On Generalized Processor Sharing With Regulated Multimedia Traffic Flows

Chaiwat Oottamakorn, *Student Member, IEEE*, Shiwen Mao, *Member, IEEE*, and Shivendra S. Panwar, *Senior Member, IEEE*

***Abstract*—Multimedia traffic is becoming an increasing portion of today's Internet traffic due to the flourishing of multimedia applications such as music/video streaming, video teleconferencing, IP telephony, and distance learning. In this paper, we study the problem of supporting multimedia traffic using a generalized processor sharing (GPS) server. By examining the sample path behavior and exploring the inherent feasible ordering of the classes, we derive tight performance bounds on backlog and delay for regulated multimedia traffic classes in a GPS system. Our approach is quite general since we do not assume any arriving traffic model or any specific traffic regulator, other than that each traffic flow is deterministically regulated. Such deterministic regulators, as well as approximations of the GPS server, are widely implemented in commercial routers. In addition, our analysis is very accurate and achieves a high utilization of the server capacity, since we exploit the independence among the traffic flows for higher statistical multiplexing gains. Numerical examples and simulation results are presented to demonstrate the accuracy and merits of our approach, which is practical and well suited for supporting multimedia applications in the Internet.**

***Index Terms*—Generalized processor sharing (GPS), multimedia, quality-of-service (QoS), scheduling, traffic regulation.**

I. Introduction

Multimedia traffic is becoming an increasing portion of today's Internet traffic due to the flourishing of multimedia applications such as music/video streaming, video teleconferencing, IP telephony, and distance learning. Applications that generate such data can have very diverse quality-of-service (QoS) requirements. One major concern in the design, implementation, and operation of the Internet is how to provide QoS guarantees for applications with diverse QoS requirements, while achieving high utilization of network resources.

QoS guarantees can be provisioned in the Internet using the architectures described in [1] or [2]. However, due to the advances in dense wavelength division multiplexing (DWDM) technology, over-provisioning in the network core has become a general practice for many service providers. Nevertheless, we argue that over-provisioning does not necessarily solve the QoS provisioning problem. This is because over-provisioning may not be applicable to *all* segments of the network, due to technical, regulatory, or capital investment limitations. This makes it difficult to guarantee over-provisioning on an end-to-end basis in order to meet QoS requirements. In order to guarantee end-to-end performance, QoS mechanisms are still needed for the relatively resource-constrained access networks (e.g., wireless access networks), while it may be possible to apply over-provisioning in the core.

Various QoS mechanisms have been developed over the years, such as traffic shaping, admission control, QoS signaling and resource reservation, QoS routing, active queue management, and packet scheduling. Leaky bucket-based traffic regulation and generalized processor sharing (GPS) are among the most successful QoS mechanisms, since both of them are not only underpinned by rigorous theoretical analysis [3]–[7], but also widely implemented in commercial routers [8].

GPS is a work-conserving scheduling discipline in which multiple traffic classes share a deterministic server. With GPS, each class is associated with a *weight* and is guaranteed a minimum service rate in proportion to its weight whenever it is backlogged. Furthermore, the residual service of the non-backlogged classes is distributed to the backlogged classes in proportion to their weights. Therefore, GPS is efficient in utilizing and sharing the server capacity (since it is work-conserving and the bandwidth is shared by all classes), while being capable of isolating the classes (since each class is guaranteed a minimum rate, it won't be affected by a misbehaving class). By assigning different weights to the classes, service differentiation can be easily achieved.

GPS has been widely studied under various traffic characterizations and under deterministic settings [3] or stochastic settings [4]–[7]. Generally, the bounds obtained by deterministic GPS analysis are very conservative, since they are derived for the worst case scenario that only occurs with a very low probability [3]. Such *hard* QoS guarantees are unnecessary for many multimedia applications, where a certain level of QoS violation is generally acceptable. On the other hand, although the existing bounds obtained by statistical GPS analysis can achieve a much higher resource utilization than deterministic bounds, the traffic characterizations used in such analysis are usually hard to measure and enforce [4]–[7].

Manuscript received July 21, 2005; revised March 8, 2006. This research was supported in part by the National Science Foundation under Grants CNS-0520054, CNS-0435303, and CNS-0435228, and by the New York State Office of Science, Technology and Academic Research (NYSTAR) through the Center for Advanced Technology in Telecommunications (CATT) at Polytechnic University. This work was previously presented at the 2005 IEEE International Conference on Communications, Seoul, Korea, May 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Simon Lucey.

C. Oottamakorn is with the Institute of Engineering and Resources, Walailak University, Nakhon Si Thammarat 80000, Thailand (e-mail: ochaiwat@wu.ac.th).

S. Mao is with the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA (e-mail: smao@ieee.org).

S. S. Panwar is with the Department of Electrical and Computer Engineering, Polytechnic University, Brooklyn, NY 11201 USA (e-mail: panwar@catt.poly.edu).

Digital Object Identifier 10.1109/TMM.2006.884613

In this paper, we investigate the behavior of a high-speed GPS server under *deterministically regulated* multimedia traffic flows, but in a *stochastic setting*. We present a practical framework for supporting multimedia traffic using GPS servers. Within this framework, multimedia traffic flows are regulated with deterministic regulators (such as the popular leaky bucket regulator or other piece-wise linear regulators), thus greatly simplifying user traffic regulation, monitoring, and enforcement, as in a deterministic GPS analysis [3]. We then derive tight statistical upper bounds on the tail distribution of buffer occupancy and delay for each class in the GPS system, which achieves a high resource utilization as in a statistical GPS analysis [4]. More specifically, we exploit the independence among the flows and derive tight moment bounds for the aggregate arrival process of each class. Such moment bounds produce an accurate statistical characterization of the aggregate traffic of each class when used with the *Chernoff bound*. Then, by exploring the inherent feasible ordering of the classes [3], [4], we examine the sample path behavior of the multimedia sessions served by a single GPS server in isolation, and derive the backlog (i.e., buffer occupancy) and delay bounds for each class in the GPS system. We also present simulation and numerical results to demonstrate the efficacy of the proposed approach.

Such a "hybrid" approach has many advantages. First, it obtains tight statistical bounds on the backlog and delay without any assumption on the traffic arrival models. Second, our approach is amenable to implementation and policing as in a deterministic GPS analysis [3], and being capable of achieving high resource utilizations as in a statistical GPS analysis. Third, our approach is computationally efficient. GPS systems with a large number of classes and hundreds or thousands of flows per class can be easily handled (see Section IV). Such an approach is quite practical and effective for supporting multimedia applications in the Internet.

The remainder of the present paper is organized as follows. In Section II, we present some preliminary results for the analytical framework. We derive the backlog and delay bounds for a GPS queue with regulated multimedia traffic in Section III. Numerical studies and simulation and numerical results are presented in Section IV. Related work is discussed in Section V. Section VI concludes this paper.

II. Preliminaries

Consider a GPS server with a transmission rate C. The server serves N classes of incoming traffic, where each Class i contains a set of flows (denoted by $\mathcal{C}_i, i = 1, 2, \cdots, N$). Each flow may itself be an aggregate of the traffic from multiple sessions. As shown in Fig. 1, arriving traffic from each flow k belonging to Class i, is regulated by a deterministic envelope $A^*_{i,k}(\tau)$. Arriving Class i traffic is buffered at the corresponding queue, i.e., Queue i. In the following analysis, we consider a fluid traffic model. It is worth noting that our results can also be applied to packetized or discrete-time traffic arrival models as well [9].

