Asymptotic Buffer Overflow Probabilities in Multiclass Multiplexers: An Optimal Control Approach

Dimitris Bertsimas, Ioannis Ch. Paschalidis, *Member, IEEE*, and John N. Tsitsiklis, *Senior Member, IEEE*

***Abstract*—We consider a multiclass multiplexer with support for multiple service classes and dedicated buffers for each service class. Under specific scheduling policies for sharing bandwidth among these classes, we seek the asymptotic (as the buffer size goes to infinity) tail of the buffer overflow probability for each dedicated buffer. We assume dependent arrival and service processes as 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 introduce a novel *optimal control* approach to address these problems. In particular, 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. We specialize our results to the *generalized processor sharing policy (GPS)* and the *generalized longest queue first policy (GLQF)*. The performance of strict priority policies is obtained as a corollary. We compare the GPS and GLQF policies and conclude that GLQF achieves smaller overflow probabilities than GPS for all arrival and service processes for which our analysis holds. Our results have important implications for traffic management of high-speed networks and can be used as a basis for an admission control mechanism which guarantees a different loss probability for each class.**

***Index Terms*—ATM-based B-ISDN, communication networks, large deviations.**

I. Introduction

HIGH-SPEED packet-switched communication networks, for example ATM-based B-ISDN networks, accommodate various types of traffic (digitized voice, encoded video, and data) and offer a variety of services. 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 probability of 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}. 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. The problem is particularly difficult since it essentially requires finding the distributions of queue lengths in a multiclass network of G/G/1 queues with correlated arrival processes (since it is needed to model bursty traffic) and nonexponentially 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 we focus on 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 (switch)* which accommodates multiple service classes. A *service class* is characterized by the statistical properties of the incoming traffic and by the QoS requirements. Different types of traffic (i.e., voice, video, data, etc.) have different statistical properties, and in addition they may have distinct QoS requirements (e.g., video may need more stringent QoS requirements than voice), thus they belong to different service classes. Moreover, sessions of the same type of traffic may belong to different service classes if they have different QoS requirements (e.g., we can consider a situation where we want to support both high- and low-quality video).

Under specific scheduling policies for sharing bandwidth among service classes, we seek the asymptotic (as the buffer size goes to infinity) tail of the buffer overflow probability that each class experiences. We focus on the *generalized processor sharing policy (GPS)* (introduced in [9] and further explored in [29] and [30]) and the *generalized longest queue first policy (GLQF)*. The GLQF policy is a generalization of the *longest queue first policy (LQF)*, under which the server allocates all of its capacity to the longest queue. 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 [16]).

Manuscript received July 8, 1996. Recommended by Associate Editor, J. B. Lasserre. This work was supported in part by a Presidential Young Investigator award DDM-9158118 with matching funds from Draper Laboratory, by the ARO under Grant DAAL-03-92-G-0115, by the NSF under Grant NCR-9706148, and by funds from the College of Engineering, Boston University.

D. Bertsimas is with the Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

I. Ch. Paschalidis is with the Department of Manufacturing Engineering, Boston University, Boston, MA 02215 USA (e-mail: yannisp@bu.edu).

J. N. Tsitsiklis is with the Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Publisher Item Identifier S 0018-9286(98)02073-X.

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 out 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. (The general case of N streams is more complicated since there is an exponential explosion of the number of overflow modes.) Our results have important implications in traffic management of high-speed networks. They can be used as a basis for an admission control mechanism which provides statistical QoS guarantees for each service class and allows for different QoS requirements for each class (see [28] where this direction is pursued).

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 (see [33] for a survey). The problem of estimating tail probabilities of rare events in a single class queue has received extensive attention in the literature [22], [20], [23], [24], [21], [15], [32]. The extension of these ideas to single-class networks, although much harder, has been treated in various versions and degrees of rigor in [1], [18], [7], [25], and [10].

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 [11] and [34]. Both papers use a large deviation result for the departure process from a G/D/1 queue [10]. Tail overflow probabilities for the GPS policy and deterministic service capacity were also reported in [26] and [7]. The authors in [7] 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 [19] the authors suggest the use of the LQF policy in high-speed networks and use a deterministic model (only the rate of each incoming stream is known) to calculate buffer sizes that guarantee no loss with probability one. Our analysis significantly extends the scope of this work by generalizing the policy (GLQF) and by taking the statistical properties of the incoming traffic into account. This leads to a more efficient utilization of the network resources. Large deviations results for the LQF policy in an M/M/1 setting are also reported in [31].

We consider the following to be some of the main contributions of the work in this paper.

- The derivation of tight asymptotic expressions for the performance of multiclass multiplexers operated under sophisticated (and of interest in practice) scheduling policies for sharing bandwidth among classes.
- The introduction of an *optimal control* approach to address the problem. Our formulation is different from the one in [7]. 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 optimal control formulation is general enough to include any scheduling policy; only the dynamics of the system are policy-dependent. Optimal control formulations are also used in [31] for large deviations results for jump Markov processes.
- The extension of some GPS results existing in the literature 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 class and use the GPS policy for the remaining two. These two remaining classes face a GPS server with stochastic capacity. Stochastic capacity significantly alters the way overflows occur. To see this, note that in deriving their results [11] and [34] use the departure process from a G/D/1 queue. The large deviations behavior of the departure process is different with deterministic and stochastic service capacity as it is pointed out in [1] and [8].
- The introduction of a new policy, the GLQF, which generalizes the LQF policy. We provide analytic performance analysis results for the GLQF policy and compare it to the GPS policy. We argue that GLQF is preferable, at least in the absence of fairness considerations.

Regarding the structure of this paper, we begin in Section II with a brief review of the large deviations results that we will use. We also state a set of assumptions to which arrival and service processes need to conform. In Section III we formally define the multiclass model that we consider, and in Subsections III-A and III-B we introduce the GPS and the GLQF policy, respectively. Moreover, in Subsection III-C we provide an outline of the methodology that we follow in proving our results. In Section IV we establish lower bounds on the overflow probability under the GLQF (Subsection IV-A) and the GPS policy (Subsection IV-B). The optimal control formulation is introduced in Section V and the results are specialized to the GPS (Subsection V-A) and the GLQF (Subsection V-B) case. In Section VI we describe the most likely modes of overflow, under both policies, obtained from the solution of the corresponding control problems. In Section VII we state the upper bound for the GPS policy (the proof is quite technical and involved and we omit it in the interest of space; we refer the interested reader to [3]). Section VIII contains the proof for the upper bound in the GLQF case. We gather our main performance analysis results in Section IX, where we also treat the special case of strict priority policies. Finally, we compare the two scheduling policies in Section X, and conclusions are in Section XI.

II. Preliminaries

In this section we review some basic results on the theory of large deviations [13], [31], [4] that will be used in the sequel.

We first state the Gärtner–Ellis theorem [17], [14] (see also Bucklew [4] and Dembo and Zeitouni [13]) which establishes a *Large Deviations Principle (LDP)* for dependent random variables in R. It is a generalization of Cramér's theorem [6] which applies to independent and identically distributed (i.i.d.) random variables.

Consider a sequence $\{S_1, S_2, \cdots\}$ of random variables, with values in 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 an 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 [13] 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 holds that

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

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* $S_n \in R$ 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 [12], under certain mild mixing conditions on the stationary sequence $\{X_i;\ i \geq 1\}$, Dembo and Zajic establish an LDP for the process $S_n(\cdot)$ in $D[[0, 1], (R, ||\cdot||_\infty)]$ (the space of right continuous functions with left limits equipped with the supremum norm topology). 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 $m \in N$, for every $\epsilon_1, \epsilon_2 > 0$, and for every scalar $a_0, \cdots, a_{m-1}$, there exists $M > 0$ such that for all $n \geq M$ and all $k_0, \cdots, k_m$ with $1 = k_0 \leq k_1 \leq \cdots \leq k_m = n$

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

A detailed discussion of this assumption, and the technical conditions under which it is satisfied, is given in [12]. In the simpler case where dependencies are not present (i.e., $S_i = \sum_{j=1}^i X_j$, where X_i's are i.i.d.), Assumption B is a consequence of Mogulskii's theorem (see [13]). 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 $a_0, \cdots, a_{m-1}$. In [12] 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.

We will be also making the following related assumption.

Assumption C: For all $m \in N$ there exists $M > 0$ and a function $\Gamma(\cdot)$ with $0 \leq \Gamma(y) < \infty$, for all $y > 0$, such that for all $n \geq M$ and all $k_0, \cdots, k_m$ with $1 = k_0 \leq k_1 \leq \cdots \leq k_m = n$

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

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

Chang [5] provides a uniform bounding condition under which Assumption B is true and verifies that the condition is satisfied by renewal, Markov-modulated, and stationary processes with mild mixing conditions. Using his uniform bounding condition it can be verified (see [5] for a proof) that Assumption C is also satisfied. This latter assumption can be viewed as the "convex dual analog" of Assumption B.

On a notational remark, in the rest of the paper we will be denoting by $S_{i,j}^X \triangleq \sum_{k=i}^j X_k,\ i \leq j$ the partial sums of the random sequence $\{X_i;\ i \in Z\}$. 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 [cf. (2) and (5)], respectively, of the process X.

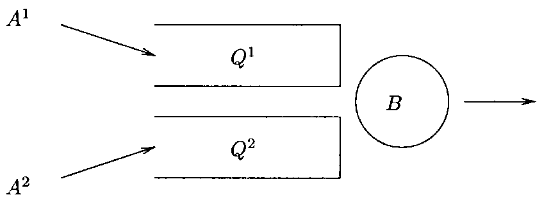


Fig. 1. A multiclass model.

III. A Multiclass Model

In this section we introduce a multiclass multiplexer model that we plan to analyze, in the large deviations regime, under two specific scheduling policies for sharing bandwidth among classes: the GPS and the GLQF. The former policy is described in Subsection III-A, and the latter one in Subsection III-B. Subsection III-C provides an outline of the approach we follow.

Consider the system depicted in Fig. 1. We assume a slotted time model (i.e., discrete time) and we let A_i^1 (respectively, A_i^2), $i \in Z$, denote the number of class 1 (respectively, 2) customers that enter queue Q^1 (respectively, Q^2) at time i. Both queues have infinite buffers and share the same server which can process B_i customers during the time interval $[i, i+1]$. We assume that the processes $\{A_i^1;\ i \in Z\}$, $\{A_i^2;\ i \in Z\}$, and $\{B_i;\ i \in Z\}$ are stationary and mutually independent. However, we allow dependencies between the number of customers at different slots in each process. 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 denote by L_i^1 and L_i^2 the queue lengths at time i (without counting arrivals at time i) in queues Q^1 and Q^2, respectively. We assume that the server allocates its capacity between queues Q^1 and Q^2 according to a work-conserving policy (i.e., the server never stays idle when there is work in the system). We also assume that the queue length processes $\{L_i^j, j = 1, 2, i \in Z\}$ are stationary (under a work-conserving policy, the system reaches 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 discrete-time "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 affect the results in the large deviations regime.

Finally, we assume that the arrival and service processes satisfy an LDP (Assumption A) as well as Assumptions B and C. As we have noted in Section II, 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.

A. The GPS Policy

The *generalized processor sharing* (GPS) policy was proposed in [9] and further explored in [29] and [30]. 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. This policy is also known as *fair queueing* because it guarantees a certain fraction of the available bandwidth to each class and thus avoids situations that occur under first come/first served (FCFS) where a bursty class can take the lion's share of the bandwidth.

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

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

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

B. The GLQF Policy

Fig. 2 depicts the operation of the GLQF policy in the $L^1 - L^2$ space. Fix the parameter of the policy $\beta \ge 0$. There is a threshold line, of slope β, which divides the positive orthant of the $L^1 - L^2$ space in two regions. The GLQF policy serves class 2 customers above the threshold line and class 1 below it. The value $\beta = 1$ corresponds to the longest queue first (LQF) policy. Intuitively, the GLQF policy tries to maintain a desirable ratio β of the queue lengths per class by attending to the class that overshoots this ratio. Since delays are due to long queues, it is also intuitive that the GLQF policy tries to balance (with a β "bias") the delay of the two classes.

More formally, we define the GLQF policy to be the work-conserving policy that at each time slot i serves class 1 customers when

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

It serves class 2 customers when

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

When

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

or when

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

then the GLQF policy allocates appropriate capacity to both classes of customers such that $L_{i+1}^2 = \beta L_{i+1}^1$. Similarly, whenever $L_i^2 = \beta L_i^1$, the GLQF policy allocates its capacity to class 1 and 2 customers so that $L_{i+1}^2 = \beta L_{i+1}^1$, if possible.

C. An Outline of Our Approach

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

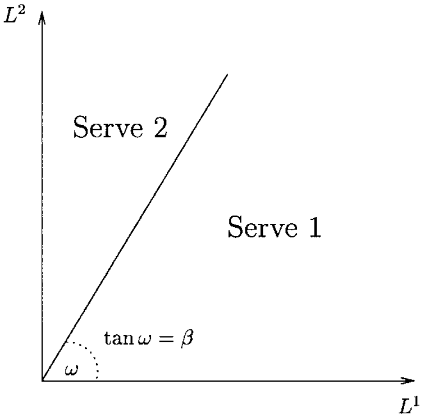


Fig. 2. The operation of the GLQF policy.

We will prove that these overflow probabilities satisfy

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

and

$$\mathbf{P}[L_i^1 > U] \sim e^{-U\theta^*_{\mathrm{GLQF}}} \tag{11}$$

asymptotically, as $U \to \infty$.

To this end, we will develop a lower bound on each overflow probability, along with a matching upper bound. Fix the scheduling policy and 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 $\mathbf{P}[L_i^1 > U]$. We select the tightest lower bound by performing the minimization $\theta^*_{\mathrm{GPS}} = \min_\omega \theta(\omega)$, in the GPS case, which amounts to solving a deterministic optimal control problem. Notice that both the function $\theta(\omega)$ and the overflow paths ω depend on the policy, hence this minimization will yield a different optimal value in the GLQF case, which we will denote by θ^*_{GLQF}. Optimal trajectories (paths) of the control problem correspond to *most likely* overflow scenarios. We will show that these must be of one out of two possible types, in both the GPS and the GLQF case. In other words, with high probability, overflow occurs in one out of two possible modes.

