Asymptotic Buffer Overflow Probabilities in Multiclass Multiplexers, Part II: The GLQF Policy [1]

Dimitris Bertsimas
dbertsim@aris.mit.edu

Ioannis Ch. Paschalidis
yannis@mit.edu

John N. Tsitsiklis
jnt@mit.edu

Laboratory for Information and Decision Systems
and
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139

June 1996

LIDS Report: LIDS-P-2343

[1] A preliminary version of these results was reported in [BPT95]. The results in this paper are included in [Pas96].
Research supported by a Presidential Young Investigator award DDM-9158118 with matching funds from Draper Laboratory, and by the ARO under grant DAAL-03-92-G-0115.

Abstract

In this paper and its companion [BPT96] we consider a multiclass multiplexer, with segregated buffers for each type of traffic, and under specific scheduling policies for sharing bandwidth we seek the asymptotic (as the buffer size goes to infinity) tail of the buffer overflow probability for each buffer. We assume correlated arrival and service processes that are usually used in modeling bursty traffic. Here we consider the *generalized longest queue first policy (GLQF)* and in [BPT96] the *generalized processor sharing policy (GPS)*. In the standard *large deviations* methodology we provide a lower and a matching (up to first degree in the exponent) upper bound on the buffer overflow probabilities. We relate the lower bound derivation to a *deterministic optimal control problem*, which we explicitly solve. Optimal state trajectories of the control problem correspond to typical congestion scenarios. We explicitly and in detail characterize the *most likely* modes of overflow. We find that the GLQF policy outperforms the GPS policy with respect to loss probabilities characteristics. Our results have important implications in traffic management of high-speed networks and can be used as a basis for an admission control mechanism which guarantees different loss probability for each type of traffic.

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

1 Introduction

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

Congestion causes packet losses, due to buffer overflows, and excessive delays, phenomena that greatly contribute to the degradation of the *quality of service (qos)* that the network delivers to its users. Since voice and video are very sensitive to such phenomena the network should have the ability to guarantee certain qos parameters to the user. We quantify qos by the probabilities of excessive delay and buffer overflow. It is desirable to operate the network in a regime where packet loss probabilities are very small, e.g., in the order of 10^{-9}. Moreover, large delays should also have a correspondingly small probability. An essential step for preventing congestion, through a variety of control mechanisms (buffer dimensioning, admission control, resource allocation) is to determine how it occurs and to estimate the probabilities of congestion phenomena, i.e., buffer overflow and delay exceedance probabilities. The problem is particularly difficult since it essentially requires finding the distributions of waiting times and queue lengths in a multiclass network of G/G/1 queues with correlated arrival processes (since it is needed to model bursty traffic) and non-exponentially distributed service times. In this light, it is natural to focus on the *large deviations regime* and obtain asymptotic expressions for the tails of congestion probabilities.

In this paper and its companion [BPT96] we focus at a simplified version of the problem which nevertheless keeps the most salient features, that is, it is multiclass and has correlated arrival and service processes. In particular, we consider a multiclass multiplexer (one node), with segregated buffers for each type of traffic, and under specific scheduling policies for sharing bandwidth, we seek the asymptotic (as the buffer size goes to infinity) tail of the buffer overflow probability for each buffer. In other words, we estimate the loss probability for each type of traffic. In this paper we consider the *generalized longest queue first policy (GLQF)* and in [BPT96] the *generalized processor sharing policy (GPS)* (introduced in [DKS90] and further explored in [PG93, PG94]). 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. Consider the case of two buffers (types of traffic). According to the GLQF policy there is a threshold level, β, and the server allocates all of its capacity to the first buffer, if the ratio of the queue length in the second buffer versus the queue length in

the first buffer is below the threshold, otherwise it allocates all of its capacity to the second buffer. For $\beta = 1$ we have the LQF policy. The LQF policy can be viewed as an attempt to reduce the variance of delay between different types of traffic.

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 two *most likely* ways (modes of overflow) and we explicitly and in detail characterize these modes. Our line of development is very similar to [BPT96]. We address the case of multiplexing two different traffic streams; for the general case of N streams our lower bound approach (which also determines the modes of overflow) can be easily extended. Proving an upper bound is still an open problem. It should however be noted that there is an exponential explosion of the number of possible overflow modes (there are 2^{N-1} modes). Our results have implications for the traffic management of high-speed networks. They can be used as a basis for an admission control which guarantees desirable loss probability, and allows us to deal with different requirements for each type of traffic. We compare the loss probabilities characteristics of the GLQF and the GPS policy and find that the first outperforms the second. However, this may be happening at the expense of greater delay. Though, since delay is due to long queues, it is intuitive that the GLQF policy tries to balance (with a β "bias") the delay of the two traffic streams. In any case, if only loss probability guarantees are needed, our results clearly suggest the use of the GLQF policy instead of the GPS.

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

Large deviations techniques have been used, recently, in a variety of problems in communications. A nice survey can be found in [Wei95]. The problem of estimating tail probabilities of rare events in a single class queue has received extensive attention in the literature [Hui88, GH91, Kel91, KWC93, GW94, EM93, TGT95]. The extension of these ideas to single class networks, although much harder, has been treated in various versions and degree of rigor in [BPT94, GA94, Cha95, O'C95, dVCW93].

Closer to the subject of this paper, [GGG+93] suggests the use of the LQF policy in high speed networks and uses a deterministic model (only the rate of each incoming stream is known) to calculate buffer sizes that guarantee no loss with probability 1. In [SW95] the authors consider the LQF policy in a system with two buffers and address the question of how one queue builds up when the other is large. They consider the M/M/1 version of the system (i.e., Poisson arrivals and exponential service times).

Our work considers the generalization of LQF, the GLQF policy, and obtains the tails of the buffer overflow probabilities for a system with correlated arrivals and stochastic capacity. Stochastic capacity makes it possible to treat more complicated service disciplines. Consider for example the case where we have a deterministic server and three types of traffic with dedicated buffers. We give priority to the first stream and use the GLQF policy for the remaining streams. These two remaining streams face a server with stochastic capacity, a model of which can be obtained using the model for the arrival process of the first stream. Moreover, we provide an *optimal control formulation* of the problem. In particular, the exponent of the overflow probability is the optimal value of a control problem, which we explicitly solve. This formulation, as it will be apparent later, motivates the selection of two overflow scenarios whose probability constitutes the lower bound, a selection which is sort of arbitrary in most of the existing literature. 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. The technique for proving the upper bound is different from the corresponding proof in [BPT96] and does not use explicitly the optimal control formulation. The optimal control formulation is general enough to include any scheduling policy. The only thing that changes with the policy is the system dynamics. Optimal control formulations are also used in [SW95] for large deviations results of jump Markov processes.

Regarding the structure of this paper, we begin in Section 2 with a brief review of large deviations results that we use in this paper. We also state a set of assumptions that arrival and service processes need to conform to. In Section 3 we formally define the multiclass model that we consider and in Section 4 we formally define the GLQF policy and the probabilities of which we seek the asymptotic tails. Moreover, in the latter section, we provide an orientation of the methodology that we follow in proving our results. In Section 5 we prove a lower bound on the overflow probability and in Section 6 we introduce the optimal control formulation and solve the control problem. In Section 7 we summarize the most likely modes of overflow obtained from the solution of the control problem and Section 8 we prove the matching upper bound. We gather our main results in Section 9. In Section 10 we compare the GPS and GLQF policy and we conclude in Section 11.

2 Preliminaries

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

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

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

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

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

Assumption A

1. *The limit*

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

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

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

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

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

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

Upper Bound: *For every closed set F*

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

Lower Bound: *For every open set G*

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

where

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

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

It is important to note that $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$ are convex duals (Legendre transforms of each other). Namely, along with (5), it also holds

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

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

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

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

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

Note that the random variable $S_n = \sum_{i=1}^n X_i$ corresponds to the terminal value (at $t = 1$) of the process $S_n(t)$, $t \in [0,1]$. In a key paper [DZ93a], under certain mild mixing conditions on the stationary sequence $\{X_i;\ i \geq 1\}$, the authors establish an LDP for the process $S_n(\cdot)$ in $D[0,1]$ (the space of right continuous functions with left limits).