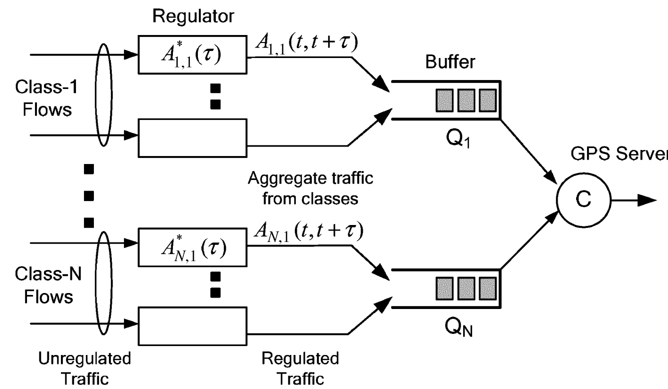


Fig. 1. A GPS system with N classes of regulated multimedia traffic flows.

A. Regulated Flows

The traffic arriving from a flow k of Class i in an interval $[t_1, t_2)$ is denoted by $A_{i,k}(t_1, t_2)$. We assume that the arriving process $A_{i,k}(t_1, t_2)$ has the following properties.

(P1) *Additivity:* For any $t_1 < t_2 < t_3$, we have $A_{i,k}(t_1, t_2) + A_{i,k}(t_2, t_3) = A_{i,k}(t_1, t_3)$.

(P2) *Subadditive Bounds:* The traffic is regulated by a *deterministic subadditive envelope* $A^*_{i,k}$, i.e., $A_{i,k}(t, t+\tau) \leq A^*_{i,k}(\tau)$, for all $t \geq 0$ and $\tau \geq 0$ and $A^*_{i,k}(\tau_1) + A^*_{i,k}(\tau_2) \geq A^*_{i,k}(\tau_1 + \tau_2)$, for all $\tau_1 > 0$ and $\tau_2 \geq 0$.

(P3) *Stationarity:* The $A_{i,k}$'s are *stationary*, i.e., for all positive t, τ, and t' $Pr[A_{i,k}(t, t+\tau) \leq x] = Pr[A_{i,k}(t', t'+\tau) \leq x]$, for all $x \geq 0$. In other words, the statistical properties of $A_{i,k}$ do not change with time.

(P4) *Independence:* All flows are statistically independent.

Remarks:

- These assumptions are quite general [3]–[7]. The class of subadditive deterministic traffic envelopes is most commonly used as traffic regulators. The stationarity and independence assumptions are also quite common, especially when the flows are multimedia streams from different users. Note that ergodicity is not assumed.
- The traffic regulators most widely used in practice are *leaky buckets* with a peak rate enforcer. As generalization, a traffic session could be regulated by $M \geq 2$ leaky buckets, each being characterized by a two-tuple, $\{\sigma^m_{i,k}, \rho^m_{i,k}\}$, $m = 1, \cdots, M$. The resulting deterministic envelope of these M leaky buckets is the minimum of themselves, i.e.,

$$A^*_{i,k}(\tau) = \min_{m \in \{1, \cdots, M\}} \left\{\sigma^m_{i,k} + \rho^m_{i,k}\tau\right\}, \quad \forall \tau \geq 0 \qquad (1)$$

which is a concave, piecewise linear envelope process. The $M = 1$ case corresponds to the single leaky bucket regulator.
- A consequence of subadditivity of A^*_i is that, for all t

$$\lim_{\tau \to \infty} \frac{A_{i,k}(t, t+\tau)}{\tau} = \rho_{i,k}. \qquad (2)$$

That is, the long-term average rate for $A_{i,k}(t, t+\tau)$ exists.

B. Generalized Processor Sharing

A scheduler at a network node determines the rates at which buffered traffic for the classes is transmitted (see Fig. 1). GPS is a *work-conserving* scheduling discipline: if there is backlog in the queue, the traffic will be served continuously with a constant rate C. Under GPS, each Class i is assigned a fixed and real-valued positive number $\phi_i, i = 1, \cdots, N$, called its *GPS weight*. As shown in Fig. 1, a GPS server serves the aggregate traffic from N classes simultaneously. More precisely, the classes are served as follows.

1) Let $S_i(t, t+\tau)$ denote the amount of Class i traffic served in the interval $[t, t+\tau)$. If Class i is backlogged during $[t, t+\tau)$, we have

$$\frac{S_i(t,t+\tau)}{S_j(t,t+\tau)} \geq \frac{\phi_i}{\phi_j}, \quad j = 1, 2, \cdots, N. \tag{3}$$

2) It has been shown that there exists an order among the classes such that, after relabeling the classes, we have

$$\rho_i < \frac{\phi_i}{\sum_{j=i}^{N} \phi_j} \left(C - \sum_{j=1}^{i-1} \rho_j \right), \quad 1 \leq i \leq N \tag{4}$$

where $\rho_j = \sum_{k=1}^{\mathcal{C}_j} \rho_{j,k}$, $\rho_{j,k}$ is defined in (2), and $\sum_{j=1}^{i-1} \rho_j$ is assumed to be zero when $i = 1$ [3], [4]. Note that we assume $\sum_{i=1}^{N} \rho_i < C$ for the system to be stable.

GPS queues are generally very difficult to analyze due to the coupled service rates and backlogs: the service rate of a class is dependent on backlogs of all the classes, while the backlogs of all the classes are in turn dependent on the service rates. By applying a sample path analysis, and exploring the inherent feasible ordering among the sessions, we can break such couplings and derive QoS bounds. It has been shown in [3] that for a stable GPS queue, such a feasible order always exists, although not being unique.

It is also worth noting that due to the fluid traffic assumption, GPS cannot be directly used in packet switched networks (such as the Internet). To adopt GPS in packet switched networks, discrete GPS approximations, such as packet-by-packet GPS (PGPS) [3] and worst-case fair weighted fair queueing ($\mathrm{WF^2Q}$) [9] have been proposed. Specifically, it has been shown in [9] that $\mathrm{WF^2Q}$ provides almost identical service to GPS with a maximum difference of one packet size, and it shares the fairness property of GPS.

III. Statistical Analysis

In this section, we will derive the upper bounds on the tail distributions of backlog and delay for each traffic class, consisting of multimedia data flows satisfying properties (P1)–(P4). We first derive the bound for the moment generating function of the aggregate traffic for each class, and then obtain upper bounds on the tail distributions of backlog and delay for each class by applying the moment generating function bounds derived in the first step. The notation used in the following analysis is summarized in Table I.

TABLE I
NOTATION

$A_i(t, t+\tau)$	aggregate traffic from Class i flows in $[t, t+\tau)$.
$S_i(t, t+\tau)$	amount of Class i traffic served in $[t, t+\tau)$.
$A_{i,k}(t, t+\tau)$	aggregate traffic from the kth flow of Class i in $[t, t+\tau)$.
$A^*_{i,k}(\tau)$	deterministic envelope of $A_{i,k}(t, t+\tau)$.
$M_{i,k}(\theta, \tau)$	moment generating function of $A_{i,k}(t, t+\tau)$.
$M_i(\theta, \tau)$	moment generating function of $A_i(t, t+\tau)$.
C	capacity of the GPS server.
ϕ_i	GPS weight of Class i.
$\rho_{i,k}$	long term average rate of $A_{i,k}(t, t+\tau)$.
ρ_i	long term average rate of $A_i(t, t+\tau)$.
$\mathcal{C}_i$	number of Class i sources.
$Q_i(t)$	Class i backlog at time t.
$D_i(t)$	delay experienced by Class i traffic arriving at time t.

