BOUNDS, APPROXIMATIONS AND APPLICATIONS FOR A TWO-QUEUE GPS SYSTEM

Francesco Lo Presti
Dipartimento di Ingegneria Elettronica
Università di Roma "Tor Vergata"
00133 Roma, Italy

Zhi-Li Zhang, Don Towsley
Computer Science Department
University of Massachusetts
Amherst, MA 01003, USA

An abridged version of this paper appeared in Proc. of INFOCOM'96

Abstract

In this paper we study *Generalized Processor Sharing* (GPS) scheduling with Markov Modulated Fluid Sources (MMFSs). We focus on a GPS system serving two classes of customers. By using a combination of the bounding approach and the approximation approach to performance analysis and by taking advantage of the specific structure of MMFSs, we are able to derive lower bound and upper bound approximation on queue length distributions for each class of the GPS system. Numerical investigation shows that the lower bound and the upper bound approximation are very tight. Hence our work greatly improves the earlier results on GPS scheduling in [YaSi94, ZTK95] which are obtained for a more general stochastic model. Applications of our performance bounds for GPS scheduling to call admission control and bandwidth sharing are also illustrated, and comparison with FIFO and strict priority in different scenarios is made and discussed. We show that the flexibility provided by GPS does not provide much better performance than FIFO and priority when the classes only have loss requirements. However, this flexibility provides better performance when the classes exhibit delay as well as loss requirements.

Keywords: *Bandwidth Sharing, Call Admission Control, Generalized Processor Sharing, Markov Modulated Fluid Model, Performance Bound and Approximation, Quality-of-Service Guarantees, Scheduling.*

1 Introduction

Future broadband ISDNs are expected to provide a variety of services to users, integrating currently separate telephone, cable and data networks. Due to the different traffic characteristics and *Quality-of-Service* (QoS) requirements of network traffic, coexistence of voice, video, and data in the same network poses new issues in packet scheduling, call admission control, and bandwidth sharing. For example, data traffic is generally considered to be very bursty, much less sensitive to delay, but loss intolerant. On the other hand, real-time traffic such as voice and video is delay-sensitive but can tolerate some loss. Among real-time traffic, voice is usually less bursty and has a smaller bit rate, whereas video is generally burstier and has a higher bandwidth requirement. This diversity of traffic characteristics and QoS requirements indicates that different classes of traffic should be treated separately according to their respective QoS requirements instead of as a monolith.

Packet scheduling and call admission control are important means for achieving this separation and while at the same time allowing bandwidth sharing among different classes. The simultaneous needs of isolation and bandwidth sharing among classes has been identified in many recent papers (see, *e.g.* [CSZ92]). Isolation among classes is important to ensure that misbehavior of traffic in one class will not affect other classes. Bandwidth sharing can be used to exploit the statistical multiplexing gain made possible by such an integrated service network.

One of the most promising packet scheduling disciplines proposed in recent years is the *Generalized Processor Sharing* (GPS) [PG93a, Parekh92] (also known as Weighted Fair Queueing [DKS89]). One important feature of GPS is its ability to provide isolation among different classes while, at the same time, allowing bandwidth sharing among classes. Moreover, the bandwidth sharing mechanism of GPS is explicitly controllable. Because of these features, GPS has been recommended in [CSZ92, SCZ93] as a scheduling discipline to support different service classes.

The performance of GPS has been studied in a variety of setting [PG93a, PG93b, Parekh92, YaSi94, ZTK95]. Most of these approaches take the so-called bounding approach [Kur93] where performance bounds on the interested metrics such as loss or delay are derived, usually based on some very general source models, *e.g.* Cruz's *Linear Bounded Arrival Process* (LBAP) traffic model [Cruz91a] in a deterministic setting, or Yaron and Sidi's Exponentially Bounded Burstiness (E.B.B.) process model [YaSi93] in a stochastic setting. Although results derived from the above approaches are applicable to many different arrival processes such as the commonly used Markov Modulated Poisson Processes (MMPPs) and Markov Modulated Fluid Sources (MMFSs), they usually lead to quite loose bounds.

In this paper, we study GPS scheduling with Markov Modulated Fluid Sources. Due to the complexity of the GPS bandwidth sharing mechanism, the study of the general multiple-queue case is extremely difficult. In this paper we principally focus on GPS scheduling with only two service classes, each of which has its own buffer. We call such a system a two-queue GPS system. We combine the bounding approach of [ZTK95] with the approximation approach using spectral analysis techniques developed in [AMS82, EM91, EM92, EM95]. More precisely, we exploit the idea of decomposition first employed in [ZTK95]. However, by taking advantage of the specific structure of MMFSs, we are able to derive lower and upper bound systems that better approximate the original GPS system. The decomposition approach leads to the study of single-queue statistical multiplexing systems with modulated service processes. Lower and upper bounds on the queue-length distribution for each queue of the GPS system are easily obtained for this systems. Comparison with simulation shows that our lower and upper bound approximations are generally very close to the simulated values. Refined effective bandwidth approximations based on these bounds can be also easily derived. Application of these bounds for GPS scheduling in the context of call admission control and bandwidth sharing is also illustrated, and comparison with FIFO (First-In First-Out) and strict priority in different scenarios is made and discussed.

FIFO is the simplest way to provide bandwidth sharing among sessions/classes. There have been a large volume of literature on the subject, in particular, the recent development of the theory of *effective bandwidth* approach to call admission control for a single-queue statistical multiplexing system with a single QoS requirement (see, e.g., [GAN91, GH91, Kel91, EM92, KWC93, Whi93, Chang94, GWh94]). The theory of *effective bandwidth* provides a simple and elegant way to perform call admission control based on asymptotic approximations.

Strict priority scheduling of Markov Modulated Fluid Sources has been studied in [EM95]. In particular, a two-priority system is thoroughly investigated, where each priority class has its own buffer. The high priority class with explicit QoS guarantee is for real-time traffic, where low class provides *engineered best-effort* service to traffic with less stringent performance requirement such as data transport. Approximations to both queue length distributions are obtained; moreover, a refined effective bandwidth approximation called the Chernoff-Dominant Eigenvalue approximation is used for call admission control purpose. Here we remark that strict priority can be regarded as a special case of GPS scheduling. Hence our work can be considered as an extension of the work in [EM95].

Call admission control and related issues under GPS scheduling have also been investigated. In [JDSZ95], based on the results of [PG93a, PG93b, Parekh92], a heuristic call admission control algorithm using traffic measurement is proposed for the so-called *predictive service* class [CSZ92]. In [ZLKT95] call admission control schemes for multiple service classes with statistical QoS guarantees using GPS scheduling are designed for various network service models. Asymptotic study of GPS system using large deviation principle has also been carried out [dVK94, Zha95a, Zha95b].

The rest of the paper is organized as follows. In Section 2, we define GPS scheduling formally and discuss related work on GPS and statistical multiplexing system with MMFSs. In Section 3, we describe the two-queue GPS system model. In Section 4, we derive lower and upper bound approximation for the two-queue GPS system based on study of a statistical multiplexing system with modulated service process. Numerical investigations are presented to show the tightness of the bounds. In Section 5, applications of the bounds to call admission control and bandwidth sharing are illustrated. In Section 6 we consider the multiple-queue GPS system and derive lower and upper bound approximation, using a generalization of the procedures presented for the two-queue GPS system. Finally, the paper is concluded in Section 7.

2 Preliminaries and Related Work

2.1 Generalized Processor Sharing

Generalized Process Sharing (GPS) is a work-conserving scheduling discipline that can be regarded as the limiting form of a weighted round robin policy, where traffic from sessions is treated as an infinitely divisible fluid (hence there is no notion of "packet" in this traffic model [PG93a, Pare92]). Assume we have n sessions sharing a GPS server with rate c. Associated with the sessions is a set of parameters $\{\phi_i\}_{1\leq i\leq n}$ (called the *GPS assignment*) which determines the minimum sharing of bandwidth of each session. Each session is guaranteed a minimum service rate of $g_i = \frac{\phi_i}{\sum_{j=1}^{n}\phi_j}c$. More generally, if the set of sessions with queued data at time t is $B(t) \subseteq \{1,\ldots,n\}$, the session $i \in B(t)$ receives service at rate $\frac{\phi_i}{\sum_{j\in S(t)}\phi_j}$ at time t.

In [PG93a, PG93b, Parekh92], Parekh and Gallager presented a thorough examination of the GPS scheduling under a deterministic setting where the source traffic of each session is characterized as a *Linear Bounded Arrival Process* (LBAP) [Cruz91a]. A LBAP has two parameters, rate ρ and maximum burst size σ such that the amount of source traffic arriving over any time interval of length t is bounded above by $\rho t + \sigma$. Given that the traffic of each session conforms to a LBAP (as would be the case when a session is regulated by a leaky bucket mechanism) and that the total arrival rate of all the sessions is smaller than the service rate, it was shown that the backlog and delay of each session are bounded from above both for a single GPS server in isolation and a broad class of GPS networks with arbitrary topology.

In [YaSi94, ZTK95], the GPS scheduling was studied when the traffic generated by sources is modeled by an *Exponentially Bounded Burstiness* (E.B.B.) process [YaSi93]. We say an (cumulative) arrival process, A_i, is a $(\rho_i, \Lambda_i, \alpha_i)$-E.B.B. process, if for any τ and t such that $\tau \leq t$ and for any $x \geq 0$,

$$Pr\{A_i(\tau, t) \geq \rho_i(t - \tau) + x\} \leq \Lambda_i e^{-\alpha_i x} \tag{1}$$

where $A_i(\tau, t)$ denotes the amount of traffic arriving during $[\tau, t]$. Here ρ_i is called the long term *upper rate* of the arrival process, Λ_i the prefactor, and α_i the decay rate. Under the E.B.B. model, performance bounds analogous to those of the deterministic model are obtained in [ZTK95]. Given that the appropriate stability conditions are satisfied, upper bounds on the backlog and delay tail distributions for each session sharing a single GPS server are obtained in a form similar to (1) and the departure process of each is shown to be an E.B.B process as well (the latter was also proved in [YaSi94]) and a broad class of GPS networks with arbitrary topology is shown to be stable (see also [YaSi94]).

The aforementioned results, both deterministic and statistical, are derived via an important concept introduced by Parekh and Gallager: the notion of *feasible ordering*. Given the rates ρ_i (either the rate of an LBAP or an upper rate of an E.B.B. process), $1 \leq i \leq n$, an ordering of the sessions, $s_1, s_2, \ldots, s_n$, is a feasible ordering with respect to $\{\rho_i\}_{1 \leq i \leq n}$ and $\{\phi_i\}_{1 \leq i \leq n}$ if for $i = 1, 2, \ldots, n$,

$$\rho_{s_i} < \frac{\phi_{s_i}}{\sum_{j=s_i}^{s_n} \phi_j}(c - \sum_{j=s_1}^{s_{i-1}} \rho_j). \tag{2}$$

Such a feasible ordering always exists as long as $\sum_{i=1}^{N} \rho_i < r$. In [ZTK95], the notion of *feasible partition* is introduced which generalizes the notion of feasible ordering. The feasible partition is a partition of the n sessions, $\mathcal{H} = \{H_l\}_{1 \leq l \leq L}$, $H_1 \cup \cdots \cup H_L = \{1, 2, \ldots, n\}$, where each H_l , $1 \leq l \leq L$, is defined recursively as follows:

$$i \in H_1 \text{ if } \frac{\rho_i}{\phi_i} < \frac{1}{\sum_{j=1}^{n} \phi_j} c, \tag{3}$$

and for $k \geq 1$, if $H^k := H_1 \cup \cdots \cup H_k \neq \{1, 2, \ldots, n\}$, then H_{k+1} is defined such that

$$i \in H_{k+1} \text{ if } \frac{\rho_i}{\phi_i} < \frac{1}{\sum_{j \notin H^k} \phi_j}(c - \sum_{j \in H^k} \rho_j). \tag{4}$$

A GPS assignment is an RPPS GPS assignment if $\mathcal{H} = \{H_1\}$ and $H_1 = \{1, 2, \ldots, n\}$, i.e., for $1 \leq i \leq n$, $\rho_i < g_i = \frac{\phi_i}{\sum_{j=1}^{n} \phi_j} c$. In particular, if $\rho_i = \phi_i$, then $\rho_i < g_i$ assuming that $\sum_{i=1}^{n} \rho_i < c$, hence the name *Rate Proportional Processor Sharing*

As we will see later, the difficulty in analyzing the GPS system in a stochastic setting lies in that the service rate each session receives depends both on the arrival rate and queue length of all sessions. To circumvent this difficulty, [ZTK95] takes a decomposition approach to derive *upper bounds* on the session backlog and delay tail distributions; an n-queue GPS system is decomposed in a way that the sessions are decoupled, and thus they can be analyzed separately as n independent G/D/1/∞ queues.

Since the E.B.B. source model used in [ZTK95] is generic, the upper bounds obtained are expected to be rather loose due, in particular, to the value of the prefactor of the bounding exponentially decaying function. This is mostly due to the fact that fine-grained structure in the arrival processes are ignored by the E.B.B. model and, therefore, the dynamics of the system are not well captured.

2.2 Statistical Multiplexing System with Constant Service Process

In this section we briefly review the analysis of a single-queue statistical multiplexing system with Markov modulated fluid sources (MMFS) and a constant service process. We assume the capacity of the buffer is infinite and the rate of the service process is a constant c. Thus we have a MMFS/ D/ 1/∞ queueing system. Suppose we have K independent Markov modulated fluid sources labelled $k = 1, 2, \ldots, K$ where source k is characterized by the pair $(\boldsymbol{M}^{(k)}, \boldsymbol{\lambda}^{(k)})$ with $\boldsymbol{M}^{(k)}$ being the generator of the Markov chain and $\boldsymbol{\lambda}^{(k)}$ being the rate vector. Let $\mathcal{S}^{(k)}$ be the state space of source k and $\boldsymbol{\Lambda}^{(k)} = \text{diag}(\boldsymbol{\lambda}^{(k)})$. When source k is in state $s \in \mathcal{S}^{(k)}$ it generates fluid at the constant rate $\lambda_s^{(k)}$. The stationary probability vector for source k is $\boldsymbol{w}^{(k)}$, where $\boldsymbol{w}^{(k)}\boldsymbol{M}^{(k)} = 0$ and $\langle \boldsymbol{w}^{(k)}, \boldsymbol{1} \rangle = 1$ ($\langle \cdot, \cdot \rangle$ denotes the inner product of vectors, and $\boldsymbol{1}$ is a unity vector of proper dimension). The average input rate for source k is $\overline{\lambda}_k = \langle \boldsymbol{w}^{(k)}, \boldsymbol{\lambda}^{(k)} \rangle$; finally, let $\hat{\lambda}_k$ denote source k peak rate, $\hat{\lambda}_k = \max_{s \in \mathcal{S}^{(k)}} \lambda_s^{(k)}$.

The aggregate source is also a Markov modulated fluid source with state space

$$\mathcal{S} = \{\boldsymbol{s} \mid \boldsymbol{s} = (s^{(1)}, \ldots, s^{(K)}), s^{(k)} \in \mathcal{S}^{(k)}, 1 \le k \le K\}$$

and is characterized by the pair $(\boldsymbol{M}, \boldsymbol{\lambda})$, where $\boldsymbol{M}$ is the K-fold Kronecker sum

$$\boldsymbol{M} = \boldsymbol{M}^{(1)} \oplus \boldsymbol{M}^{(2)} \oplus \cdots \oplus \boldsymbol{M}^{(K)} \tag{5}$$

and $\boldsymbol{\lambda}$ is given by the diagonal of the aggregate rate matrix

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^{(1)} \oplus \boldsymbol{\Lambda}^{(2)} \oplus \cdots \oplus \boldsymbol{\Lambda}^{(K)}. \tag{6}$$

Clearly, the peak rate of the aggregate source is $\hat{\lambda} = \max_{s \in \mathcal{S}} \lambda_s = \sum_{k=1}^{K} \hat{\lambda}_k$ and the average input rate is given by $\overline{\lambda} = \sum_{k=1}^{K} \overline{\lambda}_k$. In order to have a stable system, we assume that $\overline{\lambda} < c$.

Let the random variables Σ and Q represent the stationary input state and the queue length, and define the stationary distribution vector $\boldsymbol{\pi}(q) = [\pi_s]$ as follows:

$$\pi_s(q) = Pr\{\Sigma = s, Q \le q\}, \quad s \in \mathcal{S}, 0 \le q < \infty.$$

Then the system can be described by the following Kolmogorov differential equation

$$\frac{d}{dq}\boldsymbol{\pi}(q)\boldsymbol{D} = \boldsymbol{\pi}(q)\boldsymbol{M}, \quad 0 \le q < \infty \tag{7}$$

where $\boldsymbol{D} = \boldsymbol{\Lambda} - c\boldsymbol{I}$ ($\boldsymbol{I}$ is the identity matrix and $\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda})$ is the system rate matrix) is the drift matrix.

The stationary queue length tail distribution $G(q) = Pr\{Q > q\} = 1 - \langle \boldsymbol{\pi}(q), \boldsymbol{1} \rangle$ has the following spectral representation

$$G(q) = \sum_{i:\text{Re}(z_i)<0} a_i e^{z_i q}, \quad 0 \le q \le \infty \tag{8}$$

where $\{z_i\}$ is the solution of the generalized eigenvalue problem $z\phi\boldsymbol{D} = \phi\boldsymbol{M}$.