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

Assumption B

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

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

A detailed discussion of this Assumption, and the technical conditions under which it is satisfied is given by Dembo and Zajic in [DZ93a]. In the simpler case when dependencies are not present (i.e., $S_i = \sum_{j=1}^i X_j$, where X_i's are iid), Assumption B is a consequence of Mogulskii's theorem (see [DZ93b]). Intuitively, Assumption B deals with the probability of sample paths that are constrained to be within a tube around a "polygonal" path made up

with linear segments of slopes $a_0, \ldots, a_{m-1}$. In [DZ93a] it is proved that this assumption is satisfied by processes that are commonly used in modeling the input traffic to communication networks, that is, renewal processes, Markov modulated processes and correlated stationary processes with mild mixing conditions.

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

Assumption C

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

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

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

On a notational remark, in the rest of the paper we will be denoting by $S_{i,j}^X \triangleq \sum_{k=i}^{j} X_k$, $i \leq j$, the partial sums of the random sequence $\{X_i;\ i \in \mathbb{Z}\}$. We will be also denoting by $\Lambda_X(\cdot)$ and $\Lambda_X^*(\cdot)$ the limiting log-moment generating function and the large deviations rate function (see eqs. (2) and (5) for definitions), respectively, of the process X.

3 A Multiclass Model

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

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

We denote by L_i^1 and L_i^2, the queue lengths at time i (without counting arrivals at time i) in queues Q^1 and Q^2, respectively. We assume that the server allocates its capacity

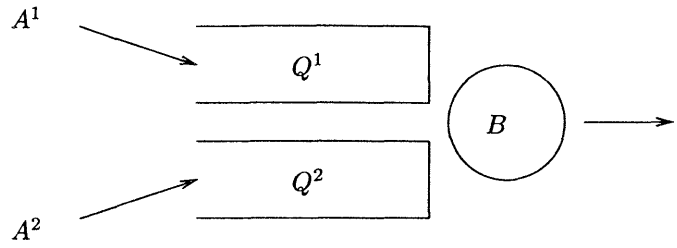

Figure 1: A multiclass model.

between queues Q^1 and Q^2 according to a work-conserving policy (i.e., the server never stays idle when there is work in the system). We also assume that the queue length processes $\{L_i^j, j = 1, 2, i \in \mathbb{Z}\}$ are stationary (under a work-conserving policy, the system reaches steady-state due to the stability condition (9) by assuming ergodicity for the arrival and service processes).

To simplify the analysis and avoid integrality issues we assume a "fluid" model, meaning that we will be treating A_i^1, A_i^2 and B_i as real numbers (the amount of fluid entering or being served). This will not change the results in the large deviations regime.

For stability purposes we assume that for all i

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

We further assume that the arrival and service processes satisfy a LDP (Assumption A), as well as Assumptions B and C. As we have noted in Section 2, these assumptions are satisfied by processes that are commonly used to model bursty traffic in communication networks, e.g., renewal processes, Markov-modulated processes and more generally stationary processes with mild mixing conditions.

4 The GLQF policy

In this section we introduce the *generalized longest queue first policy (GLQF)*.

Figure 2 depicts the operation of the GLQF policy in the L^1-L^2 space. Fix the parameter of the policy $\beta \geq 0$. There is a threshold line, of slope β, which divides the positive orthant of the $L^1 - L^2$ space in two regions. The GLQF policy serves Type 2 customers above the

threshold line and Type 1 below it. The value $\beta = 1$ corresponds to the longest queue first (LQF) policy. More formally, we define the GLQF policy to be the work-conserving policy that at each time slot i serves Type 1 customers when

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

It serves Type 2 customers when

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

When

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

or when

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

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

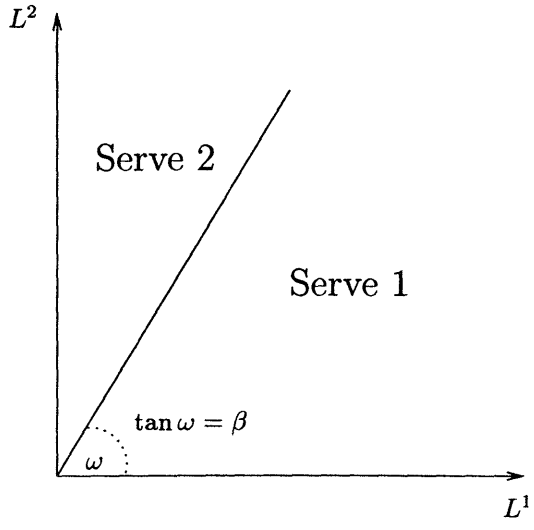

Figure 2: The operation of the GLQF policy.

As in Section 3, we assume that the queue length processes $\{L_i^j, j = 1, 2, i \in \mathbb{Z}\}$ are stationary. We are interested in estimating the overflow probability $\mathbf{P}[L_i^1 > U]$ for large values of U, at an arbitrary time slot i in steady-state. Having determined this, the overflow probability of the second queue can be obtained by a symmetrical argument.

We will prove that the overflow probability satisfies

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

asymptotically, as $U \to \infty$. Our methodology is similar to the one we used in analyzing the GPS policy [BPT96]. To this end, we will develop a lower bound on the overflow probability, along with a matching upper bound. Consider all scenarios (paths) that lead to an overflow. We will show that the probability of each such scenario ω asymptotically behaves as $e^{-U\theta(\omega)}$, for some function $\theta(\omega)$. For every ω, this probability is a lower bound on $\mathbf{P}[L_i^1 > U]$. We select the tightest lower bound by performing the minimization $\theta^*_{GLQF} = \min_\omega \theta(\omega)$. This is a deterministic optimal control problem, which we will solve. Optimal trajectories (paths) of the control problem correspond to *most likely* overflow scenarios. We show that these must be of one out of two possible types. In other words, with high probability, overflow occurs in one out of two possible modes. For the upper bound, we will consider the probability of all sample paths that lead to overflow and show that it is, asymptotically, no more that $e^{-U\theta^*_{GLQF}}$.

5 A Lower Bound

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

Proposition 5.1 ***(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*

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

where θ^*_{GLQF} is given by

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

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

$$\Lambda^{I*}_{GLQF}(a) \stackrel{\Delta}{=} \inf_{\substack{x_1-x_3=a \\ x_2\leq\beta(x_1-x_3)}} [\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)], \tag{13}$$

and

$$\Lambda^{II*}_{GLQF}(a) \stackrel{\Delta}{=} \inf_{\substack{x_1-\phi x_3=a \\ x_2-(1-\phi)x_3=\beta a \\ 0\leq\phi<1}} [\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)]. \tag{14}$$

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

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

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

$$\begin{array}{lll} \underline{\text{Scenario 1}}: & x_1 - x_3 = a & \underline{\text{Scenario 2}}: \quad x_1 - \phi x_3 = a \\ & x_2 \leq \beta(x_1 - x_3) & x_2 - (1-\phi)x_3 = \beta a \\ & & 0 \leq \phi < 1 \end{array} \tag{15}$$

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

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

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

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

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

$$\begin{aligned}
\mathbf{P}[L^1_0 \ge n(a-\epsilon_1')] \ge & \sup_{\substack{x_1 - x_3 = a \\ x_2 \le \beta(x_1 - x_3)}} \mathbf{P}[\,|S^{A^1}_{-n,-i-1} - (n-i)x_1| \le \epsilon_1 n,\ i = 0,1,\ldots,n-1] \\
& \times \mathbf{P}[\,|S^{A^2}_{-n,-i-1} - (n-i)x_2| \le \epsilon_2 n,\ i = 0,1,\ldots,n-1] \\
& \times \mathbf{P}[\,|S^{B}_{-n,-i-1} - (n-i)x_3| \le \epsilon_3 n,\ i = 0,1,\ldots,n-1] \\
\ge & \exp\left\{-n\left(\inf_{\substack{x_1 - x_3 = a \\ x_2 \le \beta(x_1 - x_3)}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)] + \epsilon\right)\right\} \\
= & \exp\{-n(\Lambda^{I*}_{GLQF}(a) + \epsilon)\}, \qquad (16)
\end{aligned}$$