To establish the tightness of the lower bounds and show (10) and (11), we will obtain an upper bound on $\mathbf{P}[L_i^1 > U]$. We will first obtain a sample path upper bound, i.e., $L_i^1 \le \tilde{L}_i^1$ (which implies $\mathbf{P}[L_i^1 > U] \le \mathbf{P}[\tilde{L}_i^1 > U]$) and then establish that $\mathbf{P}[\tilde{L}_i^1 > U]$ is at most $e^{-U\theta^*_{\mathrm{GPS}}}$ in the GPS case and $e^{-U\theta^*_{\mathrm{GLQF}}}$ in the GLQF case.

IV. A Lower Bound

In this section we establish a lower bound on the overflow probability $\mathbf{P}[L_i^1 > U]$ under each one of the two scheduling policies. We first present the lower bound in the GLQF case and then the one in the GPS case. The main idea is that we select the dominant overflow scenarios which are responsible for overflows with high probability. The optimal control formulation in Section V substantiates why the selected scenarios are the dominant ones.

A. GLQF Lower Bound

Proposition 4.1 (GLQF Lower Bound): Assuming that the arrival and service processes satisfy Assumptions A and B, and under the GLQF policy, the steady-state queue length L^1 of queue Q^1 at an arbitrary time slot satisfies

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \ge -\theta^*_{\mathrm{GLQF}} \tag{12}$$

where θ^*_{GLQF} is given by

$$\theta^*_{\mathrm{GLQF}} = \min\left[\inf_{a>0} \frac{1}{a}\Lambda^{\mathrm{I}*}_{\mathrm{GLQF}}(a), \inf_{a>0} \frac{1}{a}\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(a)\right] \tag{13}$$

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

$$\Lambda^{\mathrm{I}*}_{\mathrm{GLQF}}(a) \triangleq \inf_{\substack{x_1 - x_3 = a \\ x_2 \le \beta(x_1 - x_3)}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)] \tag{14}$$

and

$$\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(a) \triangleq \inf_{\substack{x_1 - \phi x_3 = a \\ x_2 - (1-\phi)x_3 = \beta a \\ 0 \le \phi < 1}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \tag{15}$$

Proof: Let $-n \le 0$ and $a > 0$. Fix $x_1, x_2, x_3 \ge 0$, $0 \le \phi < 1$, and $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ and consider the event

$$\mathcal{A} \triangleq \Big\{ \big|S^{A^1}_{-n,-i-1} - (n-i)x_1\big| \le \epsilon_1 n, \big|S^{A^2}_{-n,-i-1} - (n-i)x_2\big| \le \epsilon_2 n,\ \big|S^B_{-n,-i-1} - (n-i)x_3\big| \le \epsilon_3 n, \quad i = 0, 1, \cdots, n-1 \Big\}.$$

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:

$$\begin{aligned} &\text{Scenario 1)} \quad x_1 - x_3 = a \\ &\qquad\qquad\quad x_2 \le \beta(x_1 - x_3) \\ &\text{Scenario 2)} \quad x_1 - \phi x_3 = a \\ &\qquad\qquad\quad x_2 - (1-\phi)x_3 = \beta a. \end{aligned} \tag{16}$$

Under Scenario 1, even if the server always serves class 1 customers[1] in $[-n, 0]$ we have that $L_0^1 \ge na - n\epsilon_1'$, where $\epsilon_1' \to 0$ as $\epsilon_1, \epsilon_2, \epsilon_3 \to 0$.

Consider now Scenario 2, and let us for the moment ignore ϵ's (i.e., $\epsilon_1 = \epsilon_2 = \epsilon_3 = 0$). We will argue that $L_0^1 \ge na$. If $L^2_{-n} = \beta L^1_{-n}$, then both queues build up together, with the relation $L^2 = \beta L^1$ holding in the interval $[-n, 0]$. According to the GLQF policy the server arbitrarily allocates its capacity to the two queues, giving fraction ϕ to Q^1 and the remaining $1-\phi$ to Q^2, yielding $L_0^1 = na + L^1_{-n} \ge na$. If $L^2_{-n} > \beta L^1_{-n}$, then the first queue receives less capacity than $n\phi x_3$ in $[-n, 0]$, resulting also in $L_0^1 \ge na$. Finally, consider the case $L^2_{-n} < \beta L^1_{-n}$. Then at some time $-t \in [-n, 0]$ we have $L^1_{-t} = L^1_{-n} + (n-t)(x_1 - x_3)$ and $L^2_{-t} = L^2_{-n} + (n-t)x_2$.

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

Notice that $x_2 > \beta(x_1 - x_3)$, since otherwise we have a contradiction, i.e.,

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

Thus, for large enough n, there exists some t, say t^*, such that $L^2_{-t^*} = \beta L^1_{-t^*}$. This relationship, along with $L^1_{-t^*} + L^2_{-t^*} \ge (n-t^*)(1+\beta)a$ implies $L^1_{-t^*} \ge (n-t^*)a$. Now note that from t^* both queues build up together with the relation $L^2 = \beta L^1$ holding. Observing that $L^1_0 \ge L^1_{-t^*} + t^* a$, we conclude that $L^1_0 \ge na$.

When we take the ϵ's into account a similar argument holds. With $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ and with the same ϕ, there exists $\epsilon_2' > 0$ such that the queue lengths are within an ϵ_2' band of the values in the previous paragraph, resulting in $L^1_0 \ge na - n\epsilon_2'$, where $\epsilon_2' \to 0$ as $\epsilon_1, \epsilon_2, \epsilon_3 \to 0$.

The probability of Scenario 1 is a lower bound on $\mathbf{P}[L^1_0 \ge na]$. 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

$$\begin{aligned}
\mathbf{P}\big[L^1_0 \ge n(a-\epsilon_1')\big]
&\ge \sup_{\substack{x_1-x_3=a\\ x_2\le\beta(x_1-x_3)}} \mathbf{P}\big[\big|S^{A^1}_{-n,-i-1} - (n-i)x_1\big| \le \epsilon_1 n,\\
&\qquad\qquad i=0,1,\cdots,n-1\big]\\
&\quad \times \mathbf{P}\big[\big|S^{A^2}_{-n,-i-1} - (n-i)x_2\big| \le \epsilon_2 n,\ i=0,1,\cdots,n-1\big]\\
&\quad \times \mathbf{P}\big[\big|S^{B}_{-n,-i-1} - (n-i)x_3\big| \le \epsilon_3 n,\ i=0,1,\cdots,n-1\big]\\
&\ge \exp\left\{-n\left(\inf_{\substack{x_1-x_3=a\\ x_2\le\beta(x_1-x_3)}} \big[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2)\right.\right.\\
&\qquad\qquad \left.\left. + \Lambda^*_B(x_3)\big] + \epsilon\right)\right\}\\
&= \exp\big\{-n\big(\Lambda^{\mathrm{I}*}_{\mathrm{GLQF}}(a) + \epsilon\big)\big\}
\end{aligned} \tag{17}$$

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

Similarly, calculating the probability of Scenario 2, we have

$$\begin{aligned}
\mathbf{P}\big[L^1_0 \ge n(a-\epsilon_2')\big]
&\ge \sup_{\substack{x_1-\phi x_3=a\\ x_2-(1-\phi)x_3=\beta a\\ 0\le\phi<1}} \mathbf{P}\big[\big|S^{A^1}_{-n,-i-1} - (n-i)x_1\big| \le \epsilon_1 n,\\
&\qquad\qquad i=0,1,\cdots,n-1\big]\\
&\quad \times \mathbf{P}\big[\big|S^{A^2}_{-n,-i-1} - (n-i)x_2\big| \le \epsilon_2 n,\ i=0,1,\cdots,n-1\big]\\
&\quad \times \mathbf{P}\big[\big|S^{B}_{-n,-i-1} - (n-i)x_3\big| \le \epsilon_3 n,\ i=0,1,\cdots,n-1\big]\\
&\ge \exp\left\{-n\left(\inf_{\substack{x_1-\phi x_3=a\\ x_2-(1-\phi)x_3=\beta a\\ 0\le\phi<1}} \big[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2)\right.\right.\\
&\qquad\qquad \left.\left. + \Lambda^*_B(x_3)\big] + \epsilon'\right)\right\}\\
&= \exp\big\{-n\big(\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(a) + \epsilon'\big)\big\}
\end{aligned} \tag{18}$$

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

Combining (17) and (18) we obtain that for all $\epsilon, \epsilon' > 0$ there exists N such that for all $n > N$

$$\begin{aligned}
&\frac{1}{n}\log \mathbf{P}\big[L^1_0 \ge n(a-\epsilon)\big]\\
&\qquad \ge -\big(\min\big(\Lambda^{\mathrm{I}*}_{\mathrm{GLQF}}(a), \Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(a)\big) + \epsilon'\big).
\end{aligned} \tag{19}$$

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

$$\begin{aligned}
&\frac{1}{U}\log \mathbf{P}[L^1 > U]\\
&\quad = \frac{1}{n(a-\epsilon)}\log \mathbf{P}\big[L^1_0 \ge n(a-\epsilon)\big]\\
&\quad \ge -\frac{1}{a-\epsilon}\big(\min\big(\Lambda^{\mathrm{I}*}_{\mathrm{GLQF}}(a), \Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(a)\big) + \epsilon'\big)
\end{aligned}$$

which implies

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

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

$$\begin{aligned}
&\lim_{U\to\infty}\frac{1}{U}\log \mathbf{P}[L^1 > U]\\
&\quad \ge -\min\left[\inf_{a>0}\frac{1}{a}\Lambda^{\mathrm{I}*}_{\mathrm{GLQF}}(a), \inf_{a>0}\frac{1}{a}\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(a)\right]. \qquad \square
\end{aligned}$$

B. GPS Lower Bound

We next turn our attention to the GPS policy and establish a lower bound on the overflow probability. In the interest of space we provide an outline of the proof. The complete proof can be found in [3].

Proposition 4.2 (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

$$\lim_{U\to\infty}\frac{1}{U}\log \mathbf{P}[L^1 > U] \ge -\theta^*_{\mathrm{GPS}} \tag{20}$$

where θ^*_{GPS} is given by

$$\theta^*_{\mathrm{GPS}} = \min\left[\inf_{a>0}\frac{1}{a}\Lambda^{\mathrm{I}*}_{\mathrm{GPS}}(a), \inf_{a>0}\frac{1}{a}\Lambda^{\mathrm{II}*}_{\mathrm{GPS}}(a)\right] \tag{21}$$

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

$$\Lambda^{\mathrm{I}*}_{\mathrm{GPS}}(a) \triangleq \inf_{\substack{x_1+x_2-x_3=a\\ x_2\le\phi_2 x_3}} \big[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)\big] \tag{22}$$

and

$$\Lambda^{\mathrm{II}*}_{\mathrm{GPS}}(a) \triangleq \inf_{\substack{x_1-\phi_1 x_3=a\\ x_2\ge\phi_2 x_3}} \big[\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)\big]. \tag{23}$$

Proof (Outline): Let $-n \leq 0$ and $a > 0$. Let also $x_1, x_2, x_3 \geq 0$ be the empirical arrival and service rates during the interval $[-n, -1]$ (in the sense introduced in the proof of Proposition 4.1)

We focus on two particular scenarios:

$$\begin{aligned} \text{Scenario 1)} \quad & x_1 + x_2 - x_3 = a \\ & x_2 \leq \phi_2 x_3 \\ \text{Scenario 2)} \quad & x_1 - \phi_1 x_3 = a \\ & x_2 \geq \phi_2 x_3. \end{aligned} \tag{24}$$

Under both scenarios it can be established that $L_0^1 \geq na$. Calculating their probabilities we obtain a lower bound on $\mathbf{P}[L_0^1 \geq na]$. We then optimize over all the parameters involved and use arguments similar to the ones in Proposition 4.1 to arrive at (20). □

V. The Optimal Control Problem

In this section we introduce an optimal control problem for each of the two scheduling policies and show that its optimal value provides the exponents θ^*_{GPS} and θ^*_{GLQF}, respectively, of the overflow probabilities. We will first motivate the control problem formulation and establish some properties that are independent of the scheduling policy. We will subsequently specialize the results to the GLQF and the GPS policy.

To motivate the control problem, we relate it, heuristically, with the problem of obtaining an asymptotically tight estimate of the overflow probability.[2] 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)

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

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

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

where $x(t)$ are arbitrary nonnegative 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 $a_0, a_1, \cdots, a_{m-1}$ 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 and the scheduling policy employed. According to the policy, we will distinguish a number of regions of system dynamics. We do not yet specify the scheduling policy, we assume, however, that we employ a scheduling policy with *linear* dynamics. More specifically, we consider M *convex* subsets $\mathcal{R}_1, \cdots, \mathcal{R}_M$ of the positive orthant such that

$$\bigcup_{i=1}^{M} \mathcal{R}_i = \{(L^1, L^2) \mid L^1 \geq 0,\ L^2 \geq 0\},$$
$$\mathcal{R}_i \cap \mathcal{R}_j = \emptyset,\ \forall i \neq j.$$

We fix constants $\gamma^1_{\mathcal{R}_j, i}, \gamma^2_{\mathcal{R}_j, i}$ for $j = 1, \cdots, M$ and $i = 1, 2, 3$ and consider the following system dynamics.

