -T_2], \tag{21}$$

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

$$L^j(t) = a_j + \int_{-T_1}^{t} [x_j(\tau) - \phi_j x_3(\tau)] > 0, \qquad j = 1, 2,\ t \in (-T_1, -T_2). \tag{22}$$

Moreover, we also have

$$L^j(-T_2) = a_j + \int_{-T_1}^{-T_2} [x_j(\tau) - \phi_j x_3(\tau)] = b_j, \qquad j = 1, 2. \tag{23}$$

Notice now that the time-average trajectory, has the same end points (i.e., satisfies (20)), 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

$$\begin{aligned} x_i^Q(t) &= x_i^L(t/\alpha), \qquad i = 1,2,3,\ t \in [-\alpha T, 0], \\ Q^j(t) &= \alpha L^j(t/\alpha), \qquad j = 1,2,\ t \in [-\alpha T, 0], \end{aligned}$$

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 2(a), (b) and (c). If the trajectory in Figure 2(a) is optimal then so does the scaled version (by $\alpha = a_2/a_1$) in Figure 2(b) and as consequence its segment which appears in Figure 2(c) is also optimal (since we have a free time problem).

Using the homogeneity property we can make the reduction in Figure 2(e), starting from any arbitrary trajectory with constant controls as the one appearing in Figure 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(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(a) and (b). Consider now the trajectories in Figure 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 (a), (b) and (d).

Finally, consider the state trajectory in Figure 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

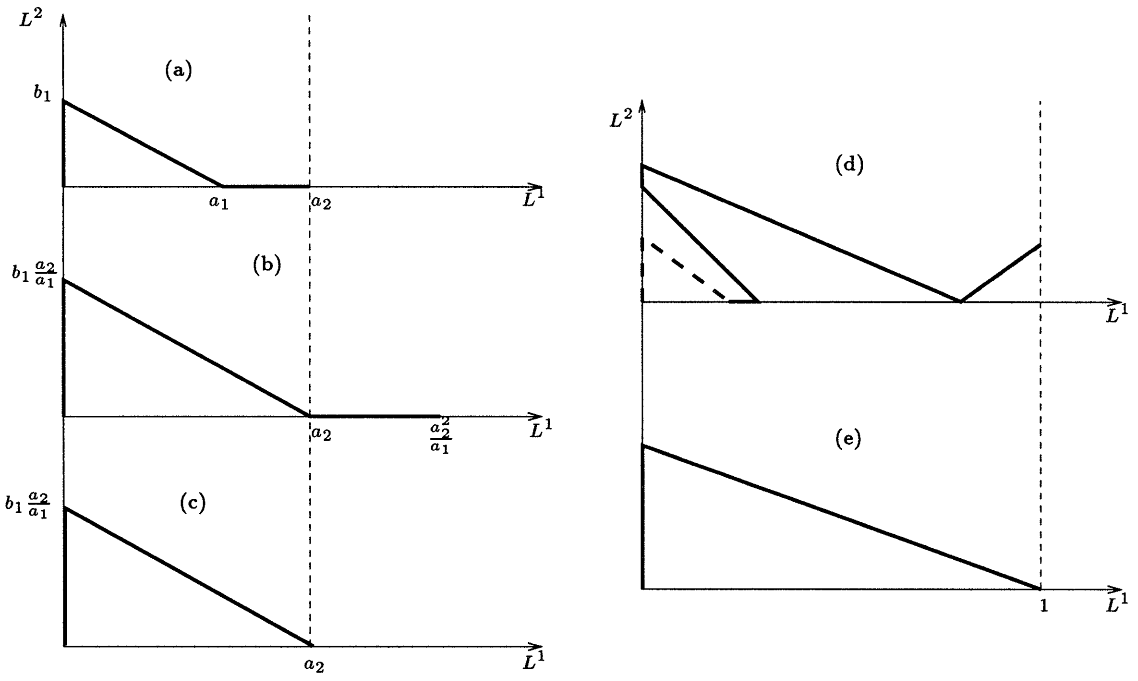

Figure 2: By the homogeneity property, optimality of the trajectory in (a) implies optimality of the trajectory in (b) which by its turn implies optimality of the trajectory in (c). Using the homogeneity property the trajectory in (d) reduces to the one in (e).