A. Moment Bounds for Aggregate Traffic

For Class i, $i \in \{1, 2, \cdots, N\}$, let $A_i(t_1, t_2)$ denote the aggregate traffic from Class i flows in an interval (t_1, t_2), i.e., $A_i(t_1, t_2) = \sum_{k=1}^{\mathcal{C}_i} A_{i,k}(t_1, t_2)$. By definition, the moment generating function of the cumulative arrival of a Class i session k, $A_{i,k}(t, t+\tau)$ is

$$M_{i,k}(\theta, \tau) = E\left[e^{\theta A_{i,k}(t, t+\tau)}\right]. \tag{5}$$

With the independence assumption (P4), we can derive the moment generating function of Class i aggregate traffic as

$$M_i(\theta, \tau) = E\left[e^{\theta A_i(t, t+\tau)}\right] = \prod_{k=1}^{\mathcal{C}_i} M_{i,k}(\theta, \tau). \tag{6}$$

We then apply Theorem 1 in [10], which presents a bound on the moment generating function of each flow, and have

$$M_{i,k}(\theta, \tau) \leq 1 + \frac{\rho_{i,k}\tau}{A^*_{i,k}(\tau)} \left(e^{\theta A^*_{i,k}(\tau)} - 1\right). \tag{7}$$

Substituting (7) into (6), we derive the bound on the moment generating function of Class i aggregate traffic as follows:

$$M_i(\theta, \tau) \leq \prod_{k=1}^{\mathcal{C}_i} \left[1 + \frac{\rho_{i,k}\tau}{A^*_{i,k}(\tau)} \left(e^{\theta A^*_{i,k}(\tau)} - 1\right)\right]. \tag{8}$$

Remarks:

- $M_i(\cdot)$ is expressed in terms of simple deterministic envelopes of arriving flows, that is, the leaky bucket or multiple leaky bucket functions [see (1)].
- If the Class i flows are homogeneous, i.e., all flows in Class i have the same envelope, $A^*_{i,1}(\tau) = A^*_{i,2}(\tau) = \cdots = A^*_{i,\mathcal{C}_i}(\tau)$, then $M_i(\cdot)$ has the following simplified form:

$$M_i(\theta, \tau) \leq \left[1 + \frac{\rho_{i,1}\tau}{A^*_{i,1}(\tau)} \left(e^{\theta A^*_{i,1}(\tau)} - 1\right)\right]^{\mathcal{C}_i}. \tag{9}$$

B. Statistical Bounds on Backlog and Delay

In this section, we study the sample path behavior of the aggregate traffic of each class in the GPS system, as shown in

Fig. 1. We apply the moment generating function bounds derived in Section III-A to obtain the upper bounds on the tail distributions of the backlog and delay for each class.

Consider a tagged Class i. Let time 0 indicate the start of a busy period of length B during which Class i is continuously backlogged. During the busy period, traffic will be served and transmitted continuously at a constant rate C. For some subinterval $[0,\tau) \in B$, since there is backlog and due to the work conserving property, the server is busy during this subinterval. Then we have

$$\sum_{j=1}^{N} S_j(0,\tau) = C\tau$$

$$\sum_{j=1}^{i-1} S_j(0,\tau) + \sum_{j=i}^{N} \frac{\phi_j}{\phi_i} S_i(0,\tau) \geq C\tau. \tag{10}$$

We can solve (10) for $S_i(0,\tau)$ and have

$$S_i(0,\tau) \geq \frac{\phi_i}{\sum_{j=i}^{N} \phi_j} \left(C\tau - \sum_{j=1}^{i-1} S_j(0,\tau) \right). \tag{11}$$

It has been shown that for regulated classes, the maximum backlog and delay are achieved when all the classes become greedy from time 0 [3]. Since the busy period starts at time 0, the amount of serviced traffic of a class is bounded by its arrival during $[0,\tau)$. The following inequality holds true for all classes j:

$$S_j(0,\tau) \leq A_j(0,\tau). \tag{12}$$

Applying (12), (11) can be rewritten as

$$S_i(0,\tau) \geq \frac{\phi_i}{\sum_{j=i}^{N} \phi_j} \left(C\tau - \sum_{j=1}^{i-1} A_j(0,\tau) \right). \tag{13}$$

Define $\psi_i = \phi_i / (\sum_{j=i}^{N} \phi_j)$ to simplify notation. As $Q_i(0) = 0$, we have $Q_i(\tau) = A_i(0,\tau) - S_i(0,\tau)$ for $\tau \leq B$. We then obtain

$$Q_i(\tau) \leq A_i(0,\tau) - \psi_i \left(C\tau - \sum_{j=1}^{i-1} A_j(0,\tau) \right). \tag{14}$$

Theorem 1: Given a GPS server with a service capacity C that serves N traffic classes $\{A_1, A_2, \cdots, A_N\}$ satisfying Properties (P1)–(P4). Assume that $\sum_{i=1}^{N} \rho_i < C$ and a feasible ordering with respect to $\{\phi_1, \phi_2, \cdots, \phi_N\}$ and $\{\rho_1, \rho_2, \cdots, \rho_N\}$. For any time instance τ in a busy period and any Class i backlog level $q_i \geq 0$, we have

$$\begin{aligned} Pr\left[Q_i(\tau) \geq q_i\right] &\leq \inf_{\theta \geq 0} \left\{ e^{-\theta(q_i + \psi_i C\tau)} \prod_{k=1}^{\mathcal{C}_i} \left[1 + \frac{\rho_{i,k}\tau}{A^*_{i,k}(\tau)} \left(e^{\theta A^*_{i,k}(\tau)} - 1 \right) \right] \right. \\ &\quad \left. \times \prod_{j=1}^{i-1} \prod_{k=1}^{\mathcal{C}_j} \left[1 + \frac{\rho_{j,k}\tau}{A^*_{j,k}(\tau)} \left(e^{\psi\theta A^*_{j,k}(\tau)} - 1 \right) \right] \right\}. \end{aligned} \tag{15}$$

Proof: We will use the moment generating function bound in (8) to construct the upper bound on the tail distribution of Class i backlog. To obtain a rigorous upper bound on $Pr[Q_i(\tau) \geq q_i]$, recall that the Chernoff bound for a random variable Y is

$$Pr[Y \geq y] \leq e^{-\theta y} E[e^{\theta Y}], \quad \forall \theta \geq 0. \tag{16}$$

In particular, for $Q_i(\cdot)$ in (14), this gives

$$\begin{aligned} &Pr\left[Q_i(\tau) \geq q_i\right] \\ &\quad \leq Pr\left[A_i(0,\tau) + \psi_i \sum_{j=1}^{i-1} A_j(0,\tau) \geq q_i + \psi_i C\tau \right] \\ &\quad \leq e^{-\theta(q_i + \psi_i C\tau)} M_i(\theta,\tau) \prod_{j=1}^{i-1} M_j(\psi_i\theta,\tau) \end{aligned} \tag{17}$$

where $M_i(\cdot)$ is given in (8) and (9). Note that (17) holds true for any positive value of θ. In order to get a tight bound on the tail distribution, we need to choose the θ that minimizes the right-hand side (RHS) of (17), i.e.,

$$\begin{aligned} &Pr\left[Q_i(\tau) \geq q_i\right] \\ &\quad \leq \inf_{\theta \geq 0} \left\{ e^{-\theta(q_i + \psi_i C\tau)} M_i(\theta,\tau) \prod_{j=1}^{i-1} M_j(\psi_i\theta,\tau) \right\}. \end{aligned} \tag{18}$$

■

We next derive the upper bound on the delay distribution for each class. Again, consider a busy period of the GPS queue starting at time 0 during which Class i is continuously backlogged. Let t denote the time interval after which the cumulative service received by Class i becomes equal to or larger than the cumulative arrivals of Class i at time τ. For traffic enqueued at time instance τ within the busy period, the corresponding queueing delay is [11]

$$D_i(\tau) = \min\left\{ t : t \geq 0 \text{ and } A_i(0,\tau) \leq S_i(0,\tau + t) \right\}. \tag{19}$$

Equation (19) applies to the class of first-come-first-serve (FCFS) systems, to which GPS queues belong.

