Inter-Class Resource Sharing using Statistical Service Envelopes*

Jing-yu Qiu, Coskun Cetinkaya, Chengzhi Li, and Edward W. Knightly
Department of Electrical and Computer Engineering
Rice University†

Abstract

Networks that support multiple services through "link-sharing" must address the fundamental conflicting requirement between *isolation* among service classes to satisfy each class' quality of service requirements, and statistical *sharing* of resources for efficient network utilization. While a number of service disciplines have been devised which provide mechanisms to both isolate flows and fairly share excess capacity, admission control algorithms are needed which exploit the effects of inter-class resource sharing. In this paper, we develop a framework of using *statistical service envelopes* to study inter-class statistical resource sharing. We show how this service envelope enables a class to over-book resources beyond its deterministically guaranteed capacity by statistically characterizing the excess service available due to fluctuating demands of other service classes. We apply our techniques to several multi-class schedulers, including Generalized Processor Sharing, and design new admission control algorithms for multi-class link-sharing environments. We quantify the utilization gains of our approach with a set of experiments using long traces of compressed video.

1 Introduction

Future integrated services networks will support heterogeneous Quality of Service (QoS) specifications and traffic demands. For example, a deterministic service [21] uses worst-case resource allocation to support applications requiring packet delivery without losses or delay bound violations; a statistical service [14] achieves a statistical multiplexing gain and provides statistical QoS guarantees with controlled "overbooking" of resources; a measurement-based service [10] supports QoS by basing admission control decisions on empirical observations of aggregate traffic behavior; best-effort services support applications with less stringent QoS requirements such as bulk data transfer. With appropriate admission control and traffic scheduling, these services and others can co-exist in a single network, as admission control limits the number of admitted traffic flows to ensure that each class' QoS requirements are met, and packet schedulers ensure that packets are assigned the priority levels needed to meet their QoS objectives.

In a link sharing environment as outlined in [7], traffic class k is allocated capacity c_k such that whenever packets from class k are backlogged, the class receives service at a rate of at least c_k. If class k is not

*An earlier version of this paper appears in the Proceedings of IEEE INFOCOM '99.
†This work was supported by NSF CAREER Award ANI-9733610, NSF Grant ANI-9730104, and Nokia Corporation.

backlogged, then class k's unused capacity is distributed fairly among backlogged sessions. Consequently, classes can be assured to meet their respective QoS requirements, regardless of the behavior of other traffic classes, allowing any number of services to co-exist in the network.

In the literature, a number of service disciplines have been designed to support such link sharing objectives [7, 2]. For example, [2] develops a class of Hierarchical Packet Fair Queueing algorithms focusing on an algorithm's fairness, complexity, and ability to provide low end-to-end deterministic delay bounds. While scheduling algorithms for efficiently and fairly allocating excess capacity to backlogged classes are an important aspect of a link-sharing network, an admission control policy that enables one class of traffic to quantify the improved QoS it will receive due to capacity unused by other classes has not been addressed.

In addition to service disciplines, a number of admission control algorithms have also been designed both for deterministic services which do not exploit statistical resource sharing [5, 17], as well as statistical [14, 4, 6, 11, 22] and measurement-based services [10] which do. However, such admission control algorithms consider traffic classes in isolation, and while a statistical multiplexing gain is achieved *within* a particular traffic class, *inter-class* resource sharing is not addressed. In particular, [6, 22] study statistical service for Generalized Processor Sharing (GPS) [17], and while the "isolation" property of GPS is exploited for multi-node analysis, inter-class statistical resource sharing is not addressed. Moreover, while several previous studies do consider inter-class sharing from the perspective of scheduling [8, 9] or video transmission [1], general inter-class link sharing environments have not been addressed.

In this paper, we address the problem of inter-class statistical resource sharing. Our contribution is to develop a theoretical framework for inter-class resource sharing, and to derive admission control algorithms for several important schedulers. Our key technique is to develop a framework of *statistical service envelopes* to study the problem. Inspired by [5], we define a statistical service envelope as a probabilistic description of the service available to a traffic class as a function of interval length. We use this service envelope to characterize the additional capacity available to a traffic class beyond the minimum deterministically guaranteed capacity set aside by the link sharing rules. In this way, we statistically capture the fluctuating excess capacity left unused by one traffic class so that another class may exploit an inter-class statistical multiplexing gain and potentially admit additional traffic flows that would not otherwise have been deemed admissible. Thus, we use the statistical service envelope as a tool for overbooking inter-class resources in a controlled manner, so that a class can probabilistically quantify the additional resources available in a link sharing environment.

We apply this framework of statistical service envelopes to three multi-class service disciplines, namely, Strict Priority (SP), Earliest Deadline First (EDF), and link-sharing GPS [7, 2]. We show that while the concept of a statistical service envelope was implicitly used in previous studies of SP [11], explicitly computing the service envelope of other traffic classes provides a simpler analysis and allows us to uniformly treat deterministic and statistical service classes.

For GPS, we conceptually partition traffic classes into *isolation* classes and *sharing* classes depending on whether or not the traffic class will exploit the effects of inter-class resource sharing in making admission control decisions. For example, a deterministic service is an isolation class as excess capacity from other traffic classes is not guaranteed in the worst case and hence a statistical envelope of excess capacity cannot

improve this class' admissible region. We then bound the total service received by all sharing classes and show how the weighted fairness property of GPS can then be used to derive each class' service envelope. In this way, admission control for each *sharing* class can characterize the capacity available beyond its guaranteed rate, incorporating the relative weights and traffic demands of all other traffic classes, and improving the class' admissible region.

We illustrate the potential utilization gains of our inter-class resource sharing scheme with a set of trace-driven simulation experiments using long traces of MPEG-compressed video. As an illustrative example with a 45 Mbps link supporting equally weighted deterministic and statistical service classes with the GPS service discipline, we find that the average utilization of the link can be improved from 47.7% to 84.6% by using the statistical service envelope to characterize the excess capacity of the deterministic class.

2 Statistical Service Envelopes: Theory and Applications

In this section, we define statistical service envelopes and develop their applications to inter-class resource sharing. In particular, we first study the delay distribution for a single class using statistical traffic envelopes and deterministic service envelopes. Next, we extend this analysis to include statistical traffic envelopes and *statistical* service envelopes. Finally, we illustrate the application of statistical service envelopes by deriving admission control tests for SP and EDF schedulers using this theory.

2.1 Multi-Class Queueing Concepts

Here, we introduce two key concepts for inter-class resource sharing. First, we define essential traffic: for a particular class n, this refers to the total class-n traffic that must be serviced in order for class i to meet its delay constraints. The second concept is available service, a characterization of the capacity available to a class as a function of interval length.

Throughout this paper, we model a multiplexer by a discrete-time infinite buffer queue in which fluid flows into and out of the buffer only at discrete time slots. For traffic class i, let $X^i(t)$ denote its aggregate arrivals in time slot t, and let $X^i(s,t)$ denote the total arrivals in time interval $[s,t]$, i.e., $X^i(s,t) = \sum_{h=s}^{t} X^i(h)$. Without loss of generality, we assume that $X^i(\cdot,\cdot)$, $i = 1, 2, \cdots$, are independent. Let $Y^i(t)$ represent the amount of fluid served for traffic class i in time slot t, and denote $Y^i(s,t)$ as the total fluid served in time interval $[s,t]$, i.e., $Y^i(s,t) = \sum_{h=s}^{t} Y^i(h)$.