It is known [AMS82] that there exists one real eigenvalue, z_1, called the *dominant eigenvalue*, that satisfies

$$0 > z_1 > \max_{i:\text{Re}(z_i)<0, i \ne 1} \{\text{Re}(z_i)\}. \tag{9}$$

From (8), it follows that

$$G(q) \approx e^{z_1 q} \quad q \to \infty. \tag{10}$$

Hence the dominant eigenvalue determines the asymptotic behavior of the queue length distribution.

The generalized eigenvalue problem $z\phi \boldsymbol{D} = \phi \boldsymbol{M}$ can be reformulated as a parameterized eigenvalue problem of the "essentially non-negative matrix" $(\boldsymbol{\Lambda} - \frac{1}{z}\boldsymbol{M})$ [EM92]. Let $g(z)$ denote the *maximum real eigenvalue* of this matrix. Then it can be shown that z_1 in (10) is the unique solution of the equation $g(z) = c$ for $c \in (\bar{\lambda}, \hat{\lambda})$. $g(z)$ is called the *effective bandwidth* of the aggregate source (described by $\boldsymbol{M}$ and $\boldsymbol{\lambda}$ [EM92]) and (10) the *effective bandwidth approximation* to the queue length tail distribution $G(q)$, since it captures the minimum service rate required to achieve an asymptotic queue length decay rate of z. Moreover, $g(z)$ is additive: let $g^{(k)}(z)$ be the effective bandwidth of source k, *i.e.*, the maximum real eigenvalue of the matrix $(\boldsymbol{\Lambda}^{(i)} - \frac{1}{z}\boldsymbol{M}^{(i)})$, then $g(z) = \sum_{k=1}^{K} g^{(k)}(z)$.

We now turn our attention to the output process characterization. Although the aggregation of Markov modulated fluid sources is still a Markov modulated fluid source, the output process from the single-queue multiplexing system is in general no longer so. A method to approximate the output process by a Markov modulated fluid source is proposed in [EM95], based on the techniques developed in [EM91]. Observe that when the system is in a busy period, the instantaneous rate of the output process is c, the rate of the service process. On the other hand, when the system is not in a busy period, the instantaneous rate of the output process is the same as the input rate of the aggregate source. The crucial idea is to represent the system busy state by a single state and to approximate the busy period by an exponential random variable.

Let $\mathcal{S}_U$ denote the set of underload states of the aggregate source, *i.e.*, $\mathcal{S}_U = \{\boldsymbol{s} \in \mathcal{S} : \lambda_s < c\}$, and let s_b denote the busy state. The output process is approximately characterized by a Markov modulated fluid source $(\boldsymbol{M}^{(o)}, \boldsymbol{\lambda}^{(o)})$ on the state space $\mathcal{S}^{(o)} = \mathcal{S}_U \cup \{s_b\}$, where the generator matrix and the rate vector are given as follows:

$$\boldsymbol{M}^{(o)} = \left[\begin{array}{c|c} \boldsymbol{M}_{U,U} & \boldsymbol{a} \\ \hline \boldsymbol{b} & -(1/\bar{b}) \end{array}\right] \text{ and } \boldsymbol{\lambda}^{(o)} = [\boldsymbol{\lambda}_U, c]\,. \tag{11}$$

The matrix $\boldsymbol{M}_{U,U}$ is a submatrix of $\boldsymbol{M}$ specifying the transition probability among the underload states. The vector $\boldsymbol{a}$ denotes the transition rates from the underload states to the busy period state s_b and the row vector $\boldsymbol{b}$ the transition rates from s_b to the underload states. Here $\bar{b}$ is the mean busy period. The bulk of the output process characterization lies in computing $\boldsymbol{a}$, $\boldsymbol{b}$ and $\bar{b}$, the details of which are left in Appendix B.

An important property of the above approximate characterization of the output process is that *all* of the stationary moments of the actual output process are identical to those of the process $(\boldsymbol{M}^{(o)}, \boldsymbol{\lambda}^{(o)})$.

2.3 Statistical Multiplexing System with Priority

In [EM95] a statistical multiplexing system with strict priority service is considered where each priority class has its own buffer and all priority classes share a channel of capacity c. Sources that belong to the same priority class are serviced according to FCFS. All sources are modeled by Markov modulated fluid source and are assumed to be independent of each other.

Let the aggregate sources for the high and low priority classes be modeled by the pairs $(\boldsymbol{M}^{(1)}, \boldsymbol{\lambda}^{(1)})$ and $(\boldsymbol{M}^{(2)}, \boldsymbol{\lambda}^{(2)})$ on state space $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$, respectively. The queue length distribution of the high priority class can be obtained by treating the two-priority multiplexing system as a single-queue multiplexing system with a channel capacity c, oblivious of the low priority class traffic. This analysis can be done as outlined in § 2.2. Let $g^{(1)}(z)$ be the effective bandwidth function of the high priority source, then the asymptotic decay rate of its queue length distribution is the solution to the equation $g^{(1)}(z) = c$

The analysis of the low priority class traffic proceeds by first approximately characterizing the output process of the high priority class traffic by a Markov modulated fluid sources, then considering a single-queue multiplexing system with a service rate modulated by this output process and the low priority class source as the arrival process. The asymptotic decay rate of the low priority queue length distribution is then determined by the solution to the equation

$$\tilde{g}^{(1)}(z) + g^{(2)}(z) = c \tag{12}$$

where $\tilde{g}^{(1)}(z)$ is the effective bandwidth functions of the high priority output process and $g^{(2)}(z)$ that of the low priority source. In [EM95], it is proposed that the low priority queue length tail distribution, $G_2(q) = Pr\{Q_2 > q\}$, be approximated by the following expression

$$G_2(q) \approx L_0^{(2)} e^{z^{(2)} q} \tag{13}$$

where $z^{(2)}$ is the solution to (12) and $L_0^{(2)}$ is the refined Chernoff large deviation approximation to the loss probability of the corresponding bufferless multiplexing system. (13) is called the Chernoff-Dominant Eigenvalue (CDE) approximation in [EM95] which provides a much better approximation to $G_2(q)$ than the usual effective bandwidth approximation (10) by adding an prefactor. We will call an approximation formula in the form of (13) the *refined effective bandwidth approximation* (REB) (see, *e.g*, [CLW93, LNT94]).

3 Model

In this paper we focus on the two-queue GPS system (Figure 1). For $i = 1, 2$, the session i source i is modeled as a Markov-modulated fluid process with an irreducible generator $\boldsymbol{M}^{(i)}$ on state space $\mathcal{S}^{(i)} = \{1, \cdots, N^{(i)}\}$, and a rate vector $\boldsymbol{\lambda}^{(i)} = \{\lambda_1^{(i)}, \cdots, \lambda_{N^{(i)}}^{(i)}\}$.

Let $\phi_i, i = 1, 2$ be the GPS assignment for the two sessions. Without loss of generality, we assume that $\phi_1 + \phi_2 = 1$. Hence each session is guaranteed a minimum service rate $g_i = \phi_i c$. For $i = 1, 2$, let $\overline{\lambda}_i$ denote the average input rate of session i source. As a necessary stability condition, we require that $\overline{\lambda}_1 + \overline{\lambda}_2 < c$.

For $i = 1, 2$, let $r_i(t) \in \{\lambda_1^{(i)}, \cdots, \lambda_{N^{(i)}}^{(i)}\}$ denote the rate of the session i arrival process at time t, and for any $\tau < t$, define $A_i(\tau, t) = \int_\tau^t r_i(u)\, du$, the (cumulative) arrival process for session i. Similarly, let $s_i(t)$ denote the service/departure rate of session i at time t, and define $S_i(\tau, t) = \int_\tau^t s_i(u)\, du$ to be the session i departure process. The session i backlog at time t denoted as $Q_i(t)$, is given by

$$Q_i(t) = \sup_{\tau \le t} \{A_i(\tau, t) - S_i(\tau, t)\}. \tag{14}$$

The delay experienced by a session i traffic arriving at time t is denoted as $D_i(t)$. If the traffic from a session is serviced by the FCFS scheduling discipline (as we assume in this paper), then $D_i(t)$ is the time that the session i backlog at time t, $Q_i(t)$, gets cleared. Finally, we assume that the system is empty before time zero.

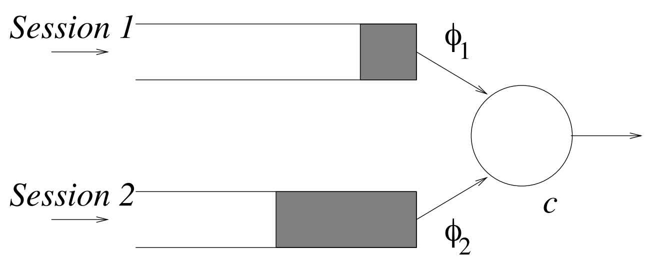

Figure 1: A two-queue GPS system.

The dynamics of the GPS systems can be described by the following sample path equations:

$$s_1(t) = \begin{cases} c - s_2(t), & \text{if } Q_1(t) > 0 \\ r_1(t), & \text{if } Q_1(t) = 0, \end{cases} \quad s_2(t) = \begin{cases} c - s_1(t), & \text{if } Q_2(t) > 0 \\ r_2(t), & \text{if } Q_2(t) = 0. \end{cases} \tag{15}$$

and

$$\frac{d}{dt}Q_1(t) = r_1(t) - [c - s_2(t)], \quad \text{if } Q_1(t) > 0 \tag{16}$$

$$\frac{d}{dt}Q_2(t) = r_2(t) - [c - s_1(t)], \quad \text{if } Q_2(t) > 0. \tag{17}$$

We note that from the definition of GPS, if both queues are not empty at time t, then $s_1(t) = g_1 = \phi_1 c$, $s_2(t) = g_2 = \phi_2 c$. If one of the sessions has an empty queue at time t, then the other session will receive the residual service from this session in addition to its guaranteed service rate. Therefore, (16) and (17) have the following equivalent formulation:

$$\frac{d}{dt}Q_1(t) = r_1(t) - \left[\phi_1 c + (\phi_2 c - r_2(t))^+ \mathbf{1}_{\{Q_2(t)=0\}}\right], \quad \text{if } Q_1(t) > 0 \tag{18}$$

$$\frac{d}{dt}Q_2(t) = r_2(t) - \left[\phi_2 c + (\phi_1 c - r_1(t))^+ \mathbf{1}_{\{Q_1(t)=0\}}\right], \quad \text{if } Q_2(t) > 0 \tag{19}$$

where $(x)^+ = \max x, 0$ and $\mathbf{1}_{\{.\}}$ is the indicator function.

4 Analysis

The difficulty in analyzing a GPS system lies in the fact that the service rate each session receives depends on the arrival rate and queue length of both sessions. As a consequence, an exact analysis of the joint queue length distribution of the two-queue GPS system involves a set of differential partial equations with coupled boundary conditions, the solution of which is very difficult to acquire, while the exact delay analysis is even more difficult.

In this section, we study the two-queue GPS systems with Markov-Modulated Fluid Sources by combining the bounding approach in [ZTK95] and the spectral analysis approximation approach used in [EM92, EM95]. As in [ZTK95], we decompose the two-queue GPS system into two independent queueing system. By taking advantage of the specific structure of the arrival processes and the system, we are able to derive sample-path upper and lower bounds on the backlog process and sample-path upper bound on the delay experienced by each session. Using the approach in [EM92, EM95], these sample-path relations can be converted to upper and lower bounds on the queue length distributions and to an upper bound on the delay distributions of the GPS system.

The rest of the section is organized as follows. In § 4.1, we study a single-queue statistical multiplexing system with a modulated service process using the techniques of [EM92, SE91, EM95], which lays the foundation for the lower and upper bound procedures proposed in § 4.2 and § 4.3. In § 4.2, we derive the lower bounds (which are the simpler of the two to derive), and in § 4.3 we describe the procedure for obtaining the upper bounds, the proofs of the proposed bounds are relegated to Appendix A. In § 4.4, further refinement and the refined effective bandwidth approximation is discussed. The numerical results obtained from these approaches are compared to simulation results in § 4.5.

4.1 Statistical Multiplexing System with Modulated Service Process

In this section we study a single-queue statistical multiplexing system with a modulated service process. The capacity of the buffer is assumed to be infinite. The aggregate arrival process is a Markovian modulated fluid source characterized by the pair $(\boldsymbol{M}^{(a)}, \boldsymbol{\lambda}^{(a)})$ on state space $\mathcal{S}^{(a)}$. The service process is modulated by another Markovian modulated fluid source characterized by the pair $(\boldsymbol{M}^{(s)}, \boldsymbol{\lambda}^{(s)})$ on state space $\mathcal{S}^{(s)}$, which is independent of the arrival process. When the arrival process is in state $s^{(a)} \in \mathcal{S}^{(a)}$, it generates fluid at constant rate $\lambda^{(a)}_{s^{(a)}}$, while when the modulating service process is in state $s^{(s)} \in \mathcal{S}^{(s)}$, the queue is served at the rate $c - \lambda^{(s)}_{s^{(s)}}$, where c is the maximum service rate (hence $0 \leq \lambda^{(s)}_{s^{(s)}} \leq c, s^{(s)} \in \mathcal{S}^{(s)}$). We note that this multiplexing system with modulated service process is very similar to the producer-consumer system studied in [Mit88].

Denote the overall system state space by

$$\mathcal{S} = \{\boldsymbol{s} \mid \boldsymbol{s} = (s^{(a)}, s^{(s)}), s^{(a)} \in \mathcal{S}^{(a)}, s^{(s)} \in \mathcal{S}^{(s)}\}$$

and by $Q(t)$ the queue length of the system at time t, then $Q(t)$ satisfies the following sample path equation:

$$\frac{d}{dt}Q(t) = \lambda^{(a)}_{\Sigma^{(a)}(t)} - [c - \lambda^{(s)}_{\Sigma^{(s)}(t)}], \text{ if } Q(t) > 0, \tag{20}$$

where $(\Sigma^{(a)}(t), \Sigma^{(s)}(t)) \in \mathcal{S}$ is a pair of random variables representing the states of the arrival process and the modulating service process at time t.

Equivalently, we can rewrite (20) as follows

$$\frac{d}{dt}Q(t) = [\lambda^{(a)}_{\Sigma^{(a)}(t)} + \lambda^{(s)}_{\Sigma^{(s)}(t)}] - c, \text{ if } Q(t) > 0. \tag{21}$$

From the above two expressions, it follows that, as far as the buffer content is concerned, the term $\lambda^{(s)}_{\Sigma^{(s)}(t)}$ can be considered either as a portion of capacity removed from the its maximum value c, or as an additional arrival rate (although the waiting time is different in the two cases) [SE91]. Therefore, the system is equivalent to a statistical multiplexing system with constant service rate c, the input of which is provided by the superposition of the Markov modulated fluid sources $(\boldsymbol{M}^{(a)}, \boldsymbol{\lambda}^{(a)})$ and $(\boldsymbol{M}^{(s)}, \boldsymbol{\lambda}^{(s)})$. Thus, the analysis of the queue length distribution resembles the one we have described in § 2.2. Therefore, the stationary tail queue length distribution, $G(q) = Pr\{Q > q\}$ has the following spectral representation

$$G(q) = \sum_{i:\mathrm{Re}(z_i)<0} a_i e^{z_i q} \quad (0 \le q < \infty) \tag{22}$$

where $\{z_i\}$ are the system eigenvalues.

The asymptotic decay rate of the tail queue length distribution is determined by the dominant eigenvalue of the system, z_1, which is the unique solution to the equation

$$g^{(a)}(z) + g^{(s)}(z) = c, \tag{23}$$

where $g^{(i)}(z)$ is the effective bandwidth function of the Markov modulated fluid source $(\boldsymbol{M}^{(a)}, \boldsymbol{\lambda}^{(a)}), i \in \{a, s\}$.

The output process characterization of a single-queue multiplexing system with modulated service process can be carried out in a similar spirit as in § 2.2 (but which is somewhat more complicated). To approximate the output process by a Markov modulated fluid source, we still model the system busy state by a single state and assume the system busy period to be exponentially distributed.

Let $(\boldsymbol{M}^{(o)}, \boldsymbol{\lambda}^{(o)})$ be the Markov modulated fluid source that approximately characterizes the output process of the system, the state space of which is $\mathcal{S}^{(o)} = \mathcal{S}_U^{(a)} \cup \{s_b\}$, where s_b is the busy state, and $\mathcal{S}_U^{(a)}$ the set of the underload input states, i.e.

$$\mathcal{S}_U^{(a)} = \{s^{(a)} \in \mathcal{S}^{(a)} \mid \exists \boldsymbol{s} = (s^{(a)}, s^{(s)}) \in \mathcal{S}, \boldsymbol{s} \in \mathcal{S}_U\},$$

The generator matrix $\boldsymbol{M}^{(o)}$ has the similar form given in (11). However, the rate vector has the form $\boldsymbol{\lambda}^{(o)} = [\boldsymbol{\lambda}_U, \bar{c}]$ where $\bar{c}$ is the average service rate when the system is in a busy period. This is because unlike in a constant service rate single-queue multiplexing system where the output rate is exactly c, the channel service rate, when the system is in a busy period, in this case the output rate depends on the state of the modulated service process. Therefore, we associate $\bar{c}$ with the single busy period. The vectors $\boldsymbol{a}, \boldsymbol{b}$, the average busy period $\bar{b}$ in $\boldsymbol{M}^{(o)}$ and the average service rate $\bar{c}$ can be computed exactly, the details of which are left to Appendix B. The above approximate characterization of the output process match all the stationary moments of the actual output process during the empty periods. Moreover, the average rate coincides with the one of the actual output process.