Region $\mathcal{R}_j$: $(L^1(t), L^2(t)) \in \mathcal{R}_j$ where

$$\begin{aligned} \dot{L}^1 &= \gamma^1_{\mathcal{R}_j,1} x_1(t) + \gamma^1_{\mathcal{R}_j,2} x_2(t) - \gamma^1_{\mathcal{R}_j,3} x_3(t) \\ \dot{L}^2 &= \gamma^2_{\mathcal{R}_j,1} x_1(t) + \gamma^2_{\mathcal{R}_j,2} x_2(t) - \gamma^2_{\mathcal{R}_j,3} x_3(t) \\ \dot{L}^1 + \dot{L}^2 &= x_1(t) + x_2(t) - x_3(t). \end{aligned}$$

Dotted variables in the above expressions denote derivatives.[3] Let (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 (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

(OVERFLOW)

$$\begin{aligned} \text{minimize} \quad & \int_{-T}^{0} [\Lambda^*_{A^1}(x_1(t)) + \Lambda^*_{A^2}(x_2(t)) + \Lambda^*_B(x_3(t))]\, dt \\ \text{subject to:} \quad & L^1(-T) = L^2(-T) = 0 \\ & L^1(0) = 1 \\ & L^2(0) : \text{free} \\ & T : \text{free} \\ & \{L^j(t) : t \in [-T, 0],\ j = 1, 2\} \in \text{(DYNAMICS)}. \end{aligned} \tag{25}$$

The first property of (OVERFLOW) that we show is that *optimal control trajectories can be taken to be constant* within each of the state dynamics regions.

[2] Such a relation can be rigorously established using the sample path LDP for the arrival and service processes, as it is defined in [12] and [5].

[3] Here we use the notion of derivative for simplicity of the exposition. Note that these derivatives may not exist everywhere. Thus, in Region $\mathcal{R}_j$ for example, the rigorous version of the statement $\dot{L}^1 + \dot{L}^2 = x_1(t) + x_2(t) - x_3(t)$ is $L^1(t_2) + L^2(t_2) = L^1(t_1) + L^2(t_1) + \int_{t_1}^{t_2} (x_1(t) + x_2(t) - x_3(t))\, dt$ for all intervals (t_1, t_2) that the system remains in Region $\mathcal{R}_j$.

Lemma 5.1: 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 $\mathcal{R}_j$. 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 $t = -T_2$.

Proof: We will focus on one region of system dynamics, say $\mathcal{R}_j$. 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

$$\begin{aligned}(L^1(-T_1), L^2(-T_1)) &= (a_1, a_2) \in \mathcal{R}_j\\ (L^1(-T_2), L^2(-T_2)) &= (b_1, b_2) \in \mathcal{R}_j \qquad (26)\end{aligned}$$

and stays in Region $\mathcal{R}_j$, i.e., $(L^1(t), L^2(t)) \in \mathcal{R}_j$ for all $t \in (-T_1, -T_2)$, where

$$\begin{aligned}&L^k(t)\\ &\quad = a_k + \int_{-T_1}^{t} \left[\gamma^k_{\mathcal{R}_j,1} x_1(\tau) + \gamma^k_{\mathcal{R}_j,2} x_2(\tau) - \gamma^k_{\mathcal{R}_j,3} x_3(\tau)\right] d\tau,\\ &\qquad k = 1, 2, t \in (-T_1, -T_2). \qquad (27)\end{aligned}$$

Moreover, we also have

$$\begin{aligned}&L^k(-T_2)\\ &\quad = a_k + \int_{-T_1}^{-T_2} \left[\gamma^k_{\mathcal{R}_j,1} x_1(\tau) + \gamma^k_{\mathcal{R}_j,2} x_2(\tau) - \gamma^k_{\mathcal{R}_j,3} x_3(\tau)\right] d\tau\\ &\quad = b_k, \qquad k = 1, 2. \qquad (28)\end{aligned}$$

We will prove that the time–average control trajectory

$$\begin{aligned}\bar{x}_i(\tau) = \frac{1}{T_1 - T_2} \int_{-T_1}^{-T_2} x_i(t)\, dt,&\\ i = 1, 2, 3, \forall \tau \in [-T_1, -T_2] \quad (29)&\end{aligned}$$

is no more costly. To this end, notice that the time–average trajectory has the same end points [i.e., satisfies (26)], moves along a straight line, and thus stays in Region $\mathcal{R}_j$ (by convexity) for $t \in (-T_1, -T_2)$. Moreover, by convexity of the rate functions we have

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

□

Given this property, to solve (OVERFLOW) it suffices to restrict ourselves to state trajectories with constant control variables in each of the regions $\mathcal{R}_j$. 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 for the problem of moving from the start state to the end state of the segment.

Consider now a control trajectory $\{x^L_i(t);\ t \in [-T, 0]\}$ 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

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

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

$$\begin{aligned}\int_{-\alpha T}^{0} &\left[\Lambda^*_{A^1}\big(x^Q_1(t)\big) + \Lambda^*_{A^2}\big(x^Q_2(t)\big) + \Lambda^*_B\big(x^Q_3(t)\big)\right] dt\\ &= \alpha \int_{-T}^{0} \left[\Lambda^*_{A^1}\big(x^L_1(t)\big) + \Lambda^*_{A^2}\big(x^L_2(t)\big) + \Lambda^*_B\big(x^L_3(t)\big)\right] dt.\end{aligned}$$

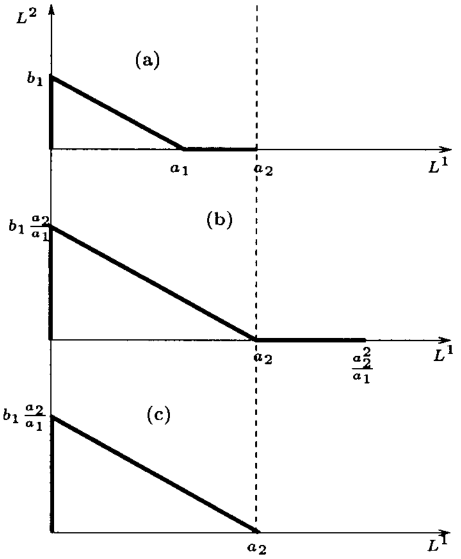


Fig. 3. By the homogeneity property, optimality of the trajectory in (a) implies optimality of the trajectory in (b) which in turn implies optimality of the trajectory in (c).

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

In the rest of this section we will specialize the optimal control formulation to the GPS and the GLQF case and use Lemma 5.1 along with the homogeneity property to obtain an optimal solution.

A. The GPS Optimal Control Problem

In the case of the GPS policy we will distinguish three regions of system dynamics, depending on which of the two queues is empty. In particular, we have:

Region $\mathcal{R}_1$: $L^1(t), L^2(t) > 0$, 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 $\mathcal{R}_2$: $L^1(t) = 0, L^2(t) > 0$, where according to the GPS policy

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

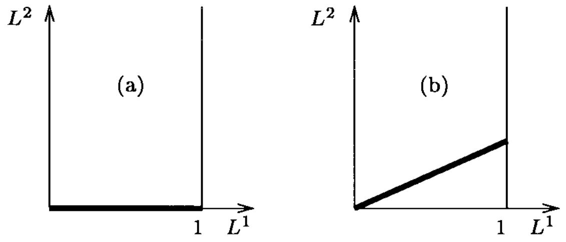


Fig. 4. In searching for optimal state trajectories of (GPS-OVERFLOW), we only need to consider trajectories of the form in (a) or (b).

Region $\mathcal{R}_3$: $L^1(t) > 0, L^2(t) = 0$, where according to the GPS policy

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

We let (GPS-DYNAMICS) denote the set of state trajectories $L^j(t),\ j = 1,2,\ t \in [-T, 0]$ that obey these dynamics. We will denote by (GPS-OVERFLOW) the special case of the problem (OVERFLOW), where state trajectories are constrained to satisfy (GPS-DYNAMICS).

The main result of this subsection is the following theorem.

Theorem 5.2: The optimal value of the problem (GPS-OVERFLOW) is given by θ^*_{GPS}, as it is defined in (21).

Due to space limitations we will skip the proof; we refer the interested reader to [3]. The proof uses Lemma 5.1 and the homogeneity property and follows an elaborate interchange argument to reduce any trajectory which is a potential candidate for optimality to one of the two trajectories that appear in Fig. 4.

B. The GLQF Optimal Control Problem

We next turn our attention to the GLQF policy. Depending on the state of the system, we distinguish the following three regions of system dynamics:

Region $\mathcal{R}_1$: $L^2(t) > \beta L^1(t)$, where according to the GLQF policy

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

Region $\mathcal{R}_2$: $L^2(t) < \beta L^1(t)$, where according to the GLQF policy

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

Region $\mathcal{R}_3$: $L^2(t) = \beta L^1(t)$, where according to the GLQF policy

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

Let (GLQF-DYNAMICS) denote the set of state trajectories $L^j(t),\ j = 1,2,\ t \in [-T, 0]$ that obey these dynamics. We will denote by (GLQF-OVERFLOW) the special case of the problem (OVERFLOW), where state trajectories are constrained to satisfy (GLQF-DYNAMICS).

This problem exhibits both the properties of constant control trajectories (cf. Lemma 5.1) within each region of system dynamics and homogeneity. Using these properties, we can make the reductions appearing in Fig. 5(a)–(c), starting from an arbitrary trajectory with piecewise constant controls. More specifically, consider first an arbitrary trajectory with linear pieces as the one in Fig. 5(a). We apply Lemma 5.1 to its initial segment (until it reaches $L^1 = \rho$), and we obtain a no more costly segment which stays in Region $\mathcal{R}_1$ and is arbitrarily close to the threshold line $L^2 = \beta L^1$. By a continuity argument, we conclude that the initial segment of the trajectory in Fig. 5(a) (until it reaches $L^1 = \rho$) reduces to the corresponding segment of the trajectory in Fig. 5(b). Using the same argument for the remaining segments of the trajectory in Fig. 5(a), it reduces to the one in Fig. 5(b). We now apply the homogeneity property to the latter trajectory to finally obtain the trajectory in Fig. 5(c). We conclude that optimal state trajectories can be reduced to having one of the forms depicted in Fig. 5(d)–(f).

The optimal trajectory of the form shown in Fig. 5(d) has value equal to $\inf_T [T\Lambda^{\text{I}*}_{\text{GLQF}}(\frac{1}{T})]$, and the optimal trajectory of the form shown in Fig. 5(e) has value equal to $\inf_T [T\Lambda^{\text{II}*}_{\text{GLQF}}(\frac{1}{T})]$, where $\Lambda^{\text{I}*}_{\text{GLQF}}(\cdot)$ and $\Lambda^{\text{II}*}_{\text{GLQF}}(\cdot)$ are defined in (14) and (15), respectively. Consider now the best trajectory of the form shown in Fig. 5(f), which has value

$$\inf_{T} \inf_{\substack{x_1 = \frac{1}{T} \\ x_2 - x_3 \geq \beta \frac{1}{T}}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_{B}(x_3)]. \tag{30}$$

The functions $\Lambda^*_{A^2}(x_2)$ and $\Lambda^*_B(x_3)$ are nonnegative, convex, and achieve their minimum value which is equal to zero at $x_2 = \mathbf{E}[A^2_0]$ and $x_3 = \mathbf{E}[B_0]$, respectively. Moreover, due to the stability condition (9) we have $\mathbf{E}[A^2_0] - \mathbf{E}[B_0] < 0$. Since $\frac{1}{T} \geq 0$ and in order to have $x_2 - x_3 \geq \beta\frac{1}{T}$, it has to be the case 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 as $x_2 - x_3 \geq \beta\frac{1}{T}$ holds. Also, if $x_3 < \mathbf{E}[B_0]$ is the case, we can increase x_3 and reduce the cost, as long as $x_2 - x_3 \geq \beta\frac{1}{T}$ holds. Thus, at optimality it is true that $x_2 - x_3 = \beta\frac{1}{T}$. Then, the expression in (30) is equal to $\inf_T [T\Lambda^{\text{II}*}_{\text{GLQF}}(\frac{1}{T})]$ with $\phi = 0$ in the definition of $\Lambda^{\text{II}*}_{\text{GLQF}}(\frac{1}{T})$. Thus, since the calculation of $\Lambda^{\text{II}*}_{\text{GLQF}}(\frac{1}{T})$ involves optimization over ϕ, we conclude that the state trajectory Fig. 5(f) is no more profitable than the one in Fig. 5(e), leaving us with only the trajectories in Fig. 5(d) and (e) as possible candidates for optimality. We summarize the above discussion in the following theorem.

Theorem 5.3: The optimal value of the problem (GLQF-OVERFLOW) is given by θ^*_{GLQF}.

VI. The Most Likely Paths

In essence, solving the control problem is equivalent to discovering scenarios of overflow that maximize the overflow probability over all feasible overflow scenarios. In this section we summarize these *most likely* ways of overflow for both policies.

A. The GPS Most Likely Paths

The two optimal state trajectories of (GPS-OVERFLOW) are the two generic most likely ways that queue Q^1 overflows, under the GPS policy. In particular, we distinguish two cases.