Denoting $Q^i(t)$ as the backlog of class i at the end of time slot t, $Q^i(t)$ is given by

$$Q^i(t) = \max_{s \le t} \{X^i(s,t) - Y^i(s,t)\}. \tag{1}$$

Class i is said to be continually backlogged in the interval $[s,t]$ if $Q^i(h) > 0, \forall\, h \in [s,t]$.

Definition 1 (Essential Traffic) *The essential traffic of class n with respect to class i is defined as*

$$X^n_{D_i}(s,t) = X^n(s, t+D_i) \cap Y^n(s, t+D_i) \tag{2}$$

The essential traffic has an important interpretation: suppose a class-i packet arrives at time t and is serviced exactly at its delay bound $t + D_i$. Then $X^n_{D_i}(s,t)$ is the class-n traffic which will be serviced before the class-i packet. As we will show below, the essential traffic is a function of the particular service discipline, and plays a key role in characterizing inter-class resource sharing.

Definition 2 (Available Service) *Let $\widetilde{X}^i(s,t)$ denote the minimal class i input such that class i is continuously backlogged in $[s,t]$. The available service of class i in $[s, t+D_i]$ is defined as the class i output $\widetilde{Y}^i_{D_i}(s,t)$ given this minimally backlogging input traffic $\widetilde{X}^i(s,t)$, and other classes' input traffic as their essential traffic $X^n_{D_i}(s,t)$, $n \neq i$.*

Note that the available service $\widetilde{Y}^i_{D_i}(s,t)$ is a function of the scheduling mechanism and the essential traffic $X^n_{D_i}(s,t)$, $n \neq i$. Notice further that $\widetilde{Y}^i_{D_i}(s,t)$ is independent of the input traffic of class i; whereas the *actual* output process $Y^i(s, t+D_i)$ is decided by *all* classes' inputs. By using this notion of available service, we decouple class i's input traffic $X^i(s,t)$ from its availabe service $\widetilde{Y}^i_{D_i}(s,t)$, making $\widetilde{Y}^i_{D_i}(s,t)$ a pure description of available network resources, separate from the traffic that is actually sent.

We next review several facts about stochastic ordering [1] that are used later in this section.

Lemma 1 *Let X_i, $i = 1, \cdots, n$, be independent random variables with distributions $G_i(\cdot)$, $i = 1, \cdots, n$, respectively and Y_i , $i = 1, \cdots, n$, be independent random variables with distributions $F_i(\cdot)$, $i = 1, \cdots, n$, respectively, if $X_i \leq_{st} Y_i$ for $i = 1, \cdots, n$, then*

1. *$\sum_{i=1}^n X_i \leq_{st} \sum_{i=1}^n Y_i$.*
2. *$f(X_1) \leq_{st} f(Y_1)$ for any increasing function f.*
3. *$c - \sum_{i=1}^n X_i \geq_{st} c - \sum_{i=1}^n Y_i$ for any real variable c.*
4. *$\overline{Y_i} = F_i^{-1}(G_i(X_i))$, $i = 1, \cdots, n$, are independent random variables with distributions $F_i(\cdot)$, $i = 1, \cdots, n$, respectively and $X_i \leq \overline{Y_i}$ for $i = 1, \cdots, n$.*

Proof: See [19] for detail. □

2.2 Statistical Service with Deterministic Service Envelopes

Deterministic service is studied in [5] using deterministic service envelopes and deterministic traffic envelopes. Here, we first study statistical service with *statistical* traffic envelopes and *deterministic* service envelopes, and later focus on *statistical* service envelopes. First, we formally define both deterministic and statistical traffic envelopes and deterministic service envelopes.

Definition 3 (Deterministic Service Envelope) *A non-decreasing non-negative function $s^i_{D_i}(t)$ is a deterministic service envelope of traffic class i, if for any backlogged interval $[u+1, u+t]$, the available service satisfies*[2]

$$\widetilde{Y}^i_{D_i}(u+1, u+t) \geq s^i_{D_i}(t).$$

[1] Throughout, $X \leq_{st} Y$ (stochastic inequality) denotes $P[X > z] \leq P[Y > z]$ for all z.

[2] Throughout, $Y \geq X$ denotes almost sure inequality, $P[Y \geq X] = 1$.

To illustrate the concept of a deterministic service envelope, note that for a GPS server, a service class with guaranteed rate g^i, satisfies $\widetilde{Y}^i_{D_i}(u+1,u+t) \geq s^i_{D_i}(t) = g^i(t+D_i)$.

Definition 4 (Deterministic Traffic Envelope) *[15] A non-decreasing non-negative function $b^i(t)$ is a deterministic traffic envelope of class i, if for any interval $[u+1,u+t]$, the input traffic satisfies*

$$X^i(u+1,u+t) \leq b^i(t).$$

Definition 5 (Statistical Traffic Envelope) *A sequence of random variables $B^i(t)$ is a statistical traffic envelope of class i, if for any interval $[u+1,u+t]$, the input traffic satisfies*

$$X^i(u+1,u+t) \leq_{st} B^i(t).$$

In other words, $b^i(t)$ describes the maximum class-i arrivals in any interval of length t, whereas $B^i(t)$ describes the distribution of arrivals in intervals of length t. Without loss of generality, we assume that $X^i(\cdot,\cdot)$ and $X^j(\cdot,\cdot)$ are independent and $B^i(\cdot)$ and $B^j(\cdot)$ are independent if $i \neq j$.

Denoting D^i_t as the virtual delay experienced by a bit of class i arriving at time slot t, the key QoS metric that we consider is the probability of (virtual) delay bound violation, $P[D^i_t > D_i]$. As long as

$$\lim_{t\to\infty} \frac{EX^i(1,t)}{t} < \lim_{t\to\infty} \frac{s^i_{D_i}(t)}{(t+D_i)}$$

(the stability condition), and $X^i(t)$ is stationary and ergodic, $P[D^i_t > D_i]$ converges to a steady state tail probability $P[D^i > D_i]$ [16].