4.2 Lower Bound for the Two-Queue GPS System

In this section, we describe the procedure for obtaining the lower bound on the queue length distribution for each session.

We decompose the two-queue GPS system into two independent one-queue systems as depicted in Figure 2 where $r_i(t)$ is the arrival rate to the ith one-queue system at time t, $\tilde{s}_i(t)$ the service rate at time t, and $\tilde{Q}_i(t)$ the backlog of the queue at time t, $i = 1, 2$. We want to choose $\tilde{s}_i(t)$ such that $\tilde{Q}_i(t) \le Q_i(t)$ for all t.

Considering (18), if we assume that queue 2 is always empty, which amounts to ignoring the impact of the queue length of session 2 on the service rate of session 1, then queue 1 is effectively decoupled from queue 2. Hence, by choosing $\tilde{s}_1(t) = \phi_1 c + (\phi_2 c - r_2(t))^+ = c - \tilde{r}_2(t)$, where $\tilde{r}_2(t) = \min\{r_2(t), \phi_2 c\}$, we have an independent one-queue system for session 1 with a modulated service process described by $\tilde{s}_1(t) = c - \tilde{r}_2(t)$.

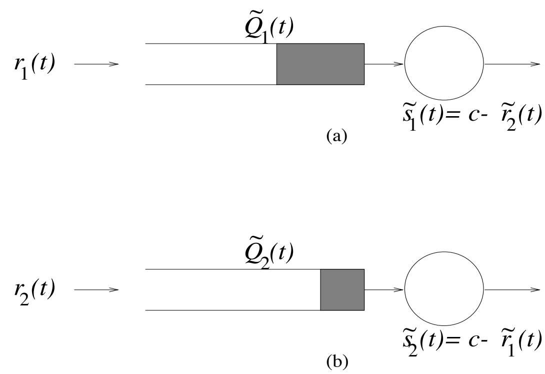

Figure 2: The decomposed lower bound system.

Similarly, by choosing $\tilde{s}_2(t) = \phi_2 c + (\phi_1 c - r_1(t))^+ = c - \tilde{r}_1(t)$, where $\tilde{r}_1(t) = \min\{r_1(t), \phi_1 c\}$, we have a one-queue system for session 2 with a modulated service process described by $\tilde{s}_2(t) = c - \tilde{r}_1(t)$.

The dynamics of the queue evolution of the two independent one-queue systems can be described as follows:

$$\frac{d\tilde{Q}_1(t)}{dt} = r_1(t) - [c - \tilde{r}_2(t)], \quad \text{if } \tilde{Q}_1(t) > 0 \tag{24}$$

$$\frac{d\tilde{Q}_2(t)}{dt} = r_2(t) - [c - \tilde{r}_1(t)] \quad \text{if } \tilde{Q}_2(t) > 0. \tag{25}$$

By comparing (18) and (19) with the above two equations, it is not too hard to see that the following is true, the proof of which is relegated to Appendix A.

Lemma 1 *For $i = 1, 2$, and for any t, the following holds*

$$Q_i(t) \geq \tilde{Q}_i(t) \tag{26}$$

It is interesting to observe that the two independent one-queue systems depicted in Figure 2 are equivalent to the two independent GPS systems in Figure 3, where one of the sessions (the session whose queue content is ignored) has no buffer, while the other has an infinite capacity buffer as usual. Lemma 1 is also evident by considering this lossy GPS system.

Using the spectral analysis technique outlined in § 4.1, the stationary distributions of $\tilde{Q}_i(t)$ can be derived. For $i = 1, 2$, the input process for session i is characterized by the pair $(\boldsymbol{M}^{(i)}, \boldsymbol{\lambda}^{(i)})$ and the modulating processes by the pair $(\tilde{\boldsymbol{M}}^{(i)}, \tilde{\boldsymbol{\lambda}}^{(i)})$. According to the expressions for service rate, the service modulating processes are related to the input processes as follows

$$\tilde{\boldsymbol{M}}^{(i)} = \boldsymbol{M}^{(i)} \tag{27}$$

$$\tilde{\boldsymbol{\lambda}}^{(i)} = \left\{\tilde{\lambda}_1^{(i)}, \cdots, \tilde{\lambda}_{N^{(i)}}^{(i)}\right\} \tag{28}$$

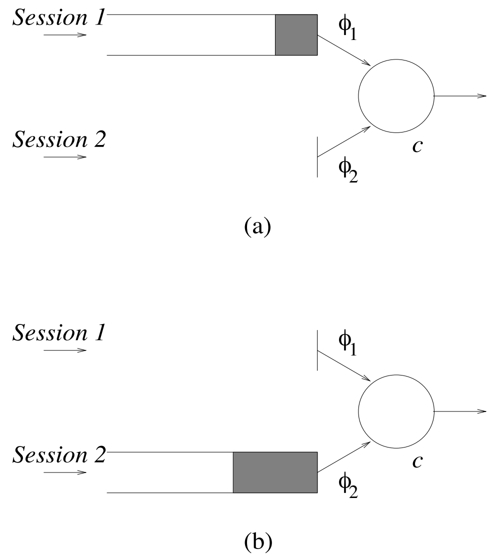

Figure 3: Equivalent lower bound system.

where $\tilde{\lambda}_s^{(i)} = \min\{\lambda_s^{(i)}, g_i\}, 1 \le s \le N^{(i)}$.

Given the stationary distribution of $\tilde{Q}_i$, from lemma 1, we have that

$$\Pr\{\tilde{Q}_i > q\} \le \Pr\{Q_i > q\},\ i = 1, 2,\ 0 \le q < \infty \tag{29}$$

where Q_i is the stationary distribution of $Q_i(t)$.

4.3 Upper Bound Approximation for the Two-Queue GPS System

In this section, we describe the procedure for obtaining the upper bound to the queue length and delay distribution for each session. For the sake of clarity, all of the proofs have been placed in Appendix A.2.

For ease of exposition, we consider each of the following two possible regimes separately:

1. $\overline{\lambda}_1 < g_1$ and $\overline{\lambda}_2 < g_2$;
2. $\overline{\lambda}_1 < g_1$ and $\overline{\lambda}_2 \ge g_2$, or $\overline{\lambda}_1 \ge g_1$ and $\overline{\lambda}_2 < g_2$.

We refer to these as the *unbiased* and *biased* regimes, respectively. In the biased regime, without loss of generality, we assume that $\overline{\lambda}_1 < g_1$ and $\overline{\lambda}_2 \ge g_2$. In the biased regime, the queue length of session 2 will not decrease on average until the session 1 queue is empty. In this case, session 2 receives the residual service rate from session 1. The strict priority case considered in [EM95] is an example of the biased regime with $\phi_1 = 1$ and $\phi_2 = 0$.

We describe the procedure for obtaining an upper bound for the unbiased regime first and, without loss of generality, we focus on session 2. The session 1 upper bound is obtained by reversing the roles of session 1 and session 2 below.

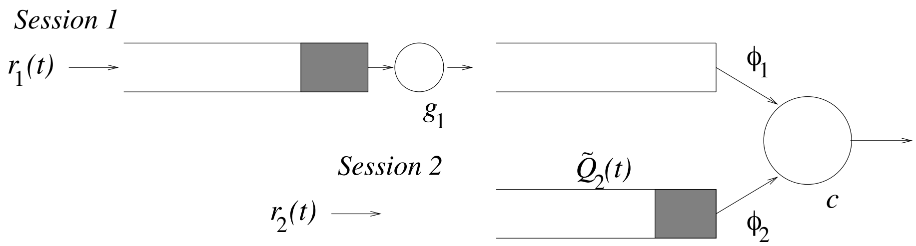

Figure 4: The Session 2 upper bound system.

In order to compute an upper bound on session 2 queue length we need to provide session 2 worse service than in the original system. Intuitively, we provide session 2 with worse service rate by making session 2 more likely to "encounter" session 1; by this, we mean to make session 1 more likely to be active when session 2 is busy. As we shall see, this can be done by "smoothing" session 1 through a queue served at rate g_1, the session 1 guaranteed rate, and replacing session 1 arrival process with this "smoother" process as illustrated in Figure 4. This results in the following two step procedure for computing an upper bound on session 2 queue length.

In the first step, we have a one-queue multiplexing system with a constant service process of rate g_1 serving session 1. Let $s_1'(t)$ denote the departure rate of this multiplexing system at time t. We have the following result relating $s_1'(t)$ to the departure rate $s_1(t)$ of session 1; see Appendix A for the proof.

Lemma 2 *Let B be a busy period of session 2 in the two-queue GPS system. Then, for any $t \in B$*

$$s_1'(t) \geq s_1(t), \tag{30}$$

In the second step, we feed the departure process specified by $s_1'(t)$ back to the two-queue GPS system as the new session 1 arrival process (Figure 4). Since $s_1'(t) \leq g_1 = \phi_1 c$, this new GPS system is equivalent to a one-queue multiplexing system with a modulated service process characterized by $\tilde{s}_2(t) = c - s_1'(t)$ (Figure 5). Let $\tilde{Q}_2(t)$ and $\tilde{D}_2(t)$ denote the queue length of this multiplexing system at time t, and the delay experienced by the traffic arriving at time t, respectively. Then it can be shown that

Lemma 3 *For any t, the following holds*

$$Q_2(t) \leq \tilde{Q}_2(t) \tag{31}$$

and

$$D_2(t) \leq \tilde{D}_2(t) \tag{32}$$

Therefore, the distribution of $\tilde{Q}_2(t)$ provides an upper bound on that of $Q_2(t)$, the queue length distribution of session 2. Similarly, the distribution of $\tilde{D}_2(t)$ provides an upper bound on that of $D_2(t)$, the delay distribution of session 2. As before, we focus only on the stationary distributions only. Let $\tilde{Q}_2$, Q_2, $\tilde{D}_2$ and D_2 be the stationary versions of $\tilde{Q}_2(t)$, $Q_2(t)$, $\tilde{D}_2(t)$ and $D_2(t)$. Then from Lemma 3, we have

$$\Pr\{Q_2 > q\} \leq \Pr\{\tilde{Q}_2 > q\},\ 0 \leq q < \infty. \tag{33}$$

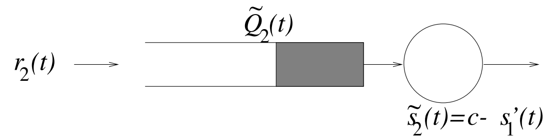

Figure 5: The decomposed Session 2 upper bound system.

and

$$\Pr\{D_2 > d\} \leq \Pr\{\tilde{D}_2 > d\} \leq \Pr\{\tilde{Q}_2 > \frac{d}{g_2}\},\ 0 \leq d < \infty, \tag{34}$$

where the last inequality in (34) follows since g_2 is the session 2 guaranteed (i.e. minimum) backlog clearing rate. The distribution of $\tilde{Q}_2$ is computed using the spectral analysis technique outlined in § 4.1. Finally, we can also compute the following upper bound on the average delay by means of Little's theorem:

$$E[D_2] \leq \frac{E[\tilde{Q}_2]}{\overline{\lambda}_2}, \tag{35}$$

where $E[\cdot]$ denotes expectation.

Unfortunely, however, we cannot explicitly compute the distributions above when, as in our case, the sessions are modeled as MMFSs. Indeed, because the output process from a single queue serving MMFSs is no longer Markovian, in the first step we can only compute an approximate characterization of the departure process specified by $s_1'(t)$. We adopt the technique developed in [EM95] as described in Appendix B.1. The departure process is approximated by a Markov modulated Fluid Source. The crucial idea is to represent the system busy state by a single state and to approximate the busy period by an exponential random variable. An important property of the above approximate characterization of the output process is that *all* of the stationary moments of the actual output process are preserved. As a consequence, in the second step we have only an approximate characterization of the modulating process; therefore, the distribution we compute must be regarded as an approximation to the upper bound distribution.

We now consider the biased regime. Recall that we have assumed that $\overline{\lambda}_1 < g_1 = \phi_1 c$ and $\overline{\lambda}_2 \geq g_2 = \phi_2 c$. Under this assumption, the upper bound approximation to the session 2 queue length and delay processes, $Q_2(t)$ and $D_2(t)$, can be obtained by using exactly the same two-step procedure outlined above for the unbiased regime. However, the upper bound approximation to the session 1 queue length and delay processes, $Q_1(t)$ and $D_1(t)$ can be obtained in just one step by considering a one-queue multiplexing system with a service process of constant rate $g_1 = \phi_1$. This is because as $\overline{\lambda}_2 \geq g_2 = \phi_2 c$, the session 2 queue is very unlikely to be empty whenever session 1 is busy. This, in turn, implies that session 1 is most likely to receive only its minimum guaranteed service g_1 to clear its backlog. Hence its queue length and delay process should be well approximated by that of the one-queue multiplexing system with a service process of constant rate g_1.

The procedure for obtaining the upper bound approximation for the biased regime results in the decomposition of the two-queue GPS system into two independent one-queue multiplexing systems as depicted in Figure 6. We remark that in the case that $\phi_1 = 1, \phi_2 = 0$, the procedure for the biased regime reduces to the strict priority case considered in [EM95].

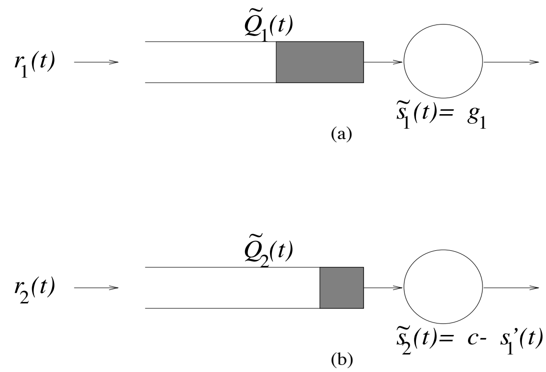

Figure 6: The decomposed upper bound system (*biased regime*).

4.4 Effective Bandwidth Approximation

The lower and upper bound procedures proposed in the previous two sections assume a complete analysis of the decomposed bounding system. As the dimension of the system grows, this becomes increasingly expensive in terms of computational cost, especially when performed on-line. To remedy the situation, as in [EM95], we use the refined effective bandwidth approximation (REB). Namely, we approximate the queue length of a session by the following formula:

$$\Pr\{Q > q\} \approx Le^{-z^* q} \tag{36}$$

where z^* is the dominant eigenvalue of the system matrix in question and L is an appropriate prefactor.

As when $q = 0$, the approximation simply yields $\Pr\{Q > 0\} \approx L$, thus L approximates the probability that the buffer is not empty. In [EM95], L is computed using the refined Chernoff large deviation approximation, which provides an upper bound on the stationary probability that the input rate exceeds the channel rate in a bufferless system. Here, we calculate L by simply computing the stationary probability that the input rate exceeds the channel capacity. This approximation underestimates the actual value of $\Pr\{Q > 0\}$, as $\Pr\{Q > 0\} \geq L$.

4.5 Numerical Investigation

In this section we present several numerical examples to compare the proposed analytical bounds with simulation results.

For $i = 1, 2$, we assume session i is modeled by a superposition of K_i on-off fluid sources each characterized by the triple $(\alpha_i, \beta_i, \lambda_i)$. When the source is in the on state it generates fluid at a constant rate λ_i; when it is in the off state, it generates no traffic. The rate at which the source changes from the off-state to the on-state is α_i, the rate from the on-state into the off state is β_i. Session i has peak rate $K_i \lambda_i$ and average rate $K_i \frac{\lambda_i \alpha_i}{\alpha_i + \beta_i}$. In the following examples, we adopt the convention used in [AMS82], namely, the unit of time corresponds to the average on period of each source and the unit of data to the amount of data generated during the average on period. The channel capacity is always assumed to be $c = 10.1$. We first investigate how well the analytical bounds perform under different GPS assignments. The sources parameters for both sessions are listed in Table

	K_i	α_i	β_i	λ_i
Session 1	10	0.4	1	1
Session 2	10	0.4	1	1

Table 1: System parameters for Example 1.

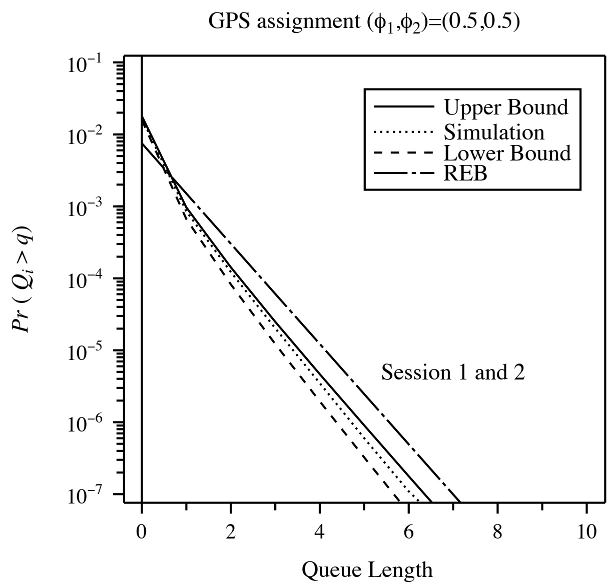

Figure 7: Bounds on queue length distributions for $(\phi_1, \phi_2) = (0.5, 0.5)$.