$1-\zeta$ 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 constraints are

$$\begin{gathered} x_1 \le \phi_1 x_3, \\ \zeta T(x_1 + x_2 - x_3) + (1-\zeta) T(y_2 - \phi_2 y_3) = 0, \\ (1-\zeta) T(y_1 - \phi_1 y_3) = 1. \end{gathered}$$

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 6.1) it is at least as profitable to have $x_2 = y_2 = \bar{x}_2$. The corresponding trajectory

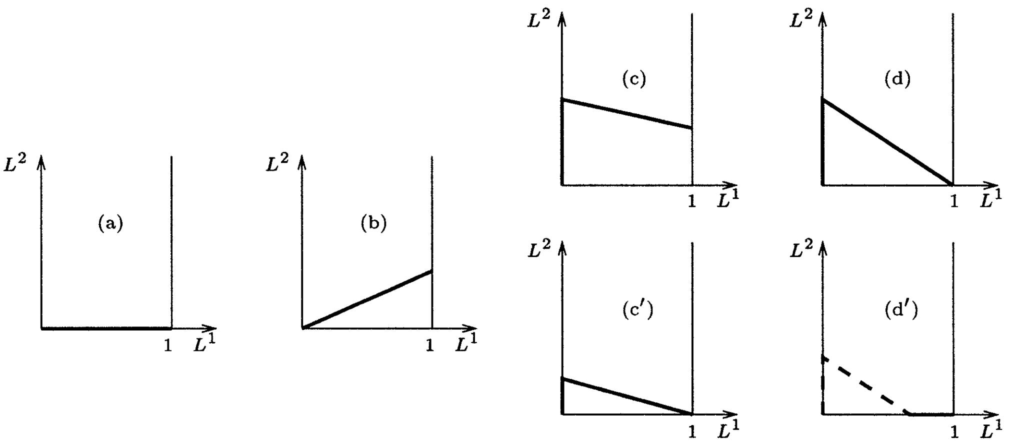

Figure 3: Candidates for optimal state trajectories are depicted in (a), (b), (c) and (d). The trajectory in (c) is reduced to the one in (c′) which has the same form as the one in (d). The trajectory in (d) is reduced to the one in (d′) which is contradicted by the time-homogeneity property. Hence,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

can either have the form in Figure 3(a) or Figure 3(d). If the latter is the case then

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

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(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(a) is at least as cheap.

Therefore, for every state trajectory of (GPS-OVERFLOW), there exists one of the form depicted in Figure 3(a) or (b) with no larger cost. We next calculate the optimal value of (GPS-OVERFLOW). The best trajectory of the form shown in Figure 3(a) has value

$$\inf_{T} \inf_{\substack{x_1+x_2-x_3=\frac{1}{T} \\ x_2 \leq \phi_2 x_3}} T[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)], \tag{24}$$

which is equal to $\inf_T[T\Lambda^{I*}_{GPS}(1/T)]$ by the definition in (13). The best trajectory of the form shown in Figure 3(b) has value

$$\inf_T \inf_{\substack{x_1-\phi_1 x_3=\frac{1}{T}\\ x_2\geq\phi_2 x_3}} T[\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)], \tag{25}$$

which is equal to $\inf_T[T\Lambda^{II*}_{GPS}(1/T)]$ by the definition in (14). 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 (12). In summary we have established the following:

Theorem 6.2 *The optimal value of the problem (GPS-OVERFLOW) is given by θ^*_{GPS}, as it is defined in (12).*

It is of interest to investigate under what conditions on the parameters of the arrival and service processes the trajectory in Figure 3(a) dominates the one in (b) and vice versa. We will distinguish two cases: $\mathbf{E}[A^2]\geq\phi_2\mathbf{E}[B]$ 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(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 6.3 *If $\mathbf{E}[A^2]\geq\phi_2\mathbf{E}[B]$ then optimal state trajectories of (GPS-OVERFLOW) can be restricted to have the form in Figure 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(a) which has optimal value given by the expression in (24). 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(b). Thus we have reduced optimal state trajectories to Figure 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.

■

7 The most likely paths

As we have explained in the Section 4 we will prove an upper bound that matches the lower bound in Proposition 5.1. This is sufficient to guarantee that the two scenarios identified in the proof of Proposition 5.1 (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 $\Lambda^{I*}_{GPS}(a^*)$ (Eq. (13)). The trajectory in L^1-L^2 space is depicted in Figure 3(a).

Case 2: Suppose $\theta^*_{GPS} = \inf_a \Lambda^{II*}_{GPS}(a)/a$ 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 $\Lambda^{II*}_{GPS}(a^*)$ (Eq. (14)). The trajectory in L^1-L^2 space is depicted in Figure 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 paper 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.

