Large deviations analysis of the generalized processor sharing policy *

Dimitris Bertsimas [a], Ioannis Ch. Paschalidis [b,**] and John N. Tsitsiklis [c]

[a] *Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
E-mail: dbertsim@mit.edu
[b] *Department of Manufacturing Engineering, Boston University, Boston, MA 02215, USA*
E-mail: yannisp@bu.edu
[c] *Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
E-mail: jnt@mit.edu

Received 23 January 1998; revised 11 January 1999

In this paper we consider a stochastic server (modeling a multiclass communication switch) fed by a set of parallel buffers. The dynamics of the system evolve in discrete-time and the *generalized processor sharing* (GPS) scheduling policy of [25] is implemented. The arrival process in each buffer is an arbitrary, and possibly autocorrelated, stochastic process. We obtain a *large deviations* asymptotic for the buffer overflow probability at each buffer. 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 view the problem of finding a *most likely* sample path that leads to an overflow as an *optimal control problem*. Using ideas from convex optimization we analytically solve the control problem to obtain both the asymptotic exponent of the overflow probability and a characterization of *most likely* modes of overflow. These results have important implications for traffic management of high-speed networks. They extend the deterministic, worst-case analysis of [25] 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.

Keywords: large deviations, communication networks

1. Introduction

In the near future, high speed, packet-switched communication networks will offer an even greater than today variety of multimedia, real-time, services accommodating various types of traffic, namely, digitized voice, encoded video, and data.

* A preliminary version of these results was reported in [2]. The results in this paper are included in [26]. Research partially supported by a Presidential Young Investigator award DDM-9158118 with matching funds from Draper Laboratory, by the NSF under grants NCR-9706148 and ACI-9873339, and by the ARO under grant DAAL-03-92-G-0115.
** Corresponding author.

Real-time services are very sensitive to congestion phenomena, such as packet losses, due to buffer overflows. As a consequence, it is widely accepted that the packet loss probability is a critical measure of *Quality of Service* (QoS). It is desirable to operate the network in a regime where this probability is very small, e.g., on the order of 10^{-9}. 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 its probability.

In this paper we model and analyze a communication switch which can support multiple *service classes*. A service class is characterized by the statistical properties of the incoming traffic and by its QoS requirements. The switch has a dedicated buffer for each service class, and employs the *generalized processor sharing* (GPS) policy which was introduced in [12] and analyzed in a deterministic setting in [25]. This policy, also known as *fair queueing*, allocates a fraction ϕ_i of the available capacity (bandwidth) to class i, such that $\sum_{i=1}^{N} \phi_i = 1$, where N is the number of classes. We seek to obtain the buffer overflow probabilities for each class, since these determine the QoS faced by each class. Typical traffic in communication networks is bursty, thus, stochastic processes with autocorrelations are needed to model it. As a result, the problem is particularly difficult since it essentially requires finding the distributions of waiting times and queue lengths in a multiclass $G/G/1$ setting with autocorrelated arrival processes and arbitrary (possibly autocorrelated) service times. In this light, we will focus on the *large deviations regime* and obtain asymptotic expressions for the tails of the overflow probabilities.

To this end, we will provide a lower and a matching (up to first degree in the exponent) upper bound on the buffer overflow probabilities. We will address the case of two classes; the general case of N classes appears to be more complicated since there is an exponential explosion of the number of overflow modes (see [27] for approximations in the general multiclass case). We view the exponent of the overflow probabilities as the optimal value of an associated *optimal 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. These results have important implications in the traffic management of high-speed networks (see [27]). They extend the deterministic, worst-case analysis of [25] to the case where statistical measures of QoS are used to achieve more efficient utilization of the available resources. They can be used as a basis for an admission control mechanism which provides class-dependent statistical QoS guarantees.

The optimal control formulation is introduced in a somewhat more general setting in [3]. The emphasis there is on the analysis of another scheduling policy for sharing bandwidth among classes, the generalized longest queue first. In [3] also, the performance of the latter policy is compared with the performance of the GPS policy, as it is established in the present paper. 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.

There is a growing literature on applications of large deviations techniques in communications (see [31] for a survey). The single class queue case has received extensive attention [16,18–20,22,23,28]. The extension of these ideas to single class networks, although much harder, has been treated in various versions and degrees of rigor in [4,6,13,17]. In [14,32] the authors obtain the asymptotic tails of the overflow probabilities for the GPS policy with deterministic service capacity. The analysis there is based on a large deviations result for the departure process from a $G/D/1$ queue [13]. Tail overflow probabilities for the GPS policy and deterministic service capacity were also reported in [8,24]. The authors in [8] view the problem as a control problem, different than ours, 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. More recently, [15] developed a Skorokhod problem formulation for the large deviations analysis of the GPS policy in a different limiting regime.

In this paper, we extend the GPS results of [8,14,24,32] to the case of a 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 classes with dedicated buffers. We give priority to the first stream and use the GPS policy for the remaining two. 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. Note that stochastic capacity significantly alters the way overflows occur. The reason is that the large deviations behaviour of the departure process from a single class queue is different with deterministic and stochastic service capacity [4,7], and this affects the overflow probabilities in our model (note that in deriving their results [14] and [32] use the departure process from a $G/D/1$ queue).

Among the main contributions of this work we consider (a) the use of the optimal control formulation of the problem because it provides a more intuitive understanding of the operation of the system when it overflows, and (b) the treatment of stochastic service capacities.

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. In section 3 we introduce a model of the switch we will analyze, formally define the GPS policy, and state the main result of the paper. In section 4 we prove a lower bound on the overflow probability and in section 5 we introduce the optimal control formulation and solve the control problem. In section 6 we prove the matching upper bound. Section 7 treats the special case of priority policies and provides an alternative way of calculating the large deviations exponent. Conclusions are given in section 8.

2. Preliminaries

In this section we review some basic results on the theory of Large Deviations [5,11,30] that will be used in the sequel.

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

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

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, $i \geqslant 1$, are identically distributed, possibly dependent random variables. We will be making the following assumption.

Assumption A.

(1) The limit

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

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., $\liminf_{\theta_n\to\theta} \Lambda(\theta_n) \geqslant \Lambda(\theta)$, for all θ.

Let us next define

$$\Lambda^*(a) \triangleq \sup_{\theta} \big(\theta a - \Lambda(\theta)\big), \tag{3}$$

which is the Legendre transform of $\Lambda(\cdot)$. It is important to note that $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$ are convex duals, namely, along with (3), it also holds that

$$\Lambda(\theta) = \sup_{a} \big(\theta a - \Lambda^*(a)\big). \tag{4}$$

The function $\Lambda^*(\cdot)$ is convex and lower semicontinuous (see [11]).

Under assumption A, the Gärtner–Ellis theorem (see [5,11]) establishes that $\{S_n\}$ satisfies a *Large Deviations Principle* (LDP) with *rate function* $\Lambda^*(\cdot)$. In particular, this theorem intuitively asserts that for large enough n and for small $\varepsilon > 0$,

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

The Gärtner–Ellis theorem generalizes Cramér's theorem [9] which applies to independent and identically distributed (iid) random variables.

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)

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

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 [10], under certain mild mixing conditions on the stationary sequence $\{X_i;\ i \geqslant 1\}$, the authors establish an LDP for the process $S_n(\cdot)$ in $D[0, 1]$ (the space of right continuous functions with left limits) equipped with the supremum norm topology. In the spirit of the sample path LDP in [10] we will be assuming the following.

Assumption B. For all $m \in \mathbb{N}$, for every $\varepsilon_1, \varepsilon_2 > 0$, and for every scalars $a_0, \ldots, a_{m-1}$, there exists $M > 0$ such that for all $n \geqslant M$ and all $k_0, \ldots, k_m$ with $1 = k_0 \leqslant k_1 \leqslant \cdots \leqslant k_m = n$,

$$\mathrm{e}^{-(n\varepsilon_2 + \sum_{i=0}^{m-1}(k_{i+1}-k_i)\Lambda^*(a_i))} \leqslant \mathbf{P}\big[\big|S_{k_{i+1}} - S_{k_i} - (k_{i+1} - k_i)a_i\big| \leqslant \varepsilon_1 n,\ i = 0, \ldots, m-1\big].$$

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 [11]). 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 $a_0, \ldots, a_{m-1}$. We will also be making the following assumption, which 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 $\Gamma(\cdot)$ with $0 \leqslant \Gamma(y) < \infty$, for all $y > 0$, such that for all $n \geqslant M$ and all $k_0, \ldots, k_m$ with $1 = k_0 \leqslant k_1 \leqslant \cdots \leqslant k_m = n$,

$$\mathbf{E}\big[\mathrm{e}^{\theta \cdot Z}\big] \leqslant \exp\left\{\sum_{j=1}^{m}\big[(k_j - k_{j-1})\Lambda(\theta_j) + \Gamma(\theta_j)\big]\right\}, \tag{5}$$

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

In [6] a uniform bounding condition is given under which assumptions B and C are satisfied. It is verified that the set of processes satisfying these assumptions is large enough to include renewal, Markov-modulated, and stationary processes with mild mixing conditions. Such processes can model "burstiness" and are commonly used in modeling the input traffic to communication networks.

On a notational remark, in the rest of the paper we will be denoting by

$$S_{i,j}^{X} \triangleq \sum_{k=i}^{j} X_k, \quad i \leqslant j,$$

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 equations (2) and (3) for definitions), respectively, of the process X.

3. A multiclass model

In this section we introduce a model for the multiclass switch operated under the GPS policy that we plan to analyze, state the main result, and provide a brief outline of the approach we plan to follow.

Consider the system depicted in figure 1. We assume a slotted time model (i.e., discrete time) and we let A_i^j, $i \in \mathbb{Z}$, denote the number of class j customers that enter queue Q^j at time i, for $j = 1, 2$. 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 A_i^j's for fixed j and different values of i.

We denote by L_i^j the queue length at time i (without counting arrivals at time i) in queue Q^j, for $j = 1, 2$. 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.

To simplify the analysis we consider a discrete-time "fluid" model, meaning that we will be treating A_i^j, L_i^j, for $j = 1, 2$, and B_i as non-negative real numbers (the amount of fluid entering, in queue, or served).