1. The average input rate for both session is $\overline{\lambda}_1 = \overline{\lambda}_2 = 2.857$, and the utilization factor of the channel is $\rho \approx 0.56$.

The analytical and simulation results for the following GPS assignments, $(\phi_1, \phi_2) = (0.5, 0.5)$, $(0.7, 0.3)$ and $(0.8, 0.2)$ are shown in Figures 7, 8 and 9, respectively; in Table 2 are reported the corresponding average queue length values. Moreover, the upper bound on the average delay are listed in Table 3. We note that the first two assignments yield two *unbiased* GPS systems, while the last one yields a *biased* system.

From the figures, we see that both lower and upper bounds are quite tight in all the cases and the simulation values always lie between the upper bounds and the lower bounds. In particular, for small values of q, they almost coincide with the simulations. As ϕ_1 increases from 0.5 to 0.8, we obtain increasingly tighter upper and lower bounds. The reason for this behavior is explained in the next paragraph when we compare the sessions' mean busy and empty period (here empty period refers to an interval of time during which the queue is empty). The bounds computed using the refined effective bandwidth (REB) approximation underestimate the loss probability in the small buffer region, but are generally looser when the buffer size is large.

To further compare the analytical bounds with the simulation results, we look at the mean busy and empty

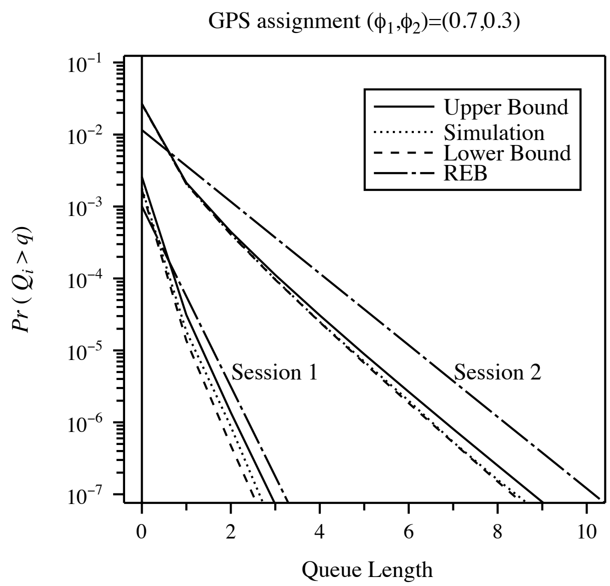

Figure 8: Bounds on queue length distributions for $(\phi_1, \phi_2) = (0.7, 0.3)$.

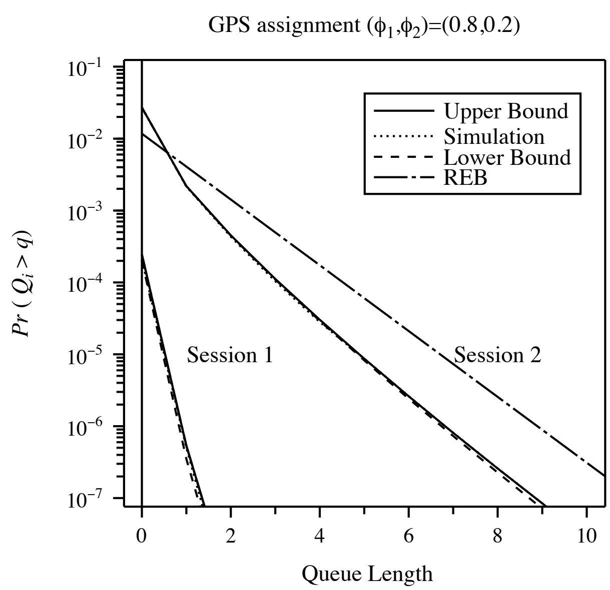

Figure 9: Bounds on queue length distributions for $(\phi_1, \phi_2) = (0.8, 0.2)$.

	$\phi_1 = 0.5$			$\phi_1 = 0.7$			$\phi_1 = 0.8$		
	Lower	Upper	Simul.	Lower	Upper	Simul.	Lower	Upper	Simul.
Session 1	3.9e-3	5e-3	4.6e-3	2.6e-4	4.6e-4	2.8e-4	2.4e-5	3.1e-5	2.6e-5
Session 2	3.9e-3	5e-3	4.6e-3	8.5e-3	9e-3	8.9e-3	9e-3	9.1e-3	9.1e-3

Table 2: Average queue length.

	$\phi_1 = 0.5$	$\phi_1 = 0.7$	$\phi_1 = 0.8$
Session 1	1.75e-3	1.61e-4	1.08e-5
Session 2	1.75e-3	3.15e-3	3.18e-3

Table 3: Upper Bounds on the average delay.

periods from the analytical lower and upper bound systems and compare them with the simulated GPS systems. The results are listed in Table 4. We see that both the analytical computed mean busy and empty periods are very close to the simulation values. In particular, we see that the lower bound system slightly underestimates the mean busy period while the upper bound system slightly overestimate it. For the mean empty period, exactly the opposite occurs. We observe that in all cases the queue for each session is empty for an overwhelming fraction of time (the probability that $\Pr\{Q_i > 0\}, i = 1, 2, i.e.$, session i is busy, lies only in the range $0.0002 \sim 0.027$). These small values of busy periods are not unexpected considering the bursty sources and extreme low loss probability with the given buffer size. From the description of the lower bound and upper bound systems in the previous sections, one can see that, when a session is in an empty period, its service rate is captured exactly, while when it is in a busy period, its service rate is either overestimated (in the case of the lower bound system) or underestimated (in the case of the upper bound system). As ϕ_1 increases and ϕ_2 decreases, session 1 is more likely to be idle while session 2 is more likely to be busy. Hence the approximation of session 2 service rate during a busy period becomes more accurate. Moreover, as the service rate for session 1 becomes large, session 1 is less dependent on the behavior of session 2. Therefore, the analytical bounds tend to be more accurate compared with the simulation bounds.

We now investigate how the busy period affects the quality of the bounds by reducing the mean off period of session 1 sources, increasing α_1 from 0.4 to 1, while leaving the other parameters unchanged, thus effectively increasing the system utilization. We fix the GPS assignment at $(\phi_1, \phi_2) = (0.5, 0.5)$. The average input rate for session 1 is 5 and the system utilization has increased to $\rho \approx 0.77$. The results are plotted in Figure 10. In this case, the mean busy period of session 1 is comparable to the mean empty period (Table 5), and as a result, the probability that the queue is not empty is now much higher ($\Pr\{Q > 0\} \approx 0.23$). The figure shows that the analytical bounds are still reasonably close to the simulation results, albeit looser than in the previous examples.

		$\phi_1 = 0.5$			$\phi_1 = 0.7$			$\phi_1 = 0.8$		
		Lower	Upper	Simul.	Lower	Upper	Simul.	Lower	Upper	Simul.
Session 1	busy	0.40	0.44	0.43	0.34	0.40	0.36	0.30	0.32	0.31
	empty	25.6	24.3	25.5	206	156	203	1528	1346	1510
Session 2	busy	0.40	0.44	0.43	0.43	0.43	0.43	0.43	0.43	0.43
	empty	25.6	24.3	25.5	15.9	15.9	15.9	15.6	15.6	15.6

Table 4: Mean busy and empty periods.

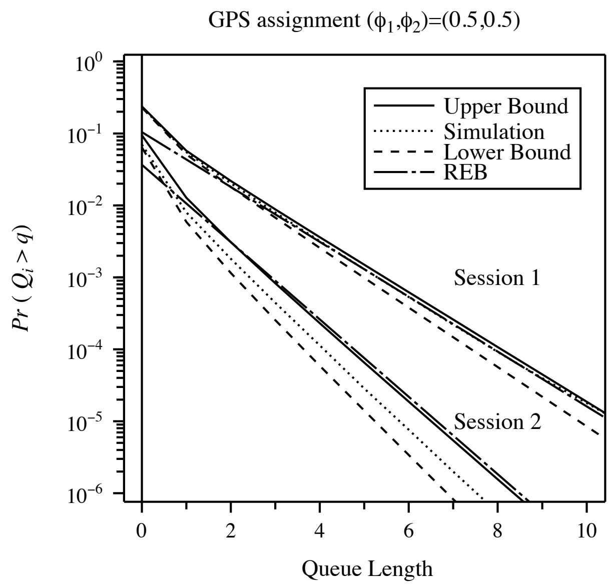

Figure 10: Bounds on queue length distributions for Example 2 ($(\phi_1, \phi_2) = (0.5, 0.5)$).

		Lower Bound	Upper Bound	Simulation
Session 1	busy	0.68	0.74	0.73
	empty	2.41	2.39	2.41
Session 2	busy	0.58	0.81	0.67
	empty	8.89	7.82	0.89

Table 5: Mean busy and empty periods of Example 2.

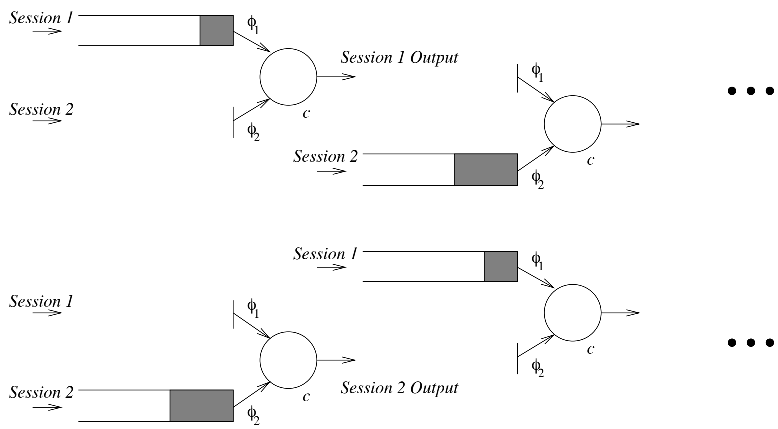

Figure 11: Iterative scheme for the lower bound.

4.6 A Further Refinement

The procedures presented in § 4.2 and § 4.3 for computing lower and upper bounds on session queue length distributions can be iterated to obtain more accurate results. Observe that the bounding procedures are based on tightly approximating the service rate each session receives; in a two-queue GPS system, as indicated by (15),(16) and (17), this simply amounts to approximate the departure process of the other session. Thus, we may expect that better results can be obtained by simply iterating the lower/upper bound procedures to obtain a better approximation of the output processes.

Consider the lower bounding system in Figure 2. Intuitively, we expect that feeding the output processes back as the arrival processes, would result in a better approximation of the service process each session receives. As a consequence, tighter lower bounds on session queue length distribution and a more accurate output process characterization could be obtained. This process can be repeated again to further improve the bounds. The resulting iterative process is schematically represented in Figure 11. The same idea can be applied for obtaining tighter upper bound approximation on the session queue length distribution. By iterating each step of the upper bound procedure, we get the iterative scheme depicted in Figure 12.

Even tough we were not able to establish that we improve the tightness of the analytical results by iterating the bounding procedures, our numerical investigation have confirmed that we do so.

Let us consider the following example. For $i = 1, 2$, we assume session i is modeled by a superposition of K_i on-off fluid sources each characterized by the triple $(\alpha_i, \beta_i, \lambda_i)$ as in §4.5. The sources parameters for both sessions are listed in Table 6. The channel capacity is assumed to be $c = 10.1$ and the GPS assignment is $(\phi_1, \phi_2) = (0.5, 0.5)$.

The analytical results and simulation results are shown in Figure 13. By carrying out one more iteration, we see from the Figure that we obtain for both sessions improved (tighter) lower bound distributions; on the contrary, no noticeable improvement is obtained for the upper bound distributions. Moreover, no considerable improvements is obtained by carrying out further iterations. Our experience with numerical investigation indicates that it is a general property of the iterative scheme to converges very fast (only few steps in all our numerical experiments). We speculate that the fast convergence is mostly due to the good approximation

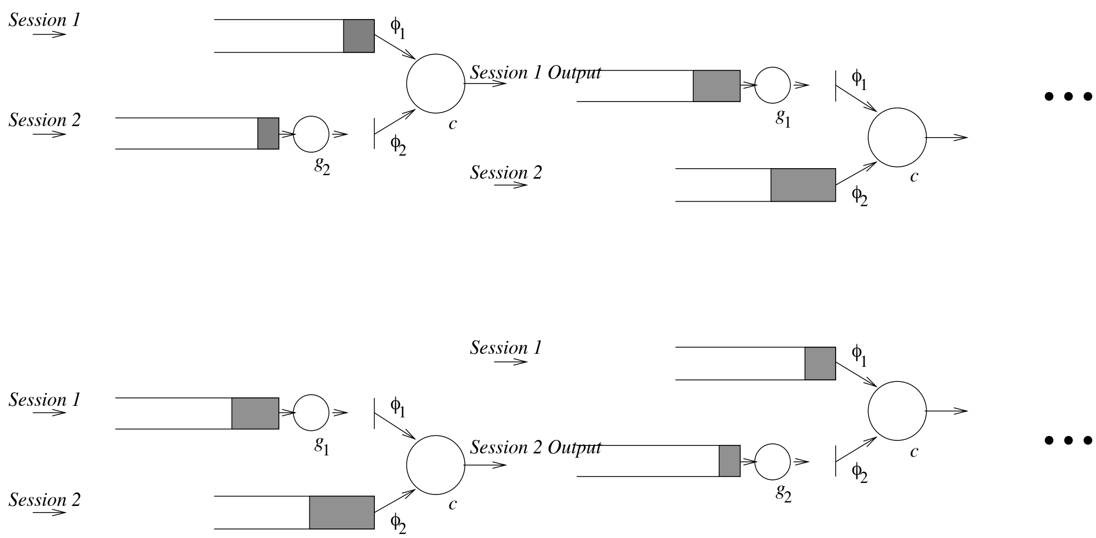

Figure 12: Iterative scheme for the upper bound.

	K_i	α_i	β_i	λ_i
Session 1	5	0.4	1	2
Session 2	5	1	1	2

Table 6: System parameters.

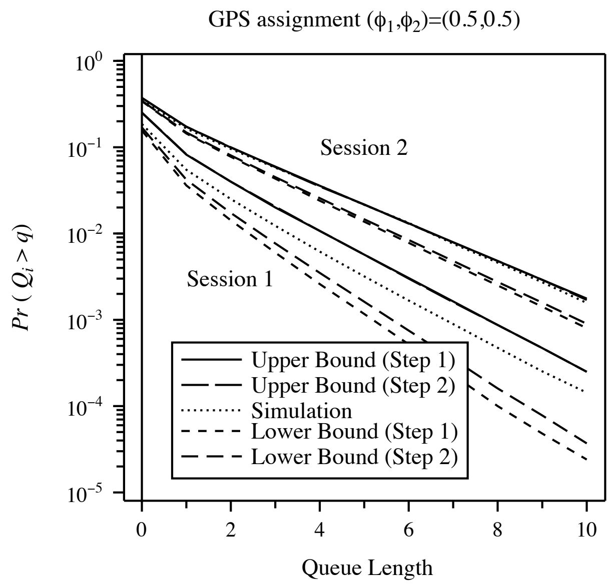

Figure 13: Bounds improvement in a two-step iteration.

resulted from the first iteration of the proposed lower and upper procedures.

5 Applications

In this section we investigate the issues that arise when performing call admission control and bandwidth allocation under GPS scheduling. By means of several examples, we study the impact of GPS scheduling on the admissible region, which is defined as the set of number of sources that can be admitted into the system without violating any QoS requirement, and we discuss the relative merits of GPS with respect to strict priority and FIFO scheduling.

5.1 Loss Probability as the Sole QoS Requirement

Consider two classes of network traffic sharing a channel of capacity c, each class having a desired loss probability bound η_i, $i = 1, 2$, as its QoS requirement. For example, one class can be real-time traffic, the other data traffic; or both classes can be real-time traffic. We are interested in comparing the admissible regions associated with the three scheduling policies, GPS, strict priority and FIFO.

For simplicity, we model sources of both classes as homogeneous on-off sources characterized by a triple $(\alpha_i, \beta_i, \lambda_i)$, the meaning of which is exactly the same as in § 4.5. The parameters for the sources are listed in Table 7. The channel capacity c is 10.1. As in § 4.5, the unit of time corresponds to the average on period of each source and the unit of data to the amount of data generated during the average on period. From the table, we see that class 1 sources are burstier than class 2 sources but that class 1 average arrival rate is smaller than

Class	α_i	β_i	λ_i	$\overline{\lambda}_i$	q_i	η_i
1	0.4	1	1	0.28	2	10^{-5}
2	0.5	0.5	1	0.5	1	10^{-7}

Table 7: System and on-off fluid source parameters for both classes

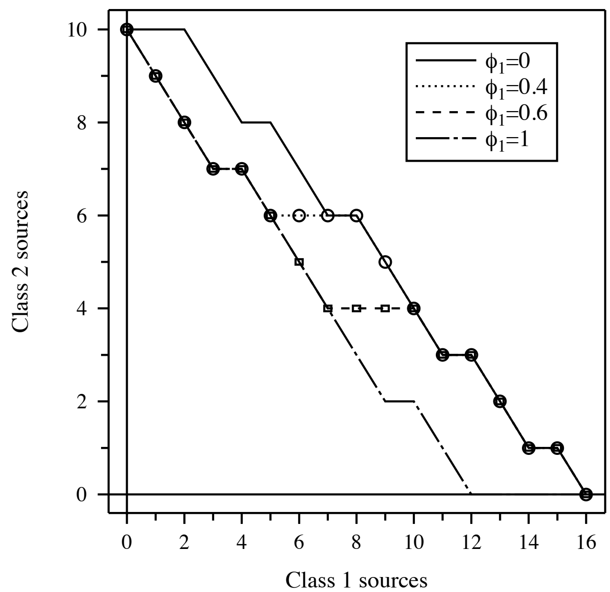

Figure 14: Admissible regions under GPS and strict priority.

that of class 2.