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

Similarly, calculating the probability of Scenario 2, we have

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

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

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

$$\frac{1}{n} \log \mathbf{P}[L_0^1 \geq n(a-\epsilon)] \geq -(\min(\Lambda_{GLQF}^{I*}(a), \Lambda_{GLQF}^{II*}(a)) + \epsilon'). \tag{18}$$

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 $U > U_0$

$$\frac{1}{U} \log \mathbf{P}[L^1 > U] = \frac{1}{n(a-\epsilon)} \log \mathbf{P}[L_0^1 \geq n(a-\epsilon)] \geq -\frac{1}{a-\epsilon}(\min(\Lambda_{GLQF}^{I*}(a), \Lambda_{GLQF}^{II*}(a)) + \epsilon'),$$

which implies

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

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

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

■

6 The optimal control problem

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

The scaling of time and fluid levels is done in exactly the same manner, as in [BPT96], therefore the resulting control problem is identical to (GPS-OVERFLOW) with the exception of the system dynamics that are different in the case of the GLQF policy. In particular, we distinguish three regions depending on the state as follows

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

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

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

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

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

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

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

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

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

This problem exhibits both the properties of constant control trajectories within each region of system dynamics, and time-homogeneity. We omit the proofs since they are similar to the GPS case. Using these properties we can make the reductions appearing in Figure 3(a), (b) and (c), starting from an arbitrary trajectory with piecewise constant controls. We conclude that optimal state trajectories can be reduced to having one of the forms depicted in Figure 3(d), (e) and (f).

The optimal trajectory of the form shown in Figure 3(d) has value equal to $\inf_T [T\Lambda^{I*}_{GLQF}(\frac{1}{T})]$ and the optimal trajectory of the form shown in Figure 3(e) has value equal to $\inf_T [T\Lambda^{II*}_{GLQF}(\frac{1}{T})]$, where $\Lambda^{I*}_{GLQF}(\cdot)$ and $\Lambda^{II*}_{GLQF}(\cdot)$ are defined in Equations (13) and (14), respectively. Consider now the trajectory in Figure 3(f) which has value

$$\inf_T \inf_{\substack{x_1 = \frac{1}{T} \\ x_2 - x_3 \geq \beta \frac{1}{T}}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \qquad (20)$$

The functions $\Lambda^*_{A^2}(x_2)$ and $\Lambda^*_B(x_3)$ are non-negative, convex, and achieve their minimum value which is equal to 0 at $x_2 = \mathbf{E}[A_0^2]$ and $x_3 = \mathbf{E}[B_0]$, respectively. Since $\frac{1}{T} \geq 0$, and due to the stability condition (9), for $x_2 - x_3 \geq \beta\frac{1}{T}$, it has to be the case that either $x_2 > \mathbf{E}[A_0^2]$ or $x_3 < \mathbf{E}[B_0]$. If the former is the case, we can decrease x_2 and reduce the cost, as long $x_2 - x_3 \geq \beta\frac{1}{T}$ holds. Also, if $x_3 < \mathbf{E}[B_0]$ is the case, we can increase x_3 and reduce the cost, as long $x_2 - x_3 \geq \beta\frac{1}{T}$ holds. Thus, at optimality it is true that $x_2 - x_3 = \beta\frac{1}{T}$. Then, the expression in (20) is equal to $\inf_T[T\Lambda^{II*}_{GLQF}(\frac{1}{T})]$ with $\phi = 0$ in the definition of $\Lambda^{II*}_{GLQF}(\frac{1}{T})$. Thus, since the calculation of $\Lambda^{II*}_{GLQF}(\frac{1}{T})$ involves optimization over ϕ, we conclude that the

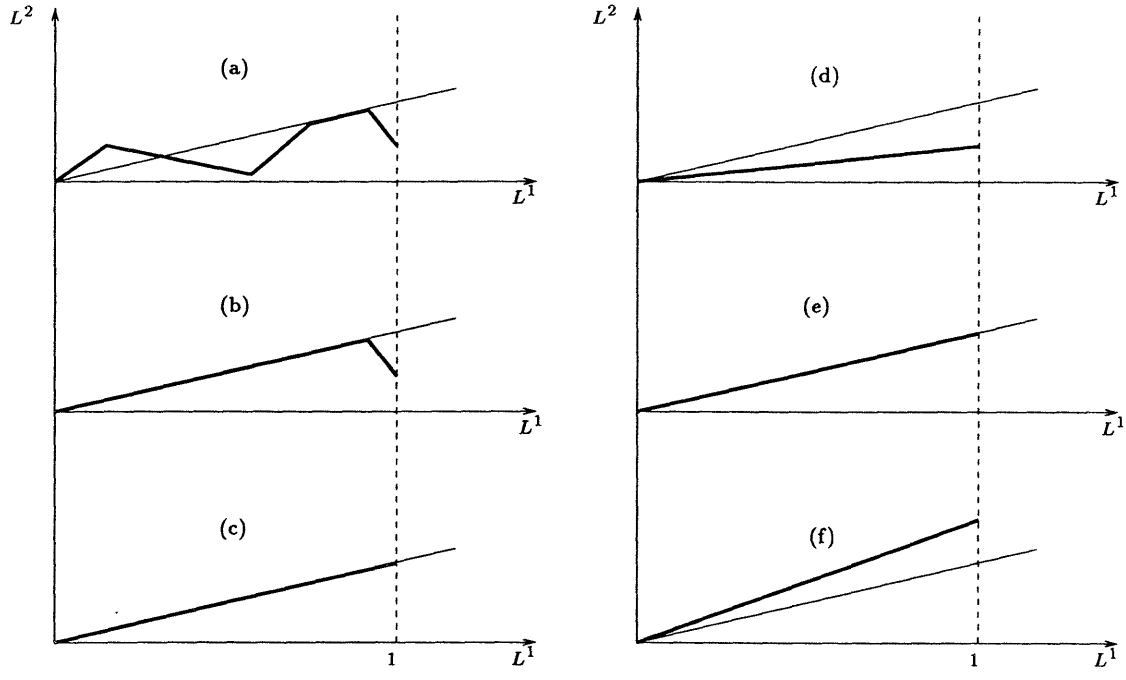

Figure 3: By the property of constant controls within each region of system dynamics the state trajectory in (b) is no more costly than the trajectory in (a). Also, by the time-homogeneity property, optimality of the state trajectory in (b) implies optimality of the trajectory in (c). Candidates for optimal state trajectories are depicted in (d), (e) and (f). The trajectory in (f) is eliminated as less profitable to the one in (e). Hence, without loss of optimality we can restrict attention to trajectories of the form in (d) and (e).

state trajectory Figure 3(f) is no more profitable than the one in Figure 3(e), leaving us

with only the trajectories in Figure 3(d) and (e) as possible candidates for optimality. We summarize the discussion of this section in the following theorem.

Theorem 6.1 *The optimal value of the problem (GLQF-OVERFLOW) is given by* θ^*_{GLQF}.

7 The most likely path

As we have explained in the Sec. 4 we will prove a matching upper bound to the one in Proposition 5.1. This is sufficient to guarantee that the two scenarios identified in the proof of Proposition 5.1 (or equivalently the two optimal state trajectories of (GLQF-OVERFLOW)) are most likely ways that queue Q^1 overflows. We summarize here these two most likely modes of overflow. We distinguish two cases:

Case 1: Suppose $\theta^*_{GLQF} = \inf_a \Lambda^{I*}_{GLQF}(a)/a$ holds. Let $a^* > 0$ the optimal solution of this optimization problem. The first queue builds up linearly with rate a^*, during a period with duration U/a^*. During this period the empirical rates of the processes A^1, A^2 and B, are roughly equal to the optimal solution (x^*_1, x^*_2, x^*_3), respectively, of the optimization problem appearing in the definition of $\Lambda^{I*}_{GLQF}(a^*)$ (Eq. (13)). 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 Figure 3(d).