Case 1) Suppose $\theta^*_{\text{GPS}} = \inf_a \Lambda^{\text{I}*}_{\text{GPS}}(a)/a$ holds. Let $a^* > 0$ be the optimal solution of this optimization

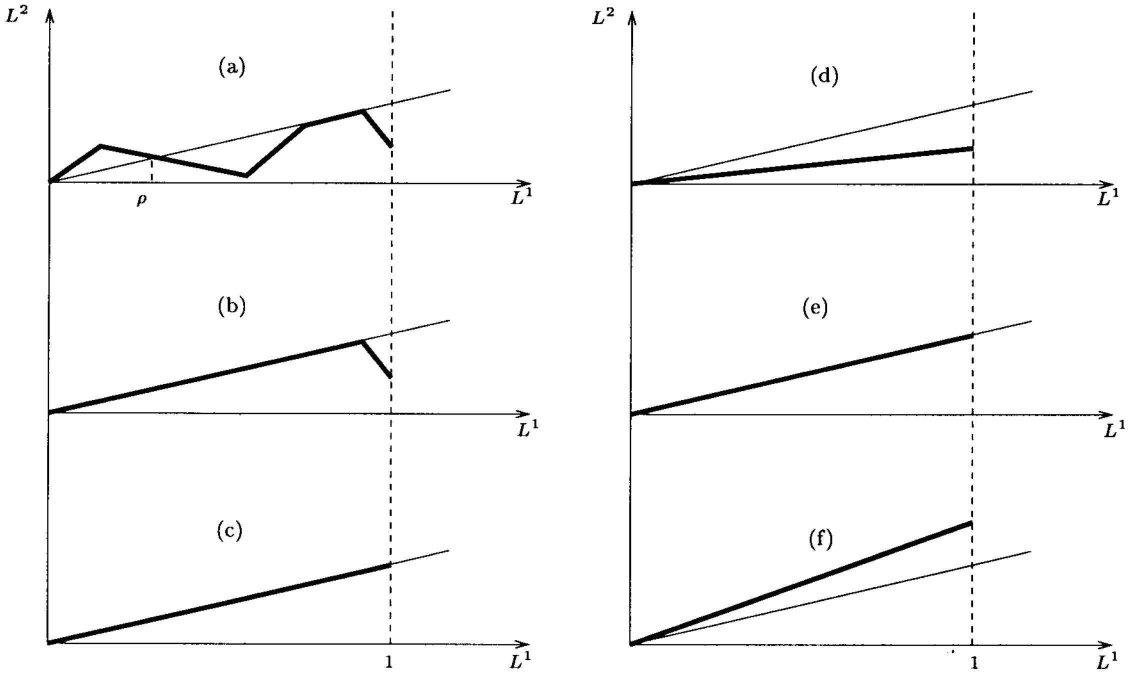


Fig. 5. By the property of constant controls within each region of system dynamics the state trajectory in (b) is no more costly than the trajectory in (a). Also, by the homogeneity property, optimality of the state trajectory in (b) implies optimality of the trajectory in (c). Candidates for optimal state trajectories are depicted in (d)–(f). The trajectory in (f) is eliminated as less profitable to the one in (e). Hence, without loss of optimality we can restrict attention to trajectories of the form in (d) and (e).

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 $\Lambda_{\mathrm{GPS}}^{\mathrm{I}*}(a^*)$ [cf. (22)]. The trajectory in L^1–L^2 space is depicted in Fig. 4(a).

Case 2) Suppose $\theta_{\mathrm{GPS}}^* = \inf_a \Lambda_{\mathrm{GPS}}^{\mathrm{II}*}(a)/a$ holds. Let $a^* > 0$ be 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 $\Lambda_{\mathrm{GPS}}^{\mathrm{II}*}(a^*)$ [cf. (23)]. The trajectory in L^1–L^2 space is depicted in Fig. 4(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 [29] 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 class 1 connections can be allowed without compromising the QoS. 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.

B. The GLQF Most Likely Paths

Considering now the GLQF policy, the two optimal state trajectories for the problem (GLQF-OVERFLOW) are most likely ways that queue Q^1 overflows. We distinguish two cases.

Case 1) Suppose $\theta_{\mathrm{GLQF}}^* = \inf_a \Lambda_{\mathrm{GLQF}}^{\mathrm{I}*}(a)/a$ holds. Let $a^* > 0$ be the optimal solution of this optimization problem. 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 $\Lambda_{\mathrm{GLQF}}^{\mathrm{I}*}(a^*)$ [cf. (14)]. In this case the first queue is building up to an $O(U)$ level, while the second queue builds up at a rate of x_2^*, in such a way that the server allocates its entire capacity to the first queue. The trajectory in L^1–L^2 space is depicted in Fig. 5(d).

Case 2) Suppose $\theta_{\mathrm{GLQF}}^* = \inf_a \Lambda_{\mathrm{GLQF}}^{\mathrm{II}*}(a)/a$ holds. Let $a^* > 0$ be the optimal solution of this optimization problem. Again, the first queue builds up linearly with rate a^*, during a period of duration U/a^*, and with the empirical rates of the processes A^1, A^2, and B being roughly equal to the optimal solution (x_1^*, x_2^*, x_3^*), respectively, of the optimization problem appearing in the definition of $\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(a^*)$ [cf. (15)]. In this case both queues are building up, the first to an $O(U)$ level and the second to an $O(\beta U)$ level. The trajectory in L^1-L^2 space is depicted in Fig. 5(e).

VII. A GPS Upper Bound

In this section we present an upper bound on the probability $\mathbf{P}[L_0^1 > U]$, in the case of the GPS policy. In particular, we have established that as $U \to \infty$ we have $\mathbf{P}[L_0^1 > U] \le e^{-\theta^*_{\mathrm{GPS}} U + o(U)}$, where $o(U)$ denotes functions with the property $\lim_{U\to\infty} \frac{o(U)}{U} = 0$. The proof is quite involved and uses the special structure of the problem which was revealed by the corresponding optimal control problem. Thus, the results in Section V are critical in establishing the upper bound.

Due to space limitations we omit the proof, which can be found in [3]. In proving the upper bound we distinguished two cases:

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

Case 2) $\mathbf{E}[A^2] \ge \phi_2 \mathbf{E}[B]$;

and established an upper bound for each one of them. The main result is the following proposition.

Proposition 7.1 (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 at an arbitrary time slot satisfies

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \le -\theta^*_{\mathrm{GPS}}. \tag{31}$$

VIII. A GLQF Upper Bound

In this section we develop an upper bound on the probability $\mathbf{P}[L_0^1 > U]$, for the GLQF case. In particular, we will prove that as $U \to \infty$ we have $\mathbf{P}[L_0^1 > U] \le e^{-\theta^*_{\mathrm{GLQF}} U + o(U)}$, where $o(U)$ denotes functions with the property $\lim_{U\to\infty} \frac{o(U)}{U} = 0$. This proof is different from the corresponding one in the GPS case in that it is independent from the GLQF optimal control formulation.

Before we proceed into the proof of the upper bound, we derive an alternative expression for θ^*_{GLQF} which will be essential in the proof. In the next theorem, we will show that the calculation of θ^*_{GLQF} is equivalent to finding the maximum root of a convex function.

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 8.1: 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. □

On a notational remark, we will be denoting by $\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\cdot)$ and $\Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\cdot)$, the convex duals of $\Lambda^{\mathrm{I}*}_{\mathrm{GLQF}}(\cdot)$ and $\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(\cdot)$, respectively. Notice that the latter are convex functions. For $\Lambda^{\mathrm{I}*}_{\mathrm{GLQF}}(a)$, convexity is implied by the fact that it is the value function of a convex optimization problem with a appearing only in the right-hand side of the constraints. For $\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(a)$, the same argument applies when we note the following reformulation:

$$\begin{aligned}\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(a) &= \inf_{\substack{x_1 - \phi x_3 = a \\ x_2 - (1-\phi) x_3 = \beta a \\ 0 \le \phi < 1}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)] \\ &= \inf_{\substack{x_1 - x_3' = a \\ x_2 - (x_3 - x_3') = \beta a \\ 0 \le x_3' \le x_3}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)].\end{aligned}$$

In preparation for the following theorem we prove the next monotonicity lemma.

Lemma 8.2 (Monotonicity): Consider a random process $\{X_i; i \in Z\}$ that satisfies Assumption A. Assume $X_i \ge 0, i \in Z$. Then for all $\theta \le \theta'$ we have $\Lambda_X(\theta) \le \Lambda_X(\theta')$.

Proof: $X_i \ge 0$, $i \in Z$ implies $S^X_{1,n} \ge 0$ which in turn implies

$$\mathbf{E}\left[e^{\theta S^X_{1,n}}\right] \le \mathbf{E}\left[e^{\theta' S^X_{1,n}}\right]$$

for all $\theta \le \theta'$. □

The above lemma clearly applies to the arrival and service processes. The next result is critical in establishing a matching upper bound on the overflow probability.

Theorem 8.3: θ^*_{GLQF} is the largest positive root of the equation

$$\Lambda_{\mathrm{GLQF}}(\theta) \triangleq \max\left[\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta), \Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta)\right] = 0 \tag{32}$$

where $\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\cdot)$ is the convex dual of $\Lambda^{\mathrm{I}*}_{\mathrm{GLQF}}(\cdot)$ and is given by

$$\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta) = \inf_{u \le 0} [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta + u\beta)] \tag{33}$$

and $\Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\cdot)$ is the convex dual of $\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(\cdot)$ and for $\theta \ge 0$ satisfies

$$\begin{aligned}\Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta) = \inf_{u \ge 0} [&\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) \\ &+ \max(\Lambda_B(-u), \Lambda_B(-\theta + u\beta))].\end{aligned} \tag{34}$$

Proof: Let us first calculate $\Lambda^{\rm I}_{\rm GLQF}(\cdot)$ and $\Lambda^{\rm II}_{\rm GLQF}(\cdot)$ by using convex duality. We have

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

Similarly

$$\begin{aligned}\Lambda^{\rm II}_{\rm GLQF}(\theta)&\\ &= \sup_a \left[\theta a - \Lambda^{\rm II*}_{\rm GLQF}(a)\right]\\ &= \sup_a \sup_{\substack{x_1-\phi x_3=a\\ x_2-(1-\phi)x_3=\beta(x_1-\phi x_3)\\ 0\le\phi<1}} [\theta a - \Lambda^*_{A^1}(x_1) - \Lambda^*_{A^2}(x_2)\\ &\qquad - \Lambda^*_B(x_3)]\\ &= \inf_u \sup_{\substack{x_1,x_2,x_3\\ 0\le\phi<1}} [\theta(x_1-\phi x_3) - \Lambda^*_{A^1}(x_1) - \Lambda^*_{A^2}(x_2)\\ &\qquad - \Lambda^*_B(x_3) + u(x_2 - \beta x_1 + (\beta\phi + \phi - 1)x_3)]\\ &= \inf_u [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \sup_{0\le\phi<1} \Lambda_B(-\theta\phi\\ &\qquad + (\beta\phi + \phi - 1)u)]\\ &= \inf_u [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u)\\ &\qquad + \max(\Lambda_B(-u), \Lambda_B(-\theta + u\beta))]\\ &= \inf_{u\ge 0} [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u)\\ &\qquad + \max(\Lambda_B(-u), \Lambda_B(-\theta + u\beta))].\end{aligned}$$

In the fifth equality above, we have used the monotonicity of $\Lambda_B(\cdot)$ (see Lemma 8.2) and the fact that the argument $-\theta\phi + (\beta\phi + \phi - 1)u$ is linear in ϕ, thus taking its maximum value at either $\phi = 0$ or $\phi = 1$. For the sixth equality above, notice that because $\Lambda_B(\cdot)$ is nondecreasing it holds

$$\begin{aligned}&\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \max(\Lambda_B(-u), \Lambda_B(-\theta + u\beta))\\ &\quad = \begin{cases}\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u), & \text{if } u < \frac{\theta}{1+\beta}\\ \Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta + u\beta), & \text{if } u \ge \frac{\theta}{1+\beta}\end{cases}\end{aligned} \tag{35}$$

since at the upper branch $-u > -\theta + u\beta$ and at the lower branch $-u \le -\theta + u\beta$. Differentiating the above expression at $u = 0$, and for $\theta \ge 0$, we obtain

$$\underbrace{-\beta\dot{\Lambda}_{A^1}(\theta)}_{\le 0} + \underbrace{\dot{\Lambda}_{A^2}(0) - \dot{\Lambda}_B(0)}_{\substack{(9)\\ \le 0}} \le 0$$

which implies (by convexity) that the infimum over unrestricted u has to be the same with the infimum over $u \ge 0$.

Using the result of Lemma 8.1, $\rho_1 \triangleq \inf_a \frac{1}{a}\Lambda^{\rm I*}_{\rm GLQF}(a)$ is the largest positive root of $\Lambda^{\rm I}_{\rm GLQF}(\theta) = 0$ (it is not hard to verify that this equation has a positive, possibly infinite root). Similarly, $\rho_2 \triangleq \inf_a \frac{1}{a}\Lambda^{\rm II*}_{\rm GLQF}(a)$ is the largest positive root of $\Lambda^{\rm II}_{\rm GLQF}(\theta) = 0$. By (13), $\theta^*_{\rm GLQF} = \min(\rho_1, \rho_2)$. This implies that $\theta^*_{\rm GLQF}$ is the largest positive root of the equation $\max[\Lambda^{\rm I}_{\rm GLQF}(\theta), \Lambda^{\rm II}_{\rm GLQF}(\theta)] = 0$. □

We next prove the upper bound for the overflow probability.

Proposition 8.4 (GLQF Upper Bound): Under the GLQF policy, assuming that the arrival and service processes satisfy Assumptions A and C, the steady-state queue length L^1 of queue Q^1 at an arbitrary time slot satisfies

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \le -\theta^*_{\rm GLQF}. \tag{36}$$

Proof: Without loss of generality we derive an upper bound for $\mathbf{P}[L^1_0 > U]$. We will restrict ourselves to sample paths with $L^1_0 > 0$ since the remaining sample paths, with $L^1_0 = 0$, do not contribute to the probability $\mathbf{P}[L^1_0 > U]$.

Consider a busy period for the system that starts at some time $-n < 0$ $(L^1_{-n} = L^2_{-n} = 0)$ and has not ended until time zero. Such a time $-n$ exists due to the stability condition (9). Note that since the system is busy in the interval $[-n, 0]$, the server works at capacity and therefore serves B_i customers at slot i, for $i \in [-n, 0]$. We will partition the set of sample paths, with $L^1_0 > 0$, in three subsets Ω_1, Ω_2, and Ω_3. The first subset, Ω_1, contains all sample paths at which only class 1 customers get serviced in the interval $[-n, 0]$. As a consequence

$$\begin{aligned}&L^1_{-k} = S^{A^1}_{-n,-k-1} - S^B_{-n,-k-1}\\ &L^2_{-k} = S^{A^2}_{-n,-k-1}, \quad \text{and} \quad \beta L^1_{-k} \ge L^2_{-k}, \qquad \forall k \in [0, n]\end{aligned}$$

which implies

$$L^1_0 = S^{A^1}_{-n,-1} - S^B_{-n,-1}$$

and

$$\beta\left(S^{A^1}_{-n,-1} - S^B_{-n,-1}\right) \ge S^{A^2}_{-n,-1}.$$

Thus we have (37), as shown at the bottom of the next page.