For GPS and strict priority scheduling, each class is assumed to have its own buffer of size q_i. We use the stationary tail distribution $Pr\{Q_i > q_i\}$ in an infinite buffer system as a slightly conservative estimation of buffer loss probability for each class. Therefore, the loss QoS requirement can be represented as the constraint $Pr\{Q_i > q_i\} \leq \eta_i$. For FIFO scheduling, the two classes share a buffer of size $q = q_1 + q_2$. However, under FIFO, the loss QoS requirement corresponds to the more stringent classes requirement, namely, the constraint is that $Pr\{Q > q\} \leq \min\{\eta_1, \eta_2\}$, where Q is the stationary aggregate queue length process. As a first example, we assume that η_1 and η_2 are different. Values of η_i and q_i for each class are listed also in Table 7.

The admissible regions under the three scheduling policies are shown in Figure 14 and Figure 15 where the admissible regions under GPS are computed via the upper bound approximation procedure. From Figure 14, we see that under the GPS scheduling, the bigger ϕ_2 is, the larger the admissible region is, which is expected as class 2 traffic has more stringent QoS requirement than class 1 traffic. Thus the strict priority policy which gives priority to class 2 (corresponding to a GPS assignment $\phi_1 = 0, \phi_2 = 1$) yields the largest admissible region, whereas reversing the priorities (corresponding to a GPS assignment $\phi_1 = 1, \phi_2 = 0$) yields the smallest. The

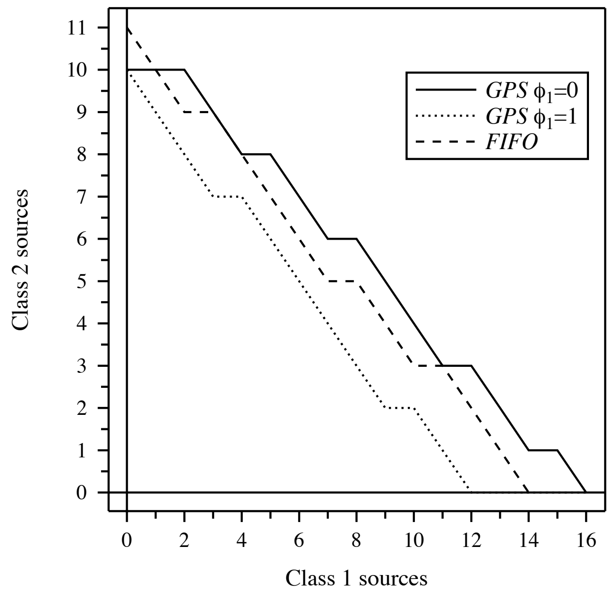

Figure 15: Admissible regions under FIFO and strict priority.

admissible region under FIFO lies between the two strict priority cases (Figure 15).

In this example where the loss requirement of the two classes are considerably different, FIFO is not really appropriate, as the buffer loss requirement is determined by the more stringent one of the two, resulting in a smaller admissible region than the strict priority scheduling with class 2 as the high priority class. GPS scheduling with explicit bandwidth sharing (*i.e.*, $\phi_1 \neq 0$ and $\phi_2 \neq 0$) does not provide any benefit over the strict priority scheduling either.

Now we consider the same two classes but with equal loss probability requirements, say, $\eta_1 = \eta_2 = 10^{-5}$. The admissible regions under FIFO and strict priority are shown in Figure 16. Clearly, by taking advantage of the shared buffer, FIFO produce the largest admissible region. Hence, in this case where the loss requirements of two classes are the same, FIFO gives the best performance.

The above examples indicate that, in the case that we have only two classes and the buffer loss probability is the sole QoS requirement, either strict priority or FIFO outperforms GPS scheduling with explicit bandwidth sharing. These observations should not be surprising.

We have performed many additional numerical investigations and found that the above observations generally hold. These results are also supported by the the recent study in [SdV95]. In [SdV95], by studying the large buffer asymptotic regime of a two-queue GPS system, it is shown that whenever the QoS requirements of the two classes are sufficiently apart, the class with the most stringent requirement should be treated as an high priority class in order to achieve the maximal admissible region.

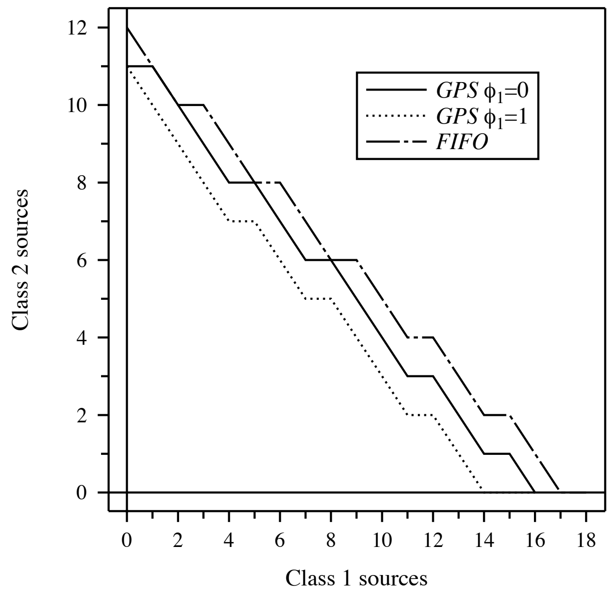

Figure 16: Admissible regions under GPS and FIFO.

5.2 Loss Probability with Maximum Delay Constraint as QoS Requirement

In this section, we investigate how the three scheduling policies perform when we add a delay constraint as an additional QoS requirement. Obviously delay is an important performance metric that should not be ignored when one or both of the classes are real-time. We consider two examples: in the first, we assume both classes have delay constraints; in the second, we assume that only class 1 has a delay constraint.

In the first example, we consider two classes of real-time traffic sharing a channel of capacity c. For each class, in addition, to a loss probability requirement there is a *maximum delay bound* $\hat{D}_i$ on the real-time traffic such that when it is not lost at the buffer, it must be serviced within $\hat{D}_i$ time units. In the case of GPS scheduling, $q_i/g_i \leq \hat{D}_i$ (where q_i/g_i is the time required to clear class i buffer at the guaranteed rate g_i) guarantees that any real-time traffic that is not lost will not experience a delay more than $\hat{D}_i$. Thus, the maximum delay constraint $\hat{D}_i$ can be translated into a lower bound $q_i/\hat{D}_i$ on the minimum guaranteed service rate, or equivalently, $\phi_i \in [q_i/\hat{D}_i c, 1]$.

For illustration purposes, we assume that the source parameters for the two classes of real-time traffic and their loss requirements are the same as listed in Table 7, and add maximum delay constraints $\hat{D}_1 = 0.5$ for class 1 and $\hat{D}_2 = 0.25$ for class 2. As a result, we have $\phi_1 \in [0.4, 0.6]$ and $\phi_2 \in [0.4, 0.6]$. This effectively eliminates strict priority scheduling as a valid choice of scheduling policy. We now determine the best choice of ϕ_1 in this range that gives the largest admissible region under the given QoS constraints.

The admissible regions computed via the upper bound method are shown in Figure 17 with three different GPS assignments $(\phi_1, \phi_2) = (0.6, 0.4)$, $(\phi_1, \phi_2) = (0.5, 0.5)$ and $(\phi_1, \phi_2) = (0.4, 0.6)$. Clearly, $(\phi_1 = 0.4, \phi_2 = 0.6)$ yields the largest admissible region, which is expected as class 2 traffic has more stringent QoS requirement than class 1 traffic. In Figure 18 the admissible regions using FIFO scheduling with buffer

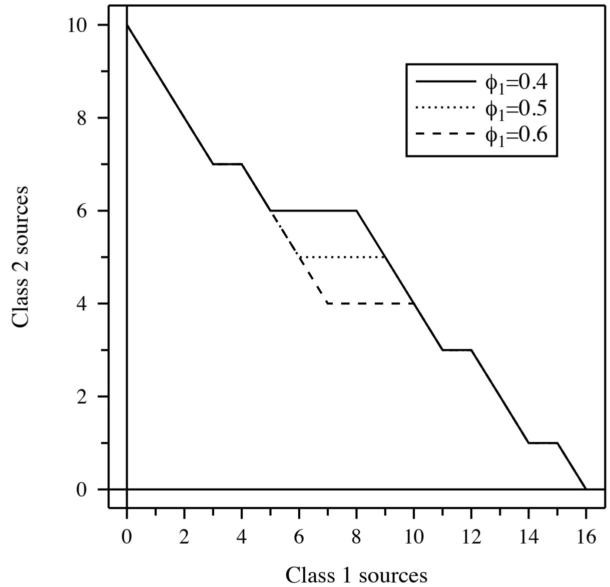

Figure 17: Admissible regions under GPS.

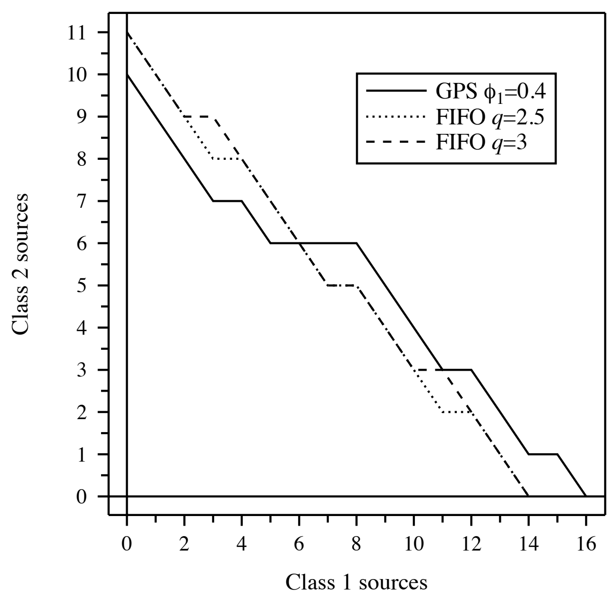

Figure 18: Admissible region under FIFO.

Class	α_i	β_i	λ_i	$\overline{\lambda_i}$	q_i	η_i
1	0.4	1	1	0.28	2	10^{-5}
2	0.1	1	1	0.09	5	10^{-9}

Table 8: System and on-off fluid source parameters for both classes .

size $q = 2.5$ and $q = 3$ are shown together with the admissible region obtained under the GPS assignment $(\phi_1 = 0.4, \phi_2 = 0.6)$. Under FIFO scheduling, the loss probability of the buffer is $\eta_2 = 10^{-7}$, the more stringent loss requirement of the two classes. In the first case of FIFO scheduling, the buffer size $q = 2.5$ is chosen so that the most stringent delay constraint $\hat{D}_2 = 0.25$ is always satisfied by the traffic that is not lost. In the second case of FIFO scheduling, the buffer $q = 3$ is the sum of two class buffers q_1 and q_2 under GPS scheduling. In this case, the maximum delay constraint of class 2 may be violated but that of class 1 is always satisfied.

From the figure, we see that the admissible region is larger under the GPS scheduling with the given GPS assignment when the number of class 2 sources is less than 6 and smaller otherwise. This is because with the larger buffer size, under FIFO more class 2 sources can be admitted when there are very few class 1 sources.

In the second example, we consider two classes of sources with different traffic characteristics and QOS requirements sharing a channel of capacity c. Class 1 sources represent real-time traffic such video or voice and class 2 sources represent non-real-time data traffic. While real-time traffic is considered to be delay-sensitive but loss-tolerant, data traffic is delay-insensitive but loss-intolerant, as data loss requires expensive retransmission. Hence the loss requirement of the data traffic is assumed to be much smaller then that for the real-time traffic, *i.e.* $\eta_2 \ll \eta_1$. We also assume that there is maximum delay bound $\hat{D}_1$ on the real-time traffic that is not lost. Given class 1 buffer size q_1 and $\hat{D}_1$ we have the constraint $g_1 \geq q_1/\hat{D}_1$ on the minimum guaranteed service rate for class 1 traffic. In the case of GPS scheduling, this means that $\phi_1 \in [q_1/\hat{D}_1 c, 1]$, while ϕ_2 is not bounded. The parameters for the sources together with the buffer sizes and loss requirements for each class are listed in Table 8. The channel capacity c is 10.1. From the table, we see that class 1 real-time sources are less bursty than class 2 data sources which has a relatively long off period. Also, the buffer size of the class 2 data traffic is considerably larger than that of class 1 real-time traffic, which is reasonable considering the delay constraint of the traffic.

We assume a maximum delay $\hat{D}_1 = 0.33$, which imposes a minimum guaranteed bandwidth of $g_1 = 0.6c$. In other words, $\phi_1 \in [0.6, 1]$. We consider the following GPS assignments: $(\phi_1, \phi_2) = (1, 0), (\phi_1, \phi_2) = (0.8, 0.2)$ and $(\phi_1, \phi_2) = (0.6, 0.4)$ and examine how different choice of ϕ_i affects the number of sources in each class that can be supported without violating the targeted loss requirements, *i.e.*, the admissible region. Note that the GPS assignment $(\phi_1, \phi_2) = (1, 0)$ corresponds to strict priority scheduling where class 1 real-time traffic has priority over class 2 data traffic. For GPS assignment $(\phi_1, \phi_2) = (1, 0)$, (0.8,0.2) and (0.6,0.4), the admissible regions computed via the upper bound approximation procedure are shown in Figure 19.

From the figure, we see that decreasing ϕ_1 from 1 to 0.6 effectively enlarges the admissible region while still satisfying the QoS requirement of each class.

For example, given the presence of 12 class 1 sources in the system, under the strict priority (*i.e.*, $(\phi_1, \phi_2) =$

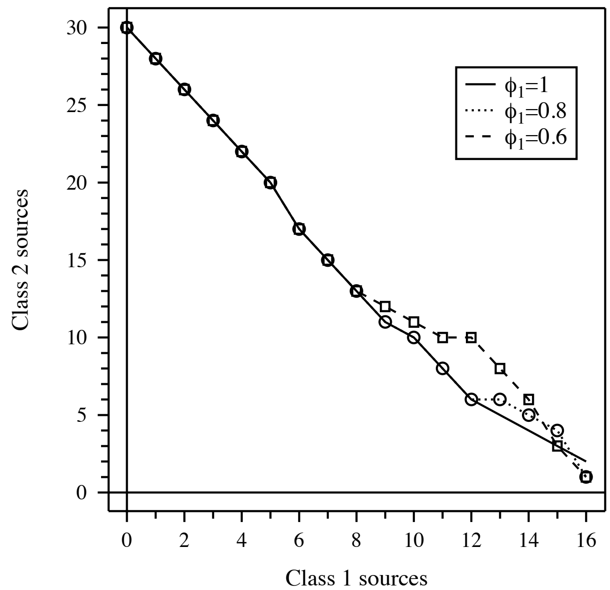

Figure 19: Admissible region via upper bounds.

$(1, 0)$), only additional 6 class 2 sources can be admitted without violating its targeted loss QoS requirement, while in the case of GPS scheduling with GPS assignment $(\phi_1, \phi_2) = (0.6, 0.4)$, 10 additional class 2 sources can be admitted, an almost double-fold increase in utilization of class 2 traffic. We remark that this increase of system utilization is gained at the expense of a possible longer delay seen by the class 1 traffic. But this delay is always kept within the specified maximum delay by the minimum bandwidth guarantee for class 1 sources. It is interesting to note that if we remove this maximum delay constraint, the policy that gives class 2 data traffic priority over class 1 real-time traffic produces the largest admissible region (*i.e.*, $(\phi_1, \phi_2) = (0, 1)$). This should come as no surprise as, from Table 8, we see that class 2 sources are burstier, have lower average rate and demand a more stringent loss QoS requirement.

In this example, FIFO scheduling is not really appropriate. With a single shared buffer for traffic of both classes, the size of the buffer will be constrained by the maximum delay constraint of the real-time traffic, while the buffer loss probability is determined by the class with most stringent loss requirement, *i.e*, class 2 data traffic in this case. As a result, we have a much smaller admissible region as shown in Figure 20.

This examples indicate that GPS scheduling provides flexibility in providing explicit bandwidth sharing, which is absent in strict priority and FIFO scheduling, when QoS guarantees include both loss and delay constraints.