For a given delay value d_i for Class i, delay violation event occurs when there exists any traffic unit experiencing a delay larger than the delay value. That is, $\min\{t : t \geq 0 \text{ and } A_i(0,\tau) \leq S_i(0,\tau + t)\} \geq d_i$, or

$$A_i(0,\tau) \geq S_i(0,\tau + d_i). \tag{20}$$

Theorem 2: Given a GPS server with a service capacity C that serves N traffic classes $\{A_1, A_2, \cdots, A_N\}$ satisfying Properties (P1)–(P4). Assume that $\sum_{i=1}^{N} \rho_i < C$ and a feasible ordering with respect to $\{\phi_1, \phi_2, \cdots, \phi_N\}$ and $\{\rho_1, \rho_2, \cdots, \rho_N\}$. For any time instance τ in a busy period and any delay value $d_i \geq 0$

$$\begin{aligned} &Pr\left[D_i(\tau) \geq d_i\right] \\ &\quad \leq \inf_{\theta \geq 0} \left\{ e^{-\theta\psi_i C(\tau + d_i)} \prod_{k=1}^{\mathcal{C}_i} \left[1 + \frac{\rho_{i,k}\tau}{A^*_{i,k}(\tau)} \left(e^{\theta A^*_{i,k}(\tau)} - 1 \right) \right] \right. \end{aligned}$$

$$\times \prod_{j=1}^{i-1} \prod_{k=1}^{\mathcal{C}_j} \left[1 + \frac{\rho_{j,k}(\tau+d_i)}{A^*_{j,k}(\tau+d_i)} \left(e^{\psi\theta A^*_{j,k}(\tau+d_i)} - 1 \right) \right] \Bigg\}. \quad (21)$$

Proof: From (20), we have

$$Pr\left[D_i(\tau) \geq d_i\right] = Pr\left[A_i(0,\tau) \geq S_i(0,\tau+d_i)\right]. \quad (22)$$

Inserting (13) into (22), we have

$$\begin{aligned} &Pr\left[D_i(\tau) \geq d_i\right] \\ &\leq Pr\left[A_i(0,\tau) + \psi_i \sum_{j=1}^{i-1} A_j(0,\tau+d_i) \geq \psi_i C(\tau+d_i) \right] \\ &\leq e^{-\theta\psi_i C(\tau+d_i)} M_i(\theta,\tau) \prod_{j=1}^{i-1} M_j(\psi_i\theta, \tau+d_i). \end{aligned} \quad (23)$$

■

C. Practical Implications

Letting Q_i be the steady state random variable for the backlog of Class i, we have $\Pr\{Q_i \geq q_i\} \leq \sup_{0<\tau<B}\{Pr[Q_i(\tau) \geq q_i]\}$. If Class i sessions require their buffer overflow probability be less than or equal to ϵ_i for the corresponding buffer size q_i, then the schedulability condition for Class i sessions is

$$\sup_{0<\tau<B} \{Pr\left[Q_i(\tau) \geq q_i\right]\} \leq \epsilon_i \quad (24)$$

where B is the busy period bound [3]. For admission control test, the above inequality and Theorem 1 can be applied to compute the corresponding loss probability for a given q_i. If the computed probability is less than ϵ_i, then the flows are admissible.

Similarly, let D_i be the steady state random variable for Class i delay. We have $\Pr\{D_i \geq d_i\} \leq \sup_{0<\tau<B}\{Pr[D_i(\tau) \geq d_i]\}$. If Class i sessions require their delay violation probability be less than or equal to ϵ_i for the corresponding delay value d_i, then the schedulability condition for Class i sessions is

$$\sup_{0<\tau<B} \{Pr\left[D_i(\tau) \geq d_i\right]\} \leq \epsilon_i. \quad (25)$$

An admission control test based on delay requirements can be conducted in a similar fashion using the above inequality and Theorem 2. For classes with both loss and delay requirements, the flows are admissible only when both requirements are satisfied.

The RHS of (15) and (21) are in a product form, which can be further simplified. Take (15) for example. By taking logarithms on both sides of (17), we can easily convert the RHS of (17) to the summation form. Therefore, increases in number of classes or in number of flows in a class only cause an increased number of terms in the summation, each having a form of $\log[1 + (\rho_{i,k}\tau/A^*_{i,k}(\tau))(e^{\theta A^*_{i,k}(\tau)} - 1)]$. Let $F_i(\theta) = \log \Pr[Q_i(\tau) \geq q_i]$. We can further show that $F_i(\theta)$ is convex with regard to θ, i.e., the second derivative of $F_i(\theta)$ is nonnegative. The optimal value for θ can be easily obtained by using, say, the *sequential quadratic programming* method, which is computationally efficient due to its super-linear convergence performance [12]. In many cases when a class consists of homogeneous sources, the computation can be further simplified [see (9)].

Our proposed scheme focuses on deterministically regulated multimedia traffic flows. There are two possible scenarios with regard to the coexistence of multimedia flows and nonmultimedia flows. *First*, if the nonmultimedia flows are also regulated and have QoS requirements, the network operator can set the GPS assignments for the classes according to the *service level agreements*, and the QoS of these classes can be guaranteed as shown in Theorems 1 and 2. Such an example is presented in Section IV. *Second*, if the nonmultimedia traffic is best-effort and does not have any QoS guarantees, the network operator could apply GPS to the QoS sessions, while letting the best-effort sessions use the residual service left by the QoS sessions (i.e., an available-bit-rate type service).

This analysis involves three key factors that determine the quality of received multimedia data (e.g., distortion of a reconstructed video): data rate, delay, and loss. The data rate of the multimedia flow is determined by the regulator used for the multimedia flow, which is subsequently determined by the service level agreement between the user and the network operator. The delay violation and buffer overflow probabilities can be computed using Theorems 1 and 2. The QoS guarantees delivered by the proposed approach can be related to the application layer multimedia quality by incorporating existing work from the multimedia research community. For example, for a video flow with average rate ρ, a distortion-rate model is developed in [13] as

$$D = D_0 + \frac{\omega}{\rho - \rho_0} + \kappa \cdot (1-p) \cdot \Pr(T > \Delta) + \kappa \cdot p \quad (26)$$

where D_0, ω, ρ_0, and κ are video sequence/codec specific constants, p is the loss probability of video packets, Δ is the decoding deadline, and $\Pr(T > \Delta)$ is the overdue probability. The backlog and delay distribution presented in this paper can be used to obtain a very good approximation for the received video distortion by using this model.