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

8.1 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 (9) 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

$$L_0^1 \leq S^{A^1}_{-n^*,-1} - \phi_1 S^B_{-n^*,-1}. \tag{26}$$

Thus,

$$\begin{aligned}\mathbf{P}[L_0^1 > U] \leq & \mathbf{P}[\exists n \geq 0 \text{ s.t. } S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1} > U] \\ \leq & \mathbf{P}[\max_{n\geq 0}(S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1}) > U]. \end{aligned} \tag{27}$$

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

$$\begin{aligned}\mathbf{E}[e^{\theta \max_{n\geq 0}(S^{A^1}_{-n,-1}-\phi_1 S^{B}_{-n,-1})}] \leq & \sum_{n\geq 0} \mathbf{E}[e^{\theta(S^{A^1}_{-n,-1}-\phi_1 S^{B}_{-n,-1})}] \\ \leq & \sum_{n\geq 0} e^{n(\Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)+\epsilon)} \\ \leq & K(\theta,\epsilon) \qquad \text{if } \Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)<0, \end{aligned} \tag{28}$$

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 (27) to obtain

$$\begin{aligned}\mathbf{P}[L_0^1 > U] \leq & \mathbf{E}[e^{\theta \max_{n\geq 0}(S^{A^1}_{-n,-1}-\phi_1 S^{B}_{-n,-1})}]e^{-\theta U} \\ \leq & K(\theta,\epsilon)e^{-\theta U} \qquad \text{if } \Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)<0. \end{aligned} \tag{29}$$

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

Proposition 8.1 *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. 6.3.

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 8.2 *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 8.1 we establish the following proposition.

Proposition 8.3 ***(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. 6.3. Due to Lemma 8.2 it suffices to prove that $\Lambda_{A^1}(\theta) + \Lambda_B(-\phi_1\theta)$ is the convex dual of $\Lambda^*(a) \triangleq \inf_{x_1-\phi_1 x_3=a}[\Lambda^*_{A^1}(x_1) + \Lambda^*_B(x_3)]$. 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 the convex dual of $\Lambda^*(a)$ is

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

■

8.2 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

$$L_0^1 \leq L_0^1 + L_0^2 = L_{-k^*}^2 + S_{-k^*,-1}^{A^1} + S_{-k^*,1}^{A^2} - S_{-k^*,1}^{B}. \tag{30}$$

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 $-n^*$ 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

$$L_{-k^*}^2 = S_{-n^*,-k^*-1}^{A^2} - \phi_2 S_{-n^*,-k^*-1}^{B}, \tag{31}$$

which when combined with (30) yields

$$L_0^1 \leq S_{-k^*,-1}^{A^1} + S_{-n^*,-1}^{A^2} - S_{-k^*,-1}^{B} - \phi_2 S_{-n^*,-k^*-1}^{B}. \tag{32}$$

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

$$L_0^1 \leq S_{-k^*,-1}^{A^1} - \phi_1 S_{-k^*,-1}^{B}. \tag{33}$$

We will use the bound in (32) when $S_{-n^*,-1}^{A^2} \leq \phi_2 S_{-n^*,-1}^{B}$ and the bound in (33) otherwise. Namely we will use

$$L_0^1 \leq \begin{cases} S_{-k^*,-1}^{A^1} + S_{-n^*,-1}^{A^2} - S_{-k^*,-1}^{B} - \phi_2 S_{-n^*,-k^*-1}^{B} & \text{if } S_{-n^*,-1}^{A^2} \leq \phi_2 S_{-n^*,-1}^{B} \\ S_{-k^*,-1}^{A^1} - \phi_1 S_{-k^*,-1}^{B} & \text{if } S_{-n^*,-1}^{A^2} \geq \phi_2 S_{-n^*,-1}^{B}. \end{cases} \tag{34}$$

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

We have

$$
\begin{aligned}
&\mathbf{P}[L_0^1 > U \text{ and } \Omega_1] \leq \\
&\leq \mathbf{P}[\exists n \geq k \geq 0 \text{ s.t. } S^{A^2}_{-n,-1} \leq \phi_2 S^B_{-n,-1} \text{ and} \\
&\qquad\qquad S^{A^1}_{-k,-1} + S^{A^2}_{-n,-1} - S^B_{-k,-1} - \phi_2 S^B_{-n,-k-1} > U] \\
&\leq \mathbf{P}[\max_{\{n\geq k\geq 0:\ S^{A^2}_{-n,-1}\leq \phi_2 S^B_{-n,-1}\}} (S^{A^1}_{-k,-1} + S^{A^2}_{-n,-1} - S^B_{-k,-1} - \phi_2 S^B_{-n,-k-1}) > U]. \qquad (35)
\end{aligned}
$$

For sample paths in Ω_2 we have

$$
\begin{aligned}
&\mathbf{P}[L_0^1 > U \text{ and } \Omega_2] \leq \\
&\leq \mathbf{P}[\exists n \geq k \geq 0 \text{ s.t. } S^{A^2}_{-n,-1} \geq \phi_2 S^B_{-n,-1} \text{ and } S^{A^1}_{-k,-1} - \phi_1 S^B_{-k,-1} > U] \\
&\leq \mathbf{P}[\max_{\{n\geq k\geq 0:\ S^{A^2}_{-n,-1}\geq \phi_2 S^B_{-n,-1}\}} (S^{A^1}_{-k,-1} - \phi_1 S^B_{-k,-1}) > U]. \qquad (36)
\end{aligned}
$$