We assume the following stability condition:

$$\mathbf{E}[B_i] > \mathbf{E}\big[A_i^1\big] + \mathbf{E}\big[A_i^2\big], \quad \forall i. \tag{6}$$

We further assume that the arrival and service processes satisfy assumptions A, 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. Note that since A_i^j, for $j = 1, 2$, and B_i represent number of arrivals and services, respectively, they are assumed to be non-negative, which implies that their rate function $\Lambda_X^*(x)$, for $X \in \{A^1, A^2, B\}$, is infinity for all $x < 0$.

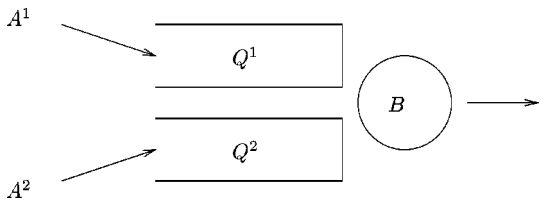

Figure 1. A multiclass model.

The switch implements the *generalized processor sharing* (GPS) policy. 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)

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

and

$$0 \leqslant L_{i+1}^j \leqslant \left[L_i^j + A_i^j - \phi_j B_i\right]^+, \quad j = 1,2,$$

where $[x]^+ \triangleq \max\{x, 0\}$. Note that L_i^j's will generally take non-integer values even if A_i^j and B_i are integers. This corresponds to the GPS policy in [25] as opposed to its "packetized" version PGPS.

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

$$\mathbf{P}\left[L_i^1 > U\right] \sim \mathrm{e}^{-U\theta^*_{\mathrm{GPS}}}, \tag{7}$$

asymptotically, as $U \to \infty$ (theorem 3.1). To this end, we will develop a lower bound on the overflow probability (proposition 4.1), along with a matching upper bound (proposition 6.7).

Theorem 3.1. 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 satisfies

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}\left[L^1 > U\right] = -\theta^*_{\mathrm{GPS}}, \tag{8}$$

where θ^*_{GPS} is given by

$$\theta^*_{\mathrm{GPS}} = \min\left[\inf_{a>0} \frac{1}{a}\Lambda^{\mathrm{I}*}_{\mathrm{GPS}}(a), \inf_{a>0} \frac{1}{a}\Lambda^{\mathrm{II}*}_{\mathrm{GPS}}(a)\right], \tag{9}$$

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

$$\Lambda^{\mathrm{I}*}_{\mathrm{GPS}}(a) \triangleq \inf_{\substack{x_1+x_2-x_3=a \\ x_2 \leqslant \phi_2 x_3}} \left[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)\right], \tag{10}$$

and

$$\Lambda^{\mathrm{II}*}_{\mathrm{GPS}}(a) \triangleq \inf_{\substack{x_1-\phi_1 x_3=a \\ x_2 \geqslant \phi_2 x_3}} \left[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)\right]. \tag{11}$$

4. A lower bound

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

Proposition 4.1 (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

$$\liminf_{U\to\infty} \frac{1}{U} \log \mathbf{P}\big[L^1 > U\big] \geqslant -\theta^*_{\mathrm{GPS}}, \tag{12}$$

where θ^*_{GPS} is defined by equations (9)–(11).

Proof. Let $-n \leqslant 0$ and $a > 0$. Fix $x_1, x_2, x_3 \geqslant 0$ and $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$ and consider the event

$$\begin{aligned}&\Big\{\big|S^{A^1}_{-n,-i-1} - (n-i)x_1\big| \leqslant \varepsilon_1 n,\ \big|S^{A^2}_{-n,-i-1} - (n-i)x_2\big| \leqslant \varepsilon_2 n,\\ &\quad \big|S^{B}_{-n,-i-1} - (n-i)x_3\big| \leqslant \varepsilon_3 n,\ i = 0,1,\ldots,n-1\Big\}.\end{aligned}$$

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

$$\begin{aligned}&\text{Scenario 1:}\quad x_1 + x_2 - x_3 = a,\\ &\qquad\qquad x_2 \leqslant \phi_2 x_3,\end{aligned}\qquad\qquad \begin{aligned}&\text{Scenario 2:}\quad x_1 - \phi_1 x_3 = a,\\ &\qquad\qquad x_2 \geqslant \phi_2 x_3.\end{aligned} \tag{13}$$

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, $L_0^1 \geqslant na - n\varepsilon_1'$, where $\varepsilon_1' \to 0$ as $\varepsilon_1, \varepsilon_2, \varepsilon_3 \to 0$. 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 $L_0^1 \geqslant na - n\varepsilon_2'$, where $\varepsilon_2' \to 0$ as $\varepsilon_1, \varepsilon_2, \varepsilon_3 \to 0$.

Now, the probability of Scenario 1 is a lower bound on $\mathbf{P}[L_0^1 \geqslant n(a - \varepsilon_1')]$. 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

$$\begin{aligned}&\mathbf{P}\big[L_0^1 \geqslant n\big(a - \varepsilon_1'\big)\big]\\ &\quad\geqslant \sup_{\substack{x_1+x_2-x_3=a\\ x_2\leqslant\phi_2 x_3}} \mathbf{P}\big[\big|S^{A^1}_{-n,-i-1} - (n-i)x_1\big| \leqslant \varepsilon_1 n,\ i = 0,1,\ldots,n-1\big]\\ &\qquad\times \mathbf{P}\big[\big|S^{A^2}_{-n,-i-1} - (n-i)x_2\big| \leqslant \varepsilon_2 n,\ i = 0,1,\ldots,n-1\big]\\ &\qquad\times \mathbf{P}\big[\big|S^{B}_{-n,-i-1} - (n-i)x_3\big| \leqslant \varepsilon_3 n,\ i = 0,1,\ldots,n-1\big]\\ &\quad\geqslant \exp\Big\{-n\Big(\inf_{\substack{x_1+x_2-x_3=a\\ x_2\leqslant\phi_2 x_3}} \big[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)\big] + \varepsilon\Big)\Big\}\\ &\quad= \exp\big\{-n\big(\Lambda^{\mathrm{I}*}_{\mathrm{GPS}}(a) + \varepsilon\big)\big\},\end{aligned} \tag{14}$$

where n is large enough, and $\varepsilon, \varepsilon_1' \to 0$ as $\varepsilon_1, \varepsilon_2, \varepsilon_3 \to 0$.

Similarly, calculating the probability of Scenario 2, we obtain

$$\mathbf{P}\big[L_0^1 \geqslant n\big(a-\varepsilon_2'\big)\big] \geqslant \exp\big\{-n\big(\Lambda_{\mathrm{GPS}}^{\mathrm{II}*}(a)+\varepsilon'\big)\big\}, \tag{15}$$

for n large enough, and with $\varepsilon', \varepsilon_2' \to 0$ as $\varepsilon_1, \varepsilon_2, \varepsilon_3 \to 0$.

Combining equations (14) and (15), we obtain that for all $\varepsilon, \varepsilon' > 0$ there exists N such that for all $n > N$

$$\frac{1}{n}\log\mathbf{P}\big[L_0^1 \geqslant n(a-\varepsilon)\big] \geqslant -\big(\min\big(\Lambda_{\mathrm{GPS}}^{\mathrm{I}*}(a), \Lambda_{\mathrm{GPS}}^{\mathrm{II}*}(a)\big)+\varepsilon'\big). \tag{16}$$

As a final step to this proof, by letting $U = n(a-\varepsilon)$ and $U_0 = N(a-\varepsilon)$, we obtain that for all $\varepsilon, \varepsilon' > 0$ and for all $U > U_0$

$$\begin{aligned}\frac{1}{U}\log\mathbf{P}\big[L^1 > U\big] &= \frac{1}{n(a-\varepsilon)}\log\mathbf{P}\big[L_0^1 \geqslant n(a-\varepsilon)\big] \\ &\geqslant -\frac{1}{a-\varepsilon}\big(\min\big(\Lambda_{\mathrm{GPS}}^{\mathrm{I}*}(a), \Lambda_{\mathrm{GPS}}^{\mathrm{II}*}(a)\big)+\varepsilon'\big),\end{aligned}$$

which implies

$$\liminf_{U\to\infty}\frac{1}{U}\log\mathbf{P}\big[L^1 > U\big] \geqslant -\frac{1}{a}\min\big(\Lambda_{\mathrm{GPS}}^{\mathrm{I}*}(a), \Lambda_{\mathrm{GPS}}^{\mathrm{II}*}(a)\big).$$

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

$$\liminf_{U\to\infty}\frac{1}{U}\log\mathbf{P}\big[L^1 > U\big] \geqslant -\min\left[\inf_{a>0}\frac{1}{a}\Lambda_{\mathrm{GPS}}^{\mathrm{I}*}(a), \inf_{a>0}\frac{1}{a}\Lambda_{\mathrm{GPS}}^{\mathrm{II}*}(a)\right]. \qquad \square$$

5. The optimal control problem

In this section we introduce an optimal control problem and show that θ_{GPS}^* is its optimal value. This interpretation of θ_{GPS}^* will be used later to establish an upper bound on the overflow probability.

To motivate the control problem, we relate it, heuristically, to 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 $-n \leqslant 0$ 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 $T = n/U$ and define the following continuous-time functions in $D[-T, 0]$ (these are right-continuous functions with left-limits):

$$\widehat{L}^j(t) = \frac{1}{U}L^j_{\lfloor Ut\rfloor}, \quad j = 1, 2, \qquad S^X(t) = \frac{1}{U}S^X_{-UT,\lfloor Ut\rfloor}, \quad X \in \big\{A^1, A^2, B\big\},$$

[1] Such a relation can be rigorously established using the sample path LDP for the arrival and service processes, as it is defined in [6,10].

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 sample path of process X has empirical rate $x(t)$ in the interval $[-T, 0]$ if for large U and small $\varepsilon > 0$ it is true that

$$\left| S^X(t) - \int_{-T}^{t} x(\tau)\,\mathrm{d}\tau \right| < \varepsilon, \quad \forall t \in [-T, 0],$$

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

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

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. Similarly, a "polygonal approximation" to $\widehat{L}^j(t)$ (see [10]; [11, section 5.1]) converges to some continuous functions $L^j(t)$, for $j = 1, 2$.