The second subset, Ω_2, contains sample paths at which class 1 customers do not receive the entire capacity, and $\beta L^1_0 \le L^2_0$. That is, there exists a $\phi \in [0, 1]$ such that class 1 customers receive only a ϕ fraction of the total capacity $(\phi S^B_{-n,-1})$. Then we have

$$\begin{aligned}&\mathbf{P}\left[L^1_0 > U \text{ and } \Omega_2\right]\\ &\quad \le \mathbf{P}\big[\exists n \ge 0, 0 \le \phi < 1, \text{ s.t. } S^{A^1}_{-n,-1} - \phi S^B_{-n,-1} > U \text{ and}\\ &\qquad\qquad \beta\left(S^{A^1}_{-n,-1} - \phi S^B_{-n,-1}\right) \le S^{A^2}_{-n,-1} - (1-\phi)S^B_{-n,-1}\big].\end{aligned}$$

Hence, we obtain an upper bound on $\mathbf{P}\left[L^1_0 > U \text{ and } \Omega_2\right]$ which is given in (38), shown at the bottom of the next page.

Finally, the third subset, Ω_3, contains sample paths at which class 1 customers do not receive the entire capacity, and $\beta L^1_0 \ge L^2_0$. Then there exists $k \in [0, n]$ such that the interval $[-k, 0]$ is the maximal interval that only class 1 customers get serviced. That is, $\beta L^1_{-i} \geq L^2_{-i}$, $i \in [0, k-1]$ and $\beta L^1_{-k} \leq L^2_{-k}$. Since class 1 customers do not receive the entire capacity, there exists $0 \leq \phi < 1$ such that $L^1_{-k} = S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1}$. Since $\beta L^1_{-k} \leq L^2_{-k}$, we have

$$\begin{aligned}&\beta\bigl(S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1}\bigr)\\&\quad \leq S^{A^2}_{-n,-k-1} - (1-\phi) S^B_{-n,-k-1}.\end{aligned} \tag{39}$$

Now, due to the way we defined k we have $L^1_{-i} = L^1_{-k} + S^{A^1}_{-k,-i-1} - S^B_{-k,-i-1}$, $i \in [0, k-1]$, and the inequality $\beta L^1_{-i} \geq L^2_{-i}$ becomes

$$\begin{aligned}&\beta\bigl(S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1} + S^{A^1}_{-k,-i-1} - S^B_{-k,-i-1}\bigr)\\&\quad \geq S^{A^2}_{-n,-k-1} - (1-\phi) S^B_{-n,-k-1} + S^{A^2}_{-k,-i-1}\end{aligned}$$

which by (39) implies

$$\beta\bigl(S^{A^1}_{-k,-i-1} - S^B_{-k,-i-1}\bigr) \geq S^{A^2}_{-k,-i-1}, \qquad i \in [0, k-1].$$

Thus we have (40), as shown at the bottom of the page.

Let us now define

$$L^{\mathrm{I}}_{\mathrm{GLQF}} \triangleq \max_{\{n \geq 0:\ \beta(S^{A^1}_{-n,-1} - S^B_{-n,-1}) \geq S^{A^2}_{-n,-1}\}} \bigl(S^{A^1}_{-n,-1} - S^B_{-n,-1}\bigr)$$

and the quantities $L^{\mathrm{II}}_{\mathrm{GLQF}}$ and $L^{\mathrm{III}}_{\mathrm{GLQF}}$, as shown at the bottom of the next page. By bringing the constraints in the objective function we obtain

$$\begin{aligned}L^{\mathrm{I}}_{\mathrm{GLQF}} \triangleq \max_{n \geq 0} \inf_{u \geq 0} \bigl[&(1 + u\beta) S^{A^1}_{-n,-1} - u S^{A^2}_{-n,-1}\\&+ (-1 - \beta u) S^B_{-n,-1}\bigr]\end{aligned} \tag{41}$$

$$\begin{aligned}L^{\mathrm{II}}_{\mathrm{GLQF}} \triangleq \max_{\substack{n \geq 0\\ 0 \leq \phi < 1}} \inf_{u \geq 0} \bigl[&(1 - u\beta) S^{A^1}_{-n,-1} + u S^{A^2}_{-n,-1}\\&+ (-\phi + u\beta\phi - u + u\phi) S^B_{-n,-1}\bigr]\end{aligned} \tag{42}$$

and

$$\begin{aligned}L^{\mathrm{III}}_{\mathrm{GLQF}} \triangleq \max_{\substack{n \geq 0\\ 0 \leq k \leq n\\ 0 \leq \phi < 1}} \Bigl\{&\inf_{u_1 \geq 0} \bigl[(1 - u_1\beta) S^{A^1}_{-n,-k-1} + u_1 S^{A^2}_{-n,-k-1}\\&+ (-\phi + u_1\beta\phi - u_1 + u_1\phi) S^B_{-n,-k-1}\bigr]\\&+ \inf_{u_2 \geq 0} \bigl[(1 + u_2\beta) S^{A^1}_{-k,-1} - u_2 S^{A^2}_{-k,-1}\\&+ (-1 - u_2\beta) S^B_{-k,-1}\bigr]\Bigr\}.\end{aligned} \tag{43}$$

Next, we will first upper bound the moment generating functions of $L^{\mathrm{I}}_{\mathrm{GLQF}}$, $L^{\mathrm{II}}_{\mathrm{GLQF}}$, and $L^{\mathrm{III}}_{\mathrm{GLQF}}$. For $L^{\mathrm{I}}_{\mathrm{GLQF}}$ and for $\theta \geq 0$ we have

$$\begin{aligned}&\mathbf{E}\bigl[e^{\theta L^{\mathrm{I}}_{\mathrm{GLQF}}}\bigr]\\&\leq \sum_{n \geq 0} \mathbf{E}\Bigl[\exp\Bigl\{\theta \inf_{u \geq 0} \bigl[(1 + u\beta) S^{A^1}_{-n,-1} - u S^{A^2}_{-n,-1}\\&\qquad + (-1 - \beta u) S^B_{-n,-1}\bigr]\Bigr\}\Bigr]\\&\leq \sum_{n \geq 0} \inf_{u \geq 0} \mathbf{E}\Bigl[\exp\Bigl\{\theta\bigl[(1 + u\beta) S^{A^1}_{-n,-1} - u S^{A^2}_{-n,-1}\\&\qquad + (-1 - \beta u) S^B_{-n,-1}\bigr]\Bigr\}\Bigr]\end{aligned}$$

$$\begin{aligned}&\mathbf{P}\bigl[L^1_0 > U \text{ and } \Omega_1\bigr]\\&\quad \leq \mathbf{P}\bigl[\exists n \geq 0,\ \text{s.t. } S^{A^1}_{-n,-1} - S^B_{-n,-1} > U \text{ and } \beta\bigl(S^{A^1}_{-n,-1} - S^B_{-n,-1}\bigr) \geq S^{A^2}_{-n,-1}\bigr]\\&\quad = \mathbf{P}\Biggl[\max_{\{n \geq 0\,:\,\beta(S^{A^1}_{-n,-1} - S^B_{-n,-1}) \geq S^{A^2}_{-n,-1}\}} \bigl(S^{A^1}_{-n,-1} - S^B_{-n,-1}\bigr) > U\Biggr]\end{aligned} \tag{37}$$

$$\mathbf{P}\bigl[L^1_0 > U \text{ and } \Omega_2\bigr] = \mathbf{P}\Biggl[\max_{\{n \geq 0,\ 0 \leq \phi < 1 : \beta(S^{A^1}_{-n,-1} - \phi S^B_{-n,-1}) \leq S^{A^2}_{-n,-1} - (1-\phi) S^B_{-n,-1}\}} \bigl(S^{A^1}_{-n,-1} - \phi S^B_{-n,-1}\bigr) > U\Biggr] \tag{38}$$

$$\begin{aligned}&\mathbf{P}\bigl[L^1_0 > U \text{ and } \Omega_3\bigr]\\&\quad \leq \mathbf{P}\bigl[\exists n \geq 0, 0 \leq k \leq n, 0 \leq \phi < 1,\\&\qquad\quad \text{s.t. } S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1} + S^{A^1}_{-k,-1} - S^B_{-k,-1} > U \text{ and } \beta\bigl(S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1}\bigr)\\&\qquad\quad \leq S^{A^2}_{-n,-k-1} - (1-\phi) S^B_{-n,-k-1} \text{ and } \beta\bigl(S^{A^1}_{-k,-1} - S^B_{-k,-1}\bigr) \geq S^{A^2}_{-k,-1}\bigr]\\&\qquad \leq \mathbf{P}\Biggl[\max_{\substack{n \geq 0, 0 \leq k \leq n, 0 \leq \phi < 1\\ \beta(S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1}) \leq S^{A^2}_{-n,-k-1} - (1-\phi) S^B_{-n,-k-1}\\ \beta(S^{A^1}_{-k,-1} - S^B_{-k,-1}) \geq S^{A^2}_{-k,-1}}} \bigl(S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1} + S^{A^1}_{-k,-1} - S^B_{-k,-1}\bigr) > U\Biggr]\end{aligned} \tag{40}$$

$$
\begin{aligned}
&\leq \sum_{n\geq 0} e^{n(\inf_{u\geq 0}[\Lambda_{A^1}(\theta+\theta u\beta)+\Lambda_{A^2}(-u\theta)+\Lambda_B(-\theta-u\beta\theta)]+\epsilon_1)}\\
&\leq K^{\mathrm{I}}(\theta,\epsilon_1) \qquad \text{if } \Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta)<0. \qquad (44)
\end{aligned}
$$

In the third inequality above we have used the LDP for the arrival and service processes. In the last inequality above, when the exponent is negative (that is, $\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta)<0$ and ϵ_1 is sufficiently small), the infinite geometric series converges to a constant $K^{\mathrm{I}}(\theta,\epsilon_1)$. Also, in the last inequality, we have made the substitution $u := -\theta u$ in the expression in the exponent and used the definition of $\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta)$ [cf. (33)].

Similarly, for $L^{\mathrm{II}}_{\mathrm{GLQF}}$ and for $\theta\geq 0$ we have (45), as shown at the bottom of the page. In the third inequality above, the expression to be maximized over ϕ is linear, thus the maximum is achieved at either $\phi=0$ or $\phi=1$, which implies that we can upper bound it by the sum of the terms for $\phi=0$ and $\phi=1$.

Also, for $L^{\mathrm{III}}_{\mathrm{GLQF}}$ and for $\theta\geq 0$ we have (46), as shown at the bottom of the next page. In the third inequality above we have used the LDP for arrival and service processes, as well as Assumption C. Concerning the maximization over ϕ, we have used the same argument as in (45). In the fifth inequality above, since the exponent is linear in k, the maximum over k is either at $k=0$ or at $k=n$. Thus, we bound the term by the sum of the terms for $k=0$ and $k=n$. Finally, for the last inequality, both series converge to a constant if both their exponents are negative, which requires $\max(\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta),\Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta))<0$.

To summarize (44)–(46), the moment generating functions of $L^{\mathrm{I}}_{\mathrm{GLQF}}$, $L^{\mathrm{II}}_{\mathrm{GLQF}}$, and $L^{\mathrm{III}}_{\mathrm{GLQF}}$ are upper bounded by some constant $K(\theta,\epsilon_1,\epsilon_2,\epsilon_3)$ if $\max(\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta),\ \Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta))<0$, where $\epsilon_1,\epsilon_2,\epsilon_3>0$ are sufficiently small. We can now apply the Markov inequality to obtain [using (37), (38), and (40)]

$$
\begin{aligned}
&\mathbf{P}\big[L_0^1>U\big]\\
&\quad\leq \mathbf{P}\big[L_0^1>U \text{ and Case 1}\big]+\mathbf{P}\big[L_0^1>U \text{ and Case 2}\big]\\
&\qquad+\mathbf{P}\big[L_0^1>U \text{ and Case 3}\big]\\
&\quad\leq \big(\mathbf{E}[e^{\theta\Lambda^{\mathrm{I}}(\theta)}]+\mathbf{E}[e^{\theta\Lambda^{\mathrm{II}}(\theta)}]+\mathbf{E}[e^{\theta\Lambda^{\mathrm{III}}(\theta)}]\big)e^{-\theta U}\\
&\quad\leq 3K(\theta,\epsilon_1,\epsilon_2,\epsilon_3)e^{-\theta U}\\
&\qquad\qquad \text{if } \max(\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta),\Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta))<0.
\end{aligned}
$$

Taking the limit as $U\to\infty$ and minimizing the upper bound with respect to $\theta\geq 0$, in order to obtain the tightest bound, we have

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

The right-hand side of the above is equal to $-\theta^*_{\mathrm{GLQF}}$ by Theorem 8.3. □

IX. Main Results

In this section we gather our main results on the performance of multiclass multiplexers.

