PERGAMON

Automatica 35 (1999) 1951–1968

www.elsevier.com/locate/automatica

Class-specific quality of service guarantees in multimedia communication networks☆

Ioannis Ch. Paschalidis*

Department of Manufacturing Engineering, Boston University, Boston MA 02215, USA

Received 30 June 1998; revised 10 February 1999; received in final form 1 June 1999

An admission control approach that can provide per class packet loss and delay Quality of Service guarantees is developed. The proposed approach is based on large deviations performance analysis results.

Abstract

We consider the problem of *quality-of-service* (*QoS*) provisioning in modern high-speed, multimedia, communication networks. We quantify QoS by the probabilities of *loss* and *excessive delay* of an arbitrary packet, and introduce the model of a *multiclass node* (*switch*) which provides network access to users that may belong to multiple *service classes*. We treat such a node as a *stochastic system* which we analyze and control. In particular, we develop an analytical approach to estimate both the delay and the buffer overflow probability per service class, based on ideas from *large deviations* and *optimal control*. We exploit these performance analysis results by devising a *call admission control algorithm* which can provide per class QoS guarantees. We compare the proposed approach to alternative worst-case and effective bandwidth-based schemes and argue that it leads to increased efficiency. Finally, we discuss extensions to the network case in order to provide *end-to-end* QoS guarantees.

Keywords: Communication networks; Large deviations; Optimal control; Asymptotic analysis; Effective bandwidth; Bandwidth allocation

1. Introduction

Quality-of-service (*QoS*) provisioning in today's communication networks is an increasingly important issue as real-time applications such as internet telephony, teleconferencing, web access of multimedia information, and interactive TV, become more heavily used. Technology to accommodate these applications exists (e.g., the *asynchronous transfer mode* (*ATM*) protocol, the Internet enhanced with RSVP mechanisms discussed in Zhang, Deering, Estrin, Shenker & Zappala, 1993); the challenge is how to manage the network resources (bandwidth) to provide several QoS grades that such a diverse set of applications requires.

To achieve this goal several approaches have been proposed. One class of mechanisms is based on worst-case analysis and provides deterministic QoS guarantees (Cruz, 1991a,b; Parekh & Gallager, 1993, 1994; Chang, 1994), that is, ensuring no packet losses and no large delays with certainty. Although such an approach is useful when no statistical description of the offered traffic is available, it can lead to substantial underutilization of the network resources. To realize statistical multiplexing gains the so called *effective bandwidth* mechanism has been proposed (Gibbens & Hunt, 1991; Guérin, Ahmadi & Naghshineh, 1991; Hui, 1988; Kelly, 1991, 1996). Briefly, effective bandwidth is a number between the peak and average rate of a connection such that when connections are allocated their effective bandwidth in an appropriately dimensioned buffer, the buffer overflow probability stays below a given small level (say on the order of 10^{-6}). Real-time applications can tolerate such small frequencies of congestion phenomena.

☆ Research partially supported by the NSF under grants NCR-9706148 and ACI-9873339. This paper was not presented at any IFAC meeting. This paper was recommended for publication in revised from by Guest Editors Venkat Anantharam and Jean Walrand.

* Corresponding author. Tel.: + 1-617-353-0434; fax: + 1-617-353-5548.

E-mail address: yannisp@bu.edu (I.Ch. Paschalidis)

0005-1098/99/$ - see front matter
PII: S0005-1098(99)00125-9

The effective bandwidth is a single class scheme: all connections are multiplexed into one buffer and, thus, they face the same QoS. Our objective in this paper is to provide class-specific QoS guarantees. We quantify QoS by the *probability of excessive delays* and the *loss probability*. It is desirable to keep them at very small levels (e.g., on the order of 10^{-6}). Determining such probabilities for non-trivial traffic models is a particularly hard problem, thus, it is natural to focus on asymptotic regimes and determine their exponential decay rate. To this end, *large deviations* theory (see Bucklew, 1990; Dembo & Zeitouni, 1993; Shwartz & Weiss, 1995) will be our main analytical tool.

We will introduce the model of a *multiclass node* (*switch*) where users belonging to multiple service classes request to be connected. A *service class* is characterized by the statistical properties of the incoming traffic (distribution of the stochastic process modeling the 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 carrying 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).

We formulate the large deviations problem of obtaining the tails of loss and delay probabilities for each class as a deterministic *optimal control* problem which we explicitly solve. We obtain "full" (i.e., asymptotically tight) large deviations results for the special case of two service classes. The more general multiclass case appears to be much harder; we provide approximations and evidence that they are fairly accurate. We exploit the performance analysis results by devising an admission control procedure that provides class-specific QoS guarantees. The admission controller tries to fully utilize the available bandwidth by investigating different bandwidth allocation policies (within a certain parametric class of policies) among service classes and denies admission only when there is no feasible allocation that guarantees QoS to *all* connected calls. We compare our approach to alternative worst-case and effective bandwidth schemes and show that it leads to more efficient use of the bandwidth resources.

Large deviations techniques have recently been applied to a variety of problems in telecommunications (see the survey paper by Weiss, 1995). The problem of estimating tail probabilities of rare events in a single-class queue has received extensive attention in the literature (Courcoubetis & Weber, 1995a; de Veciana & Walrand, 1995; Elwalid & Mitra, 1993; Gibbens & Hunt, 1991; Glynn & Whitt, 1994; Hui, 1988; Kelly, 1991; Kesidis, Walrand & Chang, 1993; Tse, Gallager & Tsitsiklis, 1995). The extension of these ideas to multiclass queues and networks appears to be a rather challenging problem and a very active area of research. In a multiclass setting, although some performance analysis results which estimate or approximate the asymptotic decay rates of buffer overflow probabilities have been obtained (Bertsimas, Paschalidis & Tsitsiklis, 1997, 1998a; Courcoubetis & Weber, 1995b; de Veciana & Kesidis, 1995; O'Connell, 1995; Zhang, 1997; Zhang, Towsley & Kurose, 1995), the implications to delay have not been considered and the applications to admission control not thoroughly investigated. Zhang, Liu, Kurose and Towsley (1997) consider the multiclass case using approximate performance analysis results which leaves room for substantial increase in efficiency.

Among the main contributions of the work in this paper we consider:

- The multiclass character of the analytical results and the admission control algorithm. As we demonstrate, the advantage is that each service class is allocated the capacity required by its QoS specifications which include both a measure of loss and one of delay. It allows, for instance, class 1 traffic to suffer less delay with a larger loss probability than class 2 traffic, something that can not be achieved with neither single class nor priority schemes (e.g., as in Elwalid & Mitra, 1995).
- The optimal control formulation of the calculation of the congestion probabilities. An advantage of this approach is that the optimal control solution also provides a complete characterization of the *most likely* way that congestion builds up, allowing us to acquire an intuitive understanding of the chain of events that lead to congestion.
- The handling of stochastic service capacities (in contrast to most of the work in the literature which focuses on deterministic capacity). As we elaborate in Section 3, this allows us to handle more sophisticated scheduling disciplines for allocating bandwidth among service classes.
- The interplay between admission control and scheduling. As it will become evident in Section 7, the proposed admission controller provides the input to the bandwidth allocation scheduler, which adjusts its parameters to accommodate the current load.

On the organization of this paper, we start in Section 2 with some preliminaries on large deviations. In Section 3 we introduce our model and formally define the problem. In Sections 4 and 5 we analytically obtain asymptotics for the loss and delay probabilities in the two-class case. In Section 6 we develop extensions to the multiclass case and refinements of the large deviations asymptotics. In Section 7 we use these results to develop the call admission control algorithm; we compare it with alternative schemes via illustrative examples. In Section 8 we discuss

extensions to the network case and indicate how these results can be applied to provide *end-to-end* QoS guarantees. Finally, in Section 9, we include some concluding remarks.

2. Preliminaries

In the form of background on large deviations and to establish some of our notation, we first review some basic results. Consider a sequence of i.i.d. random variables X_i, $i \geq 1$, with mean $\mathbf{E}[X_1] = \bar{X}$. The strong law of large numbers asserts that $\sum_{i=1}^{n} X_i/n$ converges to $\bar{X}$, as $n \to \infty$, with probability one (w.p.1). Thus, for large n the event $\sum_{i=1}^{n} X_i \geq na$, where $a > \bar{X}$, (or $\sum_{i=1}^{n} X_i \leq na$, for $a < \bar{X}$) is a rare event. In particular, its probability behaves as $e^{-nr(a)}$, as $n \to \infty$, where the function $r(\cdot)$ determines the rate at which the probability of this event is diminishing. Cramér's (1938) theorem determines $r(\cdot)$, and is considered the first large deviations statement.

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

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

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. We will be making the following assumption.

Assumption A.

(1) The limit

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

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.

Let us define

$$\Lambda^*(a) \triangleq \sup_{\theta} (\theta a - \Lambda(\theta)), \tag{1}$$

which is the Legendre transform of $\Lambda(\cdot)$. $\Lambda(\cdot)$ and $\Lambda^*(\cdot)$ are convex duals (see Rockafellar, 1970), namely, along with (1), it also holds

$$\Lambda(\theta) = \sup_{a} (\theta a - \Lambda^*(a)). \tag{2}$$

The function $\Lambda^*(\cdot)$ is convex and lower semicontinuous (see Dembo & Zeitouni, 1993).

Gärtner (1977) and Ellis (1984) have extended Cramér's theorem to cover autocorrelated processes. In particular, under Assumption A, the Gärtner-Ellis Theorem (see Bucklew, 1990; Dembo & Zeitouni, 1993) establishes that $\{S_n\}$ satisfies a *large deviations principle* (*LDP*) with *rate function* $\Lambda^*(\cdot)$. More specifically, this theorem intuitively asserts that for large enough n and for small $\varepsilon > 0$,

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

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* (to be referred as *Sample path LDP*)

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

In a key paper Dembo & Zajic (1995) under certain mild mixing conditions on the stationary sequence $\{X_i;\ i \geq 1\}$, establish an LDP for the process $S_n(\cdot)$ in $D[0,1]$ (right continuous functions with left limits) equipped with the supremum norm topology. In the spirit of the sample path LDP, we will be assuming the following.

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

$$e^{-(n\varepsilon_2 + \sum_{i=0}^{m-1} (k_{i+1} - k_i)\Lambda^*(a_i))} \leq \mathbf{P}[|S_{k_{i+1}} - S_{k_i} - (k_{i+1} - k_i)a_i| \leq \varepsilon_1 n,\ i = 0, \dots, m-1]. \tag{3}$$

A detailed discussion of this Assumption, and the technical conditions under which it is satisfied can be found in Dembo and Zajic (1995). Intuitively, Assumption B deals with the probability of sample paths that are constrained to be within a tube around a "polygonal" path made up with linear segments of slopes $a_0, \dots, a_{m-1}$. We will also be making the following assumption, which can be viewed as the "convex dual analog" of Assumption B.