Case 2: Suppose $\theta^*_{GLQF} = \inf_a \Lambda^{II*}_{GLQF}(a)/a$ holds. Let $a^* > 0$ the optimal solution of this optimization problem. Again, the first queue builds up linearly with rate a^*, during a period of duration U/a^*, and with the empirical rates of the processes A^1, A^2 and B being roughly equal to the optimal solution (x^*_1, x^*_2, x^*_3), respectively, of the optimization problem appearing in the definition of $\Lambda^{II*}_{GLQF}(a^*)$ (Eq. (14)). In this case both queues are building up, the first to an $O(U)$ level and the second to an $O(\beta U)$ level. The trajectory in L^1-L^2 space is depicted in Figure 3(e).

8 An Upper Bound

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

Before we proceed into the proof of the upper bound, we derive an alternative expression for θ^*_{GLQF} which will be essential in the proof. In the next theorem, we will show that the calculation of θ^*_{GLQF} is equivalent to finding the maximum root of a convex function. The equivalence relies mainly on [BPT96, Lemma 8.2].

In the derivation of such an equivalence we will be using the same convention for the term *infinite root* that we introduced in [BPT96, Section 8]. Namely, consider a convex function $f(u)$ with the property $f(0) = 0$. We define the *largest root* of $f(u)$ to be the solution of the optimization problem $\sup_{u:f(u)<0} u$. If $f(\cdot)$ has negative derivative at $u = 0$, there are two cases: either $f(\cdot)$ has a single positive root or it stays below the horizontal axis $u = 0$, for all $u > 0$. In the latter, case we will say that $f(\cdot)$ has a root at $u = \infty$. On a notational remark, we will be denoting by $\Lambda^I_{GLQF}(\cdot)$ and $\Lambda^{II}_{GLQF}(\cdot)$, the convex duals of $\Lambda^{I*}_{GLQF}(\cdot)$ and $\Lambda^{II*}_{GLQF}(\cdot)$, respectively. Notice, that the latter are convex functions. For $\Lambda^{I*}_{GLQF}(a)$, convexity is implied by the fact that it is the value function of a convex optimization problem with a appearing only in the right hand side of the constraints. For $\Lambda^{II*}_{GLQF}(a)$, the same argument applies when we note the following reformulation

$$\begin{aligned}\Lambda^{II*}_{GLQF}(a) &= \inf_{\substack{x_1-\phi x_3=a\\ x_2-(1-\phi)x_3=\beta a\\ 0\le\phi<1}} [\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)]\\ &= \inf_{\substack{x_1-x'_3=a\\ x_2-(x_3-x'_3)=\beta a\\ 0\le x'_3\le x_3}} [\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)].\end{aligned}$$

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

Lemma 8.1 ***(Monotonicity)*** *Consider a random process $\{X_i;\ i \in \mathbb{Z}\}$ that satisfies Assumption A. Assume $X_i \ge 0$, $i \in \mathbb{Z}$. Then for all $\theta \le \theta'$ we have $\Lambda_X(\theta) \le \Lambda_X(\theta')$.*

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

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

for all $\theta \le \theta'$.

■

This Lemma, clearly applies to the arrival and service processes.

Theorem 8.2 *θ^*_{GLQF} is the largest positive root of the equation*

$$\Lambda_{GLQF}(\theta) \triangleq \max[\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)] = 0, \tag{21}$$

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

$$\Lambda^I_{GLQF}(\theta) = \inf_{u\le 0}[\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta + u\beta)], \tag{22}$$

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

$$\Lambda^{II}_{GLQF}(\theta) = \inf_{u\ge 0}[\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \max(\Lambda_B(-u), \Lambda_B(-\theta + u\beta))]. \tag{23}$$

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

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

Similarly,

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

$$= \inf_u [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \sup_{0 \le \phi < 1} \Lambda_B(-\theta\phi + (\beta\phi + \phi - 1)u)]$$
$$= \inf_u [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \max(\Lambda_B(-u), \Lambda_B(-\theta + u\beta))]$$
$$= \inf_{u \ge 0} [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \max(\Lambda_B(-u), \Lambda_B(-\theta + u\beta))].$$

In the fifth equality above, we have used the monotonicity of $\Lambda_B(\cdot)$ (see Lemma 8.1), and the fact that the argument $-\theta\phi + (\beta\phi + \phi - 1)u$ is linear in ϕ, thus, taking its maximum value at either $\phi = 0$ or $\phi = 1$. For the sixth equality above, notice that because $\Lambda_B(\cdot)$ is non-decreasing it holds

$$\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \max(\Lambda_B(-u), \Lambda_B(-\theta + u\beta)) =$$
$$= \begin{cases} \Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u) & \text{if } u < \frac{\theta}{1+\beta} \\ \Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-\theta + u\beta) & \text{if } u \ge \frac{\theta}{1+\beta}, \end{cases} \quad (24)$$

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

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

which implies (by convexity) that the infimum is achieved at some $u \ge 0$. Thus, the infimum over unrestricted u has to be the same with the infimum over $u \ge 0$.

Using the result of [BPT96, Lemma 8.2], $\rho_1 \triangleq \inf_a \frac{1}{a}\Lambda^{I*}_{GLQF}(a)$ is the largest positive root of $\Lambda^I_{GLQF}(\theta) = 0$ (it is not hard to verify that this equation has a positive, possibly, infinite root). Similarly, $\rho_2 \triangleq \inf_a \frac{1}{a}\Lambda^{II*}_{GLQF}(a)$ is the largest positive root of $\Lambda^{II}_{GLQF}(\theta) = 0$. By Equation (12), $\theta^*_{GLQF} = \min(\rho_1, \rho_2)$. This implies that θ^*_{GLQF} is the largest positive root of the equation $\max[\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)] = 0$.

■

We next prove the upper bound for the overflow probability.

Proposition 8.3 ***(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*

$$\lim_{U \to \infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \le -\theta^*_{GLQF}. \quad (25)$$

Proof : Without loss of generality we derive an upper bound for $\mathbf{P}[L_0^1 > U]$. We will restrict ourselves to sample paths with $L_0^1 > 0$ since the remaining sample paths, with $L_0^1 = 0$, do not contribute to the probability $\mathbf{P}[L_0^1 > U]$.

Consider a busy period for the system that starts at some time $-n < 0$ ($L^1_{-n} = L^2_{-n} = 0$), and has not ended until time 0. Such a time $-n$ exists due to the stability condition (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 $L_0^1 > 0$, in three subsets Ω_1, Ω_2 and Ω_3. The first subset, Ω_1, contains all sample paths at which only Type 1 customers get serviced in the interval $[-n, 0]$. As a consequence,

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

which implies

$$L_0^1 = S^{A^1}_{-n,-1} - S^B_{-n,-1}, \quad \text{and} \quad \beta(S^{A^1}_{-n,-1} - S^B_{-n,-1}) \geq S^{A^2}_{-n,-1}.$$

Thus

$$\begin{aligned}
&\mathbf{P}[L_0^1 > U \text{ and } \Omega_1] \leq \\
&\qquad \leq \mathbf{P}[\exists n \geq 0 \text{ s.t. } S^{A^1}_{-n,-1} - S^B_{-n,-1} > U \text{ and } \beta(S^{A^1}_{-n,-1} - S^B_{-n,-1}) \geq S^{A^2}_{-n,-1}] \\
&\qquad = \mathbf{P}[\max_{\{n\geq 0:\ \beta(S^{A^1}_{-n,-1} - S^B_{-n,-1}) \geq S^{A^2}_{-n,-1}\}} (S^{A^1}_{-n,-1} - S^B_{-n,-1}) > U]. \qquad (26)
\end{aligned}$$