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.

[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 $L^2(t_2) = L^2(t_1) + \int_{t_1}^{t_2} (x_1(t) + x_2(t) - x_3(t))\,\mathrm{d}t$, for all intervals (t_1, t_2) that the system remains in region B.

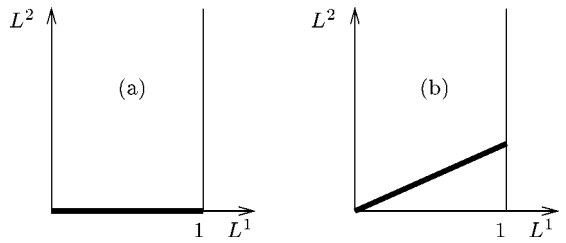

Figure 2. Trajectories for the restricted (GPS-OVERFLOW).

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.

$$\begin{aligned}
\text{(GPS-OVERFLOW)}\quad \text{minimize}\quad & \int_{-T}^{0} \big[\Lambda^*_{A^1}\big(x_1(t)\big) + \Lambda^*_{A^2}\big(x_2(t)\big) + \Lambda^*_B\big(x_3(t)\big)\big]\,\mathrm{d}t \\
\text{subject to:}\quad & L^1(-T) = L^2(-T) = 0, \\
& L^1(0) = 1, \\
& L^2(0)\text{: free}, \\
& T\text{: free}, \\
& \big\{L^j(t)\text{: } t \in [-T, 0],\ j = 1, 2\big\} \in \text{(GPS-DYNAMICS)}.
\end{aligned} \tag{17}$$

To establish that θ^*_{GPS} is the optimal value of an associated control problem, it suffices to consider a restricted version of (GPS-OVERFLOW). In particular, we will only be considering trajectories of (GPS-OVERFLOW) that have the form depicted in figure 2. We will be referring to this as the *restricted* (GPS-OVERFLOW). The choice of these trajectories is motivated by the two scenarios in the proof of the lower bound in proposition 4.1. It turns out that the trajectories in figure 2 are optimal over all feasible trajectories of (GPS-OVERFLOW). This is proved in the appendix. In this sense, these trajectories correspond to *most likely* ways that overflows occur.

Optimal value of restricted (GPS-OVERFLOW)

We next calculate the optimal value of restricted (GPS-OVERFLOW). The best trajectory of the form shown in figure 2(a) has value

$$\inf_T \inf_{\substack{x_1 + x_2 - x_3 = 1/T \\ x_2 \leqslant \phi_2 x_3}} T\big[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)\big], \tag{18}$$

which is equal to $\inf_T[T\Lambda^{\mathrm{I}*}_{\mathrm{GPS}}(1/T)]$ by the definition in (10). The best trajectory of the form shown in figure 2(b) has value

$$\inf_T \inf_{\substack{x_1 - \phi_1 x_3 = 1/T \\ x_2 \geqslant \phi_2 x_3}} T\big[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)\big], \tag{19}$$

which is equal to $\inf_T[T\Lambda^{\mathrm{II}*}_{\mathrm{GPS}}(1/T)]$ by the definition in (11). Thus, the optimal value of restricted (GPS-OVERFLOW) is equal to the minimum of the two expressions above which is identical to θ^*_{GPS} as it is defined in (9).

It is of interest (and of use in establishing the upper bound) to investigate under what conditions on the parameters of the arrival and service processes the trajectory in figure 2(a) dominates the one in figure 2(b) and vice versa. We will distinguish two cases: $\mathbf{E}[A^2] \geqslant \phi_2\mathbf{E}[B]$ and $\mathbf{E}[A^2] < \phi_2\mathbf{E}[B]$, where for $j = 1, 2$, $\mathbf{E}[A^j]$ (respectively $\mathbf{E}[B]$) denote the expected number of customers arriving from stream j (respectively expected potential number of departures). In the first case we will establish that the trajectory in figure 2(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 5.1. If $\mathbf{E}[A^2] \geqslant \phi_2\mathbf{E}[B]$ then optimal state trajectories of restricted (GPS-OVERFLOW) have the form in figure 2(b) and the optimal value θ^*_{GPS} is given by

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

Proof. Assume $\mathbf{E}[A^2] \geqslant \phi_2\mathbf{E}[B]$ and consider the state trajectory in figure 2(a) which has optimal value given by the expression in (18). Since $x_2 \leqslant \phi_2 x_3$, either $x_2 \leqslant \mathbf{E}[A^2]$ or $x_3 \geqslant \mathbf{E}[B]$. Then, since rate functions are nondecreasing above the mean and nonincreasing below the mean, we can increase x_2 and decrease x_3 until $x_2 = \phi_2 x_3$, making $x_1 + x_2 - x_3 \geqslant 1/T$. The segment of this trajectory with terminal point at $L^1 = 1/T$ is feasible (since we have a free time problem), and has the form of the state trajectory in figure 2(b). Thus, we have reduced optimal state trajectories to the one in figure 2(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 $x_2 \geqslant \phi_2 x_3$, making $x_1 - \phi_1 x_3 \geqslant 1/T$, and keeping the segment of the resulting trajectory with terminal point at $L^1 = 1/T$. Thus, it has to be the case $x_3 \leqslant \mathbf{E}[B]$. Then we can actually fix x_2 to $\mathbf{E}[A^2]$, without violating the constraint $x_2 \geqslant \phi_2 x_3$ (since $x_2 = \mathbf{E}[A^2] \geqslant \phi_2\mathbf{E}[B] \geqslant \phi_2 x_3$). This proves that the optimal value is given by the expression appearing in the statement of this theorem. □

6. A GPS upper bound

In this section we develop an upper bound on the probability $\mathbf{P}[L^1_0 > U]$, for the case of the GPS policy. In particular, we will prove that as $U \to \infty$ we have $\mathbf{P}[L^1_0 > U] \leqslant \mathrm{e}^{-\theta^*_{\mathrm{GPS}}U+\mathrm{o}(U)}$, where $\mathrm{o}(U)$ denotes functions with the property $\lim_{U\to\infty}(\mathrm{o}(U)/U) = 0$.

In proving the upper bound we will distinguish two cases:

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

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

6.1. Upper bound: Case 2

We consider a busy period of the first queue Q^1 that starts at some time $-n^* \leqslant 0$ ($L^1_{-n^*} = 0$) and has not ended until time 0. Notice that due to the stability condition (6) and the fact $\mathbf{E}[A^2] \geqslant \phi_2 \mathbf{E}[B]$, 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 $L^1_0 > U$. Note that

$$L^1_0 \leqslant S^{A^1}_{-n^*,-1} - \phi_1 S^B_{-n^*,-1}. \tag{20}$$

Thus,

$$\begin{aligned}\mathbf{P}\big[L^1_0 > U\big] &\leqslant \mathbf{P}\big[\exists n \geqslant 0 \text{ s.t. } S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1} > U\big] \\ &\leqslant \mathbf{P}\Big[\max_{n\geqslant 0}\big(S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1}\big) > U\Big].\end{aligned} \tag{21}$$

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

$$\begin{aligned}&\mathbf{E}\big[\mathrm{e}^{\theta \max_{n\geqslant 0}(S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1})}\big] \\ &\quad\leqslant \sum_{n\geqslant 0} \mathbf{E}\big[\mathrm{e}^{\theta(S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1})}\big] \\ &\quad\leqslant \sum_{n\geqslant 0} \mathrm{e}^{n(\Lambda_{A^1}(\theta) + \Lambda_B(-\phi_1\theta) + \varepsilon)} \\ &\quad= K(\theta,\varepsilon) \quad \text{if } \Lambda_{A^1}(\theta) + \Lambda_B(-\phi_1\theta) < 0,\end{aligned} \tag{22}$$

since when the exponent is negative (for sufficiently small ε), the infinite geometric series converges to some $K(\theta,\varepsilon)$. We can now apply the Markov inequality in (21) to obtain

$$\begin{aligned}\mathbf{P}\big[L^1_0 > U\big] &\leqslant \mathbf{E}\big[\mathrm{e}^{\theta \max_{n\geqslant 0}(S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1})}\big]\mathrm{e}^{-\theta U} \\ &\leqslant K(\theta,\varepsilon)\mathrm{e}^{-\theta U} \quad \text{if } \Lambda_{A^1}(\theta) + \Lambda_B(-\phi_1\theta) < 0.\end{aligned} \tag{23}$$

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

Proposition 6.1. If $\mathbf{E}[A^2] \geqslant \phi_2 \mathbf{E}[B]$ and under assumption A, for the arrival and service processes,

$$\limsup_{U\to\infty} \frac{1}{U} \log \mathbf{P}\big[L^1_0 > U\big] \leqslant - \sup_{\{\theta\geqslant 0:\ \Lambda_{A^1}(\theta) + \Lambda_B(-\phi_1\theta) < 0\}} \theta.$$

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 theorem 5.1.

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 6.2. For $\Lambda^*(\cdot)$ and $\Lambda(\cdot)$ being convex duals and assuming that $\Lambda(\theta) < 0$ for sufficiently small $\theta > 0$, it holds that

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

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

Proof.

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

In the second equality above, we have made the substitution $a' := 1/a$, and in the third one we have used duality to interchange the inf with the sup. Finally, in the last equality above we have used the convexity of $\Lambda(\theta)$ and the fact that $\Lambda(\theta) < 0$ for sufficiently small $\theta > 0$. □

We will also need the following result.

Lemma 6.3. Let $F:\mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$, $g_1, \ldots, g_m:\mathbb{R}^n \to \mathbb{R}$, and consider the following parametric optimization problem:

$$\begin{aligned} Z(a) &= \inf F(x) \\ \text{s.t. } & g_1(x) = a, \\ & g_j(x) \leqslant 0, \quad j = 2, \ldots, m, \end{aligned} \tag{24}$$

where $x \in \mathbb{R}^n$, $F(x)$ is a lower semicontinuous function that satisfies $\lim_{\|x_k\|\to\infty} F(x_k) = \infty$, and $g_j(\cdot)$ are continuous functions for all $j = 1, \ldots, m$. Assume that it has at least one feasible solution. Then its optimal value $Z(a)$ is a lower semicontinuous function of the scalar parameter a.