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

$$\mathbf{E}[e^{\boldsymbol{\theta} \cdot \mathbf{Z}}] \leq \exp\left\{ \sum_{j=1}^{m} [(k_j - k_{j-1})\Lambda(\theta_j) + \Gamma(\theta_j)] \right\}, \tag{4}$$

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

In Chang (1995) a uniform bounding condition is given under which Assumptions B and C are satisfied. It is

verified that the set of processes satisfying these assumptions is large enough to include renewal, Markov-modulated, and stationary processes with mild mixing conditions. Such processes can model "burstiness" and are commonly used in modeling the input traffic to communication networks.

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

3. The multiclass model and the GPS policy

Consider the architecture of Fig. 1 which accommodates M service classes. Calls declare their service class and request to be connected to the system. We bundle together calls of the same service class, storing them in the same buffer, which allows us to treat them identically. The QoS they will receive depends on the buffer size and the amount of bandwidth allocated to the buffer (which is a function of the employed scheduling policy).

We adopt a discrete-time model where time is divided into time slots of equal length and the state of the system is observed at the beginning of each time slot. We let $\tilde{A}_i^j$, where i is in the set of integers $\mathbb{Z}$, denote the number of bits (or packets) generated by a single class j call ($j = 1, \dots, M$) during time slot i. In fact, we do not need to distinguish between calls of the same class since we will use the same stochastic model for each of them. That is, we let N_j denote the number of admitted class j calls, and A_i^j the aggregate number of bits that enters buffer Q^j. We denote by U_j the size of buffer Q^j. All buffers share the same communication link which can accommodate B_i bits during the time slot i. We assume that the stochastic processes $\{A_i^j;\ i \in \mathbb{Z}\}$ for $j = 1, \dots, M$, and $\{B_i;\ i \in \mathbb{Z}\}$ are stationary and mutually independent. However, we allow dependencies between the number of bits at different time slots in each process, which allows us to model bursty traffic.

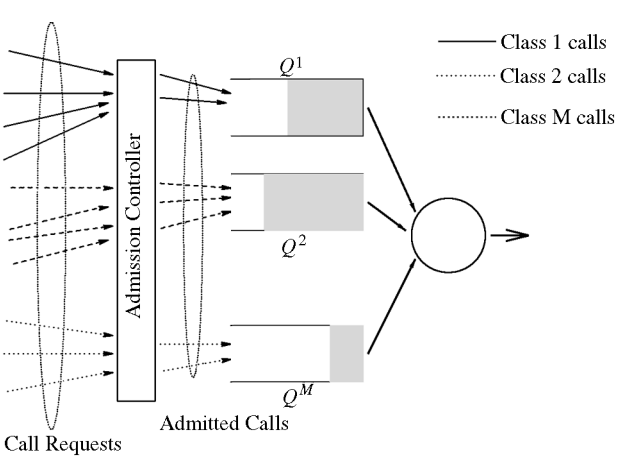


Fig. 1. A node with support for multiple service classes.

We denote by L_i^j the queue length at the beginning of time slot i (without counting arrivals during this time slot) in buffer Q^j, $j = 1, \dots, M$. Let D_i^j be the corresponding delay (the time an arbitrary class j bit spends in the buffer). Notice that both queue lengths and delays depend on the corresponding buffer size. As a general rule we will suppress this dependence in the notation, except in cases where we explicitly denote otherwise. We assume that the server (communication link) allocates its capacity between queues Q^j according to a work-conserving policy (i.e., the server never stays idle when there is work in the system). For stability purposes we assume that for all i

$$\mathbf{E}[B_i] > \sum_{j=1}^{M} \mathbf{E}[A_i^j]. \tag{5}$$

We further assume that the arrival and service processes satisfy a LDP (Assumption A), as well as Assumptions B and C.

We employ the *generalized processor sharing* (GPS) policy which was proposed in Demers, Keshav and Shenker (1990) and further explored in Parekh and Gallager (1993, 1994). It possesses certain fairness properties which are desirable in the multimedia setting we are considering. According to this policy, the server allocates a fraction $\phi_j \in [0,1]$ of its capacity to queue Q^j, where of course $\sum_{j=1}^{M} \phi_j = 1$. The policy is defined to be work-conserving, which implies that if one or more of the queues do not fully use the fraction of the capacity allocated to them, the excess is distributed to the remaining queues.

We are interested in devising an admission control algorithm which guarantees a desirable level of QoS. Let $D_{\max}^j$ be the desirable maximum allowed delay for class j, and let δ_j be scalars such that for all $j = 1, \dots, M$

$$\mathbf{P}[L_{U_j}^j \ge U_j] < \delta_j, \tag{6}$$

$$\mathbf{P}[D_{U_j}^j \ge D_{\max}^j] < \delta_j, \tag{7}$$

where the subscript U_j explicitly denotes the dependence of queue lengths and delays on the buffer size. We will refer to $D_{\max}^j$ and δ_j as *QoS parameters*, since they determine how well a particular class is treated.

To achieve this goal we will first estimate these congestion probabilities and then use the performance analysis results to develop the admission control algorithm. For analytical convenience, we will be approximating the loss probability in (6) with the level crossing probability $\mathbf{P}[L^j \ge U_j]$ in an infinite buffer system, where L^j denotes the corresponding queue length in that system. Kelly (1996) establishes that these two probabilities have the same asymptotic decay rate (same exponent). Similarly,

we upper bound the delay probability in (7) with the one in the infinite buffer system.

Before we proceed with this agenda we make a short note on the usefulness of allowing the service process $\{B_i;\ i\in\mathbb{Z}\}$ to be an arbitrary and autocorrelated stochastic process. This has to be contrasted with most of the work in the literature that assumes a deterministic service capacity. Consider for example, the case where a deterministic server, with capacity c bits per time slot, accommodates some other high priority traffic (in addition to the traffic generated by the M service classes). Let $\{H_i;\ i\in\mathbb{Z}\}$ denote the stochastic process characterizing this high priority traffic. Assuming that $c > H_i$ w.p.1., we conclude that the capacity remaining for the M service classes $\{c - H_i;\ i\in\mathbb{Z}\}$ is also stochastic. Thus, the stochasticity of the service process allows the treatment of more complicated, than the GPS, service disciplines.

4. Two-class case: overflow probabilities

As we outlined in the Introduction, our main analytical results are for two-class systems ($M = 2$). Results for the overflow probabilities in this case were obtained in Bertsimas et al. (1997) using the approach introduced in Bertsimas et al. (1998a); we will just restate them here for completeness. We will later present extensions and refinements in the general multiclass case. The next theorem summarizes the two-class overflow result and is from Bertsimas et al. (1997).

Theorem 1 (Overflows, Bertsimas et al., 1997). *In the two-class system, under the GPS policy, assuming that the arrival and service processes satisfy Assumptions A–C, the steady-state queue length in the first buffer, L^1, satisfies*

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

where $\theta^*_{L,1}$ is given by

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

and the functions $\Lambda^{\mathrm{I}*}_{\mathrm{GPS},1}(\cdot)$ and $\Lambda^{\mathrm{II}*}_{\mathrm{GPS},1}(\cdot)$ are defined as follows:

$$\Lambda^{\mathrm{I}*}_{\mathrm{GPS},1}(a)\triangleq \inf_{\substack{x_1+x_2-x_3=a\\ x_2\leq\phi_2 x_3}}[\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)], \tag{9}$$

$$\Lambda^{\mathrm{II}*}_{\mathrm{GPS},1}(a)\triangleq \inf_{\substack{x_1-\phi_1 x_3=a\\ x_2\geq\phi_2 x_3}}[\Lambda^*_{A^1}(x_1)+\Lambda^*_{A^2}(x_2)+\Lambda^*_B(x_3)]. \tag{10}$$

Intuitively, the above theorem states that for large values of the buffer size U the queue length L^1 behaves as $\mathbf{P}[L^1\geq U]\sim \mathrm{e}^{-U\theta^*_{L,1}}$.

Next, we state an alternative expression for $\theta^*_{L,1}$ (see Bertsimas et al., 1997 for a proof), which may be more convenient in computations. 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 $u = \infty$.

Theorem 2 (Bertsimas et al., 1997). *$\theta^*_{L,1}$ is the largest positive root of the equation*

$$\Lambda_{\mathrm{GPS},1}(\theta)\triangleq\Lambda_{A^1}(\theta)+\inf_{0\leq u\leq\theta}[\Lambda_{A^2}(\theta-u)+\Lambda_B(-\theta+\phi_2 u)] = 0. \tag{11}$$

In addition to the exponent $\theta^*_{L,1}$, the analysis in Bertsimas et al. (1997) also characterizes the most-likely ways (in the sense that they maximize the overflow probability) that overflow occurs. In particular, we distinguish two cases:

Case 1: Suppose $\theta^*_{L,1} = \inf_a \Lambda^{\mathrm{I}*}_{\mathrm{GPS},1}(a)/a$ holds. Let $a^* > 0$ be the optimal solution of this optimization problem. In this case, the first queue is building up linearly with rate a^*, during a period with duration U/a^*, up to an $O(U)$ level. During the same time interval, the second queue stays at an $o(U)$ level, and 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 $\Lambda^{\mathrm{I}*}_{\mathrm{GPS},1}(a^*)$ (Eq. (9)).

Case 2: Suppose $\theta^*_{L,1} = \inf_a \Lambda^{\mathrm{II}*}_{\mathrm{GPS},1}(a)/a$ 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 $\Lambda^{\mathrm{II}*}_{GPS,1}(a^*)$ (Eq. (10)).

It is interesting to reflect at this point on the implications of this result on admission control. Consider an admission control mechanism for queue Q^1 designed to guarantee a desirable level of the overflow probability. A worst-case analysis as in Parekh and Gallager (1993) 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. In contrast, due to their probabilistic nature, our results suggest that a significant (statistical multiplexing) gain can be realized by not imposing this assumption. In the overflow mode described in Case 1 above, the second queue consumes less than the fraction ϕ_2 of the total service capacity, leaving the remaining capacity for the

first queue. This implies that additional class 1 connections can be accommodated 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.

5. Two-class case: delay probabilities

We now turn our attention to the probability of large delays. We assume that the FCFS policy is implemented for customers of the same class. We first establish a general result for the delay that customers are facing in each of the queues Q^1 and Q^2. Recall that D_i^1 and D_i^2 denote the sojourn time in the system of a virtual customer arriving at time i (we assume that the virtual customer arrives at the beginning of time slot i before any other customer arrives or departs at the same slot).