The second subset, Ω_2, contains sample paths at which Type 1 customers do not receive the entire capacity, and $\beta L_0^1 \leq L_0^2$. That is, there exists a $\phi \in [0, 1]$ such that Type 1 customers receive only a ϕ fraction of the total capacity ($\phi S^B_{-n,-1}$). Then we have

$$\begin{aligned}
&\mathbf{P}[L_0^1 > U \text{ and } \Omega_2] \leq \\
&\quad \leq \mathbf{P}[\exists n \geq 0, 0 \leq \phi < 1, \text{ s.t. } S^{A^1}_{-n,-1} - \phi S^B_{-n,-1} > U \text{ and} \\
&\qquad\qquad \beta(S^{A^1}_{-n,-1} - \phi S^B_{-n,-1}) \leq S^{A^2}_{-n,-1} - (1-\phi) S^B_{-n,-1}] \\
&\quad = \mathbf{P}[\max_{\{n\geq 0,\ 0\leq\phi<1:\ \beta(S^{A^1}_{-n,-1} - \phi S^B_{-n,-1}) \leq S^{A^2}_{-n,-1} - (1-\phi) S^B_{-n,-1}\}} (S^{A^1}_{-n,-1} - \phi S^B_{-n,-1}) > U]. \\
&\qquad\qquad (27)
\end{aligned}$$

Finally, the third subset, Ω_3 contains sample paths at which Type 1 customers do not receive the entire capacity, and $\beta L_0^1 \geq L_0^2$. Then there exists $k \in [0, n]$, such that the interval $[-k, 0]$ is the maximal interval that only Type 1 customers get serviced. That is,

$\beta L^1_{-i} \geq L^2_{-i}$, $i \in [0, k-1]$, and $\beta L^1_{-k} \leq L^2_{-k}$. Since Type 1 customers do not receive the entire capacity, there exists $0 \leq \phi < 1$ such that $L^1_{-k} = S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1}$. Since $\beta L^1_{-k} \leq L^2_{-k}$, we have

$$\beta(S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1}) \leq S^{A^2}_{-n,-k-1} - (1-\phi) S^B_{-n,-k-1}. \tag{28}$$

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

$$\beta(S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1} + S^{A^1}_{-k,-i-1} - S^B_{-k,-i-1}) \geq S^{A^2}_{-n,-k-1} - (1-\phi) S^B_{-n,-k-1} + S^{A^2}_{-k,-i-1},$$

which by (28) implies

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

Thus,

$$\begin{aligned}
&\mathbf{P}[L^1_0 > U \text{ and } \Omega_3] \leq \\
&\leq \mathbf{P}[\exists n \geq 0, 0 \leq k \leq n, 0 \leq \phi < 1, \text{ s.t. } S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1} + S^{A^1}_{-k,-1} - S^B_{-k,-1} > U \\
&\qquad\qquad \text{and } \beta(S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1}) \leq S^{A^2}_{-n,-k-1} - (1-\phi) S^B_{-n,-k-1} \\
&\qquad\qquad\qquad \text{and } \beta(S^{A^1}_{-k,-1} - S^B_{-k,-1}) \geq S^{A^2}_{-k,-1}] \\
&\leq \mathbf{P}[\max_{\substack{n \geq 0, 0 \leq k \leq n, 0 \leq \phi < 1 \\ \beta(S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1}) \leq S^{A^2}_{-n,-k-1} - (1-\phi) S^B_{-n,-k-1} \\ \beta(S^{A^1}_{-k,-1} - S^B_{-k,-1}) \geq S^{A^2}_{-k,-1}}} (S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1} + S^{A^1}_{-k,-1} - S^B_{-k,-1}) > U].
\end{aligned} \tag{29}$$

Let us now define

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

$$L^{II}_{GLQF} \triangleq \max_{\{n \geq 0,\ 0 \leq \phi < 1:\ \beta(S^{A^1}_{-n,-1} - \phi S^B_{-n,-1}) \leq S^{A^2}_{-n,-1} - (1-\phi) S^B_{-n,-1}\}} (S^{A^1}_{-n,-1} - \phi S^B_{-n,-1}),$$

and

$$L^{III}_{GLQF} \triangleq \max_{\substack{n \geq 0, 0 \leq k \leq n, 0 \leq \phi < 1 \\ \beta(S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1}) \leq S^{A^2}_{-n,-k-1} - (1-\phi) S^B_{-n,-k-1} \\ \beta(S^{A^1}_{-k,-1} - S^B_{-k,-1}) \geq S^{A^2}_{-k,-1}}} (S^{A^1}_{-n,-k-1} - \phi S^B_{-n,-k-1} + S^{A^1}_{-k,-1} - S^B_{-k,-1}),$$

which after bringing the constraints in the objective function become

$$L^I_{GLQF} \stackrel{\Delta}{=} \max_{n\geq 0} \inf_{u\geq 0}[(1+u\beta)S^{A^1}_{-n,-1} - uS^{A^2}_{-n,-1} + (-1-\beta u)S^B_{-n,-1}], \tag{30}$$

$$L^{II}_{GLQF} \stackrel{\Delta}{=} \max_{\substack{n\geq 0\\ 0\leq\phi<1}} \inf_{u\geq 0}[(1-u\beta)S^{A^1}_{-n,-1} + uS^{A^2}_{-n,-1} + (-\phi+u\beta\phi-u+u\phi)S^B_{-n,-1}], \tag{31}$$

and

$$L^{III}_{GLQF} \stackrel{\Delta}{=} \max_{\substack{n\geq 0\\ 0\leq k\leq n\\ 0\leq\phi<1}} \Big\{ \inf_{u_1\geq 0}[(1-u_1\beta)S^{A^1}_{-n,-k-1} + u_1 S^{A^2}_{-n,-k-1} + (-\phi+u_1\beta\phi-u_1 +$$

$$u_1\phi)S^B_{-n,-k-1}] + \inf_{u_2\geq 0}[(1+u_2\beta)S^{A^1}_{-k,-1} - u_2 S^{A^2}_{-k,-1} + (-1-u_2\beta)S^B_{-k,-1}]\Big\}. \tag{32}$$

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

$$\begin{aligned}
&\mathbf{E}[e^{\theta L^I_{GLQF}}] \\
&\leq \sum_{n\geq 0} \mathbf{E}[\exp\{\theta \inf_{u\geq 0}[(1+u\beta)S^{A^1}_{-n,-1} - uS^{A^2}_{-n,-1} + (-1-\beta u)S^B_{-n,-1}]\}\,] \\
&\leq \sum_{n\geq 0} \inf_{u\geq 0} \mathbf{E}[\exp\{\theta[(1+u\beta)S^{A^1}_{-n,-1} - uS^{A^2}_{-n,-1} + (-1-\beta u)S^B_{-n,-1}]\}\,] \\
&\leq \sum_{n\geq 0} e^{n(\inf_{u\geq 0}[\Lambda_{A^1}(\theta+\theta u\beta)+\Lambda_{A^2}(-u\theta)+\Lambda_B(-\theta-u\beta\theta)]+\epsilon_1)} \\
&\leq K^I(\theta,\epsilon_1) \qquad \text{if } \Lambda^I_{GLQF}(\theta) < 0.
\end{aligned} \tag{33}$$

In the third inequality above we have used the LDP for the arrival and service processes. In the last inequality above, when the exponent is negative (for sufficiently small ϵ_1), the infinite geometric series converges to a constant, with respect to n, $K^I(\theta,\epsilon_1)$. Also, in the last inequality, we have made the substitution $u := -\theta u$ in the expression in the exponent and used the definition of $\Lambda^I_{GLQF}(\theta)$ (Eq. (22)).

Similarly, for L^{II}_{GLQF} and for $\theta \geq 0$ we have

$$\begin{aligned}
&\mathbf{E}[e^{\theta L^{II}_{GLQF}}] \\
&\leq \sum_{n\geq 0} \mathbf{E}[\exp\{\theta \max_{0\leq\phi<1} \inf_{u\geq 0}[(1-u\beta)S^{A^1}_{-n,-1} + uS^{A^2}_{-n,-1} + (-\phi+u\beta\phi-u+u\phi)S^B_{-n,-1}]\}\,]
\end{aligned}$$

$$\leq \sum_{n\geq 0} \inf_{u\geq 0} \mathbf{E}[\exp\{\theta \max_{0\leq\phi<1}[(1-u\beta)S^{A^1}_{-n,-1} + uS^{A^2}_{-n,-1} + (-\phi+u\beta\phi-u+u\phi)S^{B}_{-n,-1}]\}\,]$$

$$\leq \sum_{n\geq 0} \inf_{u\geq 0}\left(e^{n([\Lambda_{A^1}(\theta-\theta u\beta)+\Lambda_{A^2}(u\theta)+\Lambda_B(-\theta u)]+\epsilon_2')} + e^{n([\Lambda_{A^1}(\theta-\theta u\beta)+\Lambda_{A^2}(u\theta)+\Lambda_B(-\theta+\theta u\beta)]+\epsilon_2'')}\right)$$

$$\leq 2\sum_{n\geq 0} e^{n(\inf_{u\geq 0}[\Lambda_{A^1}(\theta-\theta u\beta)+\Lambda_{A^2}(u\theta)+\max(\Lambda_B(-\theta u),\Lambda_B(-\theta+\theta u\beta))]+\epsilon_2)}$$

$$\leq K^{II}(\theta,\epsilon_2), \qquad \text{if } \Lambda^{II}_{GLQF}(\theta) < 0. \tag{34}$$

In the third inequality above, the expression to be maximized over ϕ is linear, thus, the maximum is achieved at either $\phi = 0$ or $\phi = 1$, which implies that we can upper bound it by the sum of the terms for $\phi = 0$ and $\phi = 1$.