Proof. Let x^* be an optimal solution of (24), which exists by Weierstrass' theorem. Consider an arbitrary sequence $\{a_n\}$ converging to a, and let x_n be an optimal solution of (24) when the parameter a equals a_n. Let finally $\bar{x}$ be a finite limit point of $\{x_n\}$, if it exists. Note that

$$\liminf_{n\to\infty} Z(a_n) = \liminf_{n\to\infty} F(x_n) \geqslant F\big(\bar{x}\big) \geqslant F\big(x^*\big) = Z(a).$$

The first inequality above is due to the lower semicontinuity of $F(\cdot)$. The second inequality above is due to the continuity of $g_j(\cdot)$ which implies that $\bar{x}$ is a feasible solution for (24). If $\{x_n\}$ does not have a finite limit point, then $\|x_n\| \to \infty$, $F(x_n) \to \infty$, and the above inequalities trivially hold. □

Based on these two lemmata and proposition 6.1 we establish the following proposition.

Proposition 6.4 (GPS upper bound, Case 2). If $\mathbf{E}[A^2] \geqslant \phi_2 \mathbf{E}[B]$ and assuming that the arrival and service processes satisfy assumption A, the steady-state queue length L^1 of queue Q^1 satisfies

$$\limsup_{U\to\infty} \frac{1}{U} \log \mathbf{P}\big[L^1 > U\big] \leqslant -\theta^*_{\mathrm{GPS}}.$$

Proof. It suffices to prove that $\theta^*_{\mathrm{GPS}} = \sup_{\{\theta\geqslant 0:\ \Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)<0\}} \theta$. Since we are in Case 2, θ^*_{GPS} is given by the expression in theorem 5.1. Due to lemma 6.2 it suffices to prove that $\Lambda_{A^1}(\theta) + \Lambda_B(-\phi_1\theta)$ is the convex dual of

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

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 (see [1, exercise 6.7]). Moreover, it is lower semicontinuous by lemma 6.3, and thus, we can apply convex duality results. Finally, the stability condition (6) and the fact $\mathbf{E}[A^2] \geqslant \phi_2 \mathbf{E}[B]$ ensure that $\mathbf{E}[A^1] < \phi_1 \mathbf{E}[B]$, which implies that $\Lambda_{A^1}(\theta)+\Lambda_B(-\phi_1\theta)$ has negative right derivative at $\theta = 0$. Thus, it takes negative values for sufficiently small $\theta > 0$ and satisfies the required condition of lemma 6.2.

Indeed the convex dual of $\Lambda^*(a)$ is

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

6.2. *Upper bound: Case 1*

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

Proposition 6.5. If $\mathbf{E}[A^1] < \phi_2 \mathbf{E}[B]$ and assuming that the arrival and service processes satisfy assumptions A and C,

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

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

$$L_0^1 \leqslant L_0^1 + L_0^2 = L_{-k^*}^2 + S_{-k^*,-1}^{A^1} + S_{-k^*,-1}^{A^2} - S_{-k^*,-1}^{B}. \tag{25}$$

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 [14] and used in [32], 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 the restricted (GPS-OVERFLOW). Let $-n^* \leqslant -k^*$ be the first time (looking backwards in time from $-k^*$) that the queue length of Q^2 becomes zero in the virtual system. That is, the virtual system starts working at $-n^*$ and seizes working at $-k^*$. 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

$$\widetilde{L}_{-k^*}^2 = S_{-n^*,-k^*-1}^{A^2} - \phi_2 S_{-n^*,-k^*-1}^{B}, \tag{26}$$

where $\widetilde{L}_{-k^*}^2$ denotes the queue length of Q^2 in the virtual system at time $-k^*$. Since we argued that $\widetilde{L}_{-k^*}^2 \geqslant L_{-k^*}^2$, combining (26) with (25) yields

$$L_0^1 \leqslant S_{-k^*,-1}^{A^1} + S_{-n^*,-1}^{A^2} - S_{-k^*,-1}^{B} - \phi_2 S_{-n^*,-k^*-1}^{B}. \tag{27}$$

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

$$L_0^1 \leqslant S_{-k^*,-1}^{A^1} - \phi_1 S_{-k^*,-1}^{B}. \tag{28}$$

We will use the bound in (27) when $S_{-n^*,-1}^{A^2} \leqslant \phi_2 S_{-n^*,-1}^{B}$ and the bound in (28), otherwise. Namely, we will use

$$L_0^1 \leqslant \begin{cases} S_{-k^*,-1}^{A^1} + S_{-n^*,-1}^{A^2} - S_{-k^*,-1}^{B} - \phi_2 S_{-n^*,-k^*-1}^{B} & \text{if } S_{-n^*,-1}^{A^2} \leqslant \phi_2 S_{-n^*,-1}^{B}, \\ S_{-k^*,-1}^{A^1} - \phi_1 S_{-k^*,-1}^{B} & \text{if } S_{-n^*,-1}^{A^2} \geqslant \phi_2 S_{-n^*,-1}^{B}. \end{cases} \tag{29}$$

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

$$\begin{aligned} &\mathbf{P}\big[L_0^1 > U \text{ and } \Omega_1\big] \\ &\quad\leqslant \mathbf{P}\big[\exists n \geqslant k \geqslant 0 \text{ s.t. } S_{-n,-1}^{A^2} \leqslant \phi_2 S_{-n,-1}^{B} \text{ and} \\ &\qquad S_{-k,-1}^{A^1} + S_{-n,-1}^{A^2} - S_{-k,-1}^{B} - \phi_2 S_{-n,-k-1}^{B} > U\big] \\ &\quad\leqslant \mathbf{P}\Big[\max_{\{n\geqslant k\geqslant 0:\ S_{-n,-1}^{A^2} \leqslant \phi_2 S_{-n,-1}^{B}\}} \big(S_{-k,-1}^{A^1} + S_{-n,-1}^{A^2} - S_{-k,-1}^{B} - \phi_2 S_{-n,-k-1}^{B}\big) > U\Big]. \end{aligned} \tag{30}$$

For sample paths in Ω_2 we have

$$\begin{aligned}
&\mathbf{P}\big[L_0^1 > U \text{ and } \Omega_2\big] \\
&\quad\leqslant \mathbf{P}\big[\exists n \geqslant k \geqslant 0 \text{ s.t. } S^{A^2}_{-n,-1} \geqslant \phi_2 S^B_{-n,-1} \text{ and } S^{A^1}_{-k,-1} - \phi_1 S^B_{-k,-1} > U\big] \\
&\quad\leqslant \mathbf{P}\Big[\max_{\{n\geqslant k\geqslant 0:\ S^{A^2}_{-n,-1}\geqslant \phi_2 S^B_{-n,-1}\}} \big(S^{A^1}_{-k,-1} - \phi_1 S^B_{-k,-1}\big) > U\Big]. \qquad (31)
\end{aligned}$$

Let us now define

$$L^{\mathrm{I}}_{\mathrm{GPS},1} \triangleq \max_{\{n\geqslant k\geqslant 0:\ S^{A^2}_{-n,-1}\leqslant \phi_2 S^B_{-n,-1}\}} \big(S^{A^1}_{-k,-1} + S^{A^2}_{-n,-1} - S^B_{-k,-1} - \phi_2 S^B_{-n,-k-1}\big)$$

and

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

which after bringing the constraints in the objective function become

$$\begin{aligned}
L^{\mathrm{I}}_{\mathrm{GPS},1} = \max_{n\geqslant k\geqslant 0} \inf_{u\geqslant 0} \big[&S^{A^1}_{-k,-1} + (1-u)S^{A^2}_{-n,-1} \\
&- (1-u\phi_2)S^B_{-k,-1} - \phi_2(1-u)S^B_{-n,-k-1}\big] \qquad (32)
\end{aligned}$$

and

$$L^{\mathrm{II}}_{\mathrm{GPS},1} = \max_{n\geqslant k\geqslant 0} \inf_{u\geqslant 0} \big[S^{A^1}_{-k,-1} + uS^{A^2}_{-n,-1} + (-u\phi_2 - \phi_1)S^B_{-k,-1} - u\phi_2 S^B_{-n,-k-1}\big]. \qquad (33)$$

Next we will upper bound the moment generating functions of $L^{\mathrm{I}}_{\mathrm{GPS},1}$ and $L^{\mathrm{II}}_{\mathrm{GPS},1}$ by using assumption C for the arrival and service processes. For the moment generating function of $L^{\mathrm{I}}_{\mathrm{GPS},1}$ and $\theta \geqslant 0$ we have

$$\begin{aligned}
&\mathbf{E}\big[\mathrm{e}^{\theta L^{\mathrm{I}}_{\mathrm{GPS},1}}\big] \\
&\quad\leqslant \sum_{n\geqslant 0}\sum_{0\leqslant k\leqslant n} \inf_{u\geqslant 0} \mathbf{E}\big[\exp\big\{\theta\big[S^{A^1}_{-k,-1} + (1-u)S^{A^2}_{-n,-1} \\
&\qquad - (1-u\phi_2)S^B_{-k,-1} - \phi_2(1-u)S^B_{-n,-k-1}\big]\big\}\big] \\
&\quad\leqslant \sum_{n\geqslant 0}\sum_{0\leqslant k\leqslant n} \inf_{u\geqslant 0} \exp\big\{(n-k)\big[\Lambda_{A^2}(\theta-\theta u) + \Lambda_B\big(-\theta\phi_2(1-u)\big)\big] \\
&\qquad + k\big[\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta-\theta u) + \Lambda_B\big(-\theta(1-u\phi_2)\big)\big] + \Gamma(\theta,u)\big\} \\
&\quad\leqslant \sum_{n\geqslant 0} n \sup_{\zeta\in[0,1]} \inf_{u\geqslant 0} \exp\Big\{n\Big[\zeta\big(\Lambda_{A^2}(\theta-\theta u) + \Lambda_B\big(-\theta\phi_2(1-u)\big)\big) \\
&\qquad + (1-\zeta)\big(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta-\theta u) + \Lambda_B\big(-\theta(1-u\phi_2)\big)\big) + \frac{\Gamma(\theta,u)}{n}\Big]\Big\}, \qquad (34)
\end{aligned}$$

where we let $\zeta = (n-k)/n$. In the second inequality above we have used assumption C with $m = 2$, which implies the existence of some non-negative and bounded function $\Gamma(\theta, u)$. Let us now define

$$\Lambda^{\mathrm{I}}_{\mathrm{GPS},1}(\theta) \triangleq \sup_{\zeta\in[0,1]} \inf_{u\geqslant 0} \big[\zeta\big(\Lambda_{A^2}(\theta-\theta u) + \Lambda_B\big(-\theta\phi_2(1-u)\big)\big) + (1-\zeta)\big(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta-\theta u) + \Lambda_B\big(-\theta(1-u\phi_2)\big)\big)\big]. \quad (35)$$