A. The GPS Main Results

We first combine Propositions 4.2 and 7.1 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

$$
\lim_{U\to\infty}\frac{1}{U}\log\mathbf{P}[L^1>U]=-\theta^*_{\mathrm{GPS}} \qquad (47)
$$

where θ^*_{GPS} is given by

$$
\theta^*_{\mathrm{GPS}}=\min\left[\inf_{a>0}\frac{1}{a}\Lambda^{\mathrm{I}*}_{\mathrm{GPS}}(a),\ \inf_{a>0}\frac{1}{a}\Lambda^{\mathrm{II}*}_{\mathrm{GPS}}(a)\right] \qquad (48)
$$

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

$$
\Lambda^{\mathrm{I}*}_{\mathrm{GPS}}(a)\triangleq\inf_{\substack{x_1+x_2-x_3=a\\ x_2\leq\phi_2 x_3}}\left[\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)\right] \qquad (49)
$$

$$
\begin{aligned}
L^{\mathrm{II}}_{\mathrm{GLQF}} &\triangleq \max_{\{n\geq 0,\ 0\leq\phi<1:\beta(S^{A^1}_{-n,-1}-\phi S^B_{-n,-1})\leq S^{A^2}_{-n,-1}-(1-\phi)S^B_{-n,-1}\}}\big(S^{A^1}_{-n,-1}-\phi S^B_{-n,-1}\big)\\
L^{\mathrm{III}}_{\mathrm{GLQF}} &\triangleq \max_{\substack{n\geq 0,0\leq k\leq n,0\leq\phi<1\\ \beta(S^{A^1}_{-n,-k-1}-\phi S^B_{-n,-k-1})\leq S^{A^2}_{-n,-k-1}-(1-\phi)S^B_{-n,-k-1}\\ \beta(S^{A^1}_{-k,-1}-S^B_{-k,-1})\geq S^{A^2}_{-k,-1}}}\big(S^{A^1}_{-n,-k-1}-\phi S^B_{-n,-k-1}+S^{A^1}_{-k,-1}-S^B_{-k,-1}\big)
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{E}\big[e^{\theta L^{\mathrm{II}}_{\mathrm{GLQF}}}\big] &\leq \sum_{n\geq 0}\mathbf{E}\left[\exp\left\{\theta\max_{0\leq\phi<1}\inf_{u\geq 0}\left[(1-u\beta)S^{A^1}_{-n,-1}+uS^{A^2}_{-n,-1}+(-\phi+u\beta\phi-u+u\phi)S^B_{-n,-1}\right]\right\}\right]\\
&\leq \sum_{n\geq 0}\inf_{u\geq 0}\mathbf{E}\left[\exp\left\{\theta\max_{0\leq\phi<1}\left[(1-u\beta)S^{A^1}_{-n,-1}+uS^{A^2}_{-n,-1}+(-\phi+u\beta\phi-u+u\phi)S^B_{-n,-1}\right]\right\}\right]\\
&\leq \sum_{n\geq 0}\inf_{u\geq 0}\left(e^{n([\Lambda_{A^1}(\theta-\theta u\beta)+\Lambda_{A^2}(u\theta)+\Lambda_B(-\theta u)]+\epsilon_2')}+e^{n([\Lambda_{A^1}(\theta-\theta u\beta)+\Lambda_{A^2}(u\theta)+\Lambda_B(-\theta+\theta u\beta)]+\epsilon_2'')}\right)\\
&\leq 2\sum_{n\geq 0}e^{n(\inf_{u\geq 0}[\Lambda_{A^1}(\theta-\theta u\beta)+\Lambda_{A^2}(u\theta)+\max(\Lambda_B(-\theta u),\Lambda_B(-\theta+\theta u\beta))]+\epsilon_2)}\leq K^{\mathrm{II}}(\theta,\epsilon_2),\quad \text{if } \Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta)<0 \qquad (45)
\end{aligned}
$$

and

$$\Lambda^{\mathrm{II}*}_{\mathrm{GPS}}(a) \triangleq \inf_{\substack{x_1-\phi_1 x_3=a\\ x_2\geq\phi_2 x_3}} [\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)]. \tag{50}$$

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 [23], [21], and [1]) since under this policy, class 1 customers are oblivious to class 2 customers. We summarize the performance of priority policies in the next corollary. The discussion of Section VI-A 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 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 at an arbitrary time slot satisfies

$$\lim_{U\to\infty}\frac{1}{U}\log \mathbf{P}[L^1>U]=-\theta^*_{P_1} \tag{51}$$

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

$$\theta^*_{P_1}=\inf_{a>0}\frac{1}{a}\Lambda^*_{P_1}(a) \tag{52}$$

and where

$$\Lambda^*_{P_1}(a)\triangleq \inf_{x_1-x_3=a}[\Lambda^*_{A^1}(x_1)+\Lambda^*_B(x_3)]. \tag{53}$$

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

$$\lim_{U\to\infty}\frac{1}{U}\log \mathbf{P}[L^1>U]=-\theta^*_{P_2} \tag{54}$$

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

$$\theta^*_{P_2}=\inf_{a>0}\frac{1}{a}\Lambda^*_{P_2}(a) \tag{55}$$

and where

$$\Lambda^*_{P_2}(a)\triangleq \inf_{\substack{x_1+x_2-x_3=a\\ x_2\leq x_3}}[\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)]. \tag{56}$$

Proof: For policy P_1 apply Theorem 9.1 with $\phi_1=1$. For such ϕ_1, it is easy to verify that $\Lambda^{\mathrm{I}*}_{\mathrm{GPS}}(a)\geq\Lambda^{\mathrm{II}*}_{\mathrm{GPS}}(a)$, for all a. Thus, we define $\Lambda^*_{P_1}(a)$ to be equal to $\Lambda^{\mathrm{II}*}_{\mathrm{GPS}}(a)$ with ϕ_1 set to one.

For policy P_2 apply Theorem 9.1 with $\phi_1=0$. Application of $\phi_1=0$ to $\Lambda^{\mathrm{I}*}_{\mathrm{GPS}}(a)$ yields

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

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

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

The functions $\Lambda^*_{A^2}(x_2)$ and $\Lambda^*_B(x_3)$ are nonnegative, convex, and achieve their minimum value, which is equal to zero, 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 as $x_2\geq x_3$ holds. Also, if $x_3<\mathbf{E}[B_0]$ is the case, we can increase x_3 and reduce the cost, as long as $x_2\geq x_3$ holds. Thus, at optimality $x_2=x_3$ in (58). But, the region characterized by $x_1=a$ and $x_2=x_3$ is included in the region defined by the constraints in the

$$\begin{aligned}
&\mathbf{E}\left[e^{\theta L^{\mathrm{III}}_{\mathrm{GLQF}}}\right]\\
&\leq\sum_{n\geq 0}\sum_{0\leq k\leq n}\mathbf{E}\Bigg[\exp\Bigg\{\theta\max_{0\leq\phi<1}\inf_{u_1\geq 0}\left[(1-u_1\beta)S^{A^1}_{-n,-k-1}+u_1S^{A^2}_{-n,-k-1}+(-\phi+u_1\beta\phi-u_1+u_1\phi)S^{B}_{-n,-k-1}\right]\\
&\qquad\qquad +\theta\inf_{u_2\geq 0}\left[(1+u_2\beta)S^{A^1}_{-k,-1}-u_2S^{A^2}_{-k,-1}+(-1-u_2\beta)S^{B}_{-k,-1}\right]\Bigg\}\Bigg]\\
&\leq\sum_{n\geq 0}\sum_{0\leq k\leq n}\inf_{u_1,u_2\geq 0}\mathbf{E}\Bigg[\exp\Bigg\{\theta\max_{0\leq\phi<1}\left[(1-u_1\beta)S^{A^1}_{-n,-k-1}+u_1S^{A^2}_{-n,-k-1}+(-\phi+u_1\beta\phi-u_1+u_1\phi)S^{B}_{-n,-k-1}\right]\\
&\qquad\qquad +\theta\left[(1+u_2\beta)S^{A^1}_{-k,-1}-u_2S^{A^2}_{-k,-1}+(-1-u_2\beta)S^{B}_{-k,-1}\right]\Bigg\}\Bigg]\\
&\leq\sum_{n\geq 0}\sum_{0\leq k\leq n}\inf_{u_1,u_2\geq 0}\left[e^{(n-k)(\Lambda_{A^1}(\theta-\theta u_1\beta)+\Lambda_{A^2}(u_1\theta)+\Lambda_B(-\theta u_1)+\epsilon'_3)}+e^{(n-k)(\Lambda_{A^1}(\theta-\theta u_1\beta)+\Lambda_{A^2}(u_1\theta)+\Lambda_B(-\theta+\theta u_1\beta)+\epsilon''_3)}\right]\\
&\qquad\qquad\times e^{k(\Lambda_{A^1}(\theta+\theta u_2\beta)+\Lambda_{A^2}(-u_2\theta)+\Lambda_B(-\theta-\theta u_2\beta)+\epsilon'''_3)}\\
&\leq 2\sum_{n\geq 0}\sum_{0\leq k\leq n}e^{(n-k)(\Lambda^{\mathrm{II}}(\theta)+\hat{\epsilon}_3)}e^{k(\Lambda^{\mathrm{I}}(\theta)+\hat{\epsilon}'_3)}\leq 2\sum_{n\geq 0}ne^{n(\Lambda^{\mathrm{II}}(\theta)+\hat{\epsilon}_3)}+2\sum_{n\geq 0}ne^{n(\Lambda^{\mathrm{I}}(\theta)+\hat{\epsilon}'_3)}\\
&\leq K^{\mathrm{III}}(\theta,\epsilon_3),\qquad \text{if }\max(\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta),\Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta))<0
\end{aligned} \tag{46}$$

optimization problem in (57). Hence, for all a, and when $\phi_1 = 0$, $\Lambda^{\mathrm{I}*}_{\mathrm{GPS}}(a) \leq \Lambda^{\mathrm{II}*}_{\mathrm{GPS}}(a)$. Therefore, we define $\Lambda^*_{P_2}(a)$ to be equal to the expression in (57). □

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 8.1. Hence, using duality, we express θ^*_{GPS} as the largest root of a convex function. The result is given in the next theorem, the proof of which is omitted due to space limitations; it can be found in [3].

Theorem 9.3: θ^*_{GPS} is the largest positive root of the equation

$$\begin{aligned}&\Lambda_{\mathrm{GPS}}(\theta)\\ &\quad\triangleq \Lambda_{A^1}(\theta) + \inf_{0\leq u\leq\theta}[\Lambda_{A^2}(\theta - u) + \Lambda_B(-\theta + \phi_2 u)] = 0.\end{aligned} \tag{59}$$

Remark: Equation (59) has a positive, possibly infinite root. To establish that, notice first that $\Lambda_{\mathrm{GPS}}(\theta)$ 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.,

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

Observe now that

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

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

$$\dot{\Lambda}_{\mathrm{GPS}}(0) \leq \dot{\Lambda}_{A^1}(0) + \dot{\Lambda}_{A^2}(0) - \dot{\Lambda}_B(0) < 0$$

where the last inequality follows from the stability condition (9). The convexity of $\Lambda_{\mathrm{GPS}}(\cdot)$ is sufficient to guarantee the existence of a positive, possible infinite root.

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

$$\Lambda_{P1}(\theta) \triangleq \Lambda_{A^1}(\theta) + \Lambda_B(-\theta) = 0. \tag{60}$$

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

$$\Lambda_{P2}(\theta) \triangleq \Lambda_{A^1}(\theta) + \inf_{0\leq u\leq\theta}[\Lambda_{A^2}(\theta - u) + \Lambda_B(-\theta + u)] = 0. \tag{61}$$

We conclude this subsection 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.

B. The GLQF Main Results

Combining Propositions 4.1 and 8.4 we obtain the following main GLQF theorem. An exact characterization of the *most likely ways* that lead to overflow was discussed in Section VI-B.

Theorem 9.5 (GLQF Main): Under the GLQF 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

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

where θ^*_{GLQF} is given by

$$\theta^*_{\mathrm{GLQF}} = \min\left[\inf_{a>0}\frac{1}{a}\Lambda^{\mathrm{I}*}_{\mathrm{GLQF}}(a), \inf_{a>0}\frac{1}{a}\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(a)\right] \tag{63}$$

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

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

and

$$\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(a) \triangleq \inf_{\substack{x_1 - \phi x_3 = a\\ x_2 - (1-\phi)x_3 = \beta a\\ 0\leq\phi<1}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \tag{65}$$

It should be noted that the performance of strict priority policies, which is characterized by Corollary 9.2, can also be obtained as a corollary of the above theorem. We obtain the performance of strict priority to class 2 (P_2) when $\beta = 0$ and the performance of strict priority to class 1 (P_1) when $\beta = \infty$. It is not hard to verify that the result is identical to Corollary 9.2. The above theorem indicates that the calculation of the overflow probabilities involves the solution of a convex optimization problem. In Section VIII, and for the purposes of proving Proposition 8.4, we proved in Theorem 8.3 that the exponent of the overflow probability can also be obtained as the maximum root of a convex function. This may be easier to do in some cases. Here, we restate this latter result, simplifying the expression for $\Lambda_{\mathrm{GLQF}}(\cdot)$.

Theorem 9.6: θ^*_{GLQF} is the largest positive root of the equation

$$\begin{aligned}\Lambda_{\mathrm{GLQF}}(\theta) = \max\Big\{&\Lambda_{A^1}(\theta) + \Lambda_B(-\theta),\\ &\inf_{0\leq u\leq\frac{\theta}{1+\beta}}[\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u)\\ &+ \Lambda_B(-u)]\Big\} = 0.\end{aligned} \tag{66}$$

Proof: Due to Theorem 8.3, it suffices to prove that the expression in (66) is equal to $\max[\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta), \Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta)]$. Recall the definitions of $\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta)$ in (33) and of $\Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta)$ in (34). Recall also the expression in (35) for the objective function of the optimization problem corresponding to $\Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta)$. Now let u^* be the optimal solution of the optimization problem in the definition of $\Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta)$. We distinguish two cases.