IV. Numerical Results

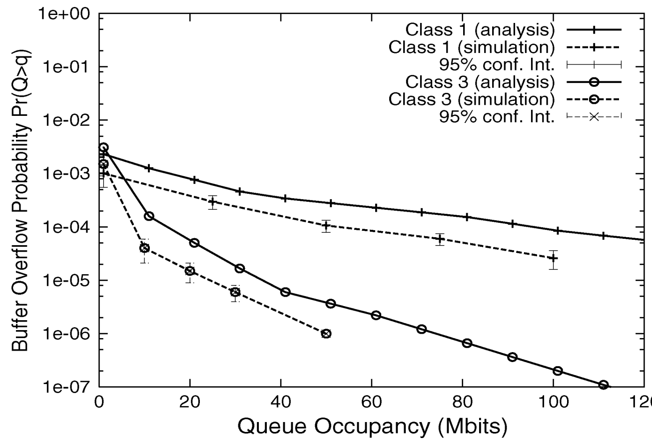
In this section, we present a set of experiments to demonstrate the performance of the proposed QoS bounds. The objective of this study is threefold: 1) to demonstrate the accuracy of the proposed bounds by comparing with OPNET simulations; 2) to demonstrate how efficient the proposed scheme is in utilizing network resources; and 3) to provide a comparison with two representative existing schemes, including a deterministic analysis-based approach [3] and a stochastic analysis-based approach [14].

A. Simulation Settings

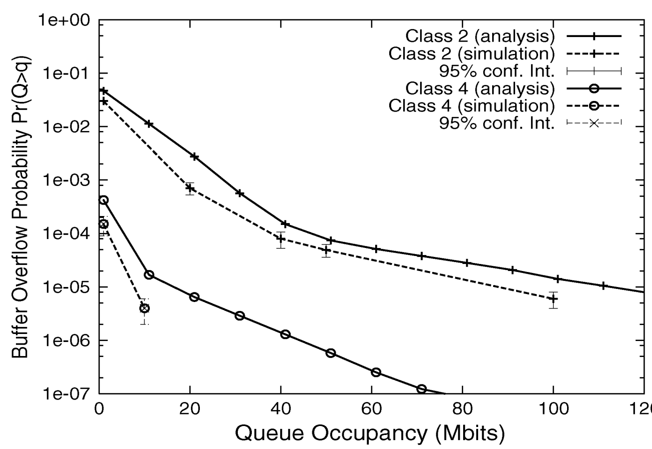
We implement a GPS system using the OPNET Modeler. In the experiments, we use four classes of sources as specified in the IETF Diffserv architecture [2]. The four types of traffic used

TABLE II
TRAFFIC SOURCE PARAMETERS

-	Avg. Rate (Kb/s)	Peak Rate (Kb/s)	GPS Weight ϕ_i
Class 1: on-off (Exponential)	230	1600	280
Class 2: on-off (Pareto)	520	2220	500
Class 3: Video Trace 1	773	2687	800
Class 4: Video Trace 2	1053	3150	1000

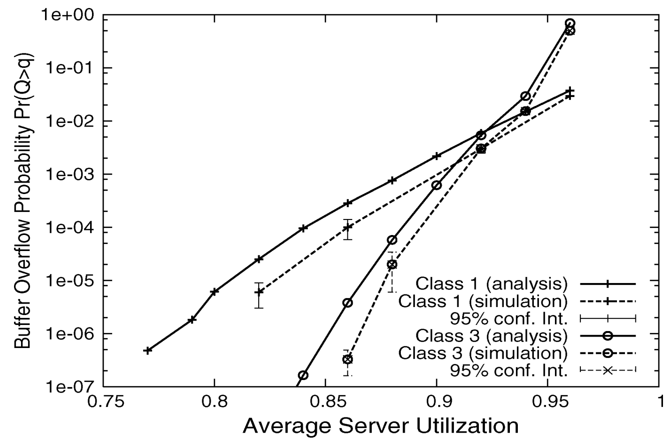


(a)

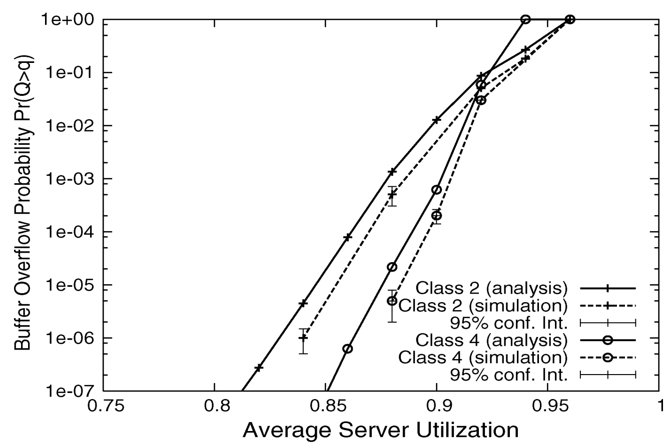


(b)

Fig. 2. Experiment 1: backlog distribution as a function of buffer size q with fixed average server utilization ($\mu = 85.7\%$). (a) Classes 1 and 3. (b) Classes 2 and 4.



(a)



(b)

Fig. 3. Experiment 1: backlog distribution as a function of average server utilization μ with fixed buffer size ($q = 50$ Mbits). (a) Classes 1 and 3. (b) Classes 2 and 4.

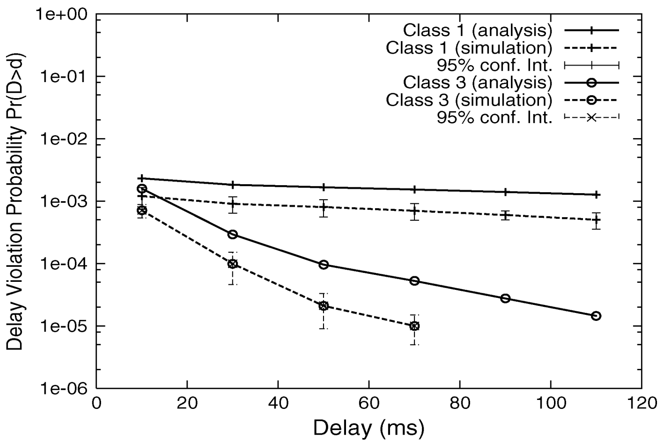
in the experiments, as well as the corresponding parameters, are presented in Table II. More specifically, Class 1 consists of on-off traffic sources. The sojourn time in the two states are exponentially distributed with average on time $= 100$ ms and average off time $= 600$ ms. Class 2 consists of on-off sources with Pareto distributed sojourn times. The mean on and off times are 100 ms and 325 ms, respectively. In the on state, these on-off sources transmit data at their peak rates; in the off state, these sources do not generate any traffic. Classes 3 and 4 sources use MPEG-4 compressed video traces from the movie *Jurassic Park* and a soccer game [15], respectively.

We choose these classes of sources to approximate a typical combination of the Internet traffic. The first class of sources generate short-range dependent (SRD) traffic, which is a good model for Voice over IP. The last three classes of sources produce long-range dependent (LRD) and self-similar traffic in the aggregate. Class 2 sources are used to model traditional Internet data applications such as FTP and web traffic, and Class 3 and 4 sources are used to model multimedia applications such as video streaming and video teleconferencing. Traffic from each of the sources is regulated by ten leaky buckets before entering the network. We implemented a high-speed GPS server with a capacity C ranging from 0.8 to 1.0 Gb/s. The rates of the sources are chosen to saturate such a high bandwidth link. In most cases examined, there are several hundred sources within each class, indicating that the proposed approach can easily handle a large number of sessions.