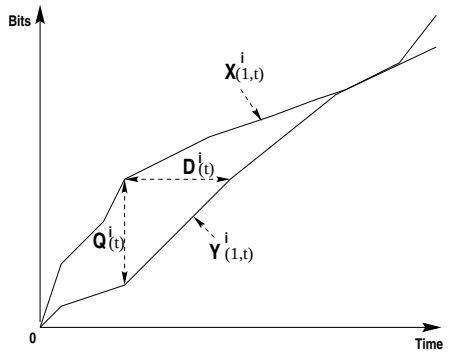

Figure 1: Delay and Buffer Occupancy

Figure 1 shows the delay and buffer occupancy of class i in terms of $X^i(1,t)$ and $Y^i(1,t)$ if the buffer is initially empty. The delay D^i_t of a class-i arrival at t is defined as [5]

$$D^i_t = \min\{\Delta : \Delta \geq 0 \text{ and } X^i(1,t) \leq Y^i(1,t+\Delta)\}. \quad (3)$$

Lemma 2 *For a delay bound D_i, the event of delay bound violation of class i at time slot t satisfies*

$$\{D_t^i > D_i\} \quad \subseteq \quad \{\max_{s\leq t}\{X^i(s,t) - \widetilde{Y}_{D_i}^i(s,t)\} > 0\}. \tag{4}$$

Proof. By definition

$$\begin{aligned}\{D_t^i > D_i\} &\equiv \{X^i(1,t) - Y^i(1,t+D_i) > 0\} \\ &\subseteq \{\max_{s\leq t}\{X^i(s,t) - Y^i(s,t+D_i)\} > 0\}.\end{aligned}$$

Observe that if $\max_{s\leq t}\{X^i(s,t) - Y^i(s,t+D_i)\} > 0$, then $\max_{s\leq t}\{X^i(s,t) - \widetilde{Y}_{D_i}^i(s,t)\} > 0$. This is because if $\max_{s\leq t}\{X^i(s,t) - Y^i(s,t+D_i)\} > 0$, there must exist an

$$s^* = \max\{s : s < t \text{ and } Q^i(s) = 0\}$$

such that

$$\max_{s\leq t}\{X^i(s,t) - Y^i(s,t+D_i)\} = X^i(s^*+1,t) - Y^i(s^*+1,t+D_i),$$

and $[s^*+1, t+D_i]$ is a backlogged interval. Furthermore, since $\widetilde{Y}_{D_i}^i(s^*+1,t)$ is the available service in $[s^*+1, t+D_i]$, we have

$$\widetilde{Y}_{D_i}^i(s^*+1,t) \leq Y^i(s^*+1,t+D_i),$$

so that

$$\{\max_{s\leq t}\{X^i(s,t) - Y^i(s,t+D_i)\} \geq 0\} \leq \{\max_{s\leq t}\{X^i(s,t) - \widetilde{Y}_{D_i}^i(s,t)\} \geq 0\}.$$

Thus

$$\{D_t^i > D_i\} \subseteq \{\max_{s\leq t}\{X^i(s,t) - \widetilde{Y}_{D_i}^i(s,t)\} > 0\}. \qquad \square$$

Theorem 1 *For service class i with deterministic service envelope $s_{D_i}^i(t)$ and statistical traffic envelope $B^i(t)$, the probability of class-i delay bound violation satisfies:*

$$P[D_t^i > D_i] \leq P[\max_{u\geq 0}\{\overline{B^i(u)} - s_{D_i}^i(u)\} > 0],\ \forall D_i \in [0,\infty), \tag{5}$$

where $\overline{B^i(u)}$ is a rondom variable with the same distribution as $B^i(u)$.

Proof. From Lemma 2,

$$P[D_t^i > D_i] \leq P[\max_{u\leq t}\{X^i(u,t) - \widetilde{Y}_{D_i}^i(u,t)\} > 0]. \tag{6}$$

Since $X^i(u,t) \leq_{st} B^i(t-u+1)$, by Lemma 1, there exists a random variable $\overline{B^i(t-u+1)}$ with the same distribution as $B^i(t-u+1)$ such that $X^i(u,t) \leq \overline{B^i(t-u+1)}$. Furthermore, by Definition 3, $\widetilde{Y}_{D_i}^i(u,t) \geq s_{D_i}^i(t-u+1)$. Thus, we have

$$\{\max_{u\leq t}\{X^i(u,t) - \widetilde{Y}_{D_i}^i(u,t)\} > 0\} \quad \subseteq \quad \{\max_{u\geq 0}\{\overline{B^i(u)} - s_{D_i}^i(u)\} > 0\}.$$

Finally, we have

$$P[\max_{u\leq t}\{X^i(u,t) - \widetilde{Y}_{D_i}^i(u,t)\} > 0] \quad \leq \quad P[\max_{u\geq 0}\{\overline{B^i(u)} - s_{D_i}^i(u)\} > 0]. \quad \square$$

Thus, the theorem provides a general multi-class statistical delay bound using the lower bound of a classes' available service $s_{D_i}^i(t)$.

2.3 Statistical Service with Statistical Service Envelopes

Theorem 1 enables us to exploit the statistical multiplexing gain of flows within a service class. This result is quite general and as we show below can be applied to a wide class of schedulers. However, while the deterministic service envelope $s^i_{D_i}(t)$ provides isolation among service classes and simplifies admission control, it precludes statistical inter-class resource sharing. In multi-class schedulers such as SP, EDF, and GPS, the utilization gains available from exploiting inter-class resource sharing can be significant. Next, we introduce a statistical service envelope to study the inter-class resource sharing problem, and develop new theory to calculate the delay bound violation probability using statistical service envelopes.

In a multi-class server, the available service for class i, $\widetilde{Y}^i_{D_i}(u,t)$, is a function of the input traffic in other classes and the particular service discipline which specifies how to schedule services among competing classes. The interference among classes is reflected in $\widetilde{Y}^i_{D_i}(u,t)$, and in some cases, it is possible that the available service is far greater than the minimally guaranteed service, i.e., $\widetilde{Y}^i_{D_i}(u,t) \gg s^i_{D_i}(t-u+1)$. Thus we define a statistical service envelope as a way to characterize the available service beyond the deterministically guaranteed $s^i_{D_i}(t)$.

Definition 6 (Statistical Service Envelope) *A sequence of random variables $S^i_{D_i}(t)$ is a statistical service envelope of class i's traffic, if for any interval $[u+1, u+t]$, the available service $\widetilde{Y}^i_{D_i}(u+1, u+t)$ satisfies*

$$\widetilde{Y}^i_{D_i}(u+1,u+t) \geq_{st} S^i_{D_i}(t).$$

Notice that while a deterministic service envelope $s^i_{D_i}(t)$ describes the service of a class in isolation, the statistical service envelope $S^i_{D_i}(t)$ describes inter-class resource sharing. We employ $S^i_{D_i}(t)$ in the delay distribution calculation with the following theorem.

Theorem 2 *For service class i with statistical service envelope $S^i_{D_i}(t)$ and statistical traffic envelope $B^i(t)$, the probability of class-i delay bound violation satisfies:*

$$P[D^i_t > D_i] \leq P[\max_{u \geq 0}\{\overline{B^i(u)} - \overline{S^i_{D_i}(u)}\} > 0],\ \forall D_i \in [0,\infty), \tag{7}$$

where $\overline{B^i(u)}$ and $\overline{S^i_{D_i}(u)}$ have the same distribution as $B^i(u)$ and $S^i_{D_i}(u)$ respectively.