Let $u^*(\theta,\zeta)$ be the optimal u in the above optimization problem for fixed ζ (it exists due to the convexity and lower-semicontinuity of the limiting log-moment generating functions). From (34) we have

$$\begin{aligned}&\mathbf{E}[\mathrm{e}^{\theta L^{\mathrm{I}}_{\mathrm{GPS},1}}]\\ &\leqslant \sum_{n\geqslant 0} n \sup_{\zeta\in[0,1]} \exp\left\{n\left[\zeta\big(\Lambda_{A^2}\big(\theta-\theta u^*\big) + \Lambda_B\big(-\theta\phi_2\big(1-u^*\big)\big)\big)\right.\right.\\ &\quad\left.\left. + (1-\zeta)\big(\Lambda_{A^1}(\theta) + \Lambda_{A^2}\big(\theta-\theta u^*\big) + \Lambda_B\big(-\theta\big(1-u^*\phi_2\big)\big)\big) + \frac{\Gamma(\theta,u^*)}{n}\right]\right\}. \quad (36)\end{aligned}$$

Now, for every $\varepsilon > 0$ and $\theta \geqslant 0$ we can take n large enough such that $\Gamma(\theta,u^*)/n < \varepsilon$. For sufficiently small ε and if $\Lambda^{\mathrm{I}}_{\mathrm{GPS},1}(\theta) < 0$ then the infinite geometric series in the right-hand side of (36) converges to some $K_1(\theta,\varepsilon)$. That is,

$$\mathbf{E}\big[\mathrm{e}^{\theta L^{\mathrm{I}}_{\mathrm{GPS},1}}\big] \leqslant K_1(\theta,\varepsilon), \quad \text{if } \Lambda^{\mathrm{I}}_{\mathrm{GPS},1}(\theta) < 0. \quad (37)$$

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

$$\begin{aligned}&\mathbf{E}\big[\mathrm{e}^{\theta L^{\mathrm{II}}_{\mathrm{GPS},1}}\big]\\ &\leqslant \sum_{n\geqslant 0}\sum_{0\leqslant k\leqslant n} \inf_{u\geqslant 0} \mathbf{E}\big[\exp\big\{\theta\big[S^{A^1}_{-k,-1} + uS^{A^2}_{-n,-1}\\ &\quad + (-u\phi_2-\phi_1)S^B_{-k,-1} - u\phi_2 S^B_{-n,-k-1}\big]\big\}\big]\\ &\leqslant \sum_{n\geqslant 0}\sum_{0\leqslant k\leqslant n} \inf_{u\geqslant 0} \exp\big\{(n-k)\big[\Lambda_{A^2}(\theta u) + \Lambda_B(-\theta\phi_2 u)\big]\\ &\quad + k\big[\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta u) + \Lambda_B\big(-\theta(\phi_1+u\phi_2)\big)\big] + \Gamma'(\theta,u)\big\}\\ &\leqslant \sum_{n\geqslant 0} n \sup_{\zeta\in[0,1]} \inf_{u\geqslant 0} \exp\left\{n\left[\zeta\big(\Lambda_{A^2}(\theta u) + \Lambda_B(-\theta\phi_2 u)\big)\right.\right.\\ &\quad\left.\left. + (1-\zeta)\big(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta u) + \Lambda_B\big(-\theta(\phi_1+u\phi_2)\big)\big) + \frac{\Gamma'(\theta,u)}{n}\right]\right\}. \quad (38)\end{aligned}$$

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

$$\Lambda^{\mathrm{II}}_{\mathrm{GPS},1}(\theta) \triangleq \sup_{\zeta\in[0,1]} \inf_{u\geqslant 0} \big[\zeta\big(\Lambda_{A^2}(\theta u) + \Lambda_B(-\theta\phi_2 u)\big) + (1-\zeta)\big(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta u) + \Lambda_B\big(-\theta(\phi_1+u\phi_2)\big)\big)\big]. \quad (39)$$

Let $\hat{u}^*(\theta,\zeta)$ be the optimal u in the above optimization problem for fixed ζ. From (38) we have

$$\mathbf{E}\big[\mathrm{e}^{\theta L^{\mathrm{II}}_{\mathrm{GPS},1}}\big] \leqslant \sum_{n\geqslant 0} n \sup_{\zeta\in[0,1]} \exp\Bigg\{ n\Bigg[\zeta\big(\Lambda_{A^2}\big(\theta\hat{u}^*\big)+\Lambda_B\big(-\theta\phi_2\hat{u}^*\big)\big) + (1-\zeta)\big(\Lambda_{A^1}(\theta)+\Lambda_{A^2}\big(\theta\hat{u}^*\big)+\Lambda_B\big(-\theta\big(\phi_1+\hat{u}^*\phi_2\big)\big)\big)+\frac{\Gamma'(\theta,\hat{u}^*)}{n}\Bigg]\Bigg\}. \quad (40)$$

Now for every $\varepsilon'>0$ and $\theta\geqslant 0$ we can take n large enough such that $\Gamma'(\theta,\hat{u}^*)/n<\varepsilon'$. For sufficiently small ε' and if $\Lambda^{\mathrm{II}}_{\mathrm{GPS},1}(\theta)<0$ then the infinite geometric series in the right-hand side of (40) converges to some $K_2(\theta,\varepsilon')$. That is,

$$\mathbf{E}\big[\mathrm{e}^{\theta L^{\mathrm{II}}_{\mathrm{GPS},1}}\big] \leqslant K_2\big(\theta,\varepsilon'\big) \quad \text{if } \Lambda^{\mathrm{II}}_{\mathrm{GPS},1}(\theta)<0. \quad (41)$$

We can now invoke the Markov inequality and by using the bounds (34) and (38) on (30) and (31) obtain

$$\begin{aligned}\mathbf{P}\big[L^1_0>U\big] &\leqslant \mathbf{P}\big[L^1_0>U \text{ and } \Omega_1\big]+\mathbf{P}\big[L^1_0>U \text{ and } \Omega_2\big]\\ &\leqslant \big(\mathbf{E}\big[\mathrm{e}^{\theta L^{\mathrm{I}}_{\mathrm{GPS},1}}\big]+\mathbf{E}\big[\mathrm{e}^{\theta L^{\mathrm{II}}_{\mathrm{GPS},1}}\big]\big)\mathrm{e}^{-\theta U}\\ &\leqslant \big(K_1(\theta,\varepsilon)+K_2\big(\theta,\varepsilon'\big)\big)\mathrm{e}^{-\theta U} \quad \text{if } \max\big(\Lambda^{\mathrm{I}}_{\mathrm{GPS},1}(\theta),\Lambda^{\mathrm{II}}_{\mathrm{GPS},1}(\theta)\big)<0. \quad (42)\end{aligned}$$

Optimizing over θ to get the tightest bound completes the proof of the proposition. □

We are now left with proving that this upper bound matches the lower bound θ^*_{GPS}. The result which is based on lemma 6.2 and convex duality is established in the next proposition.

Proposition 6.6 (GPS upper bound, Case 1). If $\mathbf{E}[A^2]<\phi_2\mathbf{E}[B]$ and assuming that the arrival and service processes satisfy assumptions A and C, the steady-state queue length L^1 of queue Q^1 satisfies

$$\limsup_{U\to\infty}\frac{1}{U}\log\mathbf{P}\big[L^1>U\big]\leqslant -\theta^*_{\mathrm{GPS}}.$$

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

$$\Lambda^{\mathrm{I}*}_{\mathrm{GPS},1}(a)\triangleq \inf_{\substack{\zeta(x_2-\phi_2x_3)+(1-\zeta)(y_1+y_2-y_3)=a\\ \zeta(x_2-\phi_2x_3)+(1-\zeta)(y_2-\phi_2y_3)\leqslant 0\\ 0\leqslant\zeta\leqslant 1}} \big[\zeta\big(\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)\big) + (1-\zeta)\big(\Lambda^*_{A^1}(y_1)+\Lambda^*_{A^2}(y_2)+\Lambda^*_B(y_3)\big)\big] \quad (43)$$

and

$$\Lambda^{\mathrm{II}*}_{\mathrm{GPS},1}(a) \triangleq \inf_{\substack{(1-\zeta)(y_1-\phi_1 y_3)=a \\ \zeta(x_2-\phi_2 x_3)+(1-\zeta)(y_2-\phi_2 y_3)\geqslant 0 \\ 0\leqslant\zeta\leqslant 1}} \big[\zeta\big(\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)\big) + (1-\zeta)\big(\Lambda^*_{A^1}(y_1)+\Lambda^*_{A^2}(y_2)+\Lambda^*_B(y_3)\big)\big], \tag{44}$$

which by a change of variables can be written as

$$\Lambda^{\mathrm{I}*}_{\mathrm{GPS},1}(a) = \inf_{\substack{(x_2-\phi_2 x_3)+(y_1+y_2-y_3)=a \\ (x_2-\phi_2 x_3)+(y_2-\phi_2 y_3)\leqslant 0}} \inf_{\zeta\in[0,1]} \left[\zeta\left(\Lambda^*_{A^2}\left(\frac{x_2}{\zeta}\right)+\Lambda^*_B\left(\frac{x_3}{\zeta}\right)\right) + (1-\zeta)\left(\Lambda^*_{A^1}\left(\frac{y_1}{1-\zeta}\right)+\Lambda^*_{A^2}\left(\frac{y_2}{1-\zeta}\right)+\Lambda^*_B\left(\frac{y_3}{1-\zeta}\right)\right)\right] \tag{45}$$

and

$$\Lambda^{\mathrm{II}*}_{\mathrm{GPS},1}(a) = \inf_{\substack{(y_1-\phi_1 y_3)=a \\ (x_2-\phi_2 x_3)+(y_2-\phi_2 y_3)\geqslant 0}} \inf_{\zeta\in[0,1]} \left[\zeta\left(\Lambda^*_{A^2}\left(\frac{x_2}{\zeta}\right)+\Lambda^*_B\left(\frac{x_3}{\zeta}\right)\right) + (1-\zeta)\left(\Lambda^*_{A^1}\left(\frac{y_1}{1-\zeta}\right)+\Lambda^*_{A^2}\left(\frac{y_2}{1-\zeta}\right)+\Lambda^*_B\left(\frac{y_3}{1-\zeta}\right)\right)\right]. \tag{46}$$