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

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

which after bringing the constraints in the objective function become

$$L^I_{GPS,1} = \max_{n\geq k\geq 0} \inf_{u\geq 0} [S^{A^1}_{-k,-1} + (1-u) S^{A^2}_{-n,-1} - (1-u\phi_2) S^B_{-k,-1} - \phi_2 (1-u) S^B_{-n,-k-1}], \qquad (37)$$

and

$$L^{II}_{GPS,1} = \max_{n\geq k\geq 0} \inf_{u\geq 0} [S^{A^1}_{-k,-1} + u S^{A^2}_{-n,-1} + (-u\phi_2 - \phi_1) S^B_{-k,-1} - u\phi_2 S^B_{-n,-k-1}]. \qquad (38)$$

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

$$
\begin{aligned}
&\mathbf{E}[e^{\theta L^I_{GPS,1}}] \leq \\
&\leq \sum_{n\geq 0} \sum_{0\leq k\leq n} \inf_{u\geq 0} \mathbf{E}[\exp\{\theta[S^{A^1}_{-k,-1} + (1-u) S^{A^2}_{-n,-1} - (1-u\phi_2) S^B_{-k,-1}
\end{aligned}
$$

$$
\begin{aligned}
&\qquad\qquad\qquad\qquad - \phi_2(1-u)S^B_{-n,-k-1}]\}] \\
&\leq \sum_{n\geq 0}\sum_{0\leq k\leq n} \inf_{u\geq 0} \exp\{(n-k)[\Lambda_{A^2}(\theta-\theta u)+\Lambda_B(-\theta\phi_2(1-u))] \\
&\qquad\qquad + k[\Lambda_{A^1}(\theta)+\Lambda_{A^2}(\theta-\theta u)+\Lambda_B(-\theta(1-u\phi_2))]+\Gamma(\theta,u)\} \\
&\leq \sum_{n\geq 0} n \sup_{\zeta\in[0,1]} \inf_{u\geq 0} \exp\{n[\zeta(\Lambda_{A^2}(\theta-\theta u)+\Lambda_B(-\theta\phi_2(1-u))) \\
&\quad + (1-\zeta)(\Lambda_{A^1}(\theta)+\Lambda_{A^2}(\theta-\theta u)+\Lambda_B(-\theta(1-u\phi_2)))+\tfrac{\Gamma(\theta,u)}{n}]\},
\end{aligned}
\tag{39}
$$

where we let $\zeta = \frac{n-k}{n}$. In the second inequality above we have used Assumption C with $m = 2$. Let us now define

$$
\begin{aligned}
\Lambda^I_{GPS,1}(\theta) \triangleq \sup_{\zeta\in[0,1]} \inf_{u\geq 0} &[\zeta(\Lambda_{A^2}(\theta-\theta u)+\Lambda_B(-\theta\phi_2(1-u))) + \\
&+ (1-\zeta)(\Lambda_{A^1}(\theta)+\Lambda_{A^2}(\theta-\theta u)+\Lambda_B(-\theta(1-u\phi_2)))].
\end{aligned}
$$

Let $u^*(\theta)$ be the optimal u in the above optimization problem. From (39) we have

$$
\begin{aligned}
&\mathbf{E}[e^{\theta L^I_{GPS,1}}] \leq \\
&\leq \sum_{n\geq 0} n \sup_{\zeta\in[0,1]} \exp\{n[\zeta(\Lambda_{A^2}(\theta-\theta u^*)+\Lambda_B(-\theta\phi_2(1-u^*))) \\
&\quad + (1-\zeta)(\Lambda_{A^1}(\theta)+\Lambda_{A^2}(\theta-\theta u^*)+\Lambda_B(-\theta(1-u^*\phi_2)))+\tfrac{\Gamma(\theta,u^*)}{n}]\}.
\end{aligned}
\tag{41}
$$

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 (41) converges to a constant, with respect to n, $K_1(\theta,\epsilon)$. That is,

$$
\mathbf{E}[e^{\theta L^I_{GPS,1}}] \leq K_1(\theta,\epsilon), \qquad \text{if } \Lambda^I_{GPS,1}(\theta) < 0. \tag{42}
$$

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

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

$$\leq \sum_{n\geq 0} n \sup_{\zeta\in[0,1]} \inf_{u\geq 0} \exp\{n[\zeta(\Lambda_{A^2}(\theta u) + \Lambda_B(-\theta\phi_2 u)))$$
$$+ (1-\zeta)(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta u) + \Lambda_B(-\theta(\phi_1 + u\phi_2))) + \tfrac{\Gamma'(\theta,u)}{n}]\}. \tag{43}$$