Theorem 3. *Assuming that customers in queue Q^1 are served in the order they arrive (FCFS policy), for each $m \in \mathbb{N}_+$ we have that*

$$\mathbf{P}[D_0^1 \geq m] = \mathbf{P}[L_m^1 \geq S_{0,m-1}^{A^1}].$$

Proof. Consider a virtual customer arriving at the beginning of time slot 0 in Q^1. If $D_0^1 \geq m$ then the customer should be in the system at time slot $m-1$, and because Q^1 operates in a FCFS fashion, the queue length at time slot m, denoted by L_m^1 (recall that this does not includes arrivals and departures during time slot m), should include all the arrivals after the virtual customer. Thus, $D_0^1 \geq m$ implies $L_m^1 \geq S_{0,m-1}^{A^1}$. Hence $\mathbf{P}[D_0^1 \geq m] \leq \mathbf{P}[L_m^1 \geq S_{0,m-1}^{A^1}]$. Similarly, $L_m^1 \geq S_{0,m-1}^{A^1}$ implies that the customer arriving at the beginning of time slot 0 is still in the system at time slot $m-1$. □

We are interested in obtaining the probability $\mathbf{P}[D_0^1 \geq m]$, up to first degree in the exponent, for large values of m. Using stationarity, the above theorem implies

$$\mathbf{P}[D_{-m}^1 \geq m] = \mathbf{P}[L_0^1 \geq S_{-m,-1}^{A^1}]; \tag{12}$$

we will be using the latter expression to calculate the probability that the delay gets large.

To this end we will employ an approach developed in Bertsimas et al. (1998a). In particular, and in the standard large deviations methodology we will establish a lower and a matching (up to first degree in the exponent) upper bound on this probability. Consider all scenarios (paths) that lead to large delays (larger than m) in the first buffer. We will show that the probability $\mathbf{P}[\omega]$ of each such scenario ω asymptotically behaves as $\mathrm{e}^{-m\theta_{D,1}(\omega)}$, for some function $\theta_{D,1}(\omega)$. For every ω, $\mathbf{P}[\omega]$ is a lower bound on $\mathbf{P}[D_0^1 \geq m]$. We select the tightest lower bound by performing the minimization

$$\theta_{D,1}^* = \min_{\omega} \theta_{D,1}(\omega).$$

This amounts to solving a deterministic optimal control problem. 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.

The derivation of the upper bound on $\mathbf{P}[D_0^1 \geq m]$ is less intuitive and more technical; we will omit it in the interest of space and refer the interested reader to Paschalidis (1996). An alternative proof to the one appearing there can be obtained by employing the techniques developed in Dupuis and Ramanan (1997b) and formulate the problem as a Skorokhod problem as in Dupuis and Ramanan (1997a). It is then shown in Dupuis and Ramanan (1997b) that the optimal control (variational) problem we will solve to obtain the lower bounds provides the answer to the large deviations problem.

5.1. Delay: The optimal control problem

Consider a virtual customer arriving at time $-m$. Due to the stability condition (5), for every possible sample path that leads to large delay (except sample paths of measure zero), there exists some time $-n \leq -m$ at which both queues are empty. Since we are interested in the asymptotics as $m \to \infty$ we scale both time and the levels of the processes A^1, A^2 and B by m. In particular, we let $T = (n-m)/m$ and define the following continuous-time functions in $D[-1-T, 0]$:

$$\hat{L}^j(t) = \frac{1}{m} L^j_{\lfloor mt \rfloor}, \quad j = 1, 2,$$

$$S^X(t) = \frac{1}{m} S^X_{-m(1+T), \lfloor mt \rfloor}, \quad X \in \{A^1, A^2, B\}.$$

Notice that the empirical rate of a process X is roughly equal to the rate of growth of $S^X(t)$. 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 $[-(1+T), 0]$ for large values of m is given (up to first degree in the exponent) by

$$\exp\left\{ -m \int_{-T-1}^{0} [\Lambda_{A^1}^*(x_1(t)) + \Lambda_{A^2}^*(x_2(t)) + \Lambda_B^*(x_3(t))] \, \mathrm{d}t \right\}. \tag{13}$$

This cost functional is a consequence of Assumption B. With the scaling introduced here, as $m \to \infty$ the sequence of slopes $a_0, a_1, \ldots, a_{m-1}$ appearing there converges to the empirical rate $x(\cdot)$ and the sum of rate

functions appearing in the exponent (see Eq. (3)) converges to an integral. Similarly, a "polygonal approximation" to $\hat{L}^j(t)$ (see Dembo & Zeitouni, 1993, Section 5.1; Dembo & Zajic, 1995) converges to some continuous functions $L^j(t)$, for $j = 1, 2$.

The empirical rates $x_1(t), x_2(t)$ and $x_3(t)$, along with T, characterize a particular scenario (path) of achieving $D_0^1 \geq m$. The probability of such a scenario is a lower bound on the delay probability $\mathbf{P}[D_0^1 \geq m]$, and is given by the expression in (13). To obtain a tight lower bound, we seek a path with maximum probability, i.e., a minimum cost path where the cost functional is given by the integral in the exponent of (13). This optimization is subject to the constraints $L^1(-1-T) = L^2(-1-T) = 0$ and

$$L^1(0) > \int_{-1}^{0} x_1(t)\,\mathrm{d}t.$$

The latter constraint guarantees (cf. Eq. (12)) that we have a large delay, i.e., $D_0^1 \geq m$. 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 on the particular scheduling policy that is implemented. In the GPS case, depending on which queue is empty, there are three regions of the state space with different set of dynamics in each region. In particular we have:

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

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

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

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

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

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

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

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

We next formally define the following deterministic optimal control problem which will be referred to as (GPS-DELAY):

$$\begin{aligned}
\text{minimize} \quad & \int_{-T-1}^{0} [\Lambda_{A^1}^*(x_1(t)) + \Lambda_{A^2}^*(x_2(t)) \\
& + \Lambda_B^*(x_3(t))]\,\mathrm{d}t \qquad (14)\\
\text{subject to} \quad & L^1(-T-1) = L^2(-T-1) = 0, \\
& L^1(0) > \int_{-1}^{0} x_1(t)\,\mathrm{d}t, \\
& L^2(0)\text{: free}, \qquad T\text{: free}, \\
& \{L^j(t)\text{: } t \in [-T-1, 0], j = 1, 2\} \\
& \in (\text{GPS-}DYNAMICS).
\end{aligned}$$

To solve the above problem we *decompose* it into the two time intervals $[-1-T, -1]$ and $[-1, 0]$. First note that for all $t \in [-1, 0]$ we have

$$\int_{-1}^{0} x_1(\tau)\,\mathrm{d}\tau < L^1(0) \leq L^1(t) + \int_{t}^{0} x_1(\tau)\,\mathrm{d}\tau \leq L^1(t) + \int_{-1}^{0} x_1(\tau)\,\mathrm{d}\tau,$$

which implies

$$L^1(t) > 0, \quad \forall t \in [-1, 0]. \qquad (15)$$

Thus, the state trajectory in the interval $[-1, 0]$ does not touch the L^2-axis in the $L^1 - L^2$ space. Let now $L^{1*}(-1)$ denote the level of buffer Q^1 at time -1 in the optimal trajectory. The problem in the time interval $[-1-T, -1]$ can be interpreted as an "overflow" problem for buffer Q^1, i.e., optimally reach the value $L^{1*}(-1)$ starting from an empty system. In particular we denote this problem by (GPS-OVERFLOW) and is formulated as

$$\begin{aligned}
\text{minimize} \quad & \int_{-T-1}^{-1} [\Lambda_{A^1}^*(x_1(t)) + \Lambda_{A^2}^*(x_2(t)) \\
& + \Lambda_B^*(x_3(t))]\,\mathrm{d}t \qquad (16)\\
\text{subject to} \quad & L^1(-T-1) = L^2(-T-1) = 0, \\
& L^1(-1) = L^{1*}(-1), \quad L^2(0)\text{: free}, \quad T\text{: free}, \\
& \{L^j(t)\text{: } t \in [-T-1, -1], j = 1, 2\} \\
& \in (\text{GPS-DYNAMICS}).
\end{aligned}$$

In fact, as shown in Bertsimas et al. (1997), the above is exactly the optimal control problem corresponding to large deviations of the queue length process, with the only exception that the final value is constrained to be

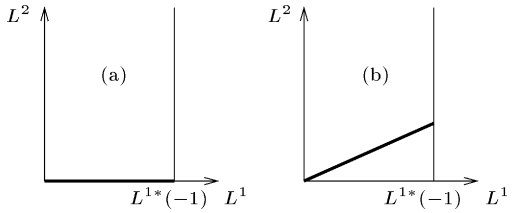


Fig. 2. Optimal state trajectories for (GPS-OVERFLOW).

1 instead of $L^{1*}(-1)$. Its optimal value is $\theta^*_{L,1}$ as defined in the statement of Theorem 1. In Bertsimas et al. (1997) it is also shown that the optimal state trajectory of (GPS-OVERFLOW) must be of one out of the two possible types depicted in Fig. 2. Moreover, there is an one-to-one correspondence between the most-likely modes of overflow described in Section 4 and these optimal state trajectories.

We next focus on the time interval $[-1,0]$. We will need the following lemma which is proved in Bertsimas et al. (1997, 1998a) based on the convexity of the large deviation rate functions $\Lambda^*_{A^1}(\cdot), \Lambda^*_{A^2}(\cdot)$ and $\Lambda^*_B(\cdot)$.

Lemma 4. *Fix a time interval* $[-T_1, -T_2]$. *Consider a segment of a control trajectory* $\{x_1(t), x_2(t), x_3(t); t \in [-T_1, -T_2]\}$, *achieving cost* V, *such that the corresponding state trajectory* $\{L^1(t), L^2(t); t \in (-T_1, -T_2)\}$ *stays in one of the regions* A, B, *or* C. *Then there exist scalars* $\bar{x}_1$, $\bar{x}_2$ *and* $\bar{x}_3$ *such that the segment of the control trajectory* $\{x_1(t) = \bar{x}_1, x_2(t) = \bar{x}_2, x_3(t) = \bar{x}_3; t \in [-T_1, -T_2]\}$ *achieves cost at most* V, *with the same corresponding states at* $t = -T_1$ *and* $t = -T_2$.

This result suggests that *optimal control trajectories can be taken to be constant* within each of the three regions of state dynamics. Thus, depending on the form of the segment of the state trajectory in $[-1-T, -1]$ we distinguish two different sets of candidates for optimality. These are depicted in Fig. 3. For candidates belonging to Set I (Set II, respectively), the segment of the state trajectory in $[-1-T, -1]$ has the form of Fig. 2(a) (Fig. 2(b), respectively).