Proof. From Equation (6),

$$P[D^i_t > D_i] \leq P[\max_{u \leq t}\{X^i(u,t) - \widetilde{Y}^i_{D_i}(u,t)\} > 0]. \tag{8}$$

Since $X^i(u,t) \leq_{st} B^i(t-u+1)$ and $\widetilde{Y}^i_{D_i}(u,t) \geq_{st} S^i_{D_i}(t-u+1)$, by Lemma 1, there exist random variables $\overline{B^i(t-u+1)}$ and $\overline{S^i_{D_i}(t-u+1)}$ with the same distribution as $B^i(t-u+1)$ and $S^i_{D_i}(t-u+1)$ respectively such that $X^i(u,t) \leq B^i(t-u+1)$ and $\widetilde{Y}^i_{D_i}(u,t) \geq \overline{S^i_{D_i}(t-u+1)}$.

Thus, we have

$$\begin{aligned} P[\max_{u \leq t}\{X^i(u,t) - \widetilde{Y}^i_{D_i}(u,t)\} > 0] &\leq P[\max_{u \leq t}\{\overline{B^i(t-u+1)} - \overline{S^i_{D_i}(t-u+1)}\} > 0] \\ &\leq P[\max_{u}\{\overline{B^i(u)} - \overline{S^i_{D_i}(u)}\} > 0]. \end{aligned} \tag{9}$$

Thus

$$P[D_t^i > D_i] \leq P[\max_u\{\overline{B^i(u)} - \overline{S_{D_i}^i(u)}\} > 0]. \qquad \square \tag{10}$$

Notice that the theorem applies to any traffic characterization $B(\cdot)$ and any inter-class relationship $S(\cdot)$. As we show below, $S(\cdot)$ is determined by the particular service discipline, as it is the service discipline which determines the manner in which multiple classes interact (with weighted fairness, strict priority, etc.) and hence the extent to which classes are strictly isolated or share system resources. Below we employ Theorem 2 and our framework of statistical service envelopes to devise admission control algorithms for multi-class servers. In this way, admission control can exploit the available inter-class statistical resource sharing that is provided by the scheduler.

2.4 Strict Priority

Admission control for strict priority schedulers was studied in [15] for deterministic service. In [11], approximate tests were developed for statistical service. Here, we approach the problem using service envelopes and obtain a general and accurate multi-class admission control test which supports multiple deterministic and statistical service classes.

Lemma 3 *Consider an SP scheduler with N priority queues, link speed C, and the aggregate traffic in class i bounded by $B^i(t)$ and $b^i(t)$, with $i = 1, \ldots, N$ denoting the priority level from higher priority to lower priority. The statistical service envelope for class i with delay bound D_i is*

$$S_{D_i}^i(t) = (C(t + D_i) - \sum_{n=1}^{i-1} B^n(t + D_i))^+ \tag{11}$$

and the deterministic service envelope for class i is

$$s_{D_i}^i(t) = (C(t + D_i) - \sum_{n=1}^{i-1} b^n(t + D_i))^+ \tag{12}$$

where $b^i(t) = \sum_j b_j^i(t)$, $B^i(t) = \sum_j B_j^i(t)$, and $b_j^i(t)$ and $B_j^i(t)$ are the respective deterministic and statistical envelopes of flow j in class i.

Proof. Consider class-i arrivals at t which have deadline $t + D_i$. Under strict priority, the essential traffic for higher priority classes consists of all traffic arriving throughout $[s, t + D_i]$, whereas lower priority classes have no effect on class i and hence have no essential traffic. Thus, we have

$$X_{D_i}^n(s, t) = \begin{cases} X^n(s, t + D_i) & n < i \\ 0 & n > i. \end{cases} \tag{13}$$

Furthermore, since the total available service in the interval $[s, t + D_i]$ is $C(t - s + D_i + 1)$ and $X^n(s, t + D_i)$, $n = 1, 2, \cdots, i - 1$, are independent and $X^n(s, t + D_i) \leq_{st} B^n(t + D_i - s + 1)$, by Lemma 1,

the remaining capacity available to a minimally backlogged class-i flow is given by

$$\begin{aligned}\widetilde{Y}^i_{D_i}(s,t) &= (C(t-s+D_i+1)-\sum_n X^n_{D_i}(s,t))^+ \\ &= (C(t-s+D_i+1)-\sum_{n=1}^{i-1} X^n(s,t+D_i))^+ \\ &\geq_{st} (C(t-s+D_i+1)-\sum_{n=1}^{i-1} B^n(t+D_i-s+1))^+. \end{aligned} \quad (14)$$

According to Definition 6, we have

$$S^i_{D_i}(t) = (C(t+D_i)-\sum_{n=1}^{i-1} B^n(t+D_i))^+. \quad (15)$$

For the deterministic service envelope, the proof is similar. □

Lemma 4 *Consider an SP scheduler with N priority queues and link speed C. For each service class, traffic is bounded by $B^i(t)$ and $b^i(t)$, with QoS parameters (D_i, P^i), where P^i is the delay bound violation probability. The QoS for all service classes in this multi-service SP scheduler is satisfied if for all deterministic service classes with $P^i = 0$,*

$$\max_t\{b^i(t)+\sum_{n=1}^{i-1} b^n(t+D_i)-C(t+D_i)\} \leq 0$$

and for all statistical service classes with $P^i > 0$,

$$P[\max_t\{\overline{B^i(t)}+\sum_{n=1}^{i-1}\overline{B^n(t+D_i)}-C(t+D_i)\}>0] \leq P^i,$$

where $\overline{B^n(t+D_i)}$ is a random variable with the same distribution as $B^n(t+D_i)$, for $n = 1, \cdots, i-1$.

Proof. For statistical service classes, according to Equation (14), we know that

$$\begin{aligned}\widetilde{Y}^i_{D_i}(s,t) &= (C(t-s+D_i+1)-\sum_n X^n_{D_i}(s,t))^+ \\ &\geq C(t-s+D_i+1)-\sum_{n=1}^{i-1} X^n(s,t+D_i).\end{aligned}$$

Since $X^n(s,t+D_i) \leq_{st} B^n(t+D_i-s+1)$ for $n = 1, \cdots, i-1$, by Lemma 1, there exist random variables $\overline{B^n(t+D_i-s+1)}$, $n = 1, \cdots, i-1$, such that $X^n(s,t+D_i) \leq \overline{B^n(t+D_i-s+1)}$, and so

$$\widetilde{Y}^i_{D_i}(s,t) \geq C(t-s+D_i+1)-\sum_{n=1}^{i-1}\overline{B^n(t+D_i-s+1)}.$$

Thus, we can use $C(t+D_i)-\sum_{n=1}^{i-1}\overline{B^n(t+D_i)}$ to replace $\overline{S^i_{D_i}(t)}$ in Theorem 2. Furthermore,

$$\overline{B^i(t)}-[C(t+D_i)-\sum_{n=1}^{i-1}\overline{B^n(t+D_i)}] = \overline{B^i(t)}+\sum_{n=1}^{i-1}\overline{B^n(t+D_i)}-C(t+D_i), \quad (16)$$

and so, if $P[\max_t\{\overline{B^i(t)}+\sum_{n=1}^{i-1}\overline{B^n(t+D_i)}-C(t+D_i)\}>0]\leq P^i$, then the statistical QoS requirement of service class i is satisfied. For deterministic service classes, the proof is similar. □

Recentely, Shakkottai and Srikant have shown that the above bound is asymptotically *exact* in a theoretical study of SP schedulers in the many-sources regime [20].