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

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

Let $\hat{u}^*(\theta)$ be the optimal u in the above optimization problem. From (43) we have

$$\mathbf{E}[e^{\theta L^{II}_{GPS,1}}] \leq$$
$$\leq \sum_{n\geq 0} n \sup_{\zeta\in[0,1]} \exp\{n[\zeta(\Lambda_{A^2}(\theta\hat{u}^*) + \Lambda_B(-\theta\phi_2\hat{u}^*)))$$
$$+ (1-\zeta)(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta\hat{u}^*) + \Lambda_B(-\theta(\phi_1 + \hat{u}^*\phi_2))) + \tfrac{\Gamma'(\theta,\hat{u}^*)}{n}]\}. \tag{45}$$

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 (45) converges to a constant, with respect to n, $K_2(\theta,\epsilon')$. That is,

$$\mathbf{E}[e^{\theta L^{II}_{GPS,1}}] \leq K_2(\theta,\epsilon'), \qquad \text{if } \Lambda^{II}_{GPS,1}(\theta) < 0. \tag{46}$$

We can now invoke the Markov inequality and by using the bounds (39) and (43) on (35) and (36) obtain

$$\begin{aligned}\mathbf{P}[L^1_0 > U] \leq & \mathbf{P}[L^1_0 > U \text{ and } \Omega_1] + \mathbf{P}[L^1_0 > U \text{ and } \Omega_2] \\ \leq & (\mathbf{E}[e^{\theta L^{I}_{GPS,1}}] + \mathbf{E}[e^{\theta L^{II}_{GPS,1}}])e^{-\theta U} \\ \leq & (K_1(\theta,\epsilon) + K_2(\theta,\epsilon'))e^{-\theta U}, \qquad \text{if } \max(\Lambda^{I}_{GPS,1}(\theta), \Lambda^{II}_{GPS,1}(\theta)) < 0.\end{aligned} \tag{47}$$

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

Proposition 8.4 *If* $\mathbf{E}[A^1] < \phi_2\mathbf{E}[B]$ *and assuming that the arrival and service processes satisfy Assumptions A and C*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1_0 > U] \leq - \sup_{\{\theta\geq 0:\ \max(\Lambda^{I}_{GPS,1}(\theta),\Lambda^{II}_{GPS,1}(\theta))<0\}} \theta.$$

We are now left with proving that this upper bound matches the lower bound, θ^*_{GPS}. The result which is based on Lemma 8.2 and convex duality is established in the next proposition.

Proposition 8.5 ***(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 C, 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

$$\Lambda^{I*}_{GPS,1}(a) \triangleq \inf_{\substack{\zeta(x_2-\phi_2 x_3)+(1-\zeta)(y_1+y_2-y_3)=a \\ \zeta(x_2-\phi_2 x_3)+(1-\zeta)(y_2-\phi_2 y_3)\leq 0 \\ 0\leq\zeta\leq 1}} [\zeta(\Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)) \\ + (1-\zeta)(\Lambda^*_{A^1}(y_1) + \Lambda^*_{A^2}(y_2) + \Lambda^*_B(y_3))], \quad (48)$$

and

$$\Lambda^{II*}_{GPS,1}(a) \triangleq \inf_{\substack{(1-\zeta)(y_1-\phi_1 y_3)=a \\ \zeta(x_2-\phi_2 x_3)+(1-\zeta)(y_2-\phi_2 y_3)\geq 0 \\ 0\leq\zeta\leq 1}} [\zeta(\Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)) \\ + (1-\zeta)(\Lambda^*_{A^1}(y_1) + \Lambda^*_{A^2}(y_2) + \Lambda^*_B(y_3))], \quad (49)$$

which by a change of variables can be written as

$$\Lambda^{I*}_{GPS,1}(a) = \inf_{\substack{(x_2-\phi_2 x_3)+(y_1+y_2-y_3)=a \\ (x_2-\phi_2 x_3)+(y_2-\phi_2 y_3)\leq 0}} \inf_{\zeta\in[0,1]} [\zeta(\Lambda^*_{A^2}(x_2/\zeta) + \Lambda^*_B(x_3/\zeta)) + \\ + (1-\zeta)(\Lambda^*_{A^1}(y_1/(1-\zeta)) + \Lambda^*_{A^2}(y_2/(1-\zeta)) + \Lambda^*_B(y_3/(1-\zeta)))], \quad (50)$$

and

$$\Lambda^{II*}_{GPS,1}(a) = \inf_{\substack{(y_1-\phi_1 y_3)=a \\ (x_2-\phi_2 x_3)+(y_2-\phi_2 y_3)\geq 0}} \inf_{\zeta\in[0,1]} [\zeta(\Lambda^*_{A^2}(x_2/\zeta) + \Lambda^*_B(x_3/\zeta)) \\ + (1-\zeta)(\Lambda^*_{A^1}(y_1/(1-\zeta)) + \Lambda^*_{A^2}(y_2/(1-\zeta)) + \Lambda^*_B(y_3/(1-\zeta)))]. \quad (51)$$