6 The Multiple-queue GPS System

In the previous sections, we have focused on the two-queue GPS system. In this section we turn our attention to the general Multiple-queue GPS System case, where n sessions, each of which has its own buffer, are serviced

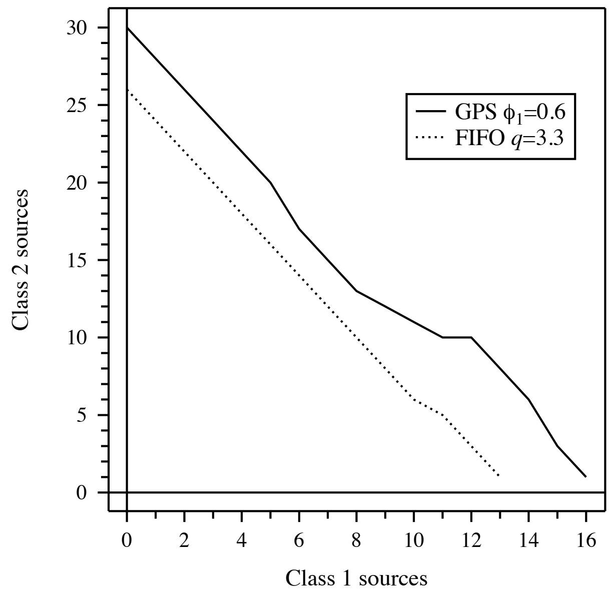

Figure 20: Admissible region under FIFO.

by a GPS server (Figure 21).

For $i = 1, \ldots, n$, the session i source is modeled as a Markov-modulated fluid process with an irreducible generator $\boldsymbol{M}^{(i)}$ on state space $\mathcal{S}^{(i)} = \{1, \cdots, N^{(i)}\}$, and a rate vector $\boldsymbol{\lambda}^{(i)} = \{\lambda_1^{(i)}, \cdots, \lambda_{N^{(i)}}^{(i)}\}$. Let $\phi_i, i = 1 \cdots, n$, be the GPS assignment for the n sessions. Without loss of generality, we assume that$\sum_{i=1}^n \phi_i = 1$. For $i = 1 \cdots, n$, let $\overline{\lambda}_i$ denote the average input rate of session i. As a necessary stability condition, we require that $\sum_{i=1}^n \overline{\lambda}_i < c$. For $i = 1, 2, \cdots, n$, let $r_i(t) \in \{\lambda_1^{(i)}, \cdots, \lambda_{N^{(i)}}^{(i)}\}$ denote the rate of session i at time t and, for any $\tau < t$, define $A_i(\tau, t) = \int_\tau^t r_i(u)\, du$, to be the cumulative arrival process for session i. Similarly, let $s_i(t)$ denote the service/departure rate of session i at time t, and $S_i(\tau, t) = \int_\tau^t s_i(u)\, du$ to be the session i departure process. The session i backlog at time t is denoted as $Q_i(t)$ and the delay experienced by a session i traffic arriving at time t is denoted as $D_i(t)$. We assume the system be empty before time zero.

Due to the complexity of the GPS bandwidth sharing mechanism, the analysis of the multiple-queue case is much more difficult than the two-queue case. Observe that in a two-queue GPS system the unused portion of bandwidth by one session is completely available to the other session. In the general case, on the other hand, the unused bandwidth by one session is shared by all the busy sessions. For example, assume that at time t a session, say session j is not busy, and let $B(t)$ denote the set of backlogged sessions at time t. If $i \in B(t)$, then the extra-bandwidth it obtains from session j at time t is

$$(g_j - r_j(t))\,\frac{\phi_i}{\sum_{k \in B(t)} \phi_k}. \tag{37}$$

From (37), it is clear that this share of bandwidth is a function of time, as $B(t)$ changes dynamically with t. This complicated dynamic coupling between sessions is extremely difficult to capture, thus making the multiple-queue GPS system much harder to analyze.

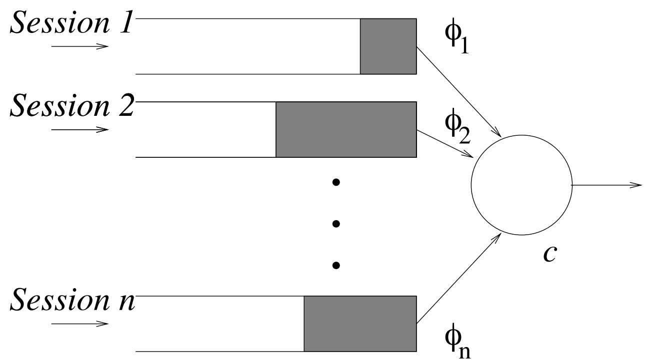

Figure 21: A Multiple-queue GPS system.

In the rest of the section, we derive lower and upper bound approximations for the multiple queue GPS system, using generalizations of the procedures presented for the two-queue GPS System. However, in light of the above observation, the bounds are expected to be looser as the number of sessions increases.

6.1 Lower Bound for the Mulitple-queue GPS System

In this section, we describe the procedure for obtaining the lower bound on the queue length distribution for each session.

We decompose a multiple n-queue GPS system into n independent one-queue systems where $r_i(t)$ is the arrival rate to the ith one-queue system at time t, $\tilde{s}_i(t)$ the service rate at time t, and $\tilde{Q}_i(t)$ the backlog of the queue at time t, $i = 1, 2, \cdots, n$. We want to choose $\tilde{s}_i(t)$ such that $\tilde{Q}_i(t) \leq Q_i(t)$ for all t. For $i = 1, \cdots, n$, as already observed in §4.2, we can easily obtain such a system by considering a system, where only session i has a buffer and all remaining sessions are bufferless (see Figure 22 for the lossy system for session 1). This lossy system can be reduced to an equivalent one-queue system serving session i with a suitable service process $\tilde{s}_i$ characterized by the pair $(\tilde{\boldsymbol{M}}^{(i)}, \tilde{\boldsymbol{\lambda}}^{(i)})$ as follows (see Figure 23 for the session 1 lower bound system). The state space is

$$\tilde{\mathcal{S}}^{(i)} = \mathcal{S}^{(1)} \cdots \times \mathcal{S}^{(i-1)} \times \mathcal{S}^{(i+1)} \cdots \times \mathcal{S}^{(n)},$$

the transition rate matrix is

$$\tilde{\boldsymbol{M}}^{(i)} = \boldsymbol{M}^{(1)} \cdots \oplus \boldsymbol{M}^{(i-1)} \oplus \boldsymbol{M}^{(i+1)} \cdots \oplus \boldsymbol{M}^{(n)}$$

and the rate vector is

$$\tilde{\boldsymbol{\lambda}}^{(i)} = \left\{ \tilde{\lambda}^{(i)}_{\tilde{s}^{(i)}} \right\}_{\tilde{s}^{(i)} \in \tilde{\mathcal{S}}^{(i)}}$$

where $\tilde{\lambda}^{(i)}_{\tilde{s}^{(i)}}$, $\tilde{s}^{(i)} = (s^{(1)}, \ldots, s^{(i-1)}, s^{(i+1)}, \ldots, s^{(n)}) \in \tilde{\mathcal{S}}^{(i)}$ is the service rate available to session i in the lossy system when session j is in state $s^{(j)} \in \mathcal{S}^{(j)}$, $j \neq i$.

The value of $\tilde{\lambda}^{(i)}_{\tilde{s}^{(i)}}$, $\tilde{s}^{(i)} \in \tilde{\mathcal{S}}^{(i)}$, can be easily expressed by using the notion of *partial feasible partition* [ZLKT95], an extension of the notions of partial feasible ordering and feasible partition. Let $N = \{1, 2, \ldots, n\}$.

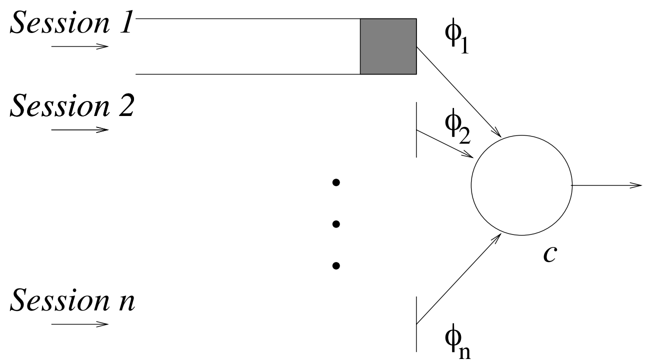

Figure 22: The lossy session 1 GPS system.

A partial feasible partition, $H_1, \ldots, H_n$, of N with respect to $\{r_j \geq 0, j = 1, \ldots, n\}$ is defined recursively as follows: for $1 \leq k \leq n$,

$$\beta_k = \frac{1}{\sum_{j \in N \setminus H^{k-1}} \phi_j}(c - \sum_{j \in H^{k-1}} r_j) \tag{38}$$

and

$$H_k = \{j \in N \setminus H^{k-1} : r_j \leq \phi_j \beta_k\}. \tag{39}$$

where $H^0 = \emptyset$ and $H^k := H_1 \cup \cdots \cup H_k$ for $k \geq 1$.

Suppose p is such that $H_p \neq \emptyset$ but $H_{p+1} = \emptyset$. Clearly, $H^p := H_1 \cup \cdots H_p \subseteq N$. The term "partial" feasible partition is used because the containment can be strict.

For each state $\tilde{s}^{(i)} = (s^{(1)}, \ldots, s^{(i-1)}, s^{(i+1)}, \ldots, s^{(n)}) \in \tilde{\mathcal{S}}^{(i)}$, it is easy to verify that the value of the service rate is

$$\tilde{\lambda}^{(i)}_{\tilde{s}^{(i)}} = \beta_p \phi_i \tag{40}$$

where $H_1, \ldots, H_n$ is the partial feasible partition with respect to

$$\{\lambda^{(1)}_{s^{(1)}}, \ldots, \lambda^{(i-1)}_{s^{(i-1)}}, c, \lambda^{(i+1)}_{s^{(i+1)}}, \ldots, \lambda^{(n)}_{s^{(n)}}\} \tag{41}$$

and p is such that $H_p \neq \emptyset$ but $H_{p+1} = \emptyset$.

Intuitively, the partial feasible partition with respect to (41) describes the behavior of the session i lossy system in state $\tilde{s}^{(i)}$. Consider session $j, j \neq i$ in state $s^{(j)} \in \mathcal{S}^{(j)}$. If $j \in H_1 = \{k : \lambda^{(k)}_{s^{(k)}} \leq g_k = \phi_k \beta_1\}$, the minimum guaranteed service rate g_j ensures that session j suffers no loss when in state $s^{(j)}$, independently of the other sessions behavior. Instead, if $j \notin H_1$, but $j \in H_q, q \leq p$, even if $g_j < \lambda^{(j)}_{s^{(j)}}$ session j suffers no loss because there is sufficient bandwidth, not used by sessions in $H^{q-1} = H_1 \cup \cdots \cup H_{q-1}$, that is available to session j. On the other hand, if session $j \notin H^p$ the available bandwidth to session j, $\phi_j \beta_p$, will not be sufficient to prevent session j from suffering loss. By using the value of c as session i rate in (41), we ensure that $i \notin H^p$ and, as a consequence, we obtain the session i available rate by (40).

Using the spectral analysis technique outlined in § 4.1, the stationary distributions of $\tilde{Q}_i(t)$ can be derived. Given the stationary distribution of $\tilde{Q}_i$, it is easy to show that (the same arguments used for the two-queue case

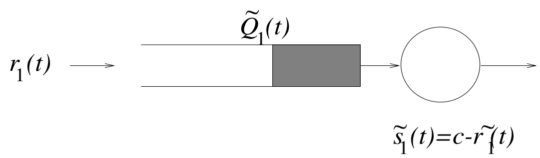

Figure 23: The lower bound system for session 1.

still apply)

$$\Pr\{\tilde{Q}_i > q\} \leq \Pr\{Q_i > q\},\ 0 \leq q < \infty \tag{42}$$

where Q_i is the stationary distribution of $Q_i(t)$. To this purpose we need to characterize the modulating process $\tilde{r}_i(t)$, such that $\tilde{s}_i(t) = c - \tilde{r}_i(t)$; the characterization can be directly derived from (43), the detail of which are omitted.

6.2 Upper Bound Approximation for the Multiple-queue GPS System

In this section, we describe the procedure for obtaining the lower bound on the queue length distribution for each session. The proofs are relegated in Appendix A.3

As in the two-queue case we consider the two following possible regimes separately:

1. For $i = 1, \cdots, n, \overline{\lambda}_i < g_i$.
2. There exists at least one $j \in \{1, \cdots, n\}$ such that $\overline{\lambda}_j > g_j$.

Both regimes can be regarded as a generalization of the regimes we have considered for the two-queue GPS system; as before, we will call the first regime *unbiased* and the second *biased*. These two different regimes can be related to the notion of *feasible partition* introduced in [ZTK95] (see §2). If we consider the *feasible partition* $\mathcal{H} = \{H_1, \cdots, H_L\}$, $1 \leq L \leq n$, $H_l \neq \emptyset$ for $l = 1, \cdots, L$, we observe that the unbiased regime corresponds to the class of RPPS GPS assignments, where $\mathcal{H} = \{H_1\}$, while the biased regime encompasses all the other possible GPS assignments, *i.e.*, $\mathcal{H} = \{H_1, \cdots, H_L\}$, $1 < L \leq n$. Note that, by definition, the session i guaranteed rate exceeds the average arrival rate only if $i \in H_1$; if session i belong to class $H_l, l > 1$, its queue length will not decrease on the average unless all session belonging to $H^{l-1} = H_1 \cup \cdots \cup H_{l-1}$ are idle.

We describe the procedure for obtaining upper bound approximation for the unbiased regime first. The procedure consists of two steps. In the first step, we characterize the departure process of each session while assuming the other sessions are always busy. It is clear that, if we assume all the other sessions as busy, session i, $i = 1, \cdots, n$, enjoys a service rate of exactly $g_i = \phi_i c$, whenever its queue is not empty. Equivalently, we have a one-queue multiplexing system with a constant service rate g. Let $s'_i(t)$ denote the departure rate of this multiplexing system at time t. Since $s'_i(t)$ is no longer Markovian, we can only compute an approximate characterization; as before, we use the technique of [EM95] as described in Appendix B.1. The departure process is *approximated* by a Markov Modulated Fluid Source $\boldsymbol{M}'^{(i)}$ and rate vector $\boldsymbol{\lambda}'^{(i)}$.

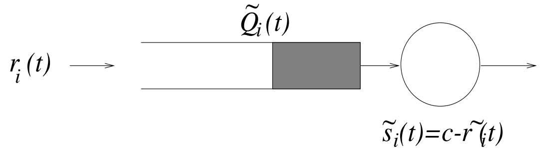

Figure 24: The session i upper bound system.

In the second step, for session $i, i = 1, \cdots, n$, we consider one-queue multiplexing system with modulated service modulated process (Figure 24) characterized by

$$\tilde{s}_i(t) = g_i + \sum_{j \neq i} \frac{\phi_i}{\sum_{k \neq j} \phi_k} \left(g_j - s'_j(t) \right). \tag{43}$$

The expression above comprises two terms: the first term is the guaranteed bandwidth session i receives; as shown in Appendix A.3, the second term represents a lower bound on the amount of extra service rate session i receives at time t. Let $\tilde{Q}_i(t)$ and $\tilde{D}_i(t)$ denote the queue length of this multiplexing system at time t, and the delay experienced by the traffic arriving at time t, respectively. Then, it can be shown that

Lemma 4 *For any t, the following holds*

$$Q_i(t) \leq \tilde{Q}_i(t) \tag{44}$$

and

$$D_i(t) \leq \tilde{D}_i(t) \tag{45}$$

Therefore, the distribution of $\tilde{Q}_i(t)$ provides an upper bound on that of $Q_i(t)$, the queue length distribution of session i of the multiple-queue GPS system. Similarly, the distribution of $\tilde{D}_i(t)$ provides an upper bound on that of $D_i(t)$, the delay distribution of session i. As before, we look at the stationary distributions only. Let $\tilde{Q}_i$, Q_i, $\tilde{D}_i$ and D_i be the stationary versions of the distributions of $\tilde{Q}_i(t)$, $Q_i(t)$, $\tilde{D}_i(t)$ and $D_i(t)$. Then from Lemma 4, we have

$$\Pr\{Q_i > q\} \leq \Pr\{\tilde{Q}_i > q\},\ 0 \leq q < \infty, \tag{46}$$

and

$$\Pr\{D_i > d\} \leq \Pr\{\tilde{D}_i > d\} \leq \Pr\{\tilde{Q}_i > \frac{d}{g_i}\},\ 0 \leq d < \infty, \tag{47}$$

where the last inequality in (47) follows since g_i is the session i guaranteed (i.e. minimum) backlog clearing rate. The distribution of $\tilde{Q}_i$ is computed using the spectral analysis technique outlined in § 4.1. To this purpose we need to characterize the modulating process $\tilde{r}_i(t)$, such that $\tilde{s}_i(t) = c - \tilde{r}_i(t)$; the characterization can be directly derived from (43), the detail of which are omitted.

Now we consider the biased regime. It is important to observe that for the biased regime, it is not possible to adopt the same approach of the unbiased regime. It is easy to verify, indeed, that an expression like (43) for session i service rate in the decomposed system, while still providing a lower bound on session i service rate, does not always ensure the stability of the decomposed system. The difficulty we encountered in defining a suitable service rate for deriving an upper bound approximation for the biased regime has prevented us from finding a satisfactory solution for the analysis of a general multiple-queue GPS system; this will be the attention of further investigation.