Note that inter-class interference in an SP scheduler is in a single direction, only from higher priority classes to lower priority ones. Note also that this strict separation of sharing classes and isolation is described by the service envelopes in Lemma 3. For EDF, we will see that every class affects every other class such that the statistical service envelope for one class becomes a function of the traffic envelopes of all other classes.

2.5 Earliest Deadline First

We now apply Theorems 1 and 2 to EDF schedulers by deriving EDF's service envelopes $s^i_{D_i}(t)$ and $S^i_{D_i}(t)$.

In an EDF scheduler, every class i is associated with a a delay bound d^i.[3] A class i packet arriving at t is assigned deadline $t+d^i$, and the EDF service discipline always selects the packet with the smallest deadline for service.

Lemma 5 *In an EDF scheduler with class i traffic bounded by $B^i(t)$ and $b^i(t)$, and EDF scheduler delay bound d^i, $i=1,2,\cdots,N$, the statistical service envelope for class i traffic is given by*

$$S^i_{D_i}(t)=(C(t+D_i)-\sum_{n\neq i}B^n(t-d^n+d^i))^+, \tag{17}$$

and the deterministic service envelope for class i is

$$s^i_{D_i}(t)=(C(t+D_i)-\sum_{n\neq i}b^n(t-d^j+d^i))^+ \tag{18}$$

where $B^i(t)$ and $b^i(t)$ are 0 if $t<0$.

Proof. Consider class-i arrivals at t which have deadline $t+d^i$. The essential traffic $X^n_{D_i}(s,t)$ $(n\neq i)$ which is serviced before the class-i arrivals at t contains only class n's traffic arriving in the interval $[s,t+d^i-d^n]$. Therefore, we have

$$X^n_{D_i}(s,t)=X^n(s,t+d^i-d^n). \tag{19}$$

As in the proof of Lemma 3, we have $C(t-s+D_i+1)$ total services in $[s,t+D_i]$ so that the available capacity to a minimally backlogged class i flow is[4]

$$\widetilde{Y}^i_{D_i}(s,t) \quad = \quad (C(t-s+D_i+1)-\sum_{n\neq i}X^n_{D_i}(s,t))^+$$

[3] We derive an expression for the entire delay distribution of class i, $P[D^i>D_i]$ for all $D_i>0$. The point of most importance is $P[D^i>d^i]$, the probability of violating the class delay constraint, which we refer to as P^i.

[4] Recall that the system is guaranteed to be backlogged during this entire interval when class i sends a minimally backlogging input.

$$
\begin{aligned}
&= \quad (C(t-s+D_i+1) - \sum_{n\neq i} X^n(s, t+d^i-d^n))^+ \\
&\geq_{st} \quad (C(t-s+D_i+1) - \sum_{n\neq i} B^n(t+d^i-d^n-s+1))^+
\end{aligned}
$$

Thus, applying Definition 6, we have

$$
S^i_{D_i}(t) = (C(t+D_i) - \sum_{n\neq i} B^n(t-d^n+d^i))^+. \tag{20}
$$

Once again the proof is similar for the deterministic service envelope. □

We now derive a multi-class admission control test for EDF schedulers that support multiple deterministic and statistical service classes.

Lemma 6 *In an EDF scheduler in which service class i has statistical traffic envelope $B^i(t)$ and delay bound d^i, the delay distribution of class i satisfies*

$$
P[D^i > D_i] \leq P[\max_{u\geq 0}\{\overline{B^i(u)} + \sum_{n\neq i} \overline{B^n(u-d^n+d^i)} - C(u+D_i)\} > 0], \quad \forall D_i \in [0,\infty), \tag{21}
$$

where $\overline{B^n(x)}$ is a random variable with the same distribution as $B^n(x)$, for $n = 1, \cdots$, and any $x \geq 0$.

Proof. Similar to the proof of Lemma 4, we have

$$
\widetilde{Y}^i_{D_i}(s,t) \quad \geq \quad C(t-s+D_i+1) - \sum_{n\neq i} \overline{B^n(t+d^i-d^n-s+1)}.
$$

Replacing $\overline{S^i_{D_i}(u)}$ by $C(t-s+D_i+1) - \sum_{n\neq i} \overline{B^n(t+d^i-d^n-s+1)}$ and Applying Theorem 2 and Lemma 5 we have

$$
\begin{aligned}
P[D^i > D_i] \quad &\leq \quad P[\max_{u\geq 0}\{\overline{B}^i(u) - [C(u+D_i) - \sum_{n\neq i} \overline{B^n(u-d^n+d^i)}]\} > 0] \\
&= \quad P[\max_{u\geq 0}\{\overline{B}^i(u) + \sum_{n\neq i} \overline{B^n(u-d^n+d^i)} - C(u+D_i)\} > 0]. \qquad \square
\end{aligned} \tag{22}
$$

For $D_i = d^i$, Inequality (21) can be simplified as

$$
P[D^i > d^i] \leq P[\max_{u\geq 0}\{\sum_n \overline{B}^n(u-d^n) - C(u)\} > 0] \tag{23}
$$

which is the delay-bound-violation probability of class i.

3 Inter-Services Resource Sharing in Link-Sharing GPS

In Section 2, we developed tools for managing multi-class services using statistical service envelopes, considering SP and EDF as specific examples. Here we study a link-sharing GPS server, again using the framework of statistical service envelopes, with a goal of increasing the total utilization of the multi-class GPS server by exploiting inter-class resource sharing.

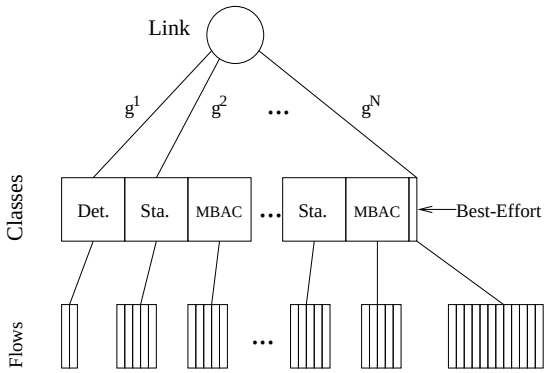

Figure 2: System Model for Admission Control

3.1 Generalized Processor Sharing

Figure 2 shows the system model for admission control in a multi-class GPS scheduler (see [7] for example). There are N service classes in the system, each allocated a weight ϕ^i. Each service class provides either deterministic, statistical, measurement-based, or best-effort services.[5] The admission control algorithm should admit a new flow only if the QoS of all classes can be satisfied. This multi-class service model can also support flow-based services, in which some service classes serve only one flow. Without considering inter-class resource sharing, one could view each service class as a FCFS server with capacity g^i, which is the guaranteed service rate $g^i = \frac{\phi^i}{\sum_m \phi^m} C$, as defined by the GPS service discipline. However, while exploiting this isolation property of GPS simplifies admission control, it does not corporate potential utilization gains due to inter-class statistical sharing.