(It is here understood that at $\zeta = 0$ or $\zeta = 1$ the expressions in (45) and (46) take the corresponding values of expressions (43) and (44), respectively.) By [29, theorem 5.8] the function

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

is convex in $(x_2, x_3, y_1, y_2, y_3)$ and therefore the functions $\Lambda^{\mathrm{I}*}_{\mathrm{GPS},1}(a)$ and $\Lambda^{\mathrm{II}*}_{\mathrm{GPS},1}(a)$ 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. Moreover, they are lower semicontinuous by lemma 6.3. We will next show that the convex duals of these functions are $\Lambda^{\mathrm{I}}_{\mathrm{GPS},1}(\theta)$ and $\Lambda^{\mathrm{II}}_{\mathrm{GPS},1}(\theta)$, respectively. Indeed, by using convex duality, we have

$$\begin{aligned}
&\sup_a \big[\theta a - \Lambda^{\mathrm{I}*}_{\mathrm{GPS},1}(a)\big] \\
&\quad = \sup_{\zeta\in[0,1]} \sup_a \sup_{\substack{\zeta(x_2-\phi_2 x_3)+(1-\zeta)(y_1+y_2-y_3)=a \\ \zeta(x_2-\phi_2 x_3)+(1-\zeta)(y_2-\phi_2 y_3)\leqslant 0 \\ 0\leqslant\zeta\leqslant 1}} \big[\theta a - \zeta\big(\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)\big) \\
&\qquad - (1-\zeta)\big(\Lambda^*_{A^1}(y_1)+\Lambda^*_{A^2}(y_2)+\Lambda^*_B(y_3)\big)\big] \\
&\quad = \sup_{\zeta\in[0,1]} \inf_{u\geqslant 0} \sup_{\substack{x_2,x_3 \\ y_1,y_2,y_3}} \big[\theta\zeta(x_2-\phi_2 x_3)+\theta(1-\zeta)(y_1+y_2-y_3) - u\zeta(x_2-\phi_2 x_3)
\end{aligned}$$

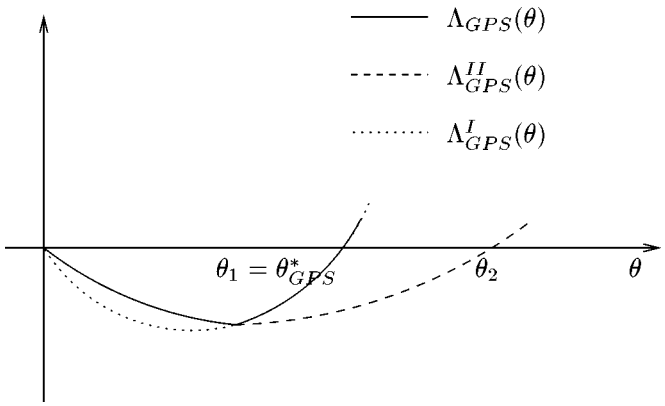

Figure 3. $\theta^*_{\text{GPS},1}$ as the largest positive root of the equation $\Lambda_{\text{GPS},1}(\theta) = 0$.

$$
\begin{aligned}
&- u(1-\zeta)(y_2 - \phi_2 y_3) - \zeta\big(\Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)\big) \\
&- (1-\zeta)\big(\Lambda^*_{A^1}(y_1) + \Lambda^*_{A^2}(y_2) + \Lambda^*_B(y_3)\big)\big] \\
&= \sup_{\zeta\in[0,1]} \inf_{u\geqslant 0} \big[\zeta\big(\Lambda_{A^2}(\theta - u) + \Lambda_B(-\theta\phi_2 + u\phi_2)\big) \\
&+ (1-\zeta)\big(\Lambda_{A^1}(\theta) + \Lambda_{A^2}(\theta - u) + \Lambda_B(-\theta + u\phi_2)\big)\big] \\
&= \Lambda^{\text{I}}_{\text{GPS},1}(\theta).
\end{aligned}
$$

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

$$\theta_{\text{I}} \triangleq \inf_{a>0} \frac{1}{a}\Lambda^{\text{I}*}_{\text{GPS},1}(a) \tag{47}$$

and

$$\theta_{\text{II}} \triangleq \inf_{a>0} \frac{1}{a}\Lambda^{\text{II}*}_{\text{GPS},1}(a). \tag{48}$$

Using the result of lemma 6.2, θ_{I} (respectively θ_{II}) is the largest positive root of $\Lambda^{\text{I}}_{\text{GPS},1}(\theta) = 0$ (respectively $\Lambda^{\text{II}}_{\text{GPS},1}(\theta) = 0$). It can be seen that $\Lambda^{\text{I}}_{\text{GPS},1}(\theta)$ satisfies the condition of lemma 6.2 (being negative for sufficiently small θ) because it takes the value zero at $\theta = 0$ and has negative right derivative at $\theta = 0$. The same is true for $\Lambda^{\text{II}}_{\text{GPS},1}(\theta)$. As figure 3 indicates, due to convexity, $\theta^*_{\text{GPS},1} \triangleq \min(\theta_{\text{I}}, \theta_{\text{II}})$ is the largest positive root of the equation $\Lambda_{\text{GPS},1}(\theta) \triangleq \max[\Lambda^{\text{I}}_{\text{GPS},1}(\theta), \Lambda^{\text{II}}_{\text{GPS},1}(\theta)] = 0$, that is, $-\theta^*_{\text{GPS},1}$ is equal to the upper bound established in proposition 6.5.

The last thing we have to show is that $\theta^*_{\text{GPS},1} = \theta^*_{\text{GPS}}$. This is based on $\theta^*_{\text{GPS},1}$ being equal to $\min(\theta_{\text{I}}, \theta_{\text{II}})$. Note, from (47), 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 4(a). Also, from (48), θ_{II} corresponds to the optimal solution of a

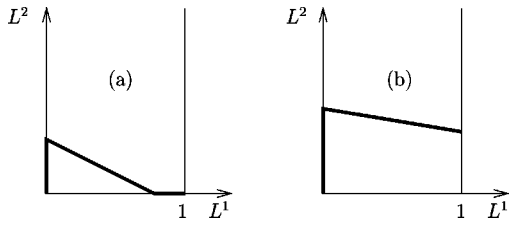

Figure 4. Trajectories for the control problems corresponding to $\theta_{\rm I}$ and $\theta_{\rm II}$.

control problem with a trajectory of the form appearing in figure 4(b).[3] This optimal control problem, whose trajectories appear in figures 4(a) and (b) is different from (GPS-OVERFLOW) in two aspects:

(i) on the L^2-axis the cost functional is $\Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)$ instead of $\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)$, and

(ii) its dynamics in region B are given by the equation $\dot{L}_2 = x_2 - \phi_2 x_3$. We will refer to this as the *modified* (GPS-OVERFLOW).

We will next argue that the trajectories in figures 4(a) and (b) are dominated by the ones in figures 2(a) and (b), respectively (equivalently, the optimal ζ in (43) and (44) is zero). Note that along the trajectories in figures 2(a) and (b) the modified (GPS-OVERFLOW) has identical cost structure and dynamics to the restricted (GPS-OVERFLOW). Thus, the above argument will establish $\theta^*_{\rm GPS,1} = \theta^*_{\rm GPS}$.

To this end, consider the trajectory in figure 4(a) with optimal value given by the expression (43). It can be seen that taking the time average over class two arrivals, i.e., setting the class two arrival rate to $\bar{x}_2 = \zeta x_2 + (1-\zeta)y_2$, we maintain feasibility and reduce the cost (by convexity). The resulting trajectory has either $\bar{x}_2 \leqslant \phi_2 x_3$ or $\bar{x}_2 > \phi_2 x_3$. In the former case, Q^2 stays empty during the first ζ fraction of its duration and it has the form appearing in figure 2(a). In the latter case, it has the form depicted in figure 4(a) but with $x_2 = y_2 = \bar{x}_2$ and $\bar{x}_2 > \phi_2 x_3$. We can now invoke the argument following equations (59) and (60) in the appendix to conclude that the trajectory of interest is dominated by the one in figure 2(a).

A similar argument applies to the trajectory in figure 4(b) with the optimal value given by the expression (44). We first shorten the time that it spends on the L_2 axis to obtain trajectories of the form appearing in figure 2(b) or figure 4(a). In the latter case, the argument outlined in the paragraph above applies. □

We summarize propositions 6.6 and 6.4 in the following proposition.

[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 (respectively y_1, y_2, y_3) the controls for the initial ζ (respectively last $1-\zeta$) fraction of the time.

Proposition 6.7 (GPS upper bound). Assuming that the arrival and service processes satisfy assumptions A and C, and under the GPS policy, the steady-state queue length L^1 of queue Q^1 satisfies

$$\limsup_{U\to\infty} \frac{1}{U} \log \mathbf{P}\big[L^1 > U\big] \leqslant -\theta^*_{\mathrm{GPS}}. \tag{49}$$

7. Reformulations and special cases

In this section we show an alternative expression for θ^*_{GPS} and specialize our results to the case of priority policies.