Case 1) $u^* \geq \frac{\theta}{1+\beta}$. Then, notice that u^* is also the minimizer of the objective function in the definition of $\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta)$. Thus, due to convexity, the constraint $u \leq 0$ is tight for the problem corresponding to $\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta)$, and

$$\max\left(\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta), \Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta)\right) = \Lambda_{A^1}(\theta) + \Lambda_B(-\theta), \quad \text{if } u^* \geq \frac{\theta}{1+\beta}. \tag{67}$$

But

$$\begin{aligned}
&\inf_{0 \leq u \leq \frac{\theta}{1+\beta}} [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)] \\
&\leq [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)]_{u=\frac{\theta}{1+\beta}} \\
&= \left[\Lambda_{A^1}\left(\frac{\theta}{1+\beta}\right) + \Lambda_{A^2}\left(\frac{\theta}{1+\beta}\right) + \Lambda_B\left(-\frac{\theta}{1+\beta}\right)\right] \\
&= [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta + u\beta)]_{u=\frac{\theta}{1+\beta}} \\
&\leq [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta + u\beta)]_{u=0} \\
&= \Lambda_{A^1}(\theta) + \Lambda_B(-\theta).
\end{aligned}$$

In the second inequality above we have used the assumption $u^* \geq \frac{\theta}{1+\beta}$ and convexity. Therefore, combining it with (67) we obtain

$$\begin{aligned}
&\max\left(\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta), \Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta)\right) \\
&= \max\left\{\Lambda_{A^1}(\theta) + \Lambda_B(-\theta), \inf_{0 \leq u \leq \frac{\theta}{1+\beta}} [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)]\right\} \\
&= \Lambda_{\mathrm{GLQF}}(\theta) \quad \text{if } u^* \geq \frac{\theta}{1+\beta}.
\end{aligned} \tag{68}$$

Case 2) $0 \leq u^* < \frac{\theta}{1+\beta}$. To conclude the proof we need to show that $\max(\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta), \Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta))$ is not $\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta)$ when the optimal solution, of the optimization problem appearing in the definition of $\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta)$, is some $\hat{u} < 0$. Let us, indeed, assume that this optimal solution is some $\hat{u} < 0$. Then, for all $u \in [0, \frac{\theta}{1+\beta})$ (hence for u^*) we have

$$\begin{aligned}
\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta) &= [\Lambda_{A^1}(\theta - \hat{u}\beta) + \Lambda_{A^2}(\hat{u}) + \Lambda_B(-\theta + \hat{u}\beta)] \\
&\leq [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta + u\beta)] \\
&\leq [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)]
\end{aligned}$$

where in the last inequality we have used the fact that $u < \frac{\theta}{1+\beta}$ which implies [see also (35)] $\Lambda_B(-u) \geq \Lambda_B(-\theta + u\beta)$.

Therefore, for $0 \leq u^* \leq \frac{\theta}{1+\beta}$ also, we have

$$\begin{aligned}
&\max\left(\Lambda^{\mathrm{I}}_{\mathrm{GLQF}}(\theta), \Lambda^{\mathrm{II}}_{\mathrm{GLQF}}(\theta)\right) \\
&= \max\{\Lambda_{A^1}(\theta) + \Lambda_B(-\theta), \inf_{0 \leq u \leq \frac{\theta}{1+\beta}} [\Lambda_{A^1}(\theta - u\beta) \\
&\quad + \Lambda_{A^2}(u) + \Lambda_B(-u)]\} = \Lambda_{\mathrm{GLQF}}(\theta). \qquad \square
\end{aligned}$$

The results of this theorem can also be specialized to the case of priority policies, to obtain the characterization of Corollary 9.4.

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

X. A COMPARISON

In this section we compare the overflow probabilities achieved by the GPS and the GLQF policy.

Let π be an arbitrary work-conserving policy used to allocate the capacity of the server to the two queues Q^1 and Q^2, and let Π be the set of all work-conserving policies π. Let L^1 and L^2 denote the queue lengths of Q^1 and Q^2, respectively, at an arbitrary time slot, when the system operates under π. Let us now define θ^π the vector $(\theta^\pi_1, \theta^\pi_2)$ where

$$\theta^\pi_1 = \lim_{U \to \infty} \frac{1}{U} \log \mathbf{P}[L^1 > U]$$

and

$$\theta^\pi_2 = \lim_{U \to \infty} \frac{1}{U} \log \mathbf{P}[L^2 > U]. \tag{69}$$

The GPS policy is a parametric policy with performance depending on the parameter ϕ_1. To make this dependence explicit we will be using the notation GPS(ϕ_1). Also, the GLQF policy is a parametric policy with performance depending on the parameter β. For the same reason we will be using the notation GLQF(β). Special cases of a work-conserving policy π are the GPS(ϕ_1) policy, the GLQF(β) policy, the strict priority to class 1 policy (P_1 policy), and the strict priority to class 2 policy (P_2 policy). Using Theorems 9.1, 9.5, and Corollary 9.2 one can readily obtain the corresponding θ^π for the policies GPS(ϕ_1), GLQF(β), P_1, and P_2.

It is intuitively obvious that

$$\theta^{P_1} = \left(\max_{\pi \in \Pi} \theta^\pi_1, \min_{\pi \in \Pi} \theta^\pi_2\right)$$

and

$$\theta^{P_2} = \left(\min_{\pi \in \Pi} \theta^\pi_1, \max_{\pi \in \Pi} \theta^\pi_2\right).$$

In Fig. 6 we plot $\theta^{\mathrm{GPS}(\phi_1)}$ as ϕ_1 varies in $[0, 1]$ and $\theta^{\mathrm{GLQF}(\beta)}$ as β varies in $[0, \infty)$. For simplicity the calculations were performed with the arrival and service processes being Bernoulli (we say that a process $\{X_i; i \in Z\}$ is Bernoulli with parameter p, denoted by $X \sim \mathrm{Ber}(p)$, when X_i are i.i.d. and $X_i = 1$ with probability p and $X_i = 0$ with probability $1 - p$). Also, for the calculations we used the expressions for θ^*_{GPS} and θ^*_{GLQF} given in Theorems 9.3 and 9.6, respectively, because they were more efficient to perform numerically than the

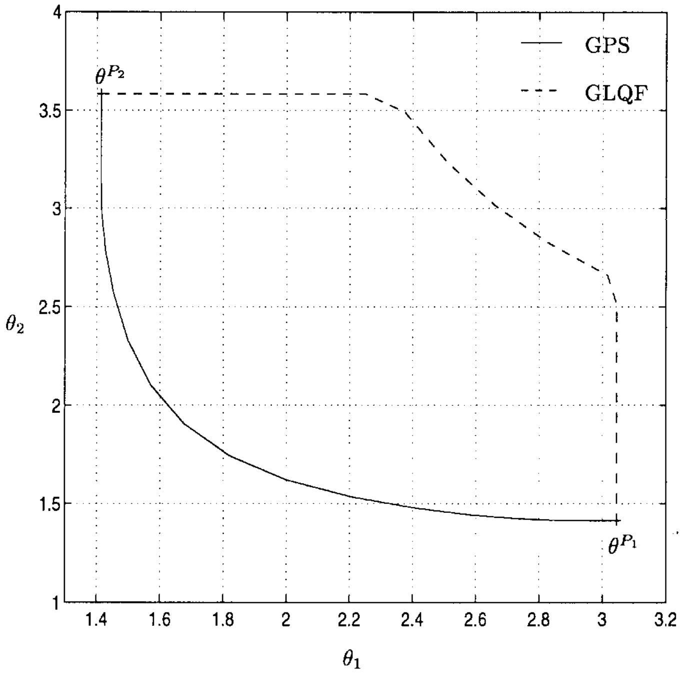


Fig. 6. The performance $\theta^{\mathrm{GPS}(\phi_1)}$ of the GPS(ϕ_1) policy as ϕ_1 varies in $[0,1]$, and the performance $\theta^{\mathrm{GLQF}(\beta)}$ of the GLQF(β) policy as β varies in $[0,\infty)$, when $A^1 \sim \mathrm{Ber}(0.3)$, $A^2 \sim \mathrm{Ber}(0.2)$, and $B \sim \mathrm{Ber}(0.9)$.

equivalent expressions in Theorems 9.1 and 9.5. Note that $\theta^{P_1} = \theta^{\mathrm{GPS}(1)} = \theta^{\mathrm{GLQF}(\infty)}$ and that $\theta^{P_2} = \theta^{\mathrm{GPS}(0)} = \theta^{\mathrm{GLQF}(0)}$.

Fig. 6 indicates that the GLQF curve dominates the GPS curve, i.e., the GLQF policy achieves smaller overflow probabilities than the GPS policy. The question that arises is whether this depends on the particular distributions and parameters chosen in the figure or is a general property. In the sequel we show that the latter is the case, that is, for all arrival and service processes that our analysis holds (processes satisfying Assumptions A, B, and C) the GLQF curve dominates the GPS curve. The intuition behind this result is that the GLQF policy, which adaptively depends on the current queue lengths, allocates capacity to the queue that builds up, thus achieving smaller overflow probabilities than the GPS policy which is static. This suggests that when one has to deal with delay insensitive traffic (i.e., when there are no delay constraints) GLQF is more suitable than GPS.

Let us first formally define the term *the GLQF curve dominates the GPS curve.*

Definition 10.1: We say that *the GLQF curve dominates the GPS curve* when there does not exist a pair of $\phi_1 \in [0,1]$ and $\beta \in [0,\infty)$ satisfying $\theta_1^{\mathrm{GPS}(\phi_1)} > \theta_1^{\mathrm{GLQF}(\beta)}$ and $\theta_2^{\mathrm{GPS}(\phi_1)} > \theta_2^{\mathrm{GLQF}(\beta)}$.

In order to establish that the GLQF curve dominates the GPS curve, we need to prove the three lemmata that follow.

Lemma 10.2: If $\phi_1 \le \phi_1'$ we have

$$\theta_1^{\mathrm{GPS}(\phi_1)} \le \theta_1^{\mathrm{GPS}(\phi_1')} \quad \text{and} \quad \theta_2^{\mathrm{GPS}(\phi_1)} \ge \theta_2^{\mathrm{GPS}(\phi_1')}.$$

Proof: We only prove the first relation. The second can be obtained by a symmetrical argument. We use the result of Theorem 9.3. Note that $\phi_1 \le \phi_1'$ implies $\phi_2' = (1-\phi_1') \le \phi_2 = (1-\phi_1)$. Thus, by Lemma 8.2, for all $u, \theta \ge 0$ we have that $\Lambda_B(-\theta + \phi_2 u) \ge \Lambda_B(-\theta + \phi_2' u)$, which by Theorem 9.3 implies $\Lambda_{\mathrm{GPS}(\phi_1)}(\theta) \ge \Lambda_{\mathrm{GPS}(\phi_1')}(\theta)$ for all θ. Therefore, by convexity, for θ^*_{GPS}, as it is defined in Theorem 9.3, we have $\theta^*_{\mathrm{GPS}(\phi_1)} \le \theta^*_{\mathrm{GPS}(\phi_1')}$. □

A similar property is proven for the GLQF policy.

Lemma 10.3: If $\beta \le \beta'$ we have

$$\theta_1^{\mathrm{GLQF}(\beta)} \le \theta_1^{\mathrm{GLQF}(\beta')} \quad \text{and} \quad \theta_2^{\mathrm{GLQF}(\beta)} \ge \theta_2^{\mathrm{GLQF}(\beta')}.$$

Proof: Again we only prove the first relation. The second can be obtained by a symmetrical argument. We use the optimal control formulation of Section V-B. We argued there that optimal trajectories have the form of Fig. 5(d) and (e), with cost $\inf_a \frac{1}{a}\Lambda^{\mathrm{I}*}_{\mathrm{GLQF}}(a)$ and $\inf_a \frac{1}{a}\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}}(a)$, respectively. Let us fix β and consider how the cost is affected by using the policy with $\beta' = \beta + \epsilon$, for small $\epsilon > 0$.

Consider first trajectories of the form in Fig. 5(e). Note that we can rewrite $\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}(\beta)}(a)$ as

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

We shall show $\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}(\beta')}(a) \ge \Lambda^{\mathrm{II}*}_{\mathrm{GLQF}(\beta)}(a)$ for all $a \ge 0$. Assume the contrary. Consider the optimal solution of the problem corresponding to β' which satisfies the feasibility constraints

$$\begin{gathered} x_1' - \phi' x_3' = a \\ x_1' + x_2' - x_3' = \beta'(1+a) \\ 0 \le \phi' < 1. \end{gathered}$$

We distinguish two cases: $\phi' > 0$ and $\phi' = 0$. We provide an argument only for the first case. The second case can be handled similarly. Since $\beta, a \ge 0$, at least one of the following holds: $x_1' > \mathbf{E}[A_0^1]$ or $x_2' > \mathbf{E}[A_0^2]$ or $x_3' < \mathbf{E}[B_0]$. Depending on which one is the case, we can decrease x_1', or x_2', or increase x_3', respectively, reducing the cost, until $x_1' + x_2' - x_3' = \beta(1+a)$. Thus, we have constructed a feasible solution of the problem corresponding to β with smaller cost than $\Lambda^{\mathrm{II}*}_{\mathrm{GLQF}(\beta')}(a)$. This contradicts our initial assumption. We conclude that by increasing β to β' we also increase the optimal cost of trajectories having the form in Fig. 5(e).

If now an optimal trajectory has the form in Fig. 5(d), then it will still be the optimal, by convexity, when β is increased to β'. Thus, in this case, the optimal cost does not change.

We summarize by considering how the cost is affected as β is increased from zero to ∞. At $\beta = 0$, possible optimal trajectories have the form of Fig. 5(e). There is a threshold value $\bar{\beta}$ such that for all $\beta \le \bar{\beta}$ optimal trajectories have the form of Fig. 5(e) with values increasing as β increases from zero to $\bar{\beta}$. For all $\beta > \bar{\beta}$, optimal trajectories have the form of Fig. 5(d) with slope $\bar{\beta}$ and do not change as β increases from $\bar{\beta}$ to ∞. □

We next prove a sufficient condition for the GLQF curve dominating the GPS curve.

Lemma 10.4: If for all $\beta \in [0, \infty)$ there exists $\phi_1 \in [0, 1)$ such that

$$\theta_1^{\mathrm{GPS}(\phi_1)} \le \theta_1^{\mathrm{GLQF}(\beta)} \quad \text{and} \quad \theta_2^{\mathrm{GPS}(\phi_1)} \le \theta_2^{\mathrm{GLQF}(\beta)}$$

then the GLQF curve dominates the GPS curve.