Also, for L^{III}_{GLQF} and for $\theta \geq 0$ we have

$$\mathbf{E}[e^{\theta L^{III}_{GLQF}}]$$

$$\leq \sum_{n\geq 0}\sum_{0\leq k\leq n} \mathbf{E}\Big[\exp\Big\{\theta \max_{0\leq\phi<1}\inf_{u_1\geq 0}[(1-u_1\beta)S^{A^1}_{-n,-k-1} + u_1 S^{A^2}_{-n,-k-1} + (-\phi+u_1\beta\phi-$$

$$u_1+u_1\phi)S^{B}_{-n,-k-1}] + \inf_{u_2\geq 0}[(1+u_2\beta)S^{A^1}_{-k,-1} - u_2 S^{A^2}_{-k,-1} + (-1-u_2\beta)S^{B}_{-k,-1}]\Big\}\Big]$$

$$\leq \sum_{n\geq 0}\sum_{0\leq k\leq n} \inf_{u_1,u_2\geq 0} \mathbf{E}\Big[\exp\Big\{\theta \max_{0\leq\phi<1}[(1-u_1\beta)S^{A^1}_{-n,-k-1} + u_1 S^{A^2}_{-n,-k-1} + (-\phi+u_1\beta\phi-$$

$$u_1+u_1\phi)S^{B}_{-n,-k-1}] + [(1+u_2\beta)S^{A^1}_{-k,-1} - u_2 S^{A^2}_{-k,-1} + (-1-u_2\beta)S^{B}_{-k,-1}]\Big\}\Big]$$

$$\leq \sum_{n\geq 0}\sum_{0\leq k\leq n} \inf_{u_1,u_2\geq 0}\Big[e^{(n-k)(\Lambda_{A^1}(\theta-\theta u_1\beta)+\Lambda_{A^2}(u_1\theta)+\Lambda_B(-\theta u_1)+\epsilon_3')} +$$

$$e^{(n-k)(\Lambda_{A^1}(\theta-\theta u_1\beta)+\Lambda_{A^2}(u_1\theta)+\Lambda_B(-\theta+\theta u_1\beta)+\epsilon_3'')}\Big]e^{k(\Lambda_{A^1}(\theta+\theta u_2\beta)+\Lambda_{A^2}(-u_2\theta)+\Lambda_B(-\theta-\theta u_2\beta)+\epsilon_3''')}$$

$$\leq 2\sum_{n\geq 0}\sum_{0\leq k\leq n} e^{(n-k)(\Lambda^{II}(\theta)+\hat{\epsilon}_3)}e^{k(\Lambda^{I}(\theta)+\hat{\epsilon}_3')}$$

$$\leq 2\sum_{n\geq 0} n e^{n(\Lambda^{II}(\theta)+\hat{\epsilon}_3)} + 2\sum_{n\geq 0} n e^{n(\Lambda^{I}(\theta)+\hat{\epsilon}_3')}$$

$$\leq K^{III}(\theta,\epsilon_3), \qquad \text{if } \max(\Lambda^{I}_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) < 0. \tag{35}$$

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 Eq. (34). In the fifth inequality above, since the exponent is linear in k, the maximum over k is either at $k = 0$ or at $k = n$. Thus, we bound the term by the sum of the terms for

$k = 0$ and $k = n$. Finally, for the last inequality, both series converge to a constant if both their exponents are negative, which requires $\max(\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) < 0$.

To summarize (33), (34) and (35), the moment generating functions of L^I_{GLQF}, L^{II}_{GLQF} and L^{III}_{GLQF} are upper bounded by some constant $K(\theta, \epsilon_1, \epsilon_2, \epsilon_3)$ if $\max(\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) < 0$, where $\epsilon_1, \epsilon_2, \epsilon_3 > 0$ are sufficiently small. We can now apply the Markov inequality to obtain (using Eqs. (26), (27) and (29))

$$\begin{aligned}
&\mathbf{P}[L^1_0 > U] \\
&\leq \mathbf{P}[L^1_0 > U \text{ and Case 1}] + \mathbf{P}[L^1_0 > U \text{ and Case 2}] + \mathbf{P}[L^1_0 > U \text{ and Case 3}] \\
&\leq \left(\mathbf{E}[e^{\theta\Lambda^I(\theta)}] + \mathbf{E}[e^{\theta\Lambda^{II}(\theta)}] + \mathbf{E}[e^{\theta\Lambda^{III}(\theta)}]\right) e^{-\theta U} \\
&\leq 3K(\theta, \epsilon_1, \epsilon_2, \epsilon_3) e^{-\theta U} \qquad \text{if } \max(\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) < 0.
\end{aligned}$$

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

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

The right hand side of the above is equal to $-\theta^*_{GLQF}$ by Theorem 8.2.

■

9 Main Results

In this section we summarize our main results for the GLQF policy.

Combining Propositions 5.1 and 8.3 we obtain the following main theorem. An exact characterization of the *most likely ways* that lead to overflow were discussed in Section 7.

Theorem 9.1 ***(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*

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

where θ^*_{GLQF} *is given by*

$$\theta^*_{GLQF} = \min\left[\inf_{a>0} \frac{1}{a}\Lambda^{I*}_{GLQF}(a), \inf_{a>0} \frac{1}{a}\Lambda^{II*}_{GLQF}(a)\right], \tag{37}$$

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

$$\Lambda^{I*}_{GLQF}(a) \stackrel{\Delta}{=} \inf_{\substack{x_1-x_3=a \\ x_2\leq\beta(x_1-x_3)}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)], \tag{38}$$

and

$$\Lambda^{II*}_{GLQF}(a) \stackrel{\Delta}{=} \inf_{\substack{x_1-\phi x_3=a \\ x_2-(1-\phi)x_3=\beta a \\ 0\leq\phi<1}} [\Lambda^*_{A^1}(x_1) + \Lambda^*_{A^2}(x_2) + \Lambda^*_B(x_3)]. \tag{39}$$

It should be noted that the performance of strict priority policies, which is characterized by [BPT96, Corollary 9.2], can be also obtained as a corollary of the above theorem. We obtain the performance of strict priority to Type 2 (P_2) when $\beta = 0$, and the performance of strict priority to Type 1 (P_1) when $\beta = \infty$. It is not hard to verify that the result is identical to [BPT96, Corollary 9.2]. The above Theorem indicates that the calculation of the overflow probabilities involves the solution of a convex optimization problem. In Section 8, and for the purposes of proving Proposition 8.3, we proved in Theorem 8.2 that the exponent of the overflow probability can also be obtained as the maximum root of a convex function. This may be easier to do in some cases. Here, we restate this latter result, simplifying the expression for $\Lambda_{GLQF}(\cdot)$.