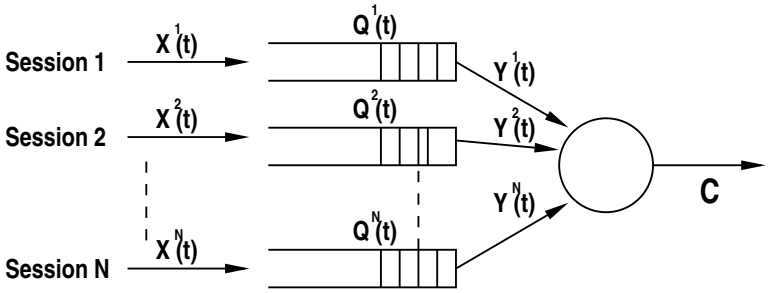

Figure 3: GPS System

Figure 3 illustrates the GPS system in the view of inputs, outputs and buffers. The aggregate traffic in each class is viewed as a session, and the notation for inputs, outputs and queues are as defined in Section 2.

For $1 \leq i \leq N$, let $Y^i(s,t)$ be the amount of class i traffic served during $[s,t]$. By definition of GPS,

$$\frac{Y^i(s,t)}{Y^m(s,t)} \geq \frac{\phi^i}{\phi^m}, m = 1, 2, \ldots, N \tag{24}$$

[5]Here, we study multiple deterministic and statistical service classes and leave study of measurement-based service to future work.

for any class i backlogged during $[s,t]$. Since each class has a guaranteed rate g^i whenever it is backlogged, the deterministic service envelope of class i is $s^i_{D_i}(t) = g^i(t + D_i)$.

3.2 Statistical Service Envelopes in GPS

In order to fully exploit inter-class resource sharing, we devise a technique for partitioning service classes. First, we illustrate an isolation/sharing model for admission control in Figure 4. In this model, some service classes will use their deterministic service envelope in admission control. These service classes may support deterministic services, in which deterministic traffic envelopes are used. Or they may support less aggressive statistical services which do not wish to exploit spare capacity from other classes. In view of service envelopes, we refer to these service classes as *isolation* classes (denoted as $\mathcal{I}$). Apart from these isolation classes, other service classes will exploit inter-class resource sharing using their statistical service envelope to admit an increased number of flows into the traffic class. We refer to these service classes as *sharing* classes (denoted as $\mathcal{S}$). Sharing classes cannot support deterministic services, but can support statistical, measurement-based, and best-effort services.

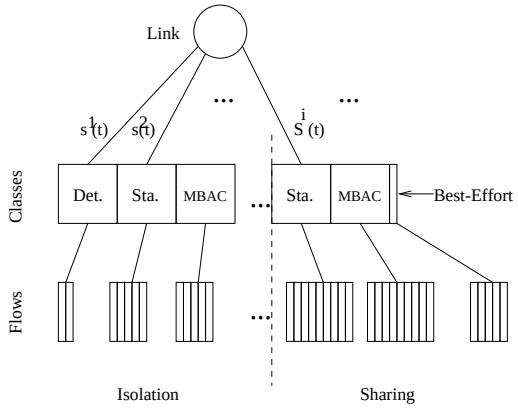

Figure 4: An Isolation/Sharing Model for Admission Control

Below, we derive an expression for the statistical service envelopes of sharing classes, which with application of Theorem 2 provides a multi-class admission control test which incorporates inter-class resource sharing.

Lemma 7 *In a GPS scheduler with class i traffic bounded by $B^i(t)$ and $b^i(t)$, the statistical service envelope for sharing class i is given by*

$$S^i_{D_i}(t) = \frac{\phi^i}{\sum_{m\in\mathcal{S}}\phi^m}[C(t+D_i) - \sum_{n\in\mathcal{I}} B^n(t+D_i)]. \tag{25}$$

Proof. Consider class-i arrivals ($i \in \mathcal{S}$) at t which have deadline $t + D_i$. The the essential traffic $X^n_{D_i}(s,t)$ ($n \in \mathcal{I}$) which gets serviced before the class-i traffic arriving at t contains at most class-n traffic arriving in the interval $[s, t + D_i]$. That is,

$$X^n_{D_i}(s,t) \le X^n(s, t + D_i). \tag{26}$$

Notice that for isolation classes such as those obtaining a deterministic service, this bound is quite tight, as such classes incur little or no queueing delays.

With the total capacity available in the interval $[s, t + D_i]$ given by $C(t - s + D_i + 1)$, according to Lemma 1 and similar to the proof of Lemma 4, the capacity for *all* sharing classes is given by

$$\begin{aligned} \widetilde{Y}^{\mathcal{S}}_{D_i}(s,t) &= (C(t-s+D_i+1) - \sum_{n\in\mathcal{I}} X^n_{D_i}(s,t))^+ \\ &\geq (C(t-s+D_i+1) - \sum_{n\in\mathcal{I}} X^n(s,t+D_i))^+ \\ &\geq_{st} (C(t-s+D_i+1) - \sum_{n\in\mathcal{I}} B^n(t+D_i-s+1))^+. \end{aligned} \tag{27}$$

For a particular class $i \in \mathcal{S}$, the GPS scheduler distributes this available service in a weighted-fair manner among classes. Thus, using Equation (24) and Definition 6, the statistical service envelope is given by

$$S^i_{D_i}(t) = \frac{\phi^i}{\sum_{m\in\mathcal{S}} \phi^m}[C(t+D_i) - \sum_{n\in\mathcal{I}} B^n(t+D_i)]. \ \square \tag{28}$$

We conclude by describing the complete admission control algorithm for a multi-class GPS server. Each class provides traffic parameters $b^i(t)$ and $B^i(t)$, and QoS parameters D_i and P^i. Each class has a weight ϕ^i and guaranteed rate g^i, with guaranteed service envelope $s^i_{D_i}(t) = g^i(t + D_i)$. For deterministic service classes, if $\max_t\{b^i(t) - s^i_{D_i}(t)\} \leq 0$, then the deterministic QoS for flows inside class i is guaranteed. For *isolation* statistical service classes, if $P[\max_t\{\overline{B^i(t)} - s^i_{D_i}(t)\} \geq 0] \leq P^i$, then the statistical QoS of class i is satisfied. For *sharing* statistical service classes, the statistical QoS is satisfied if $P[\max_t\{\overline{B^i(t)} - s^i_{D_i}(t)\} \geq 0] \leq P^i$, or if $P[\max_t\{\overline{B^i(t)} - \overline{S^i_{D_i}(t)}\} \geq 0] \leq P^i$ exists for a statistical service envelope $S^i_{D_i}(t)$.

4 Computational and Experimental Investigation

In Sections 2 and 3, we studied the delay bound violation probability using statistical traffic and service envelopes for SP, EDF, and GPS schedulers. In this section, we address the computational aspects of these admission control algorithms and perform trace-driven simulations to quantify the ability of our approach to exploit inter-class resource sharing. The workload consists of a set of 30-minute traces of MPEG compressed video from [18].