Proof: We use contradiction. Assume that the condition given in the statement holds, but the GLQF curve does not dominate the GPS curve. Then, by definition, there exist β' and ϕ_1' such that

$$\theta_1^{\mathrm{GPS}(\phi_1')} > \theta_1^{\mathrm{GLQF}(\beta')} \quad \text{and} \quad \theta_2^{\mathrm{GPS}(\phi_1')} > \theta_2^{\mathrm{GLQF}(\beta')}.$$

By Lemma 10.2 all points with $\phi_1 < \phi_1'$ have $\theta_2^{\mathrm{GPS}(\phi_1)} \ge \theta_2^{\mathrm{GPS}(\phi_1')} > \theta_2^{\mathrm{GLQF}(\beta')}$. Also, by the same lemma, all points with $\phi_1 \ge \phi_1'$ have $\theta_1^{\mathrm{GPS}(\phi_1)} \ge \theta_1^{\mathrm{GPS}(\phi_1')} > \theta_1^{\mathrm{GLQF}(\beta')}$. This contradicts our initial assumption. □

We now have all the necessary tools to prove that the GLQF curve dominates the GPS curve.

Theorem 10.5: Assuming that the arrival and service processes satisfy Assumptions A, C, and B, the GLQF curve dominates the GPS curve.

Proof: Fix an arbitrary β. We will prove that there exists ϕ_1 satisfying the condition of Lemma 10.4. It suffices to prove that for both queues and such ϕ_1, overflow with the GLQF(β) policy implies overflow with the GPS(ϕ_1) policy. Then, the overflow probability of GLQF(β) is a lower bound on the corresponding probability of GPS(ϕ_1), i.e., it holds

$$\mathbf{P}[L^j_{\mathrm{GLQF}(\beta)} > U] \le \mathbf{P}[L^j_{\mathrm{GPS}(\phi_1)} > U], \qquad j = 1, 2$$

which implies

$$\theta_1^{\mathrm{GPS}(\phi_1)} \le \theta_1^{\mathrm{GLQF}(\beta)} \quad \text{and} \quad \theta_2^{\mathrm{GPS}(\phi_1)} \le \theta_2^{\mathrm{GLQF}(\beta)}.$$

Since we have established that in both the GPS and the GLQF case the overflow probability is equal to the probability of overflowing according to one out of two scenarios, it suffices to establish the above only for these scenarios. In particular, we distinguish the following cases depending on the possible modes of overflow for GLQF(β), which are described in Section VI-B:

Case 1) Mode 1 for overflow of Q^1 and mode 1 for overflow of Q^2;

Case 2) Mode 1 for overflow of Q^1 and mode 2 for overflow of Q^2;

Case 3) Mode 2 for overflow of Q^1 and mode 1 for overflow of Q^2;

Case 4) Mode 2 for overflow of Q^1 and mode 2 for overflow of Q^2.

In Cases 1 and 2, we have

$$x_1 - x_3 = a$$
$$x_2 \le \beta a$$

where $x_j, j = 1, 2, 3, a$ solve the optimization problem corresponding to the overflow of Q^1 in mode one. Then, since $x_1 - \phi_1 x_3 \ge x_1 - x_3 = a \ \forall \phi_1$, it is clear that for all ϕ_1 the GPS policy will overflow Q^1. If we are in Case 1, then for all ϕ_1 the GPS policy will overflow Q^2. If we are in Case 2, we have

$$y_2 - \phi y_3 = a$$
$$y_1 - (1 - \phi) y_3 = a/\beta$$
$$0 \le \phi < 1$$

where $y_j, j = 1, 2, 3, a, \phi$ solve the optimization problem corresponding to the overflow of Q^2 in mode two. Then, the GPS policy with $\phi_1 \ge 1 - \phi$ will overflow Q^2.

Consider now Cases 3 and 4. We have

$$x_1 - \phi x_3 = a$$
$$x_2 - (1 - \phi) x_3 = a\beta$$
$$0 \le \phi < 1$$

where $x_j, j = 1, 2, 3, a, \phi$ solve the optimization problem corresponding to the overflow of Q^1 in mode two. Then the GPS policy with $\phi_1 \le \phi$ will overflow Q^2. In Case 3, for reasons explained in the previous paragraph, the GPS policy will overflow Q^2 for all ϕ_1. If, finally, we are in Case 4, we have

$$y_2 - (1 - \phi') y_3 = a'$$
$$y_1 - \phi' y_3 = a'/\beta$$
$$0 \le \phi' < 1$$

where $y_j, j = 1, 2, 3, a', \phi'$ solve the optimization problem corresponding to the overflow of Q^2 in mode two. Then the GPS policy with $\phi_1 \ge \phi'$ will overflow Q^2. To show that there is at least one ϕ_1 that overflows both queues we need to show $\phi = \phi'$. To see that, notice that (by making the substitution $a' := \beta a'$)

$$\inf_{a'} \frac{1}{a'} \inf_{\substack{y_2-(1-\phi')y_3=a' \\ y_1-\phi' y_3=a'/\beta \\ 0\le\phi'<1}} [\Lambda^*_{A^1}(y_1) + \Lambda^*_{A^2}(y_2) + \Lambda^*_B(y_3)]$$
$$= \frac{1}{\beta} \inf_a \frac{1}{a} \inf_{\substack{y_1-\phi' y_3=a' \\ y_2-(1-\phi')y_3=\beta a' \\ 0\le\phi'<1}} [\Lambda^*_{A^1}(y_1) + \Lambda^*_{A^2}(y_2) + \Lambda^*_B(y_3)].$$

The right-hand side is exactly the problem corresponding to the overflow of Q^1 in mode two. □

XI. Conclusion

In this paper we considered a multiclass multiplexer with segregated buffers for each service class. Under the GPS and the GLQF 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 policies 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 detailed 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.

References

[1] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis, "On the large deviations behavior of acyclic networks of G/G/1 queues," Lab. Inf. Decision Systems, Massachusetts Inst. Technol., Tech. Rep. LIDS-P-2278, Dec. 1994.

[2] ——, "On the large deviations behavior of acyclic single class networks and multiclass queues," Talk at the RSS Workshop in Stochastic Networks, Edinburgh, U.K., 1995.

[3] ——, "Large deviations analysis of the generalized processor sharing policy," Tech. Rep. MNS-97-108, Dept. Manufacturing Eng., Boston Univ., June 1996, to be published.

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

[5] C. S. Chang, "Sample path large deviations and tree networks," *Queueing Syst.*, vol. 20, pp. 7–36, 1995.

[6] H. Cramér, "Sûr un nouveau théoréme-limite de la théorie des probabilités," in *Actualités Scientifiques et Industrielles*, no. 736, in *Colloque consacré à la théorie des probabilités*, Hermann, Paris, 1938, pp. 5–23.

[7] 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.

[8] C. S. Chang and T. Zajic, "Effective bandwidths of departure process from queues with time varying capacities," in *Proc. IEEE Infocom*, Boston, MA, Apr. 1995, vol. 3, pp. 1001–1009.

[9] A. Demers, S. Keshav, and S. Shenker, "Analysis and simulation of a fair queueing algorithm," *J. Internetworking: Research and Experience*, vol. 1, pp. 3–26, 1990.

[10] G. de Veciana, C. Courcoubetis, and J. Walrand, "Decoupling bandwidths for networks: A decomposition approach to resource management," Electronics Research Lab., Univ. California, Berkeley, Memorandum, 1993.

[11] G. de Veciana and G. Kesidis, "Bandwidth allocation for multiple qualities of service using generalized processor sharing," *IEEE Trans. Inform. Theory*, vol. 42, 1995.

[12] A. Dembo and T. Zajic, "Large deviations: From empirical mean and measure to partial sums processes," to be published.

[13] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*. Jones and Bartlett, 1993.

[14] R. Ellis, "Large deviations for a general class of random vectors," *Ann. Probability*, vol. 12, pp. 1–12, 1984.

[15] A. I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, vol. 1, pp. 329–343, 1993.

[16] A. I. Elwalid and D. Mitra, "Analysis, approximations and admission control of a multiple-service multiplexing system with priorities," to be published.

[17] J. Gärtner, "On large deviations from the invariant measure," *Theory Appl. Prob.*, vol. 22, pp. 24–39, 1977.

[18] A. Ganesh and V. Anantharam, "The stationary tail probability of an exponential server tandem fed by renewal arrivals," to be published.

[19] H. R. Gail, G. Grover, R. Guérin, S. L. Hantler, Z. Rosberg, and M. Sidi, "Buffer size requirements under longest queue first," *Performance Eval.*, vol. 18, pp. 133–140, 1993.

[20] R. J. Gibbens and P. J. Hunt, "Effective bandwidths for the multi-type UAS channel," *Queueing Syst.*, vol. 9, pp. 17–28, 1991.

[21] P. W. Glynn and W. Whitt, "Logarithmic asymptotics for steady-state tail probabilities in a single-server queue," *J. Appl. Prob.*, vol. 31A, pp. 131–156, 1994.

[22] J. Y. Hui, "Resource allocation for broadband networks," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 1598–1608, 1988.

[23] F. P. Kelly, "Effective bandwidths at multi-class queues," *Queueing Syst.*, vol. 9, pp. 5–16, 1991.

[24] G. Kesidis, J. Walrand, and C. S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, pp. 424–428, 1993.

[25] N. O'Connell, "Large deviations in queueing networks," preprint, 1995.

[26] ——, "Queue lengths and departures at single-server resources," Talk at the RSS Workshop in Stochastic Networks, Edinburgh, U.K., 1995.

[27] I. Ch. Paschalidis, "Large deviations in high speed communication networks," Ph.D. dissertation, Massachusetts Inst. Technol., May 1996.

[28] ——, "Quality of service provision in multimedia communication networks," Dept. Manufacturing Eng., Boston Univ., Tech. Rep. MNS-97-101, Jan. 1997, to be published.

[29] 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*, vol. 1, pp. 344–357, 1993.

[30] ——, "A generalized processor sharing approach to flow control in integrated services networks: The multiple node case," *IEEE/ACM Trans. Networking*, vol. 2, pp. 137–150, 1994.

[31] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis*. New York: Chapman and Hall, 1995.

[32] D. Tse, R. G. Gallager, and J. N. Tsitsiklis, "Statistical multiplexing of multiple time-scale Markov streams," *IEEE J. Select. Areas Commun.*, vol. 13, 1995.

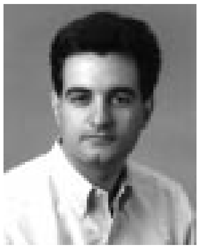
[33] A. Weiss, "An introduction to large deviations for communication networks," *IEEE J. Select. Areas in Commun.*, vol. 13, pp. 938–952, 1995.

[34] Z.-L. Zhang, "Large deviations and the generalized processor sharing scheduling: Upper and lower bounds—Part I: Two-queue systems," Computer Sci. Dept., Univ. Massachusetts, Tech. Rep., Amherst, 1995.

Dimitris Bertsimas received the B.S. degree in electrical engineering and computer science at the National Technical University of Athens, Greece, in 1985, the M.S. degree in operations research from Massachusetts Institute of Technology (M.I.T.), Cambridge, MA, in 1987, and the Ph.D. degree in applied mathematics and operations research from M.I.T. in 1988.

Since 1988, he has been with M.I.T.'s Sloan School of Management. His research interests include discrete optimization, stochastic and dynamic optimization, analysis and control of stochastic systems, and applications in manufacturing systems, finance, and transportation. He recently coauthored a graduate-level textbook, *Introduction to Linear Optimization* (Belmont, MA: Athena Scientific, 1997).

Dr. Bertsimas' awards include the Erlang Prize (1996), awarded by INFORMS for the best young applied probabilist below 35 years of age, the SIAM Prize in optimization (1996), awarded every three years for the best paper in optimization, the Presidential Young Investigator Award (1991–1996), awarded by the National Science Foundation, the Nicholson Prize (1988), awarded by ORSA for the best student paper, and the Transportation System Prize (1989), awarded by ORSA for the best doctoral dissertation in the field of transportation.



Ioannis Ch. Paschalidis (S'91–M'96) was born in Athens, Greece, in 1968. He received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Greece, in 1991, and the S.M. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (M.I.T.), Cambridge, MA, in 1993 and 1996, respectively.

During the summer of 1996 he was a Postdoctoral Associate at the Laboratory for Information and Decision Systems, M.I.T., and since September 1996 he has been with Boston University, where he is an Assistant Professor of Manufacturing Engineering. His research interests include the analysis and control of stochastic systems with main applications in manufacturing systems and communication networks.

Dr. Paschalidis has received the second prize in the 1997 George E. Nicholson paper competition by INFORMS and has been elected a full member of Sigma Xi. He is also a member of INFORMS.

John N. Tsitsiklis (S'81–M'83–SM'97) was born in Thessaloniki, Greece, in 1958. He received the B.S. degree in mathematics in 1980, and the B.S., M.S., and Ph.D. degrees in electrical engineering in 1980, 1981, and 1984, respectively, from the Massachusetts Institute of Technology (M.I.T.), Cambridge, MA.

During 1983–1984 he was an Acting Assistant Professor of Electrical Engineering at Stanford University, CA. Since 1984 he has been with M.I.T., where he is currently a Professor of Electrical Engineering. His research interests include the fields of systems, optimization, control, and operations research. He has written more than 70 journal papers in these areas and is a coauthor of *Parallel and Distributed Computation: Numerical Methods* (1989), *Neuro-Dynamic Programming* (1996), and *Introduction to Linear Optimization* (1997).

Dr. Tsitsiklis has been a recipient of an IBM Faculty Development Award (1983), an NSF Presidential Young Investigator Award (1986), an Outstanding Paper Award by the IEEE Control Systems Society, the M.I.T. Edgerton Faculty Achievement Award (1989), the Bodossakis Foundation Prize (1995), and the INFORMS/CSTS Prize (1997). He was a plenary speaker at the 1992 IEEE Conference on Decision and Control. He is a member of INFORMS, an Associate Editor of *Applied Mathematics Letters*, and has been an Associate Editor of the IEEE Transactions on Automatic Control and *Automatica*.