Let us first examine the state trajectories in Set I. Consider the trajectory in Fig. 3(b). Let y_j and x_j, $j = 1, 2, 3$, be the controls in the time intervals $[-1-T, -1]$ and $[-1, 0]$, respectively. We have

$$y_2 \le \phi_2 y_3,$$

$$x_2 \ge \phi_2 x_3,$$

$$T(y_1 + y_2 - y_3) + (x_1 - \phi_1 x_3) \ge x_1,$$

which implies

$$y_2 \le \phi_2 y_3, \tag{17}$$

$$x_2 \ge \phi_2 x_3, \tag{18}$$

$$T(y_1 + y_2 - y_3) \ge \phi_1 x_3. \tag{19}$$

We now claim that $x_3 \ge y_3$. To show this we assume that $x_3 < y_3$ and we will arrive at a contradiction. With $x_3 < y_3$, and for small $\varepsilon > 0$, we increase x_3 to $x_3 + \varepsilon$ and decrease y_3 to $y_3 - \varepsilon/T$, such that the total number of services in $[-1-T, 0]$ stays constant. Note that constraint (19) is not violated since $T(y_1 + y_2 - y_3) + \varepsilon \ge \phi_1 x_3 + \phi_1 \varepsilon$. Also, due to convexity the cost is decreased. We can keep doing this until one of constraints (17) or (18) is violated. This however contradicts the initial assumption that the trajectory has the form of Fig. 3(b). Thus, we conclude that $x_3 \ge y_3$. This implies that $y_2 \le x_2$ since $y_2 \le \phi_2 y_3 \le \phi_2 x_3 \le x_2$. For small $\varepsilon > 0$, we can now keep increasing y_2 to $y_2 + \varepsilon/T$, and decreasing x_2 to $x_2 - \varepsilon$, without violating (19), until one of constraints (17) or (18) is violated. This also contradicts the initial assumption that the trajectory has the form of Fig. 3(b). We finally conclude that we can exclude the trajectory in Fig. 3(b) from our search for optimality. The same argument also excludes the trajectory in Fig. 3(c)

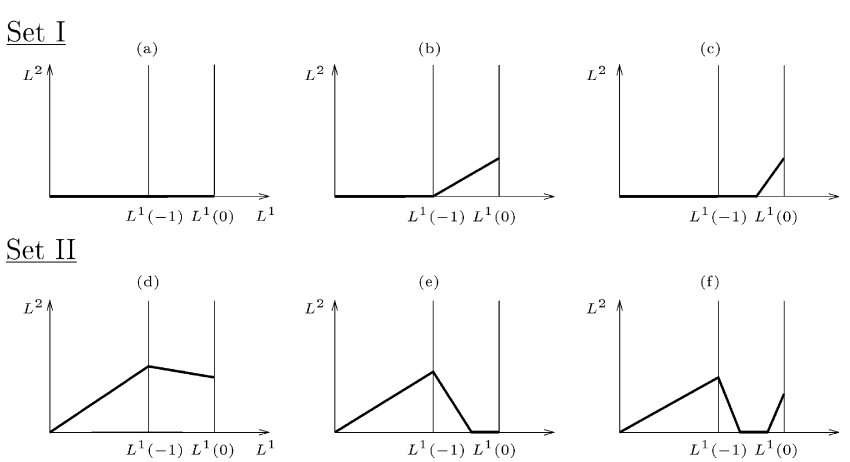


Fig. 3. Candidates for optimal state trajectories of (GPS-DELAY). From Set I, candidates for optimal trajectories are reduced to case (a). From Set II, candidates for optimal trajectories are reduced to case (d).

from this search. Hence, from trajectories in Set I, candidates for optimality are restricted to trajectories of the form of Fig. 3(a).

We next examine trajectories in Set II. Consider the trajectory in Fig. 3(e). Let $-(1-\zeta)$ the time that this trajectory hits the L^1-axis in the interval $[-1,0]$. Let y_i, $i=1,2,3$, be the rates during $[-1-T,-1]$ and x_i, $i=1,2,3$, the rates during $[-1,-(1-\zeta)]$. By taking the time average over the controls in the interval $[-1-T,-(1-\zeta)]$ we obtain constant controls during this interval. Let $\bar{y}_2$ and $\bar{y}_3$ be the arrival rate in the second buffer and the service rate, respectively, during the same interval. From the form of the trajectory we should have $(T+\zeta)(\bar{y}_2-\phi_2\bar{y}_3)=0$, which implies $\bar{y}_2=\phi_2\bar{y}_3$. Thus, the trajectory reduces to the one in Fig. 3(a). The same argument applies in the trajectory in Fig. 3(f) which reduces to the one in Fig. 3(c). Hence, from trajectories in Set II, candidates for optimality are restricted to trajectories of the form of Fig. 3(d). We summarize this discussion in the following proposition.

Proposition 5. *The state trajectories in Figs.* 3(a) *and* (d) *are optimal.*

5.1.1. Optimal value of (*GPS-DELAY*)

Next, we calculate the optimal value of the control problem (GPS-DELAY). The result of the above proposition allows us to consider only trajectories of the form of Figs. 3(a) and (d). Consider first the former. Let y_i, and x_i, $i=1,2,3$, be the rates during the time intervals $[-1-T,-1]$ and $[-1,0]$, respectively. The feasibility constraints are

$$y_2 \le \phi_2 y_3,$$

$$x_2 \le \phi_2 x_3,$$

$$T(y_1+y_2-y_3)+(x_2-x_3)\ge 0.$$

Taking the time average for x_2, y_2 (i.e., $(1+T)\bar{x}_2 = Ty_2+x_2$) and for x_3, y_3 (i.e., $(1+T)\bar{x}_3 = Ty_3+x_3$), we improve the cost and we obtain

$$\bar{x}_2 \le \phi_2 \bar{x}_3, \tag{20}$$

$$Ty_1+(1+T)(\bar{x}_2-\bar{x}_3)\ge 0. \tag{21}$$

Therefore for trajectories of the form of Fig. 3(a) the optimal cost is

$$\theta^{\mathrm{I}*}_{D,1} = \inf_T \inf_{\substack{\bar{x}_2\le\phi_2\bar{x}_3\\ Ty_1+(1+T)(\bar{x}_2-\bar{x}_3)\ge 0}} [T\Lambda^*_{A^1}(y_1)+\Lambda^*_{A^1}(x_1)+(1+T)(\Lambda^*_{A^2}(\bar{x}_2)+\Lambda^*_B(\bar{x}_3))].$$

Notice in the optimization problem above we can take $x_1=\mathbf{E}[A^1]$, making $\Lambda^*_{A^1}(x_1)=0$. We next manipulate the above expression, using convex duality, to arrive at a more compact formula. Let us first define

$$\Lambda^{\mathrm{I}}_{\mathrm{GPS},1}(\theta) \triangleq \Lambda_{A^1}(\theta) + \inf_{u\ge 0}[\Lambda_{A^2}(\theta-u)+\Lambda_B(-\theta+u\phi_2)], \tag{22}$$

and

$$\Lambda^{\mathrm{II}}_{GPS,1}(\theta) \triangleq \Lambda_{A^1}(\theta) + \inf_{u\le \theta}[\Lambda_{A^2}(\theta-u)+\Lambda_B(-\theta+u\phi_2)], \tag{23}$$

which, as it can be easily verified, are the convex duals of $\Lambda^{\mathrm{I}*}_{\mathrm{GPS},1}(\cdot)$ and $\Lambda^{\mathrm{II}*}_{\mathrm{GPS},1}(\cdot)$ (cf. Eqs. (9) and (10)), respectively. We then have

$$\begin{aligned}
\theta^{\mathrm{I}*}_{D,1} &= \inf_T \Bigg[- \sup_{\substack{(1+T)\bar{x}_2\le(1+T)\phi_2\bar{x}_3\\ Ty_1+(1+T)(\bar{x}_2-\bar{x}_3)\ge 0}} \\
&\qquad [-T\Lambda^*_{A^1}(y_1)-(1+T)(\Lambda^*_{A^2}(\bar{x}_2)+\Lambda^*_B(\bar{x}_3))]\Bigg] \\
&= \inf_T \Bigg[- \inf_{u_1,u_2\ge 0} \sup[u_1(1+T)\phi_2\bar{x}_3 - u_1(1+T)\bar{x}_2 \\
&\qquad + u_2(1+T)(\bar{x}_2-\bar{x}_3) + u_2Ty_1 - T\Lambda^*_{A^1}(y_1) \\
&\qquad - (1+T)(\Lambda^*_{A^2}(\bar{x}_2)+\Lambda^*_B(\bar{x}_3))]\Bigg] \\
&= \inf_T \Bigg[- \inf_{u_1,u_2\ge 0} [T\Lambda_{A^1}(u_2)+(1+T)(\Lambda_{A^2}(u_2-u_1) \\
&\qquad + \Lambda_B(-u_2+u_1\phi_2))]\Bigg] \\
&= \inf_T \Bigg[- \inf_{u_2\ge 0} [T\Lambda_{A^1}(u_2)+(1+T)(\Lambda^I_{\mathrm{GPS},1}(u_2) \\
&\qquad - \Lambda_{A^1}(u_2))]\Bigg] \\
&= \inf_T \sup_{u_2\ge 0} [\Lambda_{A^1}(u_2)-(1+T)\Lambda^I_{\mathrm{GPS},1}(u_2)].
\end{aligned} \tag{24}$$

We next consider the trajectory of Fig. 3(d). We again let y_i, and x_i, $i=1,2,3$, be the rates during the time intervals $[-1-T,-1]$ and $[-1,0]$, respectively. The feasibility constraints are

$$y_2 \ge \phi_2 y_3,$$

$$x_2 \ge \phi_2 x_3,$$

$$T(y_1-\phi_1 y_3)+(x_1-\phi_1 x_3)\ge x_1.$$

Taking the time average for x_2, y_2 (i.e., $(1+T)\bar{x}_2 = Ty_2 + x_2$) and for x_3, y_3 (i.e., $(1+T)\bar{x}_3 = Ty_3 + x_3$), we improve the cost and we obtain

$$\bar{x}_2 \geq \phi_2 \bar{x}_3, \tag{25}$$

$$Ty_1 \geq (1+T)\phi_1 \bar{x}_3. \tag{26}$$

Therefore for trajectories of the form of Fig. 3(d) the optimal cost is

$$\theta^{\mathrm{II}*}_{D,1} = \inf_{T} \inf_{\substack{\bar{x}_2 \geq \phi_2 \bar{x}_3 \\ Ty_1 \geq (1+T)\phi_1 \bar{x}_3}} [T\Lambda^*_{A^1}(y_1) + \Lambda^*_{A^1}(x_1) + (1+T)(\Lambda^*_{A^2}(\bar{x}_2) + \Lambda^*_B(\bar{x}_3))].$$

Notice again in the optimization problem above we can take $x_1 = \mathbf{E}[A^1]$, making $\Lambda^*_{A^1}(x_1) = 0$. As in Eq. (24), we manipulate the above expression and after some algebra we obtain

$$\theta^{\mathrm{II}*}_{D,1} = \inf_{T} \sup_{u_2 \geq 0} [\Lambda_{A^1}(u_2) - (1+T)\Lambda^{\mathrm{II}}_{\mathrm{GPS},1}(u_2)]. \tag{27}$$

Hence the optimal value of (GPS-DELAY) is $\theta^*_{D,1} = \min(\theta^{\mathrm{I}*}_{D,1}, \theta^{\mathrm{II}*}_{D,1})$ which yields

$$\begin{aligned}\theta^*_{D,1} &= \min(\theta^{\mathrm{I}*}_{D,1}, \theta^{\mathrm{II}*}_{D,1}) \\ &= \inf_{T} \sup_{u_2 \geq 0} [\Lambda_{A^1}(u_2) - (1+T)\Lambda_{\mathrm{GPS},1}(u_2)] \\ &= \sup_{u_2 \geq 0 : \Lambda_{\mathrm{GPS},1}(u_2) < 0} [\Lambda_{A^1}(u_2) - \Lambda_{GPS,1}(u_2)],\end{aligned} \tag{28}$$

by defining $\Lambda_{\mathrm{GPS},1}(\theta) \triangleq \max[\Lambda^{\mathrm{I}}_{GPS,1}(\theta), \Lambda^{\mathrm{II}}_{GPS,1}(\theta)]$. Notice that the latter is consistent with the definition given in Eq. (11). We have proved the following theorem.