An interesting observation is that strict priority policies are a special case of the GPS policy. Class 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 class 1 customers, matches the FCFS single class result (see [4,19,21]) since under this policy, class 1 customers are oblivious of class 2 customers. We summarize the performance of priority policies in the next corollary, the proof of which can be found in [3].

Corollary 7.1 (Priority policies). Under strict priority policy for class 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 satisfies

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}\big[L^1 > U\big] = -\theta^*_{P_1}, \tag{50}$$

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

$$\theta^*_{P_1} = \inf_{a>0} \frac{1}{a} \Lambda^*_{P_1}(a), \tag{51}$$

and where

$$\Lambda^*_{P_1}(a) \triangleq \inf_{x_1 - x_3 = a} \big[\Lambda^*_{A^1}(x_1) + \Lambda^*_B(x_3)\big]. \tag{52}$$

Under strict priority policy for class 2 customers (P_2), the steady-state queue length L^1 of queue Q^1 satisfies

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}\big[L^1 > U\big] = -\theta^*_{P_2}, \tag{53}$$

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

$$\theta^*_{P_2} = \inf_{a>0} \frac{1}{a} \Lambda^*_{P_2}(a), \tag{54}$$

and where

$$\Lambda^*_{P_2}(a) \triangleq \inf_{\substack{x_1+x_2-x_3=a \\ x_2 \leqslant x_3}} \left[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)\right]. \tag{55}$$

As the results of theorem 3.1 and corollary 7.1 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 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 $\Lambda^{\text{I}}_{\text{GPS}}(\cdot)$ and $\Lambda^{\text{II}}_{\text{GPS}}(\cdot)$, the convex duals of $\Lambda^{\text{I}*}_{\text{GPS}}(\cdot)$ and $\Lambda^{\text{II}*}_{\text{GPS}}(\cdot)$, respectively. Notice, that $\Lambda^{\text{I}*}_{\text{GPS}}(a)$ and $\Lambda^{\text{II}*}_{\text{GPS}}(a)$ 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 7.2. θ^*_{GPS} is the largest positive root of the equation

$$\Lambda_{\text{GPS}}(\theta) \triangleq \Lambda_{A^1}(\theta) + \inf_{0 \leqslant u \leqslant \theta} \left[\Lambda_{A^2}(\theta - u) + \Lambda_B(-\theta + \phi_2 u)\right] = 0. \tag{56}$$

Proof. The first thing to note is that $\Lambda_{\text{GPS}}(\theta)$ 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.,

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

Next we show that equation (56) has a positive, possibly infinite, root. To this end, observe that

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

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

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

where the last inequality follows from the stability condition (6). The convexity of $\Lambda_{\text{GPS}}(\cdot)$ is sufficient to guarantee the existence of a positive, possibly infinite, root. Note that this also implies that $\Lambda_{\text{GPS}}(\cdot)$ is negative for sufficiently small $\theta > 0$ as the condition in lemma 6.2 requires.

We now calculate the functions $\Lambda^{\text{I}}_{\text{GPS}}(\theta)$ and $\Lambda^{\text{II}}_{\text{GPS}}(\theta)$, using convex duality. Note that $\Lambda^{\text{I}*}_{\text{GPS}}(a)$ and $\Lambda^{\text{II}*}_{\text{GPS}}(a)$ are both lower semicontinuous by lemma 6.3. We have

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

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

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

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

Using the result of lemma 6.2, $\theta_1 \triangleq \inf_{a>0}(1/a)\Lambda^{\mathrm{I}*}_{\mathrm{GPS}}(a)$ is the largest positive root of $\Lambda^{\mathrm{I}}_{\mathrm{GPS}}(\theta) = 0$ (this equation has a positive, possibly, infinite root by the argument used to establish that $\Lambda_{\mathrm{GPS}}(\theta) = 0$ does). Similarly, $\theta_2 \triangleq \inf_{a>0}(1/a)\Lambda^{\mathrm{II}*}_{\mathrm{GPS}}(a)$ is the largest positive root of $\Lambda^{\mathrm{II}}_{\mathrm{GPS}}(\theta) = 0$. By equation (9), $\theta^*_{\mathrm{GPS}} = \min(\theta_1, \theta_2)$. The situation is exactly the same as in figure 3, that is, θ^*_{GPS} is the largest positive root of the equation $\max[\Lambda^{\mathrm{I}}_{\mathrm{GPS}}(\theta), \Lambda^{\mathrm{II}}_{\mathrm{GPS}}(\theta)] = 0$.

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

$$
\begin{aligned}
\max\bigl(\Lambda^{\mathrm{I}}_{\mathrm{GPS}}(\theta), \Lambda^{\mathrm{II}}_{\mathrm{GPS}}(\theta)\bigr) &= \max\Bigl(\Lambda_{A^1}(\theta) + \inf_{u \geqslant 0} \bigl[\Lambda_{A^2}(\theta-u) + \Lambda_B(-\theta+u\phi_2)\bigr], \\
&\qquad \Lambda_{A^1}(\theta) + \inf_{u \leqslant \theta} \bigl[\Lambda_{A^2}(\theta-u) + \Lambda_B(-\theta+u\phi_2)\bigr]\Bigr) \\
&= \Lambda_{A^1}(\theta) + \inf_{0 \leqslant u \leqslant \theta} \bigl[\Lambda_{A^2}(\theta-u) + \Lambda_B(-\theta+u\phi_2)\bigr] \\
&\overset{(56)}{=} \Lambda_{\mathrm{GPS}}(\theta). \qquad \square
\end{aligned}
$$

Again, as it was the case with theorem 3.1, the result of theorem 7.2 can be specialized to the case of priority policies.

Corollary 7.3. θ^*_{P1} is the largest positive root of the equation

$$\Lambda_{P1}(\theta) \triangleq \Lambda_{A^1}(\theta) + \Lambda_B(-\theta) = 0. \tag{57}$$

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

$$\Lambda_{P2}(\theta) \triangleq \Lambda_{A^1}(\theta) + \inf_{0 \leqslant u \leqslant \theta} \left[\Lambda_{A^2}(\theta - u) + \Lambda_B(-\theta + u)\right] = 0. \tag{58}$$

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.

8. Conclusions

In this paper we considered a multiclass switch, with dedicated buffers for each service class. Under the GPS policy, we have obtained the asymptotic 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. This formulation provides particular insight into the problem, as it yields an explicit characterization of the most likely modes of overflow. We have addressed the case of multiplexing two streams. The general case of N streams remains an open problem.

Acknowledgement

We thank Kurt Majewski for spotting a gap in an earlier version of this paper.

Appendix

We will show that the trajectories in figure 2 are optimal over all feasible trajectories of (GPS-OVERFLOW).

The first property of (GPS-OVERFLOW) that we establish is that *optimal control trajectories can be taken to be constant* within each of the three regions of state dynamics. The result is stated in the next lemma, the proof of which is given in a somewhat more general context in [3]. It is based on the convexity of the large deviations rate functions of the arrival and service processes.

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

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

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

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

$$\int_{-\alpha T}^{0} \left[\Lambda_{A^1}^*\left(x_1^Q(t)\right) + \Lambda_{A^2}^*\left(x_2^Q(t)\right) + \Lambda_B^*\left(x_3^Q(t)\right)\right] \mathrm{d}t$$
$$= \alpha \int_{-T}^{0} \left[\Lambda_{A^1}^*\left(x_1^L(t)\right) + \Lambda_{A^2}^*\left(x_2^L(t)\right) + \Lambda_B^*\left(x_3^L(t)\right)\right] \mathrm{d}t.$$

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 figures 5(a), (b) and (c). If the trajectory in figure 5(a) is optimal then so is the scaled version (by $\alpha = a_2/a_1$) in figure 5(b) and as consequence its segment which appears in figure 5(c) is also optimal (since we have a free time problem).

We next proceed with the solution of (GPS-OVERFLOW) using an elaborate interchange argument, which is mainly based on convexity considerations. Starting from any arbitrary trajectory with piecewise constant controls as the one appearing in figure 6(a), we use the homogeneity property (by appropriately scaling the dashed segment) to reduce it to the one in figure 6(b). 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 figures 7(a) and (b). 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 figures 2(a) and (b).

Consider now the trajectories in figures 7(a) and (a′). The segment of (a) and (a′) that is in region A has the same slope, thus the same controls, which implies that the

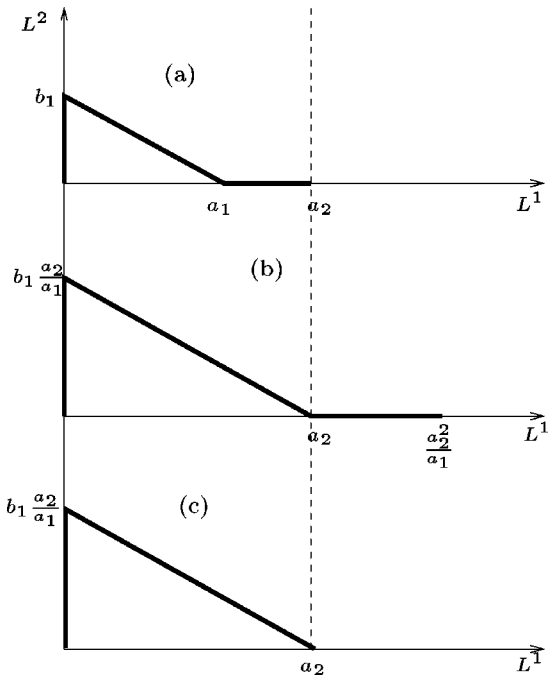

Figure 5. By the homogeneity property, optimality of the trajectory in (a) implies optimality of the trajectory in (b) which by its turn implies optimality of the trajectory in (c).