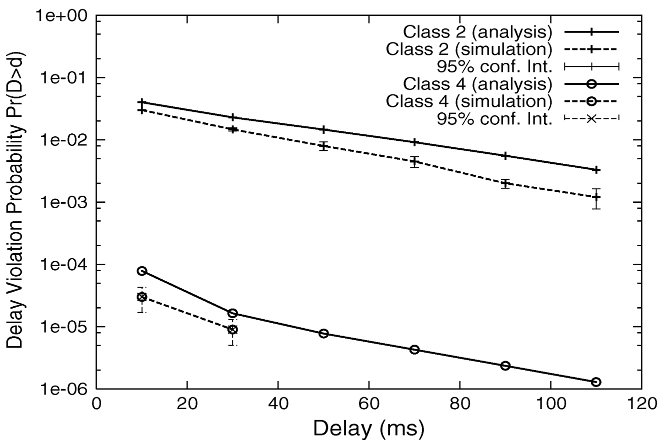
Each simulation lasts for 200 simulated seconds. During the simulations, each source randomly chooses a starting time and begins to generate traffic after that. For the figures presented in this section, each point is the average of ten simulation runs with different random traffic scenarios (e.g., random session starting time). We provide 95% confidence intervals for the points in Figs. 2–5, which are plotted as error bars in these figures.

B. Experiment 1: Backlog and Delay Bounds

In this experiment, we compute the analytical bounds on the tail distributions of backlog and delay as given in (15) and (21)

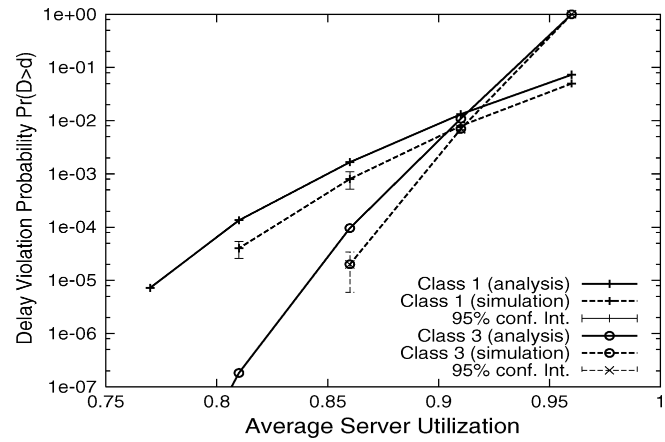


(a)

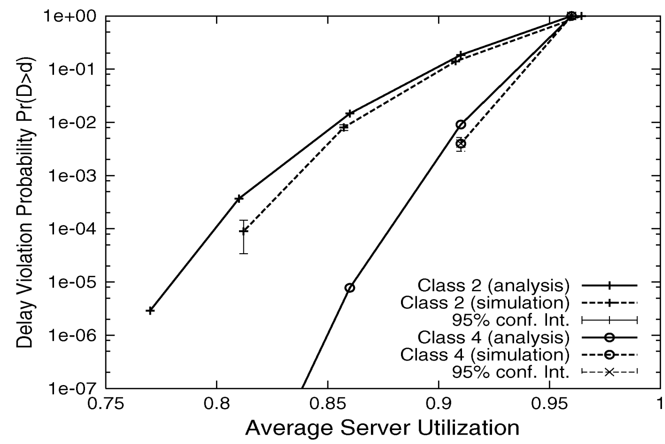


(b)

Fig. 4. Experiment 1: delay distribution as a function of delay requirement d with fixed average server utilization ($\mu = 85.7\%$). (a) Class 1 and Class 3. (b) Class 2 and Class 4.



(a)



(b)

Fig. 5. Experiment 1: delay distribution as a function of server utilization μ with fixed delay requirement ($d = 50$ ms). (a) Class 1 and Class 3. (b) Class 2 and Class 4.

and compare them with OPNET simulation results. For the results reported in this subsection, we consider a GPS system with four classes, each having 300 flows. The source parameters and the GPS weights for the classes $\{\phi_i\}_{i=1,2,3,4}$ are given in Table II. The average server utilization is defined as $\mu = \sum_{i=1}^{4} \rho_i/C$, where $\rho_i = \sum_{k=1}^{\mathcal{C}_i} \rho_{i,k}$ is the long-term average rate of Class i aggregate traffic.

The results of these experiments are plotted in Figs. 2–5. Fig. 2 shows the tail distribution of buffer occupancy as a function of buffer size q, where the average server utilization is fixed at $\mu = 85.7\%$. Fig. 3 plots the tail distribution of buffer occupancy as a function of the average server utilization μ, which is varied by changing the the server capacity C (note that a large C will give a smaller μ since the number of sources in each class is fixed). The buffer size is fixed at $q = 50$ Mbits for all the classes. These allow us to examine the tightness of the analytical bounds and their performance in utilizing the server capacacity.

From Figs. 2 and 3, we make the following observations. *First*, our derived bounds on backlog distribution given in (15) are tight upper bounds of those obtained from the simulations. For example, in Fig. 2(a), when $q = 50$ Mbits, the Class 1 backlog bound computed from (15) and that obtained from simulation are 2.78×10^{-4} and 1×10^{-4}, respectively. The optimal values that achieve the delay bound are $\theta^* = 2 \times 10^{-8}$ and $\tau^* = 21.7058$ s.

Second, all classes yield better performance as buffer size is increased (in Figs. 2) or average server utilization is decreased (in Figs. 3), i.e., as more network resources, either being buffer or bandwidth, are allocated. More precisely, as more buffer is allocated, the buffer overflow probabilities in Fig. 2 first decrease quickly; as the buffer size further increases, for classes 2 to 4 the decreases in buffer overflow probabilities get smaller and the curves becomes more flat (indicating heavy tails, a distinguishing trait of LRD traffic). As the server capacity increases in Fig. 3, on the other hand, all the buffer overflow probability curves drop drastically, except for extremely high server utilizations (around 95%) when the server is saturated. This clearly demonstrates that for LRD traffic, bandwidth provisioning is more effective in improving its queueing performance than buffer provisioning. *Third*, in Fig. 3, when $\mu \geq 88\%$, Classes 2, 3, and 4 perform poorly as compared with Class 1. This is because these LRD sources are highly bursty, and more importantly, the burstiness does not decrease with aggregation. On the other hand, when the number of Class 1 sources increases, the aggregate Class 1 traffic gets smoothed out due to the SRD property. Thus, Class

1 performance is better than the remaining three LRD classes in the high-bandwidth utilization region.

To demonstrate the performance of Theorem 2, we present the analytical and simulation results on the delay bound in Figs. 4 and 5, as functions of each class' delay requirement d_i and the average server utilization μ, respectively. Similar observations can be made from these figures. *First*, the analytical delay distribution upper bounds the corresponding simulation curve for all the classes. In addition, these delay bounds are tight. In Fig. 4(a), when the server utilization is $\mu = 85.7\%$ and the delay requirement is $d = 50$ ms, the Class 1 delay violation probability is 1.66×10^{-3} and the simulated delay violation probability is 8×10^{-4}. The optimal values for θ and τ are $\theta^* = 2 \times 10^{-8}$ and $\tau^* = 21.7058$ s, respectively.