Theorem 6. *The optimal value, $\theta^*_{D,1}$, of the control problem (GPS-DELAY) is given by the following expression*:

$$\theta^*_{D,1} = \sup_{u \geq 0 : \Lambda_{\mathrm{GPS},1}(u) < 0} [\Lambda_{A^1}(u) - \Lambda_{\mathrm{GPS},1}(u)].$$

As outlined in the beginning of this section, the solution to the control problem provides a lower bound on the probability of large delay and the optimal trajectories identify the *most likely* ways that large delays occur. This lower bound along with a matching upper bound proved in Paschalidis (1996) establish the following theorem which is the main result of this Section.

Theorem 7 (Delay). *In the two-class system under the GPS policy, assuming that the arrival and service processes satisfy Assumption A–C, the steady-state delay, D^1, of queue Q^1 satisfies*

$$\lim_{m \to \infty} \frac{1}{m} \log \mathbf{P}[D^1 \geq m] = -\theta^*_{D,1},$$

where

$$\theta^*_{D,1} = \sup_{u_2 \geq 0 : \Lambda_{\mathrm{GPS},1}(u_2) < 0} [\Lambda_{A^1}(u_2) - \Lambda_{\mathrm{GPS},1}(u_2)]$$

and where $\Lambda_{\mathrm{GPS},1}(\cdot)$ is as defined in Theorem 2.

We conclude our analysis of the two-class case by noting that due to symmetry the results can be easily adapted to cover overflows and delays in buffer Q^2 as well. To this end, it suffices to substitute $\phi_1 := 1 - \phi_1$ in the above formulas and swap A^1 with A^2.

6. Extensions to the multiclass case and refinements

The general multiclass problem is particularly hard since there is an exponential explosion of the number of overflow (or delay) modes. As a consequence, no full large deviations results have been obtained for this case (special cases have been addressed in Bertsimas et al., 1997; de Veciana & Kesidis, 1995; Courcoubetis & Weber, 1995b; Massoulié, 1998; O'Connell, 1995; Zhang, 1997). We will therefore, resort to an approximation of the overflow and delay probabilities, which is based on our two-class results. We will provide some analytical and numerical evidence that the approximation is a good one.

Let us focus on the overflow and delay probability in the first buffer Q^1. To approximate the performance of the multiclass system we consider a corresponding two-class system where the input to the first buffer is identical to the process $\{A^1; i \in \mathbb{Z}\}$, and the input to the second buffer, to be denoted by $\hat{A}^2$, is equal to the aggregate of the processes $A^2, \ldots, A^M$. By defining all the input processes $A^1, \ldots, A^M$ and the service process B in the two systems on a common probability space, we can assume that the actual arrivals and services are the same in the two systems. Consider a busy period of the multiclass system that leads to a Q^1 overflow. If during this busy period all buffers are nonempty, then buffer Q^1 receives just its allocated fraction ϕ_1 of the capacity; the same is true in the corresponding two-class system, since the total number of bits in $Q^2, \ldots, Q^M$ will also be nonzero. If during this busy period of overflow all buffers $Q^2, \ldots, Q^M$ remain empty in the multiclass system, then Q^1 gets the excess capacity; in the corresponding two-class system the evolution of Q^1 is identical since it receives exactly the same amount of capacity. If however, during this busy period of overflow only some of the buffers $Q^2, \ldots, Q^M$ remain empty in the multiclass system, buffer Q^1 receives a fraction of the excess capacity. Under the same conditions, in the two class-system it might be the case that the total number of bits in $Q^2, \ldots, Q^M$ is nonzero, which implies that Q^1 receives only its fraction ϕ_1 of the capacity. On the other hand, in the latter system the second buffer receives more capacity than all the buffers

$Q^2, \ldots, Q^M$ in the multiclass system, which implies that it has a shorter busy period and when it gets empty Q^1 will receive all the excess capacity. That is, there are times that Q^1 receives more capacity in the two-class system. Intuitively, in the two-class system Q^1 receives capacity in a more bursty manner. The following proposition compares the overflow probability of buffer Q^1 in the two systems in some special cases.

Proposition 8. *Assume that the arrival and service processes satisfy Assumption A–C. Let L^1 denote the steady-state queue length in the first buffer of the multiclass system, and $\hat{L}^1$ the same quantity in the corresponding two-class system.*

(1) *If $\sum_{j=2}^{M} \mathbf{E}[A^j] \geq \sum_{j=2}^{M} \phi_j \mathbf{E}[B]$ then*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 \geq U] \leq \lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[\hat{L}^1 \geq U]. \quad (29)$$

(2) *If the capacity is deterministic, i.e., B_i is equal to some constant almost surely (a.s.) for all i, and $\mathbf{E}[A^j] \leq \phi_j \mathbf{E}[B]$ for all $j = 1, \ldots, M$ then*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 \geq U] \leq \lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[\hat{L}^1 \geq U]. \quad (30)$$

(3) *If the capacity is deterministic, $\mathbf{E}[A^j] \leq \phi_j \mathbf{E}[B]$ for all $j = 1, \ldots, M$, $\phi_2 = \cdots = \phi_M$, and $\Lambda^*_{A^2}(x) = \cdots = \Lambda^*_{A^M}(x)$ for all $x \in \mathbb{R}$, then*

$$\lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[L^1 \geq U] = \lim_{U\to\infty} \frac{1}{U} \log \mathbf{P}[\hat{L}^1 \geq U]. \quad (31)$$

Proof. To show Part 1, let us focus on the queue length L^1_0 at time 0 in the multiclass system. We consider a busy period of the first queue, Q^1, that starts at some time $-n^* \leq 0$ ($L^1_{-n^*} = 0$) and has not ended until time 0. Notice that due to the stability condition (5) and the fact $\sum_{j=2}^{M} \mathbf{E}[A^j] \geq \sum_{j=2}^{M} \phi_j \mathbf{E}[B]$, it is true that $\mathbf{E}[A^1] < \phi_1 \mathbf{E}[B]$, which implies that such a time $-n^*$ exists w.p.1. We will focus on sample paths of the multiclass system in $[-n^*, 0]$ that lead to $L^1_0 > U$. Note that

$$L^1_0 \leq S^{A^1}_{-n^*,-1} - \phi_1 S^B_{-n^*,-1}. \quad (32)$$

Thus,

$$\mathbf{P}[L^1_0 > U] \leq \mathbf{P}[\exists n \geq 0 \text{ s.t. } S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1} > U]$$

$$\leq \mathbf{P}\left[\max_{n\geq 0} (S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1}) > U\right]. \quad (33)$$

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

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

$$\leq \sum_{n\geq 0} e^{n(\Lambda_{A^1}(\theta) + \Lambda_B(-\phi_1\theta) + \varepsilon)}$$

$$\leq K(\theta, \varepsilon) \quad \text{if } \Lambda_{A^1}(\theta) + \Lambda_B(-\phi_1\theta) < 0, \quad (34)$$

since when the exponent is negative (for sufficiently small ε), the infinite geometric series converges to some $K(\theta, \varepsilon)$. We can now apply the Markov inequality in (33) to obtain

$$\mathbf{P}[L^1_0 > U] \leq \mathbf{E}[e^{\theta \max_{n\geq 0}(S^{A^1}_{-n,-1} - \phi_1 S^B_{-n,-1})}] e^{-\theta U}$$

$$\leq K(\theta, \varepsilon) e^{-\theta U} \quad \text{if } \Lambda_{A^1}(\theta) + \Lambda_B(-\phi_1\theta) < 0. \quad (35)$$

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

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

We are now left with proving that the right-hand side of the above is equal to to the exponent of the overflow probability in the two-class system, denoted $-\theta^*_{L,1}$, which can be obtained from Theorem 1 when A^2 is replaced by the superposition of $A^2, \ldots, A^M$. This is done in Bertsimas et al. (1997). Parts 2 and 3 are shown in Massoulié (1998). □

The performance of the corresponding two-class system can be easily obtained by applying the two-class results, that is, Theorems 1 and 7. The statistics of the aggregate process $\hat{A}^2$ are characterized by

$$\Lambda_{\hat{A}^2}(\theta) = \sum_{j=2}^{M} \Lambda_{A^j}(\theta), \quad (36)$$

and

$$\Lambda^*_{\hat{A}^2}(x) = \inf_{\sum_{j=2}^{M} x_j = x} \sum_{j=2}^{M} \Lambda^*_{A^j}(x_j). \quad (37)$$

The first expression is easily obtained from the definition of the limiting log-moment generating function, and the latter one by standard convex duality properties (see Rockafellar, 1970).

6.1. Refinements of the asymptotics

It has been observed that the asymptotic $\exp(-U\theta^*_{L,j})$ (resp. $\exp(-m\theta^*_{D,j})$) might not always yield a very accurate approximation of the overflow probability (resp. delay probability) for class j. A refinement of this asymptotic can be obtained by introducing a constant in front

of the exponential, that is, for all j using the following expressions

$$\mathbf{P}[L^j \geq U] \sim \alpha_{L,j}\, \mathrm{e}^{-U\theta^*_{L,j}}, \tag{38}$$

$$\mathbf{P}[D^j \geq m] \sim \alpha_{D,j}\, \mathrm{e}^{-m\theta^*_{D,j}} \tag{39}$$

for the overflow and delay probabilities, respectively.

An estimate of the constants $\alpha_{L,j}$ and $\alpha_{D,j}$ can be obtained by using an idea from Abate, Choudhry and Whitt (1995) and assuming that the above expressions provide the *exact* distribution of the queue length and delay, respectively. Matching the expectation of the distributions in (38) and (39) with $\mathbf{E}[L^j]$ and $\mathbf{E}[D^j]$, respectively, we obtain

$$\alpha_{L,j} = \theta^*_{L,j}\mathbf{E}[L^j], \tag{40}$$

and

$$\alpha_{D,j} = \theta^*_{D,j}\mathbf{E}[D^j]. \tag{41}$$

Thus, to find the asymptotic constant we need the expectations of the queue length and delay processes, which can be obtained by simulation or direct measurements. As it will become apparent in Section 7 to implement the approach proposed in this paper one needs to obtain a model for the demand and service processes from real data (e.g., via on-line estimation). Thus, simulating or observing in the actual system the expectations $\mathbf{E}[L^j]$ and $\mathbf{E}[D^j]$ incurs no substantial additional computational cost. The interesting point here is that although small tail probabilities are very hard to reliably simulate or observe, it is computationally easy to simulate or observe expectations of queue lengths and delays.