Theorem 9.2 θ^*_{GLQF} *is the largest positive root of the equation*

$$\Lambda_{GLQF}(\theta) = \max\{\Lambda_{A^1}(\theta) + \Lambda_B(-\theta), \inf_{0\leq u\leq\frac{\theta}{1+\beta}} [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)]\} = 0. \tag{40}$$

Proof : Due to Theorem 8.2 it suffices to prove that the expression in (40) is equal to $\max[\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)]$. Recall the definitions of $\Lambda^I_{GLQF}(\theta)$ in (22) and of $\Lambda^{II}_{GLQF}(\theta)$ in (23). Recall also the expression in (24) for the objective function of the optimization problem corresponding to $\Lambda^{II}_{GLQF}(\theta)$. Let now u^* be the optimal solution of the optimization problem in the definition of $\Lambda^{II}_{GLQF}(\theta)$. We distinguish two cases:

<u>Case 1:</u> where $u^* \geq \frac{\theta}{1+\beta}$. Then, notice that u^* is also the minimizer of the objective

function in the definition of $\Lambda^I_{GLQF}(\theta)$. Thus, due to convexity, the constraint $u \leq 0$ is tight for the problem corresponding to $\Lambda^I_{GLQF}(\theta)$, and

$$\max(\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) = \Lambda_{A^1}(\theta) + \Lambda_B(-\theta), \qquad \text{if } u^* \geq \tfrac{\theta}{1+\beta}. \tag{41}$$

But,

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

In the second inequality above we have used the assumption $u^* \geq \frac{\theta}{1+\beta}$ and convexity. Therefore, combining it with (41) we obtain

$$\begin{aligned}
&\max(\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) = \max\{\Lambda_{A^1}(\theta) + \Lambda_B(-\theta), \\
&\quad \inf_{0\leq u\leq \frac{\theta}{1+\beta}} [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)] \} = \Lambda_{GLQF}(\theta) \qquad \text{if } u^* \geq \tfrac{\theta}{1+\beta}.
\end{aligned} \tag{42}$$

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

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

where in the last inequality we have used the fact that $u < \frac{\theta}{1+\beta}$ which implies (see

also (24)) $\Lambda_B(-u) \geq \Lambda_B(-\theta + u\beta)$. Therefore, for $0 \leq u^* \leq \frac{\theta}{1+\beta}$ also, we have

$$\max(\Lambda^I_{GLQF}(\theta), \Lambda^{II}_{GLQF}(\theta)) = \max\{\Lambda_{A^1}(\theta) + \Lambda_B(-\theta), \inf_{0 \leq u \leq \frac{\theta}{1+\beta}} [\Lambda_{A^1}(\theta - u\beta) + \Lambda_{A^2}(u) + \Lambda_B(-u)] \} = \Lambda_{GLQF}(\theta).$$

■

The results of this Theorem can be also specialized to the case of priority policies, to obtain the characterization of [BPT96, Corollary 9.4].

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

10 A Comparison

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

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

$$\theta^\pi_1 = \lim_{U \to \infty} \frac{1}{U} \log \mathbf{P}[L^1 > U] \qquad \text{and} \qquad \theta^\pi_2 = \lim_{U \to \infty} \frac{1}{U} \log \mathbf{P}[L^2 > U]. \tag{43}$$

The GPS policy analyzed in [BPT96] is a parametric policy with performance depending on the parameter ϕ_1. To make this dependence explicit we will be using the notation GPS(ϕ_1). Also, the GLQF policy analyzed in Section 4 is a parametric policy with performance depending on the parameter β. For the same reason we will be using the notation GLQF(β). Special cases of a work-conserving policy π are the GPS(ϕ_1) policy, the GLQF(β) policy, the strict priority to Type 1 policy (P_1 policy), and the strict priority to Type 2 policy (P_2 policy). Using [BPT96, Theorem 9.1], and [BPT96, 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

$$\theta^{P_1} = (\max_{\pi\in\Pi} \theta_1^\pi, \min_{\pi\in\Pi} \theta_2^\pi) \qquad \text{and} \qquad \theta^{P_2} = (\min_{\pi\in\Pi} \theta_1^\pi, \max_{\pi\in\Pi} \theta_2^\pi).$$

In Figure 4 we plot $\theta^{GPS(\phi_1)}$ as ϕ_1 varies in $[0,1]$, and $\theta^{GLQF(\beta)}$ as β varies in $[0,\infty)$. For simplicity the calculations were performed with the arrival and service processes being Bernoulli (we say that a process $\{X_i;\ i \in \mathbb{Z}\}$ is Bernoulli with parameter p, denoted by $X \sim$Ber(p), when X_i are i.i.d. and $X_i = 1$ with probability p and $X_i = 0$ with probability $1-p$). Also, for the calculations we used the expressions for θ^*_{GPS} and θ^*_{GLQF} given in

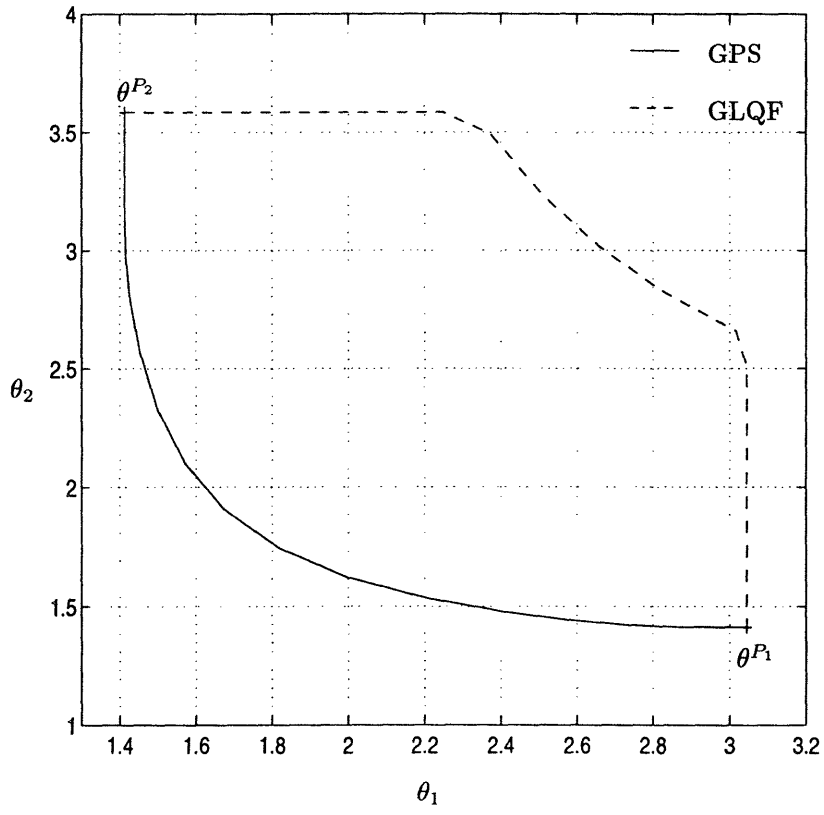

Figure 4: The performance $\theta^{GPS(\phi_1)}$ of the GPS(ϕ_1) policy as ϕ_1 varies in $[0,1]$, and the performance $\theta^{GLQF(\beta)}$ of the GLQF(β) policy as β varies in $[0,\infty)$, when $A^1 \sim$Ber(0.3), $A^2 \sim$Ber(0.2) and $B \sim$Ber(0.9).

[BPT96, Thm. 9.3] and Thm. 9.2, respectively, because they were more efficient to perform numerically than the equivalent expressions in [BPT96, Thm. 9.1] and Thm. 9.1. Note that $\theta^{P_1} = \theta^{GPS(1)} = \theta^{GLQF(\infty)}$ and that $\theta^{P_2} = \theta^{GPS(0)} = \theta^{GLQF(0)}$.

Figure 4 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 than when one has to deal with delay insensitive traffic (i.e., when there are no delay constraints) GLQF is more suitable than GPS. On the other hand, GLQF does not have the fairness property of GPS, that is it may allow a bursty class of traffic to be using all the available capacity until the backlog of the other class reaches the level of the bursty one.

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

Definition 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* $\theta_1^{GPS(\phi_1)} > \theta_1^{GLQF(\beta)}$ *and* $\theta_2^{GPS(\phi_1)} > \theta_2^{GLQF(\beta)}$.