In addition, better performance is observed as the delay bound is increased as in Fig. 4 (i.e., the delay requirement is relaxed), or as the average server utilization is decreased as in Fig. 5 (i.e., more capacity is available for each class). We also observe that bandwidth provisioning is more effective in improving the delay performance of the classes than buffer provisioning, since the curves in Fig. 5 are much more steep than those in Fig. 4. *Finally*, when the server is saturated, the delay performance of Class 1 is much better than that of the other three LRD classes (see Fig. 5). For network operation, it is important to adopt effective admission control schemes to prevent the server from being overloaded.

C. Experiment 2: Admissible Region

The admissible region of a GPS system is defined as the numbers of admissible flows of the classes whose buffer overflow and delay requirements are satisfied. By definition, the admissible region indicates the effectiveness of traffic classes in exploiting the GPS resource sharing capability.

Consider a GPS server with a capacity of 900 Mb/s serving four traffic classes (as listed in Table II) with a mixture of buffer overflow and delay requirements. For an easier presentation, we fixed the number of Class 1 sources at 300, which serves as background traffic. All the classes require statistical QoS guarantees as follows.

- Class 1 requires statistical service with a loss requirement $Pr[Q_1 \geq 30 \text{ Mbits}] \leq 10^{-2}$.
- Class 2 requires statistical service with a loss requirement $Pr[Q_2 \geq 50 \text{ Mbits}] \leq 10^{-4}$.
- Class 3 requires statistical service with a delay requirement $Pr[D_3 \geq 30 \text{ ms}] \leq 10^{-3}$.
- Class 4 requires statistical service with a delay requirement $Pr[D_4 \geq 50 \text{ ms}] \leq 10^{-4}$.

We determine the admissible region for the 4-class GPS system using Theorems 1 and 2. For comparison, we also obtain the admissible region using the following four schemes.

- *Peak Rate Allocation:* admission control based on the peak rate of each source. It is well-known that peak rate allocation provides deterministic QoS guarantees, while being an "overkill" for multimedia applications. The number of sources that can be supported with peak rate allocation serves as a lower bound for any practical admission control scheme.

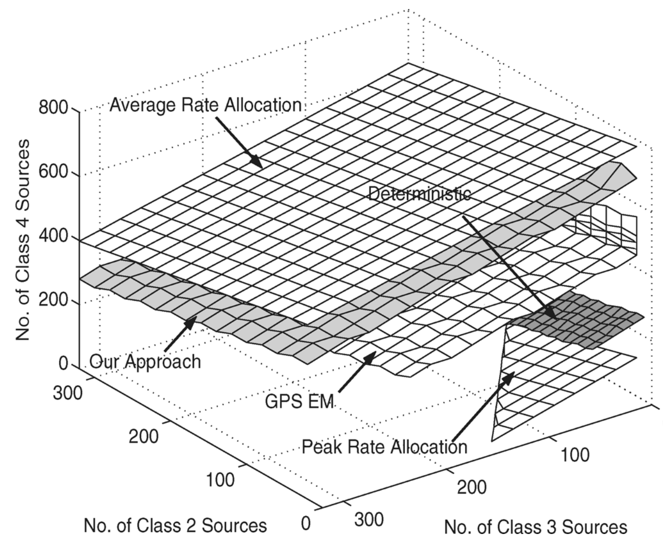


Fig. 6. Experiment 2: the admissible region for Classes 2, 3, and 4 in a four-class GPS system. The number of Class 1 sources is fixed at 300.

- *Average Rate Allocation:* admission control based on the average rate of each source. Except for constant bit rate traffic, average rate allocation leads to infinite delays and instability. The admissible region achieved by average rate allocation is an upper bound for any practical admission control scheme.
- *A Deterministic Approach:* we perform admission control tests for deterministic QoS guarantees using (14) and (20) with deterministic envelopes $A^*(\cdot)$ (instead of $A(\cdot)$). This is equivalent to the deterministic analysis in [3].
- *A Stochastic Approach:* we perform admission control test based on the analysis in [14] (called GPS-EM throughout this paper). In this method, we calculate the required bandwidth for each source and then apply the single-node admission control test as described in Sections III and IV of this paper.

The resulting admissible regions are plotted in Fig. 6. As expected, the admissible regions of peak rate allocation and average rate allocation bound the admissible regions of the other three schemes. Specifically, the average rate allocation always achieves a 100% server utilization. We also observe that peak rate allocation is highly conservative, with extremely low server utilizations (around 20% for all the points examined). This is because statistical multiplexing gain is not exploited at all.

The admissible region obtained by our analysis is the largest among the remaining three schemes, as shown in Fig. 6. It is very close to that obtained by average rate allocation. The server utilization is between 85%–89% for all the points examined. This demonstrates the tightness of our backlog and delay bounds, especially when the number of sources is large. Consequently, our approach is able to fully exploit the statistical multiplexing gain of multiclass GPS sharing. The GPS-EM admissible region is much larger than that of the deterministic approach, due to the fact the statistical multiplexing gain is explioted in GPS-EM.

Many more Class 4 sources are admitted using Theorems 1 and 2, as compared with GPS-EM and the deterministic approach. For example, our approach admits 300 Class 1 sources, 180 Class 2 sources, 60 Class 3 sources, and 540 Class 4 sources in the experiment, achieving a server utilization of 87%. The deterministic approach admits 300 Class 1 sources, 180 Class 2 sources, 50 Class 3 sources, and 60 Class 4 sources, achieving a server utilization of 29%. The GPS-EM scheme admits 300 Class 1 sources, 180 Class 2 sources, 60 Class 3 sources, and 320 Class 4 sources, achieving a server utilization of 61%. The proposed scheme achieves a 58% improvement over the deterministic approach, and a 26% improvement over GPS-EM in server utilization.

V. Related Work

The exact analysis of a GPS system is nontrivial, since the service rate a class receives is coupled with the backlog status and instant arriving rates of all other classes. Over the years, GPS has been studied under various traffic characterizations, such as leaky bucket regulated sources [3], [14], exponential bounded burstiness (EBB) sources [4], Markov modulated fluid process (MMFP) sources [7], Gaussian traffic sources [5], heavy-tailed sources [6], [16], [17], and mixed light- and heavy-tailed sources [18]. In addition to the Internet, GPS has been used in QoS provisioning in CDMA cellular networks [19]. These papers provide great insights into the behavior of GPS servers under bursty traffic flows. One of the most widely used technique in GPS analysis is the notion of *feasible ordering* [3] (with its extension of *feasible partitioning* [4]). With this technique, a GPS system can be decomposed into a set of separate FIFO queues from which performance bounds can be derived. We also used this technique in the present paper.

As discussed, the bounds obtained by deterministic GPS analysis are very conservative, since worst-case analysis is employed [3], while the existing statistical GPS bounds are not amenable for traffic regulation, monitoring, and enforcement [4]–[7]. In addition, the asymptotic performance bounds found in prior work (e.g., [6]) are only accurate for very large buffer sizes. Although shedding great insight on the GPS behavior for long-tailed sources, such bounds may not be applicable for multimedia traffic, where the generally tight end-to-end delay requirements prohibit the use of very large buffers in the intermediate routers. Finally, the analysis in [7] has the "state-explosion" problem, making it unsuitable for handling a large number of sources.

These observations motivated us to investigate the behavior of a GPS server under regulated multimedia traffic flows, but in a stochastic setting. As a result, our approach has the advantage of being as amenable to implementation and policing as deterministic GPS, and is capable of achieving the high resource utilizations achievable under statistical GPS analysis. In addition, our analytical approach is scalable. That is, it can easily handle a large number of sessions and achieve tight bounds, as illustrated in Section IV. It is therefore a practical and efficient approach for supporting regulated multimedia traffic in the Internet.

