in Proceedings of IASTED Int Conf. on Wireless and Optical Comm., 2002, pp. 410-415

# Call Admission Control and Class Assignment in Hybrid IntServ/DiffServ Networks

Shahrokh Valaee
The Edward Rogers Sr. Department of
Electrical and Computer Engineering,
University of Toronto
10 King's College Road, Toronto,
Ontario, Canada M5S 3G4
Email: valaee@comm.utoronto.ca

**ABSTRACT**
This paper introduces a call admission controller for integrated services (IntServ) and differentiated services (DiffServ) networks. The approach is based on the worst-case service curve provisioning. The service curve, in general, is a stochastic function that moves with the fluctuation of input traffics. We define a deterministic universal service curve that is independent of the traffic fluctuations. The paper instruments an elegant call admission controller to guarantee that the true service curve is always lower bounded by the universal service curve. The quality-of-service (QoS) parameters — maximum delay and maximum backlog — are then calculated with respect to the deterministic bound.

**KEY WORDS**
Call Admission Control, Quality-of-Service, Service Curve Provisioning, IntServ, DiffServ, High-speed Networks, Virtual Network Partitioning

## 1 Introduction

The next generation Internet will be heavily dependent on optical infrastructure. In an optical environment, the task of network management should be instantaneous so that it can follow the dynamics of traffic. In such networks, per flow resource allocation is not usually practiced. An approach would be to aggregate a number of single flows into a class and impose the management instructions on the aggregated traffic. This is an approach prescribed by the *differentiated services* (DiffServ) standard [1]. Unlike *integrated services* (IntServ) [2], DiffServ assigns connections into a number of hierarchical classes where each class may include several subclasses. Resources are then allocated based on priority or fair sharing.

Although Diffserv solves the scalability problem of IntServ, it suffers from the so-called *resource stealing* drawback. Flows sharing a common class will compete inside the class for the resources available to all members and in some occasions might reduce the performance of their competitors — in terms of quality-of-service (QoS) measures — by stealing the resources that were initially used by their rivals. Unfortunately, the DiffServ standard does not propose a technique to alleviate the problem.

IntServ tackles the resource stealing problem by employing the resource reservation protocol (RSVP). In the resource reservation approach, each flow is assigned a fair amount of resources by employing a call admission controller (CAC). The status of each flow is registered inside the network and is used to guarantee a certain level of QoS. In a network with a large number of connections, per-flow resource management will become a tedious task.

In the present paper, we propose an approach to network provisioning in the presence of both IntServ and DiffServ domains. We assume that the DiffServ standard is utilized in backbone and IntServ is used in access network (see Fig. 1). Translation between the two standards should be performed in the boundary of the backbone network. An end-to-end QoS support is feasible if an appropriate mapping of the resources in the two domains are fulfilled. We assume that the QoS parameters inside the backbone network are solely identified by maximum delay and maximum backlog. Maximum delay reflects the timeliness of transmission and maximum backlog is an index that is related to traffic loss. Maximum backlog, in fact, indicates the maximum buffer consumption.

Our approach is to devise an algorithm, in the boundary of the two QoS supporting networks, so that it can be used to map an incoming flow in an IntServ network into an appropriate class in the DiffServ standard. The mapping should perform under the following restrictions: (1) the QoS of incoming traffic should be satisfied, (2) the QoS of all ongoing traffics should stay within the anticipated bounds, (3) the network should be utilized efficiently.

In this paper, we introduce a new methodology for call admission control and network provisioning. We will use a *worst-case* approach to network dimensioning. The *worst-case* network design has been employed in the literature as a means to restrict the network operation to a range of circumscribed behaviors [3] [4], [5], [6]. The main results, reported in the literature, quantify the worst-case network performance by monitoring the maximum delay

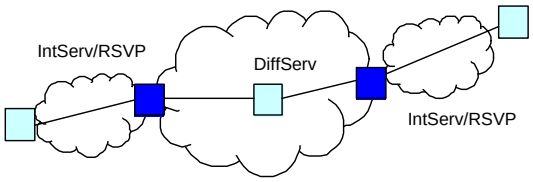


Figure 1. Access and backbone networks in an infrastructure for QoS support using the IntServ/DiffServ protocols.

and/or the maximum backlog, the scheduling strategy in the nodes, and the parameters of prescribed *regulators*, usually located at network boundaries for the *shaping* and/or *filtering* of input traffic.

The present paper focuses on assuring a guaranteed QoS for input traffics passing through a single node with a scheduler using a *generalized processor sharing* (GPS) discipline [4]. The DiffServ network will be substituted by a single node. The network is provisioned to guarantee the maximum delay and maximum backlog of input traffics. We use these parameters to specify the requested QoS and to measure the performance in the presence of *greedy* sources.

Our point of departure is to make use of the concept of service curve. Service curve provides a guaranteed lower bound to the amount of traffic handled by server [4] [6] [7]. Several approaches to service curve provisioning have been reported in the literature. Service curve, in general, is a function of the scheduler, input traffic fluctuation, and the volume of data infused by other sources as a crossing traffic. Hence, service curve is a multivariable functional moving over a range of behaviors circumscribed by some unknown time-varying parameters.

We propose using a deterministic lower bound to the service curve which we call it the *universal service curve*. A new traffic is admitted into the network if the true service curve is secured below by the prescribed lower boundary. This boundary is used to capture the worst-case behaviour of the projected traffic and to give a framework for a call admission procedure.

We do not provide proof for the theorems and lemmas in the present paper and refer to [8] for a more detailed discussion of the technique.

## 2 Background

### 2.1 Preliminaries

Let the instantaneous rate of the input traffic of the $i$th connection at time $t$ be indicated by $r_i(t)$. The aggregate input traffic is denoted by

$$A_i(t) = \int_0^t r_i(t')\,dt' \tag{1}$$

where we have assumed that the traffic flow starts at time zero. Note that $A_i(t)$ is a non-negative non-decreasing function of $t \geq 0$ and $A_i(0) = 0$. Represent the set of all non-decreasing functions over $\mathbb{R}^+$, where $\mathbb{R}^+$ is the set of all nonnegative real numbers, by

$$\mathcal{J} \triangleq \Big\{A(t)\,|\,0 \leq A(t_1) \leq A(t_2),\ \text{for } 0 \leq t_1 \leq t_2,\ \text{and } A(0) = 0\Big\}. \tag{2}$$

Further assume that the aggregate input traffic is upper bounded, for all $0 \leq s \leq t$, by a deterministic function $a_i(t)$, as

$$A_i(t) - A_i(s) \leq a_i(t-s). \tag{3}$$

$a_i(t)$ is called the *upper envelope* of the input traffic $A_i(t)$ [9] For instance, for leaky-bucket regulated traffics with the parameters $(\sigma_i, \rho_i)$, we have

$$A_i(t) \leq \sigma_i + \rho_i t \tag{4}$$

$$A_i(t) - A_i(s) \leq (t-s)\rho_i \tag{