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

$$\begin{aligned}
&\sup_a[\theta a-\Lambda^{I*}_{GPS,1}(a)]=\\
&\quad=\sup_{\zeta\in[0,1]}\sup_a\sup_{\substack{\zeta(x_2-\phi_2x_3)+(1-\zeta)(y_1+y_2-y_3)=a\\ \zeta(x_2-\phi_2x_3)+(1-\zeta)(y_2-\phi_2y_3)\le 0\\ 0\le\zeta\le 1}}[\theta a-\zeta(\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3))\\
&\qquad\qquad-(1-\zeta)(\Lambda^*_{A^1}(y_1)+\Lambda^*_{A^2}(y_2)+\Lambda^*_B(y_3))]\\
&\quad=\sup_{\zeta\in[0,1]}\inf_{u\ge 0}\sup_{\substack{x_2,x_3\\ y_1,y_2,y_3}}[\theta\zeta(x_2-\phi_2x_3)+\theta(1-\zeta)(y_1+y_2-y_3)-u\zeta(x_2-\phi_2x_3)\\
&\qquad\qquad-u(1-\zeta)(y_2-\phi_2y_3)-\zeta(\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3))\\
&\qquad\qquad-(1-\zeta)(\Lambda^*_{A^1}(y_1)+\Lambda^*_{A^2}(y_2)+\Lambda^*_B(y_3))]\\
&\quad=\sup_{\zeta\in[0,1]}\inf_{u\ge 0}[\zeta(\Lambda_{A^2}(\theta-u)+\Lambda_B(-\theta\phi_2+u\phi_2))\\
&\qquad\qquad+(1-\zeta)(\Lambda_{A^1}(\theta)+\Lambda_{A^2}(\theta-u)+\Lambda_B(-\theta+u\phi_2)))]\\
&\quad=\Lambda^I_{GPS,1}(\theta).
\end{aligned}$$

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

$$\theta_I\triangleq\inf_{a>0}\frac{1}{a}\Lambda^{I*}_{GPS,1}(a), \tag{52}$$

and

$$\theta_{II}\triangleq\inf_{a>0}\frac{1}{a}\Lambda^{II*}_{GPS,1}(a). \tag{53}$$

Using the result of Lemma 8.2, 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 4 indicates, due to convexity, $\theta^*_{GPS,1}\triangleq\min(\theta_I,\theta_{II})$ is the largest positive root of the equation $\Lambda_{GPS,1}(\theta) \triangleq \max[\Lambda^{I}_{GPS,1}(\theta), \Lambda^{II}_{GPS,1}(\theta)] = 0$, that is $-\theta^*_{GPS,1}$ is equal to the upper bound established in Prop. 8.4. The last thing we have

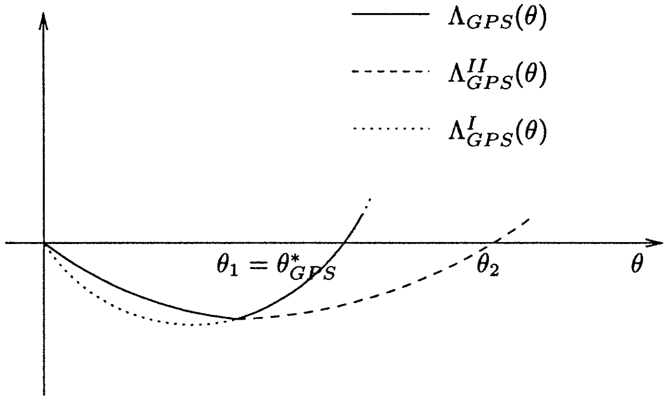

Figure 4: $\theta^*_{GPS,1}$ **as the largest positive root of the equation** $\Lambda_{GPS,1}(\theta) = 0$.

to show is that $\theta^*_{GPS,1} = \theta^*_{GPS}$. This is based on $\theta^*_{GPS,1}$ being equal to $\min(\theta_I, \theta_{II})$. Note, from (52), 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 5(a). Also, from (53), θ_{II} corresponds to the optimal solution of a control problem with a trajectory of the form appearing in Figure 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 6, 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 5(a) and (b) can be reduced to the ones in Figure 3(a) and (b), respectively. This establishes the desired result $\theta^*_{GPS,1} = \theta^*_{GPS}$ and concludes the proof of the theorem.

■

We summarize Propositions 8.5 and 8.3 in the following proposition.

Proposition 8.6 ***(GPS Upper Bound)*** *Assuming that the arrival and service processes*

[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.