We next present a numerical example indicating that the multiclass approximation proposed in this section combined with the above refinements of the asymptotics is accurate. We simulated a three-class system and compared the simulated overflow probabilities in the three buffers with the analytical approximations as presented in this section (i.e., two-class approximation of the multiclass system combined with the proposed refinements of the asymptotics). We used the GPS policy with parameters $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3) = (0.3, 0.3, 0.4)$. All arrival and service processes are Markov-modulated processes (see Fig. 4). The results are reported in Table 1. The analytical asymptotic appears to be capturing the order of magnitude of the probability very accurately; the first

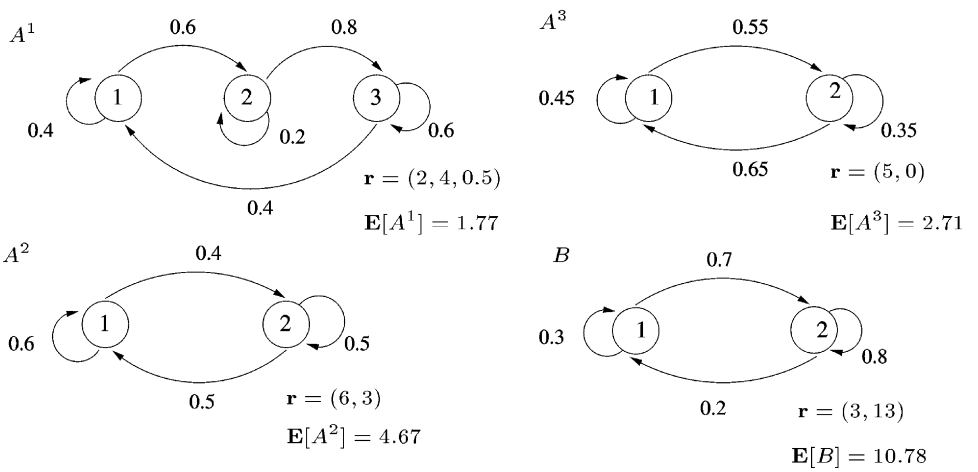


Fig. 4. The models for arrival and service processes in the three-class example. By $\mathbf{r}$ we denote the vector of bits arriving (or departing) per time slot at each state of the corresponding Markov chain. The Markov chains make one transition per time slot.

Table 1
Comparing the analytical results (cf. Eq. (38)) with simulation results for the three-class example of Fig. 4 and for various buffer sizes U

	Class 1			Class 2			Class 3	
U	Anal.	Simul.	U	Anal.	Simul.	U	Anal.	Simul.
7	1.8×10^{-3}	1.5×10^{-3}	60	1.7×10^{-3}	1.2×10^{-3}	14	1.6×10^{-3}	1.8×10^{-3}
9	4.4×10^{-4}	3.3×10^{-4}	70	6.6×10^{-4}	4.5×10^{-4}	18	3.0×10^{-4}	3.9×10^{-4}
11	8.7×10^{-5}	7.4×10^{-5}	80	2.5×10^{-4}	1.6×10^{-4}	22	6.3×10^{-5}	8.4×10^{-5}
13	2.0×10^{-5}	1.6×10^{-5}	90	9.8×10^{-5}	6.3×10^{-5}	26	1.3×10^{-5}	1.8×10^{-5}
15	4.3×10^{-6}	3.6×10^{-6}	110	1.4×10^{-5}	9.0×10^{-6}	30	2.8×10^{-6}	3.9×10^{-6}
17	9.1×10^{-7}	7.9×10^{-7}	120	5.6×10^{-6}	3.3×10^{-6}	34	5.7×10^{-7}	8.6×10^{-7}
18	4.8×10^{-7}	3.7×10^{-7}	140	8.4×10^{-7}	5.2×10^{-7}	38	1.2×10^{-7}	1.8×10^{-7}

significant digit also appears to be close to the simulated value. Since the asymptotic for the delay probability is similar in nature we expect that it has similar accuracy.

7. Admission control

In this section we propose an admission control approach based on the performance analysis results developed so far. The objective is to develop a *call admission algorithm* that provides both *loss* and *delay* per class QoS guarantees.

Consider the architecture of Fig. 1 and recall the notation introduced in Section 3. With N_j class j calls admitted, the aggregate arrival process A^j in the jth buffer Q^j is characterized by $\Lambda_{A^j}(\theta) = N_j \Lambda_{\tilde{A}^j}(\theta)$, for all θ and j. To make things simpler, let us assume that the service process is deterministic with rate c bits/s (b/s), although our analytical results allow us to handle stochastic capacities. Notice that $\Lambda_B(\theta) = c\theta$ for all θ. The objective is to satisfy the QoS constraints in Eqs. (6) and(7).

The admission controller has the freedom to:

(1) select the appropriate buffer sizes U_j, and
(2) restrict the number of admitted calls,

in order to guarantee these QoS constraints. We next argue that the appropriate buffer sizes are $U_j = cD^j_{\max}$. Consider first setting U_j to a value larger than $cD^j_{\max}$. This implies that the system will be accepting bits for transmission that require more than $D^j_{\max}$ to clear their corresponding buffer j. Such bits severely degrade the performance, therefore, they are of no use in the receiving end and could be discarded up front. Consider next setting the buffer to a value smaller than $cD^j_{\max}$. This implies that the system will be "blindly" discarding bits that have a possibility of clearing their corresponding buffer j within $D^j_{\max}$ of their arrival,[2] i.e., without checking whether the given QoS constraints are satisfied or not. We therefore, choose to set buffer sizes to $U_j = cD^j_{\max}$, for all j, and leave to the admission controller to enforce the QoS constraints. Of course, if the cost to accommodate these buffer requirements is an issue, one can alternatively select a fraction of the proposed buffer sizes at some potential efficiency cost.

Hereafter, as we have indicated in Section 3 we will use our analytical results developed for the infinite buffer system, although we are in reality dealing with a finite buffer system. We have argued in Section 3 that the level crossing probability in the infinite buffer system, which we have analytically estimated, closely approximates the loss probability in the finite buffer system (in the exponential scale we are considering in this paper). Moreover, the delay probability in the infinite buffer system is a tight upper bound (in the exponential scale) on the delay probability in the finite buffer system, which implies that we can guarantee the latter by guaranteeing the former.

A useful observation is that with buffer sizes at $U_j = cD^j_{\max}$ the event $L^j \geq U_j$ implies the event $D^j \geq D^j_{\max}$, thus,

$$\mathbf{P}[L^j \geq U_j] \leq \mathbf{P}[D^j \geq D^j_{\max}], \quad j = 1, 2, \ldots, M. \tag{42}$$

Hence, to provide the QoS guarantees[3] we only need to guarantee the delay probability constraint (Eq. (7)). Using the asymptotic in (39), we can ensure the delay QoS constraint in (7) if and only if $\theta^*_{D,j} \geq \delta^j_D$, where $\delta^j_D \triangleq -\log(\delta_j/\alpha_{D,j})/D^j_{\max}$.

We proceed with defining the notion of admission region. Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_M)$, and $\mathbf{N} = (N_1, \ldots, N_M)$.

Definition 9. We will call *admission region* for the system of Fig. 1 operated under the GPS policy the set

$$\mathscr{A} \triangleq \Big\{ (\boldsymbol{\phi}, \mathbf{N}) \,|\, \phi_j \in [0,1],\ \sum_{j=1}^{M} \phi_j = 1,\ N_j \in \mathbb{N}_+,\ \theta^*_{D,j} \geq \delta^j_D,\ j = 1, \ldots, M \Big\}$$

If a vector $(\boldsymbol{\phi}, \mathbf{N}) \in \mathscr{A}$, we can ensure that the delay QoS constraint (7) is satisfied, and by virtue of (42), the loss QoS constraint (6) is satisfied as well.

Based on this definition, the proposed *admission control algorithm* takes the following form: (assume without loss of generality a class 1 call request)

if there exists $\boldsymbol{\phi}$ such that $(\boldsymbol{\phi}, \mathbf{N} + \mathbf{e}_1) \in \mathscr{A}$
  then *accept*;
  else *reject*;
end,

where $\mathbf{e}_1 = (1, 0, \ldots, 0)$.

An issue of practical interest is whether admission decisions can be taken in a very short period of time (e.g., on the order of seconds) from the time admission requests are placed (we will refer to this as *real-time* operation). We expect the calculations required to determine the admission region to be computationally burdensome, depending on the complexity of the arrival model (recall a nonlinear optimization problem has to be solved to determine $\theta^*_{D,j}$). When an adequate traffic model for the arrival processes is available these calculations can be

[2] Note that this depends on what happens in the other buffers.

[3] If smaller buffer sizes are used then one needs to enforce both the overflow and delay QoS constraints, which can be done by using the overflow analytical result in addition to the one for delay which is used here.

performed off-line. Then, the admission control algorithm reduces to a simple look-up-table operation, which can be performed in real-time. However, such a traffic model is rarely available a priori, and has to be estimated from the actual traffic data in an on-line fashion. In this case, if we were to recalculate the admission region with every call request we would not have a real-time implementation. To remedy this we propose to avoid repeating these calculations so frequently. Instead, the system can store an appropriate description of the admission region and update it periodically. More specifically, admission decisions will be made in real-time by look-up-table operations based on the most current version of the admission region. The traffic model and the admission region can be updated in a longer time scale (e.g., on the order of minutes) than the one in which admission decisions will be taken.

We note here that traffic statistics are typically non-stationary in practice. We implicitly assume (and this can be validated from real observations in Duffield, Lewis, O'Connell, Russel & Toomey, 1994) that these statistics change in a much longer time scale (e.g., on the order of hours) than the one considered above. This justifies our approach of using steady-state analytical results for quantifying QoS.

7.1. *The admission region: an example*

Next, we provide an example of the admission region for two classes of traffic with traffic and QoS parameters given by Table 2.We will use this example to illustrate some of the advantages of the proposed approach, namely, that providing class-specific QoS guarantees leads to significant efficiency gains when compared to simpler "effective bandwidth" rules.

In this example, both classes of traffic conform to an *ON-OFF model*. Traffic is generated according to a continuous-time Markov process, with embedded Markov chain depicted in Fig. 5. In the ON state, traffic is produced with a constant rate of p bits/sec (b/s). We refer to this as the *peak rate*. In the OFF state no traffic is generated. The traffic source stays in the ON state a fraction $a/(a+b)$ of the time and for an expected number of $1/b$ transitions of the embedded Markov chain. It generates traffic with an average rate of $pa/(a+b)$ b/s. The capacity of the server is 135 Mb/s.