In the biased case, however, it is still possible to derive an upper bound approximation for session in class H_1 (*i.e.*, sessions the guaranteed rate of which exceeds the average arrival rate). The procedure is readily obtained from the unbiased case by considering only sessions in H_1. The intuition is that H_1 sessions are most likely to access only their guaranteed service rate plus only the minimum portion of service from idle H_1 sessions, since it is very unlikely the sessions not in H_1 are idle (unless all the session in H_1 are also idle).

7 Conclusion

In this paper, we study GPS scheduling with Markov Modulated Fluid Sources. For simplicity, we focus on a two-queue GPS system with two classes. We combined the bounding approach of [ZTK95] with the approximation approach based on spectral analysis techniques developed in [AMS82, EM91, EM92, EM95]. By employing the key idea of decomposition [ZTK95] and by taking advantage of the specific structure of MMFSs, we are able to derive lower and upper bound systems that produce very tight approximation to the original GPS system, as shown by numerical examples. Application of these bounds for GPS scheduling in the context of call admission control and bandwidth sharing is also illustrated, and comparison with FIFO and strict priority in different scenarios is made and discussed. We show that GPS scheduling can provide flexibility in bandwidth allocation and sharing which is absent in strict priority and FIFO scheduling, while still satisfying the QoS requirement of each class. By providing a minimum guaranteed service rate to each class, delay constraints can be taken into consideration simultaneously with loss probability requirements. By removing the delay constraint used in our examples, our experience shows that strict priority is generally the best choice in providing loss probability guarantees when the loss probability requirements for the two classes are very different; whereas FIFO performs the best when the two classes have very similar loss probability requirements. These observations should not be surprising.

The analysis can be extended to a multiple-queue GPS system with more than two classes, although the interaction among different classes are much harder to capture as also evidenced by the large deviation study in [Zha95b]. An other interesting direction for future research is the characterization of output processes of GPS system, this will be useful when we study a network of GPS servers. Besides the simplistic bound on delay using minimum guaranteed service rate, more elaborate bounds on the delay distribution that accounts for the dynamics of Markov-Modulated Fluid Sources is also desirable, and will be the attention of further investigation.

Acknowledgment
The second author would like to thank Anwar Elwalid and Debasis Mitra for many inspiring conversations on the subject while he was a summer intern at AT&T Bell Lab, Murray Hill, New Jersey. The simulator we use in the paper is based on the simulator originally written by Anwar Elwalid. The authors would also like to thank Jim Kurose for many helpful comments.

References

[AMS82] D. Anick, D. Mitra and M. M. Sondhi, Stochastic Theory of a Data Handling System with Multiple Sources, *Bell System Technical Journal*, Vol. 61, No. 8, Oct. 1982, pp. 1871-1894.

[Chang94] C. S. Chang, Stability, Queue Length and Delay of Deterministic and Stochastic Queueing Networks, *IEEE Transaction on Automatic Control*, May 1994.

[CLW93] G. L. Choundry, D. M. Lucantoni and W. Whitt, Squeezing the Most out of ATM, Preprint, 1993.

[Cruz91a] R. L. Cruz, A Calculus for Network Delay, Part I: Network Elements in Isolation, *IEEE Transaction on Information Theory*, Vol. 37, No. 1, Jan. 1991, pp. 114-131.

[Cruz91b] R. L. Cruz, A Calculus for Network Delay, Part II: Network Analysis, *IEEE Transaction on Information Theory*, Vol. 37, No. 1, Jan. 1991, pp. 132-141.

[CSZ92] D. Clark, S. Shenker and L. Zhang, Supporting Real-Time Applications in an Integrated Service Packet Network: Architecture and Mechanism. In *Proceedings of ACM SIGCOMM'92*, pp. 14-26, 1992.

[dVK94] G. de Veciana and G. Kesidis, "Bandwidth Allocation for Multiple Qualities of Service Using Generalized Processor Sharing", Revised Version, Preprint. An earlier version appeared in *Proceedings of IEEE GLOBECOM'94*, 1994.

[DKS89] A. Demers, S. Keshav and S. Shenker, Analysis and Simulation of a Fair Queueing Algorithm, *Journal of Internetworking: Research and Experience*, 1, pp. 3-26, 1990. Also in *Proceedings of ACM SIGCOMM '89*, pp. 3-12.

[EM91] A. Elwalid and D. Mitra, Analysis and Design of Rate-Based Congestion Control of High Speed Networks, I: Stochastic Fluid Models, Access Regulation, *Queueing Systems*, Vol. 9, 1993, pp. 29-64.

[EM92] A. Elwalid and D. Mitra, Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks, *IEEE/ACM Trans. on Networking*, Vol. 1, No. 3, June 1993, pp. 329-357.

[EM95] A. Elwalid and D. Mitra, Analysis, Approximation and Admission Control of a Multi-Service Multiplexing System with Priorities, INFOCOM'95, 1995.

[GAN91] R. Guérin, H. Ahmadi and M. Naghshineh, Equivalent Capacity and Its Application to Bandwidth Allocation in High-Speed Networks, *IEEE Journal on Selected Areas in Communications*, Vol. 9, No. 7, Sept.,1991, pp. 968-981.

[GH91] R. J. Gibbens and P. J. Hunt, Effective Bandwidth for Multi-Type UAS Channel, *QUESTA*, No. 9, pp. 17-28, 1991.

[GWh94] P. W. Glynn and W. Whitt, Logarithmic Asymptotics for Steady-State Tail Probabilities in a Single Server Queue,*J. Appl. Prob.*, Vol. 31, 1994.

[How71] R. A. Howard, Dynamic Probabilistic Systems, vol. 1: Markov Models, Wiley, New York, 1971.

[JDSZ95] S. Jamin, P. B. Danzig, S. Shenker, L. Zhang, A Measurement-based Admission Control Algorithm for Integrated Services Packet Network, In proceedings of *ACM SIGCOMM'95*. An earlier version of the paper with the authors S. Jamin, S. Shenker, L. Zhang and D. D. Clark and the title "An Admission Control Algorithm for Predicative Real-Time Service (Extended Abstract)" appeared in *Third International Workshop on Network and Operating System Support for Digital Audio and Video*, pp. 308-315, Nov. 1992.

[Kel91] F. P. Kelly, Effective Bandwidths for Multi-Class Queues, *QUESTA*, Vol. 9, pp. 5-16, 1991.

[Kei79] J. Keilson, Markov Chain Models - Rarity and Exponentiality, Springer, New York, 1979.

[KWC93] G. Kesidis, J. Walrand and C. S. Chang, Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources, *IEEE/ACM Trans. Networking*, Aug. 1993.

[Kur93] J. F .Kurose, Open Issues and Challenges in Providing Quality of Service Guarantees in High-Speed Networks, *ACM Computer Communication Review*, pp. 6-13, Jan. 1993.

[LNT94] Z. Liu, P. Nain and D. Towsley, Exponential Bounds for a Class of Stochastic Processes with Application to Call Admission Control in Networks, to appear in the Proc. of the 33rd *Conference on Decision and Control (CDC'93)*, February, 1994.

[Mit88] D. Mitra, Stochastic Theory of a Fluid Model of Producers and Consumers Coupled by a Buffer, *Adv. Applied Probability*, Vol. 20, 1988, pp. 646-676.

[Parekh92] A. K. Parekh, A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks, Ph.D Thesis, Department of Electrical Engineering and Computer Science, MIT, February 1992.

[PG93a] A. K. Parekh and R. G. Gallager, A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single Node Case, em IEEE/ACM Transaction on Networking, Vol. 1, No. 3, June 1993, pp. 344-357.

[PG93b] A. K. Parekh, R. G. Gallager, A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Multiple Node Case, *Proceedings of IEEE INFOCOM '93*, pp.521-530. To appear in *IEEE/ACM Transaction on Networking*.

[SCZ93] S. Shenker, D. Clark and L. Zhang, A Scheduling Service Model and a Scheduling Architecture for an Integrated Services Packet Network, *Preprint*, 1993.

[SE91] T. Stern, A. Elwalid, Analysis of Separable Markov-Modulated Rate Models for Information-Handling Systems, *Adv. Applied Probability*, Vol. 23, pp. 105-139, 1991.

[SdV95] C. Su and G. de Veciana, On the Capacity of Multi-service Networks, *U.T. Austin Technical Report*, 1995.

[Whi93] W. Whitt, Tail Probabilities with Statistical Multiplexing and Effective Bandwidths in Multi-Class Queues, *Telecommunication Systems*, No. 2, pp. 71-107, 1993.

[YaSi93] O. Yaron, M. Sidi, Performance and Stability of Communication Networks via Robust Exponential Bounds, *IEEE/ACM Trans. on Networking*, Vol. 1, No. 3, pp. 372-385, 1993.

[YaSi94] O. Yaron, M. Sidi, Generalized Processor Sharing Networks with Exponentially Bounded Burstiness Arrivals, In *Proceedings of IEEE INFOCOM '94*, pp.628-634, June 1994.

[Zha95a] Z.-L. Zhang, Large Deviations and Generalized Processor Sharing: Upper and Lower Bounds, Part I Two-Queue Systems, *Technical Report* UM-CS-95-96, Computer Science Department, University of Massachusetts, November 1995.

[Zha95b] Z.-L. Zhang, Large Deviations and Generalized Processor Sharing: Upper and Lower Bounds, Part II Multiple-Queue Systems, *Technical Report* UM-CS-95-97, Computer Science Department, University of Massachusetts, November 1995.

[ZLKT95] Z.-L. Zhang, Z. Liu, J. Kurose and D. Towsley, Call Admission Control Schemes under the Generalized Processor Sharing Scheduling Discipline, *Technical Report*, Computer Science Department, University of Massachusetts, March 1995.

[ZTK95] Z.-L. Zhang, D. Towsley and J. Kurose, Statistical Analysis of Generalized Processor Sharing Scheduling Discipline, *IEEE Journal on Selected Area in Communications*, Vol. 13, No. 6, pp. 1-10, August 1995. See also *Technical Report* UM-CS-95-10, Computer Science Department, University of Massachusetts, February 1995. Available via FTP from `gaia.cs.umass.edu` in pub/Zhan95:TR95-10.ps.Z.

A Proof of the Bounds

In this Appendix we prove, by analyzing the sample path behavior of the sessions, that the described systems allow us to compute lower and upper bounds of the session backlogs. In this section, let A_1 and A_2 denote the sample paths (or realizations) of the random arrival process A_1 and A_2; $A_i(\tau, t)$, $i = 1, 2$ denotes the amount of session i traffic during the time interval $[\tau, t]$ on this sample path. We assume that A_i is a right continuous function with left limit, therefore $A_i(t, t) = 0$ for any t. Similarly, let S_i and Q_i denote the corresponding sample paths of the corresponding random process S_i and Q_i, $i = 1, 2$.

A.1 Proof of the Lower Bound (Two-queue GPS System)

For $i = 1, 2$, let $\tilde{Q}_i(t)$ denote session i queue length at time t in the decomposed system (Figure 2); moreover, let $\tilde{S}_i(\tau, t)$ denote the amount of service received by session i during the time interval $[\tau, t]$ and $\tilde{s}_i(t)$, $i = 1, 2$ the service rate at time t. We show that, for $i = 1, 2$, session i backlog $Q_i(t)$ at any time t is bounded from below by the session i backlog $\tilde{Q}_i(t)$ in the decomposed system. Let us first show the following

Lemma 5 *For $i = 1, 2$, and any t the following holds*

$$s_i(t) \geq \tilde{r}_i(t) = \min\{r_i(t), g_i\} \tag{48}$$

Proof: We have the following two cases:

1. Session i input rate exceeds its guaranteed service rate ($r_i(t) > g_i$). In this case $\tilde{r}_i(t) = g_i$ whereas the service rate $s_i(t)$, session i receives in the GPS system, exceeds g_i if the other session is not busy at time t.

2. Session 1 input rate is less than or equal to the guaranteed capacity ($r_i(t) \leq g_i$). In this case $\tilde{r}_i(t) = r_i(t)$, while the service rate $s_i(t)$, session i receives in the GPS server, exceeds the input rate if session i is busy at time t.

■

We now prove

Lemma 6 *For $i = 1, 2$, let $[t_1, t_2]$ a time interval contained in a session i busy period B in the decomposed system. Then,*

$$\tilde{S}_i(t_1, t_2) \geq S_i(t_1, t_2) \tag{49}$$

Proof: Without loss of generality, we provide the proof for session 2. To obtain the proof for session 1, we simply reverse the role of session 1 and session 2 below.

Since $[t_1, t_2]$ is in a session 2 busy period, the amount of service session 2 receives during the time interval $[t_1, t_2]$ in the decomposed system is

$$\tilde{S}_2(t_1, t_2) = c(t_2 - t_1) - \int_{t_1}^{t_2} \tilde{r}_1(u)\, du. \tag{50}$$

By using lemma 5 and by observing that $S_1(t_1, t_2) + S_2(t_1, t_2) \leq c(t_2 - t_1)$ for any time interval $[t_1, t_2]$ we have

$$\begin{aligned} \tilde{S}_2(t_1, t_2) &\geq c(t_2 - t_1) - \int_{t_1}^{t_2} s_1(u)\, du \\ &\geq c(t_2 - t_1) - S_1(t_1, t_2) \\ &\geq S_2(t_1, t_2) \end{aligned}$$

■

This lemma states that, for $i = 1, 2$, in the decomposed system, the amount of service session i receives during any time interval, contained in a session i busy period, bounds from above the amount of service received in the GPS system during the same time interval. We can now provide the proof for lemma 1 which we state again

Lemma 1 *For $i = 1, 2$, and for any t, the following holds*

$$Q_i(t) \geq \tilde{Q}_i(t) \tag{51}$$

Proof: If session i is not busy at time t in the decomposed system, then $\tilde{Q}_i(t) = 0$ and (51) holds. If session i is busy at time t, let τ denote the beginning of the session i busy period that contains t. Since $[\tau, t]$ is in a session i busy period we have

$$\tilde{Q}_i(t) = A_i(\tau, t) - \tilde{S}_i(\tau, t). \tag{52}$$

From lemma 6, and using (14) we have

$$\begin{aligned} \tilde{Q}_i(t) &\leq A_i(\tau, t) - S_i(\tau, t) \\ &\leq Q_i(t). \end{aligned}$$

■

A.2 Proof of the Upper Bound (Two-queue GPS System)

We first consider the *unbiased* case. Without loss of generality, we provide the proof for session 2. Let $\tilde{Q}_2(t)$ and $\tilde{D}_2(t)$ denote session 2 queue length at time t and the delay experienced by session 2 traffic that arrives at time t in the decomposed system, respectively (Figure 5b); moreover, let $\tilde{s}_2(t)$, denote the service rate at time t and $\tilde{S}_2(\tau, t)$ the amount of service received by session 2 during the time interval $[\tau, t]$. Finally, let $Q_1'(t)$ denote session 1 queue length in the queue served at fixed rate g_1, $D_1'(t)$ the delay experienced by session 1 traffic, $s_1'(t)$ session 1 departure rate from such a queue and $S_1'(\tau, t)$ the amount of service received during the time interval $[\tau, t]$.

We show that session 2 backlog $Q_2(t)$ at any time t and the delay $D_2(t)$, experienced by session 2 traffic that arrives at time t, are bounded form above from above by the session 2 backlog $\tilde{Q}_2(t)$ and delay $D_2(t)$ in the decomposed system, respectively.

We first state the following result the proof of which can be found in [ZTK95].

Lemma 7 *At any time t*

$$Q_1(t) \leq Q_1'(t) \tag{53}$$

and

$$D_1(t) \leq D_1'(t) \leq \frac{Q_1'(t)}{g_1} \tag{54}$$

Now we provide the proof of lemma 2 we state again below

Lemma 2 *Let B be a busy period of session 2 in the two-queue GPS system. Then, for any $t \in B$*

$$s_1'(t) \geq s_1(t) \tag{55}$$

Proof: If at time t session 1 is busy then $s_1(t) = g_1$; moreover, by lemma 7, if at time t session 1 is busy in the GPS system it is busy too if it is served at fixed rate g_1; thus, $s_1'(t) = s_1(t) = g_1$. On the contrary, if at time t session 1 is not busy, then the service rate equals the input rate, $s_1(t) = r_1(t) \leq g_1$, whereas $s_1'(t)$ can exceed the input rate if session 1 is busy when it receives service at the fixed rate g_1. ∎

We can now prove

Lemma 8 *Let $[t_1, t_2]$ a time interval contained in a session 2 busy period B in the GPS system. Then,*

$$\tilde{S}_2(t_1, t_2) \leq S_2(t_1, t_2) \tag{56}$$

Proof: Since $[t_1, t_2]$ is in a session 2 busy period the amount of service session 2 receives during the time interval $[t_1, t_2]$ is

$$S_2(t_1, t_2) = c(t_2 - t_1) - S_1(t_1, t_2). \tag{57}$$

By using lemma 2 and by observing that $\tilde{S}_2(t_1, t_2) \leq c(t_2 - t_1) - \int_{t_1}^{t_2} s_1'(u)\, du$ for any time interval $[t_1, t_2]$ we have

$$\begin{aligned} S_2(t_1, t_2) &\geq c(t_2 - t_1) - \int_{t_1}^{t_2} s_1'(u)\, du \\ &\geq \tilde{S}_2(t_1, t_2) \end{aligned}$$

∎

This lemma states that, if session 2 is busy in the GPS system, the amount of service it receives during any time interval, contained in the busy period, bounds from above the amount of service received in the decomposed system during the same time interval. We can now provide the proof for lemma 3, we state again below

Lemma 3 *For any t, the following holds*

$$Q_2(t) \leq \tilde{Q}_2(t) \tag{58}$$

and

$$D_2(t) \leq \tilde{D}_2(t) \tag{59}$$

Proof: If session 2 is not busy at time t, then $Q_2(t) = 0$ and (58) holds. If session 2 is busy at time t, let τ denote the beginning of the session 2 busy period that contains t. Since $[\tau, t]$ is in a session 2 busy period we have

$$Q_2(t) = A_2(\tau, t) - S_2(\tau, t). \tag{60}$$

From lemma 8, and using (14) we have

$$\begin{aligned} Q_2(t) &\leq A_2(\tau, t) - \tilde{S}_2(\tau, t) \\ &\leq \tilde{Q}_2(\tau, t). \end{aligned}$$

From lemma 8 and (58), (59) immediately follows. ∎

This completes the proof of the bounds in the *unbiased* case.