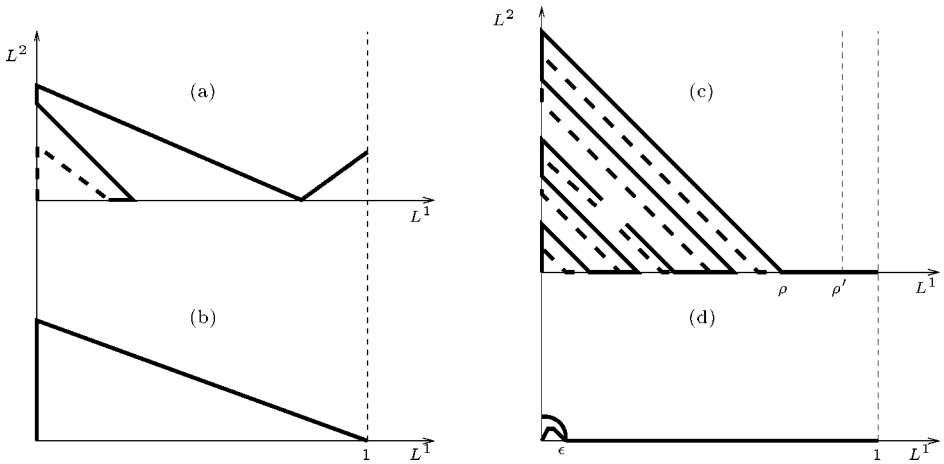

Figure 6. Using the homogeneity property the trajectory in (a) reduces to the one in (b). The same property is used to exclude trajectories with an infinite number of linear pieces such as the one in (c), and reduce them to the one in (d) which is "ε-close" to the trajectory in figure 2(a).

trajectory in (a′) 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 figures 2(a), (b), and 7(b).

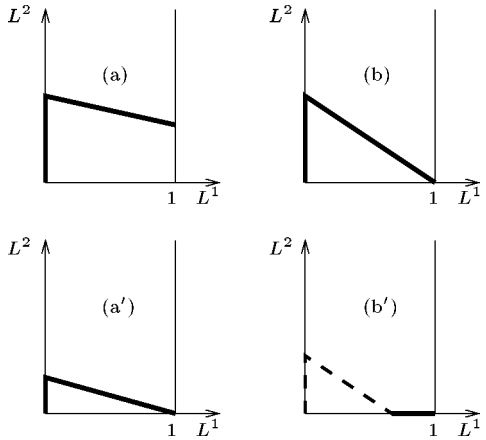

Figure 7. Candidates for optimal state trajectories are depicted in (a), (b). The trajectory in (a) is reduced to the one in (a′) which has the same form as the one in (b). The trajectory in (b) is reduced to the one in (b′) which is contradicted by the time-homogeneity property. Hence, optimal state trajectories have only the form in figures 2(a) and (b).

Finally, consider the state trajectory in figure 7(b). 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 constraints are

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

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 it is at least as profitable to have $x_2=y_2=\bar{x}_2$. The corresponding trajectory can either have the form in figure 2(a) or figure 7(b). If the latter is the case then

$$\bar{x}_2 > \phi_2 x_3, \tag{59}$$

$$\bar{x}_2 < \phi_2 y_3. \tag{60}$$

Consider the trajectory with $x_3'=x_3+\varepsilon/\zeta$ and $y_3'=y_3-\varepsilon/(1-\zeta)$ for some small $\varepsilon>0$. This latter trajectory serves the same total number of customers as the former one in the interval $[-T,0]$ (equal to $\zeta T x_3+(1-\zeta)Ty_3$) and it is at least as cheap by convexity of the rate functions. It is depicted in figure 7(b′). We can now apply the same argument to its dashed segment. If we keep doing that we conclude that the trajectory in figure 2(a) is at least as cheap.

Therefore, for every state trajectory of (GPS-OVERFLOW), there exists one of the forms depicted in figures 2(a) and (b) with no larger cost. Note that to arrive at this conclusion we have not considered trajectories with an infinite number of linear pieces

accumulating near the origin, such as the one appearing in figure 6(c). We next argue that such a trajectory is dominated by the one in figure 2(a). To see that let us consider an optimal trajectory such as the one in figure 6(c) with minimal final segment on the horizontal axis, i.e., an optimal trajectory with minimum $\|(\rho, 0) - (1, 0)\|$. We apply the homogeneity property to obtain the dashed (optimal) trajectory in the same figure with terminal state $(\rho', 0)$. Since we have a free time problem an optimal trajectory with terminal state $(1, 0)$ can be constructed by following the dashed one until state $(\rho', 0)$, and then switching to the solid one until state $(1, 0)$. Applying inductively the same construction we end up with a trajectory that stays on the horizontal axis except possibly when $\|(L^1, L^2)\| \leqslant \varepsilon$ (in the vicinity of the origin); see figure 6(d). This is a trajectory that follows the trajectory in figure 2(a) from $(\varepsilon, 0)$ to $(1, 0)$. Let J_ε denote its optimal value, and J^* denote the optimal value of (GPS-OVERFLOW). The above argument establishes

$$J^* \leqslant J_\varepsilon + \mathrm{O}(\varepsilon),$$

for all $\varepsilon > 0$. This suffices to exclude trajectories with infinite number of pieces. Note that if an optimal trajectory with infinite number of linear pieces does not have a final segment on the horizontal axis, it will have a segment with infinite number of linear pieces terminating on the vertical axis, thus, a similar argument holds in this case.

In summary, in this appendix we established the following:

Theorem A.2. The optimal value of the problem (GPS-OVERFLOW) is given by θ^*_{GPS}, as it is defined in (9).

References

[1] M.S. Bazaraa, H.D. Sherali and C.M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 2nd ed. (Wiley, New York, 1993).

[2] D. Bertsimas, I.C. Paschalidis and J.N. Tsitsiklis, On the large deviations behaviour of acyclic single class networks and multiclass queues, in: *RSS Workshop in Stochastic Networks*, Edinburgh, UK (1995).

[3] D. Bertsimas, I.C. Paschalidis and J.N. Tsitsiklis, Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach, IEEE Trans. Automat. Control 43(3) (1998) 315–335.

[4] D. Bertsimas, I.C. Paschalidis and J.N. Tsitsiklis, On the large deviations behaviour of acyclic networks of $G/G/1$ queues, Ann. Appl. Probab. 8(4) (1998) 1027–1069.

[5] J.A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation* (Wiley, New York, 1990).

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

[7] C.S. Chang and T. Zajic, Effective bandwidths of departure process from queues with time varying capacities, in: *Proc. IEEE Infocom '95*, Vol. 3, Boston, MA (April 1995) pp. 1001–1009.

[8] C. Courcoubetis and R. Weber, Estimation of overflow probabilities for state-dependent service of traffic streams with dedicated buffers, in: *RSS Workshop in Stochastic Networks*, Edinburgh, UK (1995).

[9] H. Cramér, Sûr un nouveau théorème-limite de la théorie des probabilités, in: *Colloque Consacré à la Théorie des Probabilités*, Actualités Scientifiques et Industrielles, Vol. 736 (Hermann, Paris, 1938) pp. 5–23.
[10] A. Dembo and T. Zajic, Large deviations: From empirical mean and measure to partial sums processes, Stochastic Process. Appl. 57 (1995) 191–224.
[11] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications* (Jones and Bartlett, 1993).
[12] A. Demers, S. Keshav and S. Shenker, Analysis and simulation of a fair queueing algorithm, J. Internetworking: Res. Experience 1 (1990) 3–26.
[13] 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, CA (1993).
[14] G. de Veciana and G. Kesidis, Bandwidth allocation for multiple qualities of service using generalized processor sharing, IEEE Trans. Inform. Theory 42(1) (1995).
[15] P. Dupuis and K. Ramanan, A Skorokhod problem formulation and large deviation analysis of a processor sharing model, Technical Report, Division of Applied Mathematics, Brown University (1997).
[16] A.I. Elwalid and D. Mitra, Effective bandwidth of general Markovian traffic sources and admission control of high speed networks, IEEE/ACM Trans. Networking 1(3) (1993) 329–343.
[17] A. Ganesh and V. Anantharam, Stationary tail probabilities in exponential server tandems with renewal arrivals, Queueing Systems 22 (1996) 203–248.
[18] R.J. Gibbens and P.J. Hunt, Effective bandwidths for the multi-type UAS channel, Queueing Systems 9 (1991) 17–28.
[19] P.W. Glynn and W. Whitt, Logarithmic asymptotics for steady-state tail probabilities in a single-server queue, J. Appl. Probab. A 31 (1994) 131–156.
[20] J.Y. Hui, Resource allocation for broadband networks, IEEE J. Selected Areas Commun. 6(9) (1988) 1598–1608.
[21] F.P. Kelly, Effective bandwidths at multi-class queues, Queueing Systems 9 (1991) 5–16.
[22] F.P. Kelly, Notes on effective bandwidths, in: *Stochastic Networks: Theory and Applications*, Vol. 9, eds. S. Zachary, I.B. Ziedins and F.P. Kelly (Oxford University Press, Oxford, 1996) pp. 141–168.
[23] G. Kesidis, J. Walrand and C.S. Chang, Effective bandwidths for multiclass Markov fluids and other ATM sources, IEEE/ACM Trans. Networking 1(4) (1993) 424–428.
[24] N. O'Connell, Queue lengths and departures at single-server resources, in: *RSS Workshop in Stochastic Networks*, Edinburgh, UK (1995).
[25] A.K. Parekh and R.G. Gallager, A generalized processor sharing approach to flow control in integrated services networks: The single node case, IEEE/ACM Trans. Networking 1(3) (1993) 344–357.
[26] I.C. Paschalidis, Large deviations in high speed communication networks, Ph.D. thesis, Massachusetts Institute of Technology (1996).
[27] I.C. Paschalidis, Class-specific quality of service guarantees in multimedia communication networks, Technical Report, Department of Manufacturing Engineering, Boston University (June 1998); to appear in Automatica (Special Issue on Control Methods for Communication Networks).
[28] A. Puhalskii, Large deviation analysis of the single server queue, Queueing Systems 21 (1995) 5–66.
[29] R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970.
[30] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis* (Chapman and Hall, New York, 1995).
[31] A. Weiss, An introduction to large deviations for communication networks, IEEE J. Selected Areas Commun. 13(6) (1995) 938–952.
[32] Z.-L. Zhang, Large deviations and the generalized processor sharing scheduling for a two-queue system, Queueing Systems 26(3/4) (1997) 229–264.