A few comments about the traffic and QoS parameters are in order. Class 1 traffic has parameters which are typical of a video-conferencing call which consists of the transmission of relatively low activity scenes (people sitting around a table). As a consequence the peak rate is close to the average rate. Class 2 traffic is more typical of a bursty video call (e.g., action movie). To put $D_{\max}$ into perspective, with a packet size of 53 bytes (size of an ATM cell) and with a 64 Kb/s rate for voice, the packetization delay is about 6 ms (for a discussion of typical QoS parameters see Hsu & Walrand, 1995).

Fig. 6 depicts the admission region for this particular example. For every fixed ϕ_1 and N_1 we plot the maximum allowed number of class 2 calls, N_2, such that $(\boldsymbol{\phi}, \mathbf{N}) \in \mathscr{A}$, i.e., as long as we operate the system under the plotted surface, the QoS constraints are satisfied. In Fig. 8 we show "waterfall" plots of the admission region to better depict the shape of the region for N_1 constant and for ϕ_1 constant (first and second plot in Fig. 8, respectively).

7.1.1. *Comparisons with alternative approaches*

Let us now compare the admission region generated here with (a) peak rate-based allocation, and (b) effective bandwidth-based allocation. In (a) we take a worst-case view and treat each source as if it is always transmitting with its peak rate. Hence, to guarantee absolutely no losses and no delays, the system should admit calls to satisfy $2N_1 + 10N_2 \leq 135$ (recall that the peak rates for the two service classes are 2 and 10 Mb/s, and the capacity 135 Mb/s). In (b) both classes are fed to a single buffer, hence, the QoS they receive is identical and is determined by the overflow and delay probability in this buffer. As a result, we need to enforce the most stringent QoS guarantees among the two classes, that is, $D_{\max} = 10$ ms (due to class 1) and $\delta = 10^{-9}$ (due to class 2). With these parameters, the single-class version of our admission control approach reduces to the effective bandwidth-based allocation. All three approaches are compared in Fig. 7. It is evident that peak-rate allocation can be dramatically inefficient. Moreover, providing class-specific guarantees leads to significant gain in

Table 2
Traffic Parameters for the ON-OFF model. $\mathbf{E}[t_{ON}]$ denotes the expected amount of time that the traffic source stays in the ON state. For both classes of traffic it can be easily verified that the embedded Markov chain makes one transition every 1 msec (ms)

	Traffic parameters					QoS parameters	
	Peak (Mb/s)	Avg. (Mb/s)	$\mathbf{E}[t_{ON}]$ (ms)	a	b	$D_{\max}$ (ms)	δ
Type 1	2	1	25	0.04	0.04	10	10^{-6}
Type 2	10	2	5	0.05	0.2	30	10^{-9}

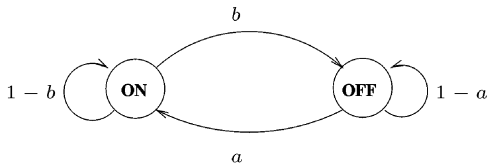


Fig. 5. The ON-OFF source model.

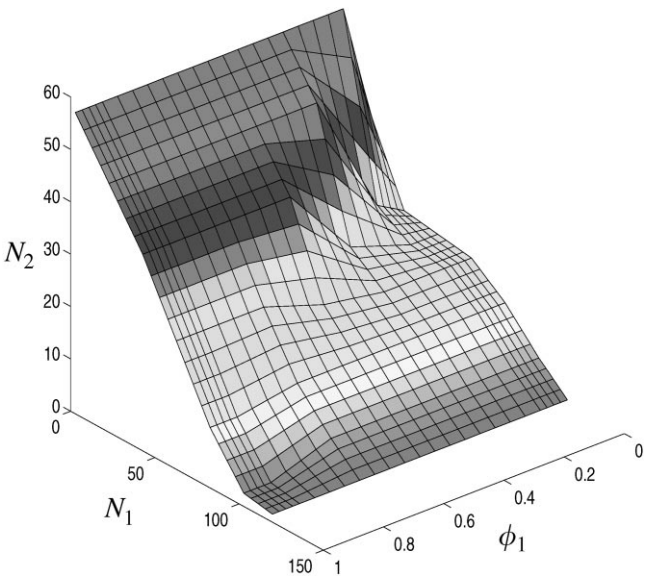


Fig. 6. The admission region for the traffic model and parameters of Section 7.1.

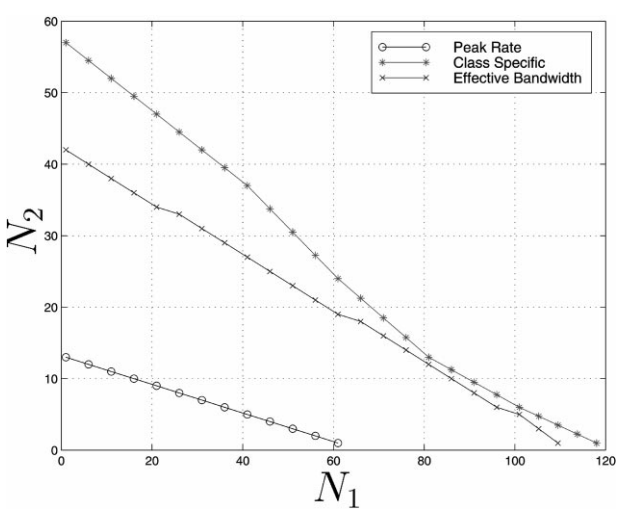


Fig. 7. Comparison with alternative approaches. The curve referred to as "class specific" in the figure is obtained by considering the surface in Fig. 6 and maximizing N_2 over ϕ_1 for each value of N_1.

efficiency over the effective bandwidth rule. The price to pay is additional complexity. Note, that the "class specific" curve in Fig. 6 is *no longer linear*, as the "effective bandwidth" curve.

Some further observations on the structure of the admission region are in order. Recall that the admission region is defined to satisfy both constraints

$$\theta^*_{D^1} \geq \delta^1_D, \tag{43}$$

$$\theta^*_{D^2} \geq \delta^2_D. \tag{44}$$

Consider the first plot of Fig. 8. Notice that for small values of N_1, the maximum allowed N_2 is non-decreasing as ϕ_1 increases in [0,1]. To explain this, notice that for large ϕ_1 we favor class 1 calls and since these are few constraint (43) is not tight. The maximum allowed N_2 is set such that (44) is tight. The situation stays the same (i.e., maximum allowed N_2 stays constant) as we decrease ϕ_1 until some threshold value ϕ_1^* at which (43) gets tight. For smaller ϕ_1 than ϕ_1^*, and since we keep N_1 fixed, to accommodate class 1 calls (i.e., satisfy (43)) we need to decrease N_2. An antipodal phenomenon, in the same plot, is occurring for large values of N_1. For small values of ϕ_1, (43) is tight while (44) is not. Increasing ϕ_1 more than some threshold value ϕ_1^* makes (44) tight and thus we can guarantee the QoS constraints only by dropping the maximum allowed N_2.

Let us now turn our attention to the second plot in Fig. 8, which depicts cross sections of the admission region for ϕ_1 fixed. Consider cross sections around $\phi_1 = 0.2$ to make the discussion clearer. We can distinguish roughly three regions: (a) small values of N_1, (b) moderate values of N_1, and (c) large values of N_1. In region (a) the maximum allowed N_2 drops almost linearly as we increase N_1. This is occurring because in this region (43) is not tight while (44) is tight and the only way to increase N_1, without compromising the quality of class 2 calls, is to decrease N_2. The decrease is roughly linear for the following reason: in this region the dominant congestion event is large delays in the second buffer, which are occurring according to the scenario depicted in Fig. 3(a). Recall that large delays are occurring because the second buffer builds up in the first part of that path (interval $[-1-T,-1]$ with the notation there). Since according to that path the first buffer stays roughly empty the second buffer gets capacity of $c - N_1 y_1$, where y_1 is the most likely arrival rate of class 1 calls during periods of congestion in the second buffer (y_1 is the solution of an optimization problem similar to the one appearing in Eq. (24)). To have (44) tight, this capacity should be equal to $N_2 y_2$, where y_2 is the most likely arrival rate of class 2 calls during periods of congestion in the second buffer. Thus $N_2 = (c - N_1 y_1)/y_2$ which is linear in N_1. Now, from region (a), as we keep increasing N_1 we enter region (b). Still (43) is not tight, however the most likely way that the second buffer generates large delays becomes the scenario of Fig. 3(d), that is by building up the first buffer also. This means that the first buffer requires capacity $\phi_1 c$ and to accommodate type 2 calls we have decreased their number such that they are

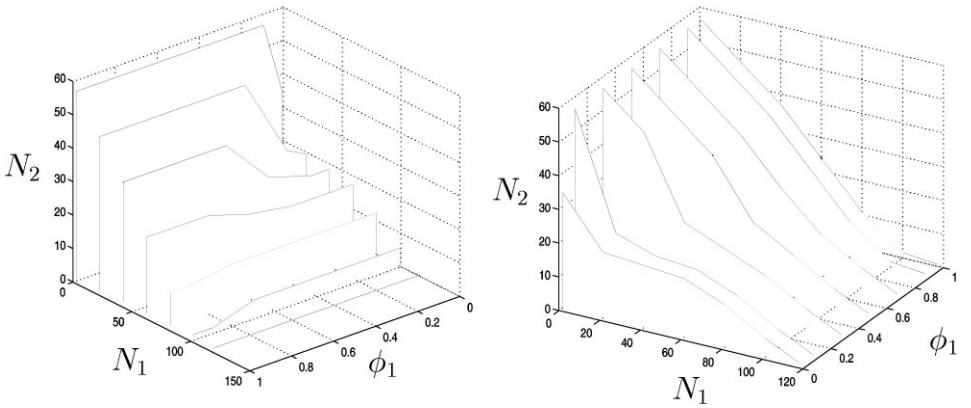


Fig. 8. Waterfall plots of the admission region for the traffic model and parameters of Section 7.1. The first (left) plot depicts how N_2 is changing with ϕ_1 for several values of N_1. The second (right) plot depicts how N_2 is changing with N_1 for several values of ϕ_1.

satisfied with capacity $\phi_2 c$. We can therefore increase N_1 even more, until we make (43) tight, without having to decrease N_2 (notice that in region (b) N_2 is roughly constant). When (43) becomes tight we enter region (c) and the only way to further increase N_1 is to drop N_2. The drop is roughly linear for a similar reason to the one explained above. The discussion extends to other values of ϕ_1 away from $\phi_1 = 0.2$, with the three regions mentioned above degenerating to two (see that for ϕ_1 around 1 we can distinguish only regions (a) and (b)).

8. End-to-end QoS guarantees

So far, in this paper, we have been considering the case of a single node. It is of course of interest to apply these control mechanisms in a *network environment* to provide *end-to-end* QoS guarantees.