4.1 Computing the Delay Tail Probability

The admission control tests of Sections 2 and 3 are expressed as sums of independent random variables (from arrivals of different flows) with temporal correlation within each flow. As the expression for the queue length distribution of a FCFS takes a similar functional form [13], we can apply previous queueing theoretic techniques in computing probabilities such as those in Equation (23). Possible techniques include include large deviations theory and maximum-variance approaches and others reviewed in [13]. Here, we

apply the maximum-variance technique of [4] due to its computational simplicity and experimental accuracy when multiplexing a large number of sources. Regardless, if higher moments than the first and second of the traffic envelopes are available, the results below can be refined using large deviations techniques using an approach such as [3].

To approximate each flow's traffic descriptor $B_j(t)$, we use the rate variance envelopes $RV_j(t)$ and mean rate m_j, i.e., $E\{B_j(t)\} = m_j t$ and var$\{B_j(t)\} = t^2 RV_j(t)$. The details of computation for $RV_j(t)$ and m_j are given in [11]. When flows are multiplexed, the aggregate traffic envelope for class i approaches a Gaussian envelope with $B^i(t)$ having mean $\sum_{j \in C_i} t m_j$, and variance $\sum_{j \in C_i} t^2 RV_j(t)$ [11]. In practice, traffic flows can specify policing parameters, and use [12] to compute such statistical traffic envelopes from the deterministic parameters, e.g., standard dual leaky bucket traffic descriptors can be applied.[6]

To calculate $P[\max_t\{\overline{B^i(t)} - \overline{S^i_{D_i}(t)}\} > 0]$ in Equation (7) we utilize the maximum variance approach of [4] which computes the normalized excess arrivals at the dominant time scale as

$$\begin{aligned} \sigma_t^2 &= \text{var}\{\overline{B^i(t)} - \overline{S^i_{D_i}(t)}\} = \text{var}\{B^i(t) - S^i_{D_i}(t)\}, \\ \alpha_t &= \frac{0 - E\{\overline{B^i(t)} - \overline{S^i_{D_i}(t)}\}}{\sigma_t} = \frac{0 - E\{B^i(t) - S^i_{D_i}(t)\}}{\sigma_t}, \\ \alpha &:= \inf_t \alpha_t. \end{aligned} \tag{29}$$

Approximating $\{\overline{B^i(t)} - \overline{S^i_{D_i}(t)}\}$ as Gaussian, under conditions (C1)-(C2) in [4],

$$P[\max_t\{\overline{B^i(t)} - \overline{S^i_{D_i}(t)}\} > 0] \geq \max_t P[\overline{B^i(t)} - \overline{S^i_{D_i}(t)} > 0] = \max_t P[B^i(t) - S^i_{D_i}(t) > 0] = \phi(\alpha)$$

and

$$P[\max_t\{\overline{B^i(t)} - \overline{S^i_{D_i}(t)}\} > 0] \leq e^{-\frac{\alpha^2}{2}} \tag{30}$$

where $\phi(\alpha) = \frac{1}{\sqrt{2\pi}} \int_\alpha^\infty e^{-\frac{x^2}{2}} dx$. Roughly, the lower bound replaces the probability of the maximum with the maximum probability, and the upper bound uses the dominant time scale, the value of t minimizing Equation (29), to derive the exponential asymptotic upper bound of Equation (30). Proofs of these two bounds are given in [4], and we utilize both in the experiments below.

4.2 Admissible Regions in Multi-Class GPS

The scenario we consider is a link sharing GPS server with a total capacity of 45 Mbps. Different weights ϕ^i are given to different classes, which require either deterministic or statistical services. In the experiments, some classes will exploit inter-class resource sharing, while others will not.

In general, the benefits of inter-class resource sharing increase with the number of classes, as in the limit of one flow per class, a system without inter-class sharing behaves as a deterministic service. In these first experiments, we consider two classes to obtain insights into the *minimum* performance gains of inter-class sharing.

[6] The particular traffic model chosen and the method by which users obtain their traffic parameters is addressed in [13] and elsewhere and is beyond the scope of this work.

In each experiment, we calculate the admissible region for each class according to the flows' traffic characterizations and QoS requirements using the admission control algorithm described in Section 3. For example, in the first experiment, we let ϕ^1 vary from 0 to 1 and $\phi^2 = 1 - \phi^1$. The maximum admissible numbers of flows in each service class are evaluated for different weight assignment (ϕ^1, ϕ^2). The admissible region for each class is the area in two dimensional space bounded by the curve determined by these maximum admissible numbers of flows. We then perform trace-driven simulations using a GPS scheduler with each flow having a randomly shifted initial phase. Many simulations are run for each combination of numbers of class one and class two flows and the average delay bound violation probability is measured. The maximum numbers of class one and class two flows subject to this measured QoS constraint is then the experimental admissible region.

In the first experiment, we consider a GPS server with two service classes. Class 1 requires deterministic service, with $D_1 = 10$ msec, class 2 requires statistical service, with $D_2 = 20$ msec and $P^2 = 10^{-4}$. In the admission control tests, we use both the lower bound of Equation (30) and the upper bound of Equation (30) to approximate the deadline violation probability.

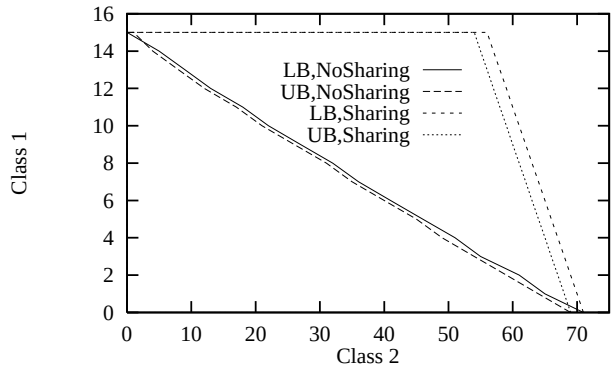

Figure 5: Admissible Regions for Deterministic and Statistical Services

Figure 5 shows the admissible regions for class 1 and 2 under four different conditions: with and without inter-class sharing for class 2, and upper and lower bounds for the deadline violation probability. Notice the significant increase in the admissible region due to exploiting inter-class resource sharing using our framework of statistical service envelopes. For example, using the lower bound and setting $g^1 = g^2 = C/2$, i.e., $\phi_1 = \phi_2$, without inter-class sharing, the admissible region is $(0 \leq$ class 1 flows $\leq 7, 0 \leq$ class 2 flows $\leq 31)$ and the total utilization is 45.3%. In contrast, with inter-class sharing, the admissible region is $(0 \leq$ class 1 flows $\leq 7, 0 \leq$ class 2 flows $\leq 62)$ and the total utilization is 82.2%, an increase of 81%. We also observe that the differences in the admissible regions using the lower and upper bounds are merely 1 or 2 flows. We next perform trace-driven simulations and measure the experimental delay bound violation rates using the admissible region calculated from the "sharing" tests. For the lower bound, the mean delay bound violation probability for class 2 is 5×10^{-4}, while for the upper bound, the mean violation probability for class 2 is approximately 5×10^{-5}. Since the QoS parameter is $P^2 = 10^{-4}$, we observe that the actual admissible region boundary must be between the LB and UB sharing curves, and that the admissible regions calculated using both bounds are very close to the true ones.