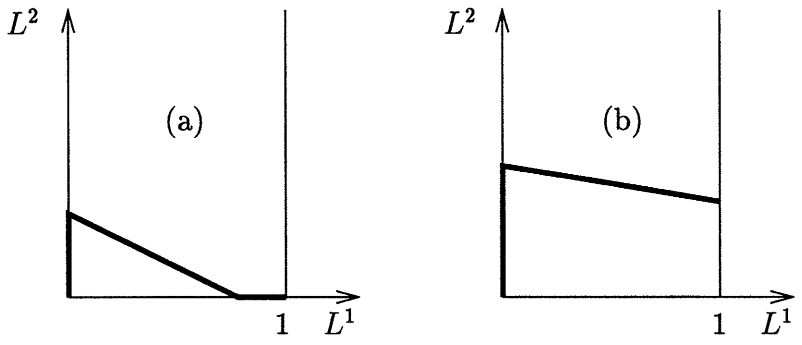

Figure 5: Trajectories for the control problems corresponding to θ_I and θ_{II}.

satisfy Assumptions A and C, and under the GPS policy, the steady-state queue length, L^1, of queue Q^1, at an arbitrary time slot satisfies

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \leq -\theta^*_{GPS}. \tag{54}$$

9 Main Results

In this section we combine Propositions 5.1 and 8.6 and summarize our main results for the GPS policy. As a corollary we obtain results for priority policies.

Theorem 9.1 ***(GPS Main)*** *Under the GPS policy, assuming that the arrival and service processes satisfy Assumptions A, B, and C the steady-state queue length, L^1, of queue Q^1, at an arbitrary time slot satisfies*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] = -\theta^*_{GPS}, \tag{55}$$

where θ^*_{GPS} is given by

$$\theta^*_{GPS} = \min\left[\inf_{a>0} \frac{1}{a}\Lambda^{I*}_{GPS}(a), \inf_{a>0} \frac{1}{a}\Lambda^{II*}_{GPS}(a)\right], \tag{56}$$

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

$$\Lambda^{I*}_{GPS}(a) \triangleq \inf_{\substack{x_1+x_2-x_3=a \\ x_2\leq\phi_2 x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)], \tag{57}$$

and

$$\Lambda^{II*}_{GPS}(a) \triangleq \inf_{\substack{x_1-\phi_1 x_3=a \\ x_2\geq\phi_2 x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \tag{58}$$

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 7 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 9.2 ***(Priority policies)*** *Under strict priority policy for Type 1 customers (P_1), assuming that the arrival and service processes satisfy Assumptions A, B, and C the steady-state queue length, L^1, of queue Q^1, at an arbitrary time slot satisfies*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] = -\theta^*_{P_1}, \tag{59}$$

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

$$\theta^*_{P_1} = \inf_{a>0} \frac{1}{a} \Lambda^*_{P_1}(a), \tag{60}$$

and where

$$\Lambda^*_{P_1}(a) \triangleq \inf_{x_1-x_3=a} [\Lambda^*_{A^1}(x_1) + \Lambda^*_B(x_3)]. \tag{61}$$

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

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] = -\theta^*_{P_2}, \tag{62}$$

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

$$\theta^*_{P_2} = \inf_{a>0} \frac{1}{a} \Lambda^*_{P_2}(a), \tag{63}$$

and where

$$\Lambda^*_{P_2}(a) \triangleq \inf_{\substack{x_1+x_2-x_3=a \\ x_2 \leq x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)], \tag{64}$$

Proof : For policy P_1 apply Theorem 9.1 with $\phi_1 = 1$. For such ϕ_1, it is easy to verify that $\Lambda^{I*}_{GPS}(a) \geq \Lambda^{II*}_{GPS}(a)$, 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 9.1 with $\phi_1 = 0$. Application of $\phi_1 = 0$ to $\Lambda^{I*}_{GPS}(a)$ yields

$$\Lambda^{I*}_{GPS}(a) = \inf_{\substack{x_1+x_2-x_3=a \\ x_2 \leq x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \tag{65}$$

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

$$\Lambda^{II*}_{GPS}(a) = \inf_{\substack{x_1=a \\ x_2 \geq x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \tag{66}$$

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 $x_2 \geq x_3$ 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 $x_2 \geq x_3$ holds. Also, if $x_3 < \mathbf{E}[B_0]$ is the case, we can increase x_3 and reduce the cost, as long $x_2 \geq x_3$ holds. Thus, at optimality $x_2 = x_3$ in (66). 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 (65). Hence, for all a, and when $\phi_1 = 0$, $\Lambda^{I*}_{GPS}(a) \leq \Lambda^{II*}_{GPS}(a)$. Therefore, we define $\Lambda^*_{P_2}(a)$ to be equal to the expression in (65).