Towards this goal, one possible direction is to develop performance analysis results for multiclass networks and to incorporate them in admission control decisions. There are both technical and practical problems with this approach. On the technical side, the network problem appears to be particularly hard, since in essence it is needed to obtain distributions of queue lengths and delays in a multiclass network of G/G/1 queues. This has been accomplished in single-class acyclic networks (Bertsimas, Paschalidis & Tsitsiklis, 1998b; Chang, 1995). Related work is reported in de Veciana, Courcoubetis and Walrand (1993) and Ganesh and Anantharam (1996). The multiclass case, however, appears much harder and no LDP results exist (in fact, some negative results have been reported in Ganesh & O'Connell, 1998). On the practical side, a network mechanism has to scale to the full range of speeds and administrative domains that communication networks (such as the Internet) span. Moreover, to ensure reliability and high-speed forwarding minimal "state" information should be kept at the internal nodes. This seems to suggest that sophisticated control mechanisms should be pushed at the edges (Jacobson, 1998).

In this light, the capacity of the node (equal to a constant c in the development of the admission control mechanism in Section 7) can be viewed as the capacity of a fixed bandwidth "pipe" (*virtual path —VP—* in ATM terminology) from origin to destination.[4] Buffering and admission control are done only at the edge of the network. This scheme, where control is pushed at the edge of the network, is *simple* to implement since it only requires from the network the ability to allocate fixed bandwidth "pipes".[5]

9. Conclusions

We have proposed a large deviations-based approach to QoS provisioning in multimedia communication networks. These networks carry a diverse set of applications at a variety of QoS grades. Since, "per-flow" QoS provisioning and accounting is not scalable in large network environments, applications are aggregated in a number of service classes. The emphasis of the proposed approach is to attend to the particular QoS needs of each class and provide class-dependent QoS guarantees.

We formulated the performance analysis problem of estimating overflow and delay probabilities as optimal control problems. The solution to these problems provided both the asymptotic exponent and a

[4] In fact, the capacity of the "pipe" can vary with time but in a longer time scale than call admission decisions, hence can be taken constant for these decisions.

[5] This capability is available both in ATM networks and can be done in the Internet (see Jacobson, 1998 on implementing "virtual leased lines").

characterization of the most-likely path for overflows or delays. The proposed admission control approach uses these performance analysis results. We demonstrated through examples that providing class-specific QoS guarantees can lead to significant gains in efficiency compared with worst-case and effective-bandwidth schemes. This comes at the expense of increased complexity. In summary: diversity is efficient but harder to achieve.

Acknowledgements

I would like to thank Dimitris Bertsimas and John Tsitsiklis for some helpful discussions on this topic.

References

Abate, J., Choudhury, G. L., & Whitt, W. (1995). Exponential approximations for tail probabilities in queues, I: Waiting times. *Operations Research*, *43*(5), 885–901.

Bertsimas, D., Paschalidis, I. Ch., & Tsitsiklis, J. N. (1997). *Large deviations analysis of the generalized processor sharing policy.* Technical Report MNS-97-108, Department of Manufacturing Engineering, Boston University, June 1997; Revised 1999. *Queueing Systems*, in press.

Bertsimas, D., Paschalidis, I. Ch., & Tsitsiklis, J. N. (1998a). Asymptotic buffer overflow probabilities in multiclass multiplexers: An optimal control approach. *IEEE Transactions on Automatic Control*, *43*(3), 315–335.

Bertsimas, D., Paschalidis, I. Ch., & Tsitsiklis, J. N. (1998b). On the large deviations behaviour of acyclic networks of G/G/1 queues. *The Annals of Applied Probability*, *8*(4), 1027–1069.

Bucklew, J. A. (1990). *Large deviation techniques in decision, simulation, and estimation.* New York: Wiley.

Chang, C. S. (1994). Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control*, *39*(5), 913–931.

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

Courcoubetis, C., & Weber, R. (1995a). Effective bandwidths for stationary sources. *Probability in the Engineering and Information Sciences*, *9*, 285–296.

Courcoubetis, C., & Weber, R. (1995b). *Estimation of overflow probabilities for state-dependent service of traffic streams with dedicated buffers.* Talk at the RSS Workshop in Stochastic Networks, Edinburgh, UK.

Cramér, H. (1938). Sûr un nouveau théorème-limite de la théorie des probabilités. In *Actualités Scientifiques et Industrielles, Colloque consacré à la théorie des probabilités*, vol. 736 (pp. 5–23). Paris: Hermann.

Cruz, R. L. (1991a). A calculus for network delay, Part I: Network elements in isolation. *IEEE Transactions on Information Theory*, *37*(1), 114–131.

Cruz, R. L. (1991b). A calculus for network delay, Part II: Network analysis. *IEEE Transactions on Information Theory*, *37*(1), 132–141.

de Veciana, G., Courcoubetis, C., & Walrand, J. (1993). Decoupling bandwidths for networks: A decomposition approach to resource management. Memorandum, Electronics Research Laboratory, University of California, Berkeley.

de Veciana, G., & Kesidis, G. (1995). Bandwidth allocation for multiple qualities of service using generalized processor sharing. *IEEE Transactions on Information Theory*, *42*(1).

de Veciana, G., & Walrand, J. (1995). Effective bandwidths: Call admission, traffic policing & filtering for ATM networks. *Queueing Systems*, *20*, 37–59.

Dembo, A., & Zajic, T. (1995). Large deviations: From empirical mean and measure to partial sums processes. *Stochastic Processes and Applications*, *57*, 191–224.

Dembo, A., & Zeitouni, O. (1993). *Large deviations techniques and applications.* Jones and Bartlett.

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

Duffield, N. G., Lewis, J. T., O'Connell, N., Russell, R., & Toomey, F., (1994). Statistical issues raised by the Bellcore data. *Proceedings of 11th IEE UK teletraffic Symposium*, Cambridge, UK, March 1994. London: IEE.

Dupuis, P., & Ramanan, K. (1997a). *Convex duality and the Skorokhod problem.* Technical Report, Division of Applied Mathematics, Brown University.

Dupuis, P., & Ramanan, K. (1997b), *A Skorokhod problem formulation and large deviation analysis of a processor sharing model.* Technical Report, Division of Applied Mathematics, Brown University.

Ellis, R. (1984). Large deviations for a general class of random vectors. *Annals of Probability*, *12*, 1–12.

Elwalid, A. I., & Mitra, D. (1993). Effective bandwidth of general Markovian traffic sources and admission control of high speed networks. *IEEE/ACM Transactions* on Networking, *1*(3), 329–343.

Elwalid, A. I., & Mitra, D. (1995). Analysis, approximations and admission control of a multiple-service multiplexing system with priorities. *Proceedings of the INFOCOM.*

Ganesh, A., & Anantharam, V. (1996). Stationary tail probabilities in exponential server tandems with renewal arrivals. *Queueing Systems*, *22*, 203–248.

Ganesh, A., & O'Connell, N. (1998). The linear geodesic property is not generally preserved by a FIFO queue. *The Annals of Applied Probability*, *8*(1), 98–111.

Gärtner, J. (1977). On large deviations from the invariant measure. *Theory of Applied Probability*, *22*, 24–39.

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

Glynn, P. W., & Whitt, W. (1994). Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability*, *31A*, 131–156.

Guérin, R., Ahmadi, H., & Naghshineh, M. (1991). Equivalent capacity and its applications to bandwidth allocation in high-speed networks. *IEEE Journal on Selected Areas in Communications*, *9*, 968–981.

Hsu, I., & Walrand, J. (1995). Admission control for ATM networks. In *Proceedings IMA workshop on stochastic networks*, *Volumes in mathematics and its applications*, vol. 71 (pp. 411–427). IMA, Berlin: Springer.

Hui, J. Y. (1988). Resource allocation for broadband networks. *IEEE Journal on Selected Areas in Communications*, *6*(9), 1598–1608.

Jacobson, V. (1998). *Differentiated services for the internet.* Talk given at the Internet2 Joint Applications/Engineering QoS Workshop, Santa Clara, CA, May 1998.

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

Kelly, F. P. (1996). Notes on effective bandwidths. In S. Zachary, I. B. Ziedins, & F. P. Kelly, *Stochastic networks: Theory and applications*, vol. 9 (pp. 141–168). Oxford: Oxford University Press.

Kesidis, G., Walrand, J., & Chang, C. S. (1993). Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking*, *1*(4), 424–428.

Massoulié, L. (1998). *Large deviations estimates for polling and weighted fair queueing service systems.* Technical Report, France Télécom-CNET.

O'Connell, N. (1995). *Queue lengths and departures at single-server resources.* Talk at the RSS Workshop in Stochastic Networks, Edinburgh, UK.

Parekh, A. K., & Gallager, R. G. (1993). A generalized processor sharing approach to flow control in integrated services networks: The single node case. *IEEE/ACM Transactions on Networking*, *1*(3), 344–357.

Parekh, A. K., & Gallager, R. G. (1994). A generalized processor sharing approach to flow control in integrated services networks: The multiple node case. *IEEE/ACM Transactions on Networking*, *2*(2), 137–150.

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

Rockafellar, R. T. (1970). *Convex analysis*. Princeton, NJ: Princeton University Press.

Shwartz, A., & Weiss, A. (1995). *Large deviations for performance analysis*. New York: Chapman & Hall.

Tse, D., Gallager, R. G., & Tsitsiklis, J. N. (1995). Statistical multiplexing of multiple time-scale Markov streams. *IEEE Journal on Selected Areas in Communications*, *13*(6).

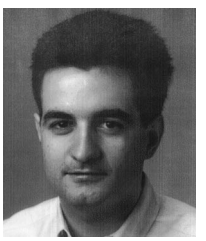
Weiss, A. (1995). An introduction to large deviations for communication networks. *IEEE Journal on Selected Areas in Communications*, *13*(6), 938–952.

Zhang, L., Deering, S., Estrin, D., Shenker, S., & Zappala, D. (1993). RSVP: A new resource ReSerVation protocol. *IEEE Network*.

Zhang, Z. -L. (1997). Large deviations and the generalized processor sharing scheduling for a two-queue system. *Queueing Systems: Theory and Applications*, *26*(3–4), 229–264.

Zhang, Z. -L., Liu, Z., Kurose, J., & Towsley, D. (1997). Call admission control schemes under the generalized processor sharing scheduling discipline. *Telecommunication Systems*, *7*(1–3), 125–152.

Zhang, Z. -L., Towsley, D., & Kurose, J. (1995). Statistical analysis of the generalized processor sharing scheduling discipline. *IEEE Journal on Selected Areas in Communications*, *13*(6), 1071–1080.



Ioannis Ch. Paschalidis was born in Athens, Greece, in 1968. He received the Diploma degree in Electrical and Computer Engineering from the National Technical University of Athens, 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, Massachusetts, in 1993 and 1996, respectively.

During the summer of 1996 he was a Post-Doctoral 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 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.

He has received the second prize in the 1997 George E. Nicholson paper competition by INFORMS and has been elected full member of Sigma Xi (M.I.T. Chapter). He is also a member of IEEE and INFORMS.