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

Lemma 10.2 *If* $\phi_1 \leq \phi_1'$ *we have*

$$\theta_1^{GPS(\phi_1)} \leq \theta_1^{GPS(\phi_1')} \qquad \text{and} \qquad \theta_2^{GPS(\phi_1)} \geq \theta_2^{GPS(\phi_1')}.$$

Proof : We only prove the first relation. The second can be obtained by a symmetrical argument. We use the result of [BPT96, Theorem 9.3]. Note that $\phi_1 \leq \phi_1'$, implies $\phi_2' = (1-\phi_1') \leq \phi_2 = (1-\phi_1)$. Thus, by Lemma 8.1, for all $u, \theta \geq 0$ we have that $\Lambda_B(-\theta+\phi_2 u) \geq \Lambda_B(-\theta+\phi_2' u)$, which by [BPT96, Theorem 9.3] implies $\Lambda_{GPS(\phi_1)}(\theta) \geq \Lambda_{GPS(\phi_1')}(\theta)$ for all θ. Therefore, by convexity, for θ^*_{GPS}, as it is defined in [BPT96, Theorem 9.3], we have $\theta^*_{GPS(\phi_1)} \leq \theta^*_{GPS(\phi_1')}$.

■

A similar property is proven for the GLQF policy.

Lemma 10.3 *If $\beta \leq \beta'$ we have*

$$\theta_1^{GLQF(\beta)} \leq \theta_1^{GLQF(\beta')} \qquad \textit{and} \qquad \theta_2^{GLQF(\beta)} \geq \theta_2^{GLQF(\beta')}.$$

Proof: Again we only prove the first relation. The second can be obtained by a symmetrical argument. We use the optimal control formulation of Section 6. We argued there that optimal trajectories have the form of Figure 3(d) and (e), with cost $\inf_a \frac{1}{a}\Lambda^{I*}_{GLQF}(a)$ and $\inf_a \frac{1}{a}\Lambda^{II*}_{GLQF}(a)$, respectively. Let us fix β and consider how the cost is affected by using the policy with $\beta' = \beta + \epsilon$, for small $\epsilon > 0$.

Consider first trajectories of the form in Figure 3(e). Note that we can rewrite $\Lambda^{II*}_{GLQF(\beta)}(a)$ as

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

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

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

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 $\beta, a \geq 0$, 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 $\Lambda^{II*}_{GLQF(\beta')}(a)$. This contradicts our initial assumption. We conclude that by increasing β to β' we also increase the optimal cost of trajectories having the form in Figure 3(e).

If now, an optimal trajectory has the form in Figure 3(d), then it will still be the optimal, by convexity, when β is increased to β'. Thus, in this case, the optimal cost does not change.

We summarize by considering how the cost is affected as β is increased from 0 to ∞. At $\beta = 0$, possible optimal trajectories have the form of Figure 3(e). There is a threshold

value $\bar{\beta}$ such that for all $\beta \leq \bar{\beta}$ optimal trajectories have the form of Figure 3(e) with values increasing as β increases from 0 to $\bar{\beta}$. For all $\beta > \bar{\beta}$ optimal trajectories have the form of Figure 3(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*

$$\theta_1^{GPS(\phi_1)} \leq \theta_1^{GLQF(\beta)} \qquad \text{and} \qquad \theta_2^{GPS(\phi_1)} \leq \theta_2^{GLQF(\beta)},$$

then the GLQF curve dominates the GPS curve.

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

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

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

■

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

Theorem 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

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

which implies

$$\theta_1^{GPS(\phi_1)} \leq \theta_1^{GLQF(\beta)} \qquad \text{and} \qquad \theta_2^{GPS(\phi_1)} \leq \theta_2^{GLQF(\beta)}.$$

Since we have established that both in the GPS and the GLQF case, the overflow probability is equal to the probability of overflowing according to one out of two scenarios, it suffices to establish the above only for these scenarios. In particular, we distinguish the following cases depending on the possible modes of overflow for GLQF(β), which are described in Section 7.

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

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

Case 3: Mode 2 for overflow of Q^1 and mode 1 for overflow of Q^2.

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

In Case 1 and 2, we have

$$x_1 - x_3 = a,$$
$$x_2 \leq \beta a,$$

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

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

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

Consider now Cases 3 and 4. We have

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

where x_j, $j = 1, 2, 3$, a, ϕ, solve the optimization problem corresponding to the overflow of Q^1 in mode 2. Then the GPS policy with $\phi_1 \leq \phi$ will overflow Q^2. In Case 3, for reasons

explained in the previous paragraph, the GPS policy will overflow Q^2 for all ϕ_1. If, finally, we are in Case 4, we have

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

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

11 Conclusions

In this paper we considered a multiclass multiplexer, with segregated buffers for each type of traffic, and under 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 have explicitly and in detail characterized the most likely modes of overflow. We formulated the problem of calculating the maximum overflow probability (over all scenarios that lead to overflow) as an optimal control problem, general enough to include any work conserving policy. This provides particular insight into the problem. We have addressed the case of multiplexing two streams. Our lower bound proof extends to the general case of N streams, the proof of an upper bound is an open problem.

References

[BPT94] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis, *On the large deviations behaviour of acyclic networks of G/G/1 queues*, Tech. Report LIDS-P-2278, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, December 1994.

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

[BPT96] D. Bertsimas, I. Ch. Paschalidis, and J. N. Tsitsiklis, *Asymptotic buffer overflow probabilities in multiclass multiplexers, Part I: The GPS policy*, Tech. Report LIDS-P-2341, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, June 1996.

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

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

[DKS90] A. Demers, S. Keshav, and S. Shenker, *Analysis and simulation of a fair queueing algorithm*, Journal of Internetworking: Research and Experience **1** (1990), 3–26.

[dVCW93] G. de Veciana, C. Courcoubetis, and J. Walrand, *Decoupling bandwidths for networks: A decomposition approach to resource management*, Memorandum, Electronics Research Laboratory, University of California Berkeley, 1993.

[DZ93a] A. Dembo and T. Zajic, *Large deviations: From empirical mean and measure to partial sums processes*, Preprint, 1993.

[DZ93b] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, Jones and Bartlett, 1993.

[EM93] A. I. Elwalid and D. Mitra, *Effective bandwidth of general Markovian traffic sources and admission control of high speed networks*, IEEE/ACM Transactions on Networking **1** (1993), no. 3, 329–343.

[GA94] A. Ganesh and V. Anantharam, *The stationary tail probability of an exponential server tandem fed by renewal arrivals*, Preprint, 1994.

[GGG+93] H.R. Gail, G. Grover, R. Guérin, S.L. Hantler, Z. Rosberg, and M. Sidi, *Buffer size requirements under longest queue first*, Performance Evaluation **18** (1993), 133–140.

[GH91] R.J. Gibbens and P.J. Hunt, *Effective bandwidths for the multi-type UAS channel*, Queueing Systems **9** (1991), 17–28.

[GW94] P.W. Glynn and W. Whitt, *Logarithmic asymptotics for steady-state tail probabilities in a single-server queue*, J. Appl. Prob. **31A** (1994), 131–156.

[Hui88] J. Y. Hui, *Resource allocation for broadband networks*, IEEE Journal on Selected Areas in Communications **6** (1988), no. 9, 1598–1608.

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

[KWC93] G. Kesidis, J. Walrand, and C.S. Chang, *Effective bandwidths for multiclass Markov fluids and other ATM sources*, IEEE/ACM Transactions on Networking **1** (1993), no. 4, 424–428.

[O'C95] N. O'Connell, *Large deviations in queueing networks*, Preprint, 1995.

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

[PG93] A.K. Parekh and R.G. Gallager, *A generalized processor sharing approach to flow control in integrated services networks: The single node case*, IEEE/ACM Transactions on Networking **1** (1993), no. 3, 344–357.

[PG94] A.K. Parekh and R.G. Gallager, *A generalized processor sharing approach to flow control in integrated services networks: The multiple node case*, IEEE/ACM Transactions on Networking **2** (1994), no. 2, 137–150.

[SW95] A. Shwartz and A. Weiss, *Large deviations for performance analysis*, Chapman and Hall, New York, 1995.

[TGT95] D. Tse, R.G. Gallager, and J.N. Tsitsiklis, *Statistical multiplexing of multiple time-scale Markov streams*, IEEE Journal on Selected Areas in Communications **13** (1995), no. 6.

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