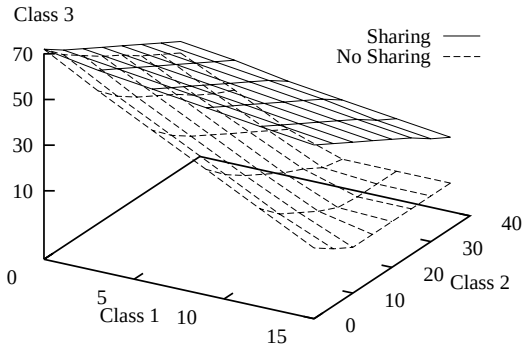

Figure 6: Admissible Regions for a Three-Class GPS Server

In the next experiment, we consider a three-class GPS scheduler. Class 1 requires deterministic service with $D_1 = 20$ msec, class 2 requires statistical service with $D_2 = 20$ msec and $P^2 = 10^{-4}$, and class 3 requires statistical service with $D_3 = 30$ msec and $P^3 = 10^{-4}$. Class 1 and 2 are isolation classes. We perform admission tests with and without class 3 exploiting inter-class sharing, and use the lower bound of Equation (30) to approximate the deadline violation probability. The admissible region is shown in Figure 6, which also illustrates the significant utilization gain of the approach.

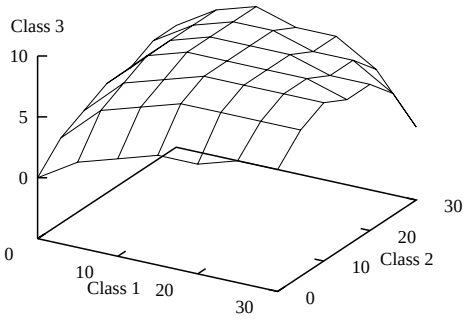

Figure 7: Increase in Admissible Regions for 3 Statistical Classes

In the above two experiments, the deterministic service class is exploited by the statistical service class to allow inter-class sharing. In the next experiment, we show that our approach is also able to exploit inter-class sharing among statistical service classes. We consider a three class GPS server with each class providing statistical services with the same QoS: $D = 20$ msec and $P = 10^{-4}$. Class 1 and 2 are set to isolation classes. In Figure 7, we show the difference in the admissible regions by allowing class 3 to exploit

inter-class sharing.

From Figure 7, observe that ignoring inter-class sharing leads to as many as 8 fewer flows admitted in class 3, for a loss of approximately 10% of the resource utilization. In this scenario, the intra-class statistics are fully exploited, and the gain comes solely from the inter-class statistics. In a high-speed GPS server, even if each class provides statistical service, when the number of service classes is large, the inter-class resource sharing gain can be significant.

5 Conclusions

In this paper, we developed multi-class admission control algorithms that exploit inter-class statistical resource sharing. We developed a framework of statistical service envelopes to study the problem and showed how such envelopes characterize the excess capacity available to a traffic class due to varying resource demands of other classes. We applied the approach to Strict Priority, Earliest Deadline First, and Generalized Processor Sharing schedulers and experimentally demonstrated that our admission control algorithms are able to extract a significant utilization gain from inter-class resource sharing.

References

[1] A. Adas and A. Mukherjee. Providing heterogeneous QoS bounds for correlated video traffic at a mutiplexor. *Performance Evaluation Journal*, 38(1):45–65, September 1999.

[2] J. Bennett and H. Zhang. Hierarchical packet fair queueing algorithms. *IEEE/ACM Transactions on Networking*, 5(5):675–689, October 1997.

[3] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Effective envelopes: Statistical bounds on multiplexed traffic in packet networks. In *Proceedings of IEEE INFOCOM 2000*, Tel Aviv, Israel, March 2000.

[4] J. Choe and N. Shroff. A central limit theorem based approach to analyze queue behavior in ATM networks. *IEEE/ACM Transactions on Networking*, 6(5):659–671, October 1998.

[5] R. Cruz. Quality of service guarantees in virtual circuit switched networks. *IEEE Journal on Selected Areas in Communications*, 13(6):1048–1056, August 1995.

[6] G. de Veciana and G. Kesidis. Bandwidth allocation for multiple qualities of service using generalized processor sharing. *IEEE Transactions on Information Theory*, 42(1):268–272, January 1995.

[7] S. Floyd and V. Jacobson. Link-sharing and resource management models for packet network. *IEEE/ACM Transactions on Networking*, 3(4):365–386, August 1995.

[8] S. Golestani. Duration-limited statistical multiplexing of delay-sensitive traffic in packet networks. In *Proceedings of IEEE INFOCOM'91*, pages 323–332, Bal Harbour, FL, April 1991.

[9] J. Hyman, A. Lazar, and G. Pacifici. Real-time scheduling with quality of service constraints. *IEEE Journal on Selected Areas of Communications*, 9(7):1052–1063, September 1991.

[10] S. Jamin, P. Danzig, S. Shenker, and L. Zhang. A measurement-based admission control algorithm for integrated services packet networks. *IEEE/ACM Transactions on Networking*, 5(1):56–70, February 1997.

[11] E. Knightly. Second moment resource allocation in multi-service networks. In *Proceedings of ACM SIGMETRICS '97*, pages 181–191, Seattle, WA, June 1997.

[12] E. Knightly. Enforceable quality of service guarantees for bursty traffic streams. In *Proceedings of IEEE INFOCOM '98*, San Francisco, CA, March 1998.

[13] E. Knightly and N. Shroff. Admission control for statistical QoS: Theory and practice. *IEEE Network*, 13(2):20–29, March 1999.

[14] J. Kurose. On computing per-session performance bounds in high-speed multi-hop computer networks. In *Proceedings of ACM SIGMETRICS '92*, pages 128–139, Newport, RI, June 1992.

[15] J. Liebeherr, D. Wrege, and D. Ferrari. Exact admission control for networks with bounded delay services. *IEEE/ACM Transactions on Networking*, 4(6):885–901, December 1996.

[16] R. Loynes. The Stability of a Queue with Non-independent Inter-arrival and Service Times. *Proc. Cambridge Philos. Soc.*, 58:497–520, 1962.

[17] A. Parekh and R. Gallager. A generalized processor sharing approach to flow control in integrated services networks: the single-node case. *IEEE/ACM Transactions on Networking*, 1(3):344–357, June 1993.

[18] O. Rose. Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems. In *Proceedings of IEEE Conference on Local Computer Networks*, pages 397–406, Minneapolis, MN, October 1995.

[19] Sheldon M. Ross. *Stochastic Processes*. Wiley, 1983.

[20] S. Shakkottai and R. Srikant. Many-sources delay asymptotics with applications to priority queues. In *Proceedings of ACM SIGMETRICS 2000*, June 2000.

[21] D. Wrege, E. Knightly, H. Zhang, and J. Liebeherr. Deterministic delay bounds for VBR video in packet-switching networks: Fundamental limits and practical tradeoffs. *IEEE/ACM Transactions on Networking*, 4(3):352–362, June 1996.

[22] Z. Zhang, D. Towsley, and J. Kurose. Statistical analysis of generalized processor sharing scheduling discipline. *IEEE Journal on Selected Areas in Communications*, 13(6):368–379, August 1995.