■

As the results of Theorem 9.1 and Corollary 9.2 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 8.2. 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 9.3 *θ^*_{GPS} is the largest positive root of the equation*

$$\Lambda_{GPS}(\theta) \triangleq \Lambda_{A^1}(\theta) + \inf_{0\le u\le\theta}[\Lambda_{A^2}(\theta - u) + \Lambda_B(-\theta + \phi_2 u)] = 0. \tag{67}$$

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 (67) 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 (9). 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

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

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

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

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

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

Using the result of Lemma 8.2, $\theta_1 \triangleq \inf_{a>0} \frac{1}{a}\Lambda^{I*}_{GPS}(a)$ 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, $\theta_2 \triangleq \inf_{a>0} \frac{1}{a}\Lambda^{II*}_{GPS}(a)$ is the largest positive root of $\Lambda^{II}_{GPS}(\theta) = 0$. By Equation (56), $\theta^*_{GPS} = \min(\theta_1, \theta_2)$. The situation is exactly the same as in Figure 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

$$\begin{aligned}
\max(\Lambda^I_{GPS}(\theta), \Lambda^{II}_{GPS}(\theta)) =& \max(\Lambda_{A^1}(\theta) + \inf_{u\geq 0}[\Lambda_{A^2}(\theta - u) + \Lambda_B(-\theta + u\phi_2)], \\
& \Lambda_{A^1}(\theta) + \inf_{u\leq\theta}[\Lambda_{A^2}(\theta - u) + \Lambda_B(-\theta + u\phi_2)]) \\
=& \Lambda_{A^1}(\theta) + \inf_{0\leq u\leq\theta}[\Lambda_{A^2}(\theta - u) + \Lambda_B(-\theta + u\phi_2)] \\
\overset{(67)}{=}& \Lambda_{GPS}(\theta).
\end{aligned}$$

■

Again, as it was the case with Theorem 9.1, the result of Theorem 9.3 can be specialized to the case of priority policies.

Corollary 9.4 *θ^*_{P1} is the largest positive root of the equation*

$$\Lambda_{P1}(\theta) \stackrel{\triangle}{=} \Lambda_{A^1}(\theta) + \Lambda_B(-\theta) = 0. \tag{68}$$

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

$$\Lambda_{P2}(\theta) \stackrel{\triangle}{=} \Lambda_{A^1}(\theta) + \inf_{0 \le u \le \theta} [\Lambda_{A^2}(\theta - u) + \Lambda_B(-\theta + u)] = 0. \tag{69}$$

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.

10 Conclusions

In this paper we considered a multiclass multiplexer, with segregated buffers for each type of traffic. Under the GPS policy, we have obtained the asymptotic (as the buffer size goes to infinity) tail of the overflow probability for each buffer. In the standard *large deviations* methodology we provided a lower and matching (up to first degree of the exponent) upper bound on the buffer overflow probabilities. We formulated the problem of calculating the maximum overflow probability (over all scenarios that lead to overflow) as an optimal control problem. The specifics of the GPS policy enter in the formulation of the control problem only through the system dynamics. Therefore, this approach can potentially be used to obtain the performance of other scheduling policies as well. The optimal control formulation provides particular insight into the problem, as it yields an explicit and in detailed characterization of the most likely modes of overflow. We have addressed the case of multiplexing two streams. Our lower bound proof extends to the general case of N streams, the proof of a matching upper bound is an open problem.

References

[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, Massachusetts Institute of Technology, December 1994.

[BPT95] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis, *On the large deviations behaviour of acyclic single class networks and multiclass queues*, Talk at the RSS Workshop in Stochastic Networks, Edinburgh, U.K., 1995.

[BPT96] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis, *Asymptotic buffer overflow probabilities in multiclass multiplexers, Part II: The GLQF policy*, Tech. Report LIDS-P-2342, Laboratory for Information and Decision System, Massachusetts Institute of Technology, June 1996.

[Buc90] J. A. Bucklew, *Large deviation techniques in decision, simulation, and estimation*, Wiley, New York, 1990.

[Cha95] C.S. Chang, *Sample path large deviations and intree networks*, Queueing Systems **20** (1995), 7–36.

[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.

[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.

[dVCW93] G. de Veciana, C. Courcoubetis, and J. Walrand, *Decoupling bandwidths for networks: A decomposition approach to resource management*, Memorandum, Electronics Research Laboratory, University of California 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.

[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.

[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.

[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.

[Kel91] F. P. Kelly, *Effective bandwidths at multi-class queues*, Queueing Systems **9** (1991), 5–16.

[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.

[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.

[Pas96] I. Ch. Paschalidis, *Large deviations in high speed communication networks*, Ph.D. thesis, Massachusetts Institute of Technology, May 1996.

[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.

[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*, IEEE Journal on Selected Areas in Communications **13** (1995), no. 6.

[Wei95] A. Weiss, *An introduction to large deviations for communication networks*, IEEE Journal on Selected Areas in Communications **13** (1995), no. 6, 938–952.

[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.