For the *biased* case, assume, without loss of generality, that $\overline{\lambda}_1 < g_1$ and $\overline{\lambda}_2 \geq g_2$ and consider the decomposed system in Figure 6. Clearly, the proof we have provided for the *unbiased* system applies for session 2 as well. The proof of the bound for session 1 is provided by lemma 7.

A.3 Proof of the Upper Bound (Multiple-queue GPS System)

For $i = 1, \cdots, n$, let $\tilde{Q}_i(t)$ and $\tilde{D}_i(t)$ denote session i queue length and delay at time t in the decomposed system (Figure 24), respectively; moreover, let $\tilde{S}_i(\tau, t)$ denote the amount of service received by session i during the time interval $[\tau, t]$ and $\tilde{s}_i(t)$, $i = 1, 2$, the service rate at time t. Let $Q_i'(t)$ denote the backlog of session i at time t in a single queue system served at a fixed rate $g = \frac{\phi_i}{\sum_{j=1}^{n} \phi_j} c$, $D_i'(t)$ the delay experienced by session i traffic in such a system, and $s_i'(t)$ the session i departure rate from such a queue. Finally, let $S_1'(\tau, t)$ denote the amount of service received during the time interval $[\tau, t]$.

We show that, for $i = 1, 2$, session i backlog $Q_i(t)$ at any time t is bounded from below by the session i backlog $\tilde{Q}_i(t)$ in the decomposed system, and that session i delay $D_i(t)$ at any time t is bounded from below by the session i delay $\tilde{D}_i(t)$ in the decomposed system.

We first state the following

Lemma 9 *For $i = 1, \cdots, n$, let $s_i(t)$ denote the session i service rate at time t in the GPS System. For any t, the following holds*

$$\tilde{s}_i(t) \leq s_i(t) \tag{61}$$

Proof: Observe that session i service rate $s_i(t)$ satisfies the following inequality

$$s_i(t) \geq g_i + \sum_{j \neq i} \frac{\phi_i}{\sum_{k \neq j} \phi_k} (g_j - r_j(t)) \, \mathbf{1}_{\{Q_j(t)=0\}}. \tag{62}$$

In (62), the first term in the RHS is the guaranteed session i service rate; the second term accounts for the unused portion of service rate of the other sessions that session i can access. Clearly, for $j \neq i$, $\frac{\phi_i}{\sum_{k \neq j} \phi_k} (g_j - r_j(t))$ represents a lower bound on the amount of the unused portion of service rate of session j that is available to session i at time t. From Lemma 7, we have

$$\begin{aligned} s_i(t) &\geq g_i + \sum_{j \neq i} \frac{\phi_i}{\sum_{k \neq j} \phi_k} (g_j - r_j(t)) \, \mathbf{1}_{\{Q_j(t)=0\}} \\ &\geq g_i + \sum_{j \neq i} \frac{\phi_i}{\sum_{k \neq j} \phi_k} (g_j - r_j(t)) \, \mathbf{1}_{\{Q_j'(t)=0\}}. \end{aligned}$$

Inequality (61) follows from (62) and the observation that, for $j = 1, \cdots, n$, $(g_j - r_j(t)) \, \mathbf{1}_{\{Q_j'(t)=0\}} = \Big(g_j - s_j'(t)\Big)$. ∎

This lemma states that, for $i = 1, \cdots, n$, the amount of service session i receives in the GPS system at any time t, $i = 1, \cdots, n$, bounds from above the amount of service received in the decomposed system at time t.

The following result (analogous to Lemma 8) follows immediately.

Lemma 10 *Let $[t_1, t_2]$ a time interval contained in a session i busy period B in the GPS system. Then,*

$$\tilde{S}_i(t_1, t_2) \leq S_i(t_1, t_2) \tag{63}$$

The proof for Lemma 4, we state again below, follows from Lemma 10 by using the same argument used in the proof of Lemma 3, and thus omitted.

Lemma 4 *For any t, the following holds*

$$Q_i(t) \leq \tilde{Q}_i(t) \tag{64}$$

and

$$D_i(t) \leq \tilde{D}_i(t) \tag{65}$$

B Output Process Characterization

B.1 Statistical Multiplexing System with Constant Service Process

Here we provide the procedure for obtaining an approximate characterization of the departure process from a statistical multiplexing system with a constant service process we reviewed in section 2.2.

The output process is approximated by a Markov modulated fluid source defined by the pair $(\boldsymbol{M}^{(o)}, \boldsymbol{\lambda}^{(o)})$ as follows. The state space is

$$\mathcal{S}^{(o)} = \mathcal{S}_U \cup \{s_b\} \tag{66}$$

where s_b denotes the *busy period* state. The generator matrix and the rate vector are

$$\boldsymbol{M}^{(o)} = \left[\begin{array}{c|c} \boldsymbol{M}_{U,U} & \boldsymbol{a} \\ \hline \boldsymbol{b} & -(1/\bar{b}) \end{array}\right] \tag{67}$$

$$\boldsymbol{\lambda}^{(o)} = [\boldsymbol{\lambda}_U, c]. \tag{68}$$

The submatrix $\boldsymbol{M}_{U,U}$ corresponds to the underload states. The column vector $\boldsymbol{a} = [a_i]_{i \in \mathcal{S}_U}$ denotes the transition rates from the underload states to the busy period state s_b; its elements are defined as follows,

$$a_i = \sum_{j \in \mathcal{S}_O} M_{ij} \quad i \in \mathcal{S}_U. \tag{69}$$

The row vector $\boldsymbol{b} = [b_i]_{i \in S_U}$ denotes the transition rates from the busy period state; its elements are defined as follows

$$b_i = q_i / \overline{b} \quad i \in \mathcal{S}_U, \tag{70}$$

where $\boldsymbol{q} = [q_i]_{i \in \mathcal{S}}$, defines the stationary probability vector that the source state is i at the onset of the empty period and $\overline{b}$ is then mean busy period. The expression for $\boldsymbol{q}$ is [EM91]

$$\boldsymbol{q} = \frac{1}{C} \boldsymbol{\pi}_U(0) \boldsymbol{M}_{U,U}, \tag{71}$$

where C is a normalization constant, and $\boldsymbol{\pi}_U(0) = \{\pi_s(0)\}_{s \in \mathcal{S}_U}$ is the stationary probability vector that the queue is empty with session 1 in state $s \in \mathcal{S}_U$. Finally, the expression $\overline{b}$ is [EM91]

$$\overline{b} = -\frac{1}{C} \left(1 - \langle \boldsymbol{\pi}_U(0), \mathbf{1} \rangle\right). \tag{72}$$

B.2 Statistical Multiplexing System with Modulated Service Process

In this appendix we present the procedure for obtaining an approximate characterization of the departure process from a statistical multiplexing system with a modulated service process we discussed in section 4.1.

The main objective of our construction is to derive an approximate characterization of the output process in a way that the state space size is of the order of the input process size. We achieve this goal by lumping the busy period into one state, and by aggregating the system underload states corresponding to the same input state into one state.

Note that in a statistical multiplexing system with a modulated service process, when the system is in a busy period, the output process departure rate depends on the state of the modulating service process; to simplify, we associate the average output rate during the busy period, we denote $\bar{c}$, with the single busy period state. On the other hand, the departure rates associated with the aggregate underload states are directly given by the corresponding input rates. The distributions of the sojourn time in the various states are approximated by exponential distributions, the mean of which are exactly computed. Moreover, we compute the exact transition rates between states.

Observe that the resulting approximate characterization of the output process has the same conditional stationary rate distribution of the actual output process whenever the system is in the empty period. However, when the system is in the busy period only the average rate is preserved.

Let $\boldsymbol{M} = \boldsymbol{M}^{(a)} \oplus \boldsymbol{M}^{(s)}$ denote the overall system generator matrix and $\boldsymbol{\lambda}^{(I)} = \boldsymbol{\lambda}^{(a)} \otimes \mathbf{1}$ and $\boldsymbol{\lambda}^{(S)} = \mathbf{1} \otimes \boldsymbol{\lambda}^{(s)}$ (where $\mathbf{1}$ is a unit vector of proper dimension) the overall system input and modulating service rate vector, respectively. Moreover, let $\boldsymbol{w}$ and $\boldsymbol{\pi}(q)$ denote the stationary probability and distribution vector, respectively. Finally, let $\mathcal{S}_U = \{\boldsymbol{s} \in \mathcal{S} : \lambda_s^{(I)} < c - \lambda_s^{(S)}\}$ and $\mathcal{S}_O = \{\boldsymbol{s} \in \mathcal{S} : \lambda_s^{(I)} \geq c - \lambda_s^{(S)}\}$ denote the set of underload and overload overall states.

We proceed in two steps in order to obtain the approximate output process characterization. In the first step, we derive an approximate characterization of the output process, the state space of which consists of the overall underload system states and the busy state. The construction is carried out in a way similar to the procedure presented in Appendix B.1. In the second step we aggregate the underload states corresponding to the same input state. This yields an approximate characterization with a state space size of the same order of the input process space size.

In the first step, the approximate characterization $(\tilde{\boldsymbol{M}}^{(o)}, \tilde{\boldsymbol{\lambda}}^{(o)})$ has a state space $\tilde{\mathcal{S}}^{(o)} = \mathcal{S}_U \cup \{s_b\}$, where s_b is the busy state and a stationary probability vector $\tilde{\boldsymbol{w}}^{(o)}$. The transition matrix has the following structure:

$$\tilde{\boldsymbol{M}}^{(o)} = \left[\begin{array}{c|c} \boldsymbol{M}_{U,U} & \boldsymbol{a} \\ \hline \boldsymbol{b} & -(1/\overline{b}) \end{array}\right] \text{ and } \tilde{\boldsymbol{\lambda}}^{(o)} = \left[\boldsymbol{\lambda}_U^{(I)}, \overline{c}\right]. \tag{73}$$

The matrix $\boldsymbol{M}_{U,U}$ is a submatrix of $\boldsymbol{M}$ specifying the transition probability among the underload states. The vector $\boldsymbol{a}$ denotes the transition rates from the underload states to the busy period state s_b and the row vector $\boldsymbol{b}$ vice versa. $\overline{b}$ is the mean busy period. These quantities can be computed as shown in Appendix B.1. $\overline{c}$ is the average service rate during the busy period given by

$$\overline{c} = \frac{\sum_{s\in\mathcal{S}}(c - \lambda_s^{(S)})(w_s - \pi_s(0))}{1 - \sum_{s\in\mathcal{S}} \pi_s(0)}. \tag{74}$$

We need to characterize the embedded discrete time Markov chain defined at the transition points of this output process characterization. Let $\tilde{\boldsymbol{P}}^{(o)} = [\tilde{P}_{s,s'}^{(o)}]_{s,s'\in\tilde{\mathcal{S}}^{(o)}}$ denote its one-step transition probability matrix and $\tilde{\boldsymbol{u}}^{(o)} = [\tilde{u}_s^{(o)}]_{s\in\tilde{\mathcal{S}}^{(o)}}$ the stationary probability vector. The following hold (see, *e.g.*, [Kei79]):

$$\tilde{p}_{s,s'}^{(o)} = \delta_{s,s'} + \frac{\tilde{M}_{s,s'}^{(o)}}{\tilde{\mu}_s^{(o)}},\ s, s' \in \tilde{\mathcal{S}}^{(o)}, \tag{75}$$

and

$$\tilde{u}_s^{(o)} = \frac{\tilde{w}_s^{(o)}\tilde{\mu}_s^{(o)}}{\langle\tilde{\boldsymbol{w}}^{(o)}, \tilde{\boldsymbol{\mu}}^{(o)}\rangle},\ s \in \tilde{\mathcal{S}}^{(o)}, \tag{76}$$

where $\delta_{s,s'}$ is the Kronecker's delta function and $\tilde{\mu}_s^{(o)} = -\tilde{M}_{s,s}^{(o)}$, $s \in \tilde{\mathcal{S}}^{(o)}$, is the transition rate out of state s and $\tilde{\boldsymbol{\mu}}^{(o)} = [\tilde{\mu}_s^{(o)}]_{s\in\tilde{\mathcal{S}}^{(o)}}$.

In the second step, we aggregate the underload states corresponding to the same input state into one state. The resulting approximate characterization $(\boldsymbol{M}^{(o)}, \boldsymbol{\lambda}^{(o)})$ has state space $\mathcal{S}^{(o)} = \mathcal{S}_U^{(a)} \cup \{s_b\}$, where $\mathcal{S}_U^{(a)}$ denotes the set of the underload input state, defined as follows

$$\mathcal{S}_U^{(a)} = \{s^{(a)} \in \mathcal{S}^{(a)} \mid \exists \boldsymbol{s} = (s^{(a)}, s^{(s)}) \in \mathcal{S}, \boldsymbol{s} \in \mathcal{S}_U\}. \tag{77}$$

For $i \in \mathcal{S}_U^{(a)}$, the diagonal entry of $M_{i,i}^{(o)}$ is equal to $-\mu_i$, where $\frac{1}{\mu_i}$ is the average residence time in the input state i conditioned on the event that the system is in an empty period. For the busy state s_b, the diagonal entry is $\frac{1}{\overline{b}}$. The off-diagonal entries are of the form $M_{i,j}^{(o)} = \mu_i p_{i,j}^{(o)}, i, j \in \tilde{\mathcal{S}}$, where $p_{i,j}^{(o)}$ gives the probability of entering state j upon a transition from state i. To complete the characterization we need to specify, for each state, the average residence time and the transition probabilities. For $i \in \mathcal{S}_U^{(a)}$, let $\mathcal{S}_U^i$ denote the set of the overall system underload states that contain i, *i.e.*

$$\mathcal{S}_U^i = \{(s^{(a)}, s^{(s)}) \in \mathcal{S}_U : s^{(a)} = i\}.$$

and let $\mathcal{S}_U^{s_b} = \{s_b\}$.

From [How71], we have

$$\frac{1}{\mu_i} = \tilde{\boldsymbol{u}}^{(o)i}(\tilde{\boldsymbol{M}}^{(o)}{}_{i,i})^{-1} \tag{78}$$

where the submatrix $\tilde{\boldsymbol{M}}^{(o)}{}_{i,i}$ corresponds to the states in $\mathcal{S}_U^i$, and $\tilde{\boldsymbol{u}}^{(o)i} = [\tilde{u}^{(o)i_s}]_{s\in\mathcal{S}_U^i}$ is the stationary probability vector just after a transition into $\mathcal{S}_U^i$. By simple probabilistic arguments, it is easy to show that

$$\tilde{u}_s^{(o)i} = \frac{\sum_{s'\notin\mathcal{S}_U^i}\tilde{P}_{s',s}^{(o)}\tilde{u}_{s'}^{(o)}}{\sum_{s''\in\mathcal{S}_U^i}\sum_{s'\notin\mathcal{S}_U^i}\tilde{P}_{s',s''}^{(o)}\tilde{u}_{s'}^{(o)}},\ s\in\mathcal{S}_U^i. \tag{79}$$

Similarly, the transition probabilities have the following expressions

$$p_{i,j}^{(o)} = \frac{\sum_{s\in\mathcal{S}_U^i}\sum_{s'\in\mathcal{S}_U^j}\tilde{P}_{s,s'}^{(o)}\tilde{u}_s^{(o)}}{\sum_{s\in\mathcal{S}_U^i}\sum_{s'\notin\mathcal{S}_U^i}\tilde{P}_{s,s'}^{(o)}\tilde{u}_s^{(o)}},\ i,j\in\mathcal{S}^{(o)} \tag{80}$$

The rate vector is $\boldsymbol{\lambda}^{(o)} = [\boldsymbol{\lambda}^{(o)}{}_U, \overline{c}]$, where $\boldsymbol{\lambda}^{(o)}{}_U = [\lambda_i^{(o)}]_{i\in\mathcal{S}_U^{(a)}}$ and $\overline{c}$ is the average service rate during the busy period. The vector $\boldsymbol{\lambda}^{(o)}{}_U$ is defined as follows

$$\lambda_i^{(o)} = \lambda_i^{(a)},\ i\in\mathcal{S}_U^{(a)}. \tag{81}$$

This complete the characterization.