VI. Conclusions

In this paper, we studied the problem of QoS provisioning for regulated multimedia applications using a GPS server. Based on the general assumptions that the flows are independent and that each flow is deterministically regulated, we derived tight backlog and delay bounds for the GPS system via examining the sample path behavior of the classes and explore the inherent feasible order among the classes. We also demonstrated the accuracy and merits of our approach via experiments and simulations with MPEG-4 video traces and synthesized SRD/LRD traffic. The derived bounds are very close to the simulation results, and achieve larger admissible regions as compared with a deterministic analysis-based approach and a stochastic analysis-based approach. The framework presented in this paper is quite general and practical for supporting multimedia applications in the Internet.

References

[1] R. Braden, D. Clark, and S. Shenker, Integrated Services in the Internet Architecture: An Overview, Jun. 1994, IETF RFC 1633.

[2] S. Blake *et al.*, An Architecture for Differentiated Service, Dec. 1998, IETF RFC 2475.

[3] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case," *IEEE/ACM Trans. Netw.*, vol. 1, no. 3, pp. 344–357, Jun. 1993.

[4] Z.-L. Zhang, D. Towsley, and J. Kurose, "Statistical analysis of the generalized processor sharing scheduling discipline," *IEEE J. Select. Areas Commun.*, vol. 13, no. 6, pp. 1071–1080, Aug. 1995.

[5] P. Mannersalo and I. Norros, "GPS schedulers and gaussian traffic," in *Proc. IEEE INFOCOM 2002*, New York, Jun. 2002, pp. 1660–1667.

[6] S. Borst, O. Boxma, and P. Jelenkovie, "Generalized processor sharing with long-tailed traffic sources," in *Proc. ITC-16*, Edinburgh, U.K., Jun. 1999, pp. 345–354.

[7] F. L. Presti, Z.-L. Zhang, and D. Towsley, "Bounds, approximations and applications for a two-queue GPS system," in *Proc. IEEE INFOCOM*, San Francisco, CA, Mar. 1996, pp. 1310–1317.

[8] Cisco Systems, Inc., QC: Cisco IOS Release 12.0 Quality of Service Solutions Configuration Guide, Cisco IOS Documentation.

[9] J. C. R. Bennett and H. Zhang, " $\mathrm{WF}^2\mathrm{Q}$: worst-case fair weighted fair queueing," in *Proc. IEEE INFOCOM'96*, San Francisco, CA, Mar. 1996, pp. 120–128.

[10] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn, "Statistical service assurances for traffic scheduling algorithms," *IEEE J. Select. Areas Commun.*, vol. 18, no. 12, pp. 2651–2664, Dec. 2000.

[11] H. Sariowan, R. Cruz, and G. Polyzos, "Scheduling for quality of service guarantees via service curves," in *Proc. IEEE ICCCN 1995*, Las Vegas, NV, Sep. 1995, pp. 512–520.

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

[13] K. Stuhlmuller, N. Farberand, M. Link, and B. Girod, "Analysis of video transmission over lossy channels," *IEEE J. Select. Areas Commun.*, vol. 18, no. 6, pp. 1012–1032, Jun. 2000.

[14] A. Elwalid and D. Mitra, "Design of generalized processor sharing schedulers which statistically multiplex heterogeneous QoS classes," in *Proc. IEEE INFOCOM'99*, New York, Mar. 1999, pp. 1220–1230.

[15] F. Fitzek and M. Reisslein, MPEG-4 and H.263 Traces for Network Performance Evaluation (Extended Version), Dept. Elect. Eng., Germany, Techn. Univ. Berlin, Berlin, Germany, Tech. Rep. TKN-00-06, Oct. 2000.

[16] F. M. Pereira, N. L. S. Fonseca, and D. S. Arantes, On the Performance of Generalized Processor Sharing Servers Under Long-Range Dependent Traffic. Inst. Comput. Sci., State Univ. Campinas, Tech. Rep. IC-01-11, Oct. 2001.

[17] X. Yu, I. L.-J. Thng, Y. Jiang, and C. Qiao, "Queueing processes in GPS and PGPS with LRD traffic inputs," *IEEE/ACM Trans. Netw.*, vol. 13, no. 3, pp. 676–689, Jun. 2005.

[18] S. Borst, M. Mandjes, and M. van Uitert, "Generalized process sharing with light-tailed and heavy-tailed input," *IEEE/ACM Trans. Netw.*, vol. 11, no. 5, pp. 821–834, Oct. 2003.

[19] L. Xu, X. Shen, and J. W. Mark, "Fair resource allocation with guaranteed statistical QoS for multimedia traffic in wideband CDMA cellular networks," *IEEE Trans. Mobile Comput.*, vol. 4, no. 2, pp. 166–177, Mar./Apr. 2005.

Chaiwat Oottamakorn (S'98) received the B.S. degree in electrical engineering from Chulalongkorn University, Bangkok, Thailand, in 1993, and the M.S. and Ph.D. degrees in electrical engineering from Polytechnic University, New York, NY, in 1995 and 2001 respectively.

Since 2001, he has been with the Institute of Industrial and Resources Technology, Walailak University, Nakhon Si Thammarat, Thailand, where he is currently an Assistant Professor. His areas of research encompass Internet technologies, mobile communications, quality of services, network servers, and multimedia systems.

Shiwen Mao (S'99-M'04) received the B.S. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1994 and 1997, respectively, the M.S. degree in system engineering in 2000, and the Ph.D. degree in electrical and computer engineering in 2004, both from Polytechnic University, Brooklyn, NY.

He was a Research Member at the IBM China Research Lab from 1997 to 1998. He spent the summer of 2001 as a Research Intern at Avaya Labs-Research and has been a Research Scientist in the Department of Electrical and Computer Engineering (ECE) at Virginia Polytechnic Institute and State University (Virginia Tech), Blacksburg, for over two years. Currently, he is an Assistant Professor in the Department of ECE at Auburn University, Auburn, AL. His research interests include cross-layer design, optimization, and cooperative networking in multi-hop wireless networks, and multimedia communications in wired and wireless networks. He is the coauthor of *TCP/IP Essentials: A Lab-Based Approach* (Cambridge, U.K.: Cambridge University Press, 2004).

Dr. Mao is a co-recipient of the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems.

Shivendra S. Panwar (S'82–M'85–SM'00) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1981 and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Massachusetts, Amherst, in 1983 and 1986, respectively.

He is a Professor in the Electrical and Computer Engineering Department, Polytechnic University of New York (now Polytechnic University), Brooklyn, NY. He is currently the Director of the New York State Center for Advanced Technology in Telecommunications (CATT). He spent the summer of 1987 as a Visiting Scientist at the IBM T. J. Watson Research Center, Yorktown Heights, NY, and has been a Consultant to AT&T Bell Laboratories, Holmdel, NJ. His research interests include the performance analysis and design of networks. He is a co-editor of two books, *Network Management and Control, Vol. II* (New York: Plenum, 1994) and *Multimedia Communications and Video Coding* (New York: Plenum, 1996) and co-author of a textbook, *TCP/IP Essentials: A Lab-Based Approach* (Cambridge, U.K.: Cambridge University Press, 2004). His current work includes video systems over peer-to-peer networks, switch performance, and wireless networks.

Dr. Panwar has served as the Secretary of the Technical Affairs Council of the IEEE Communications Society (1992-1993) and is a member of the Technical Committee on Computer Communications. He is a co-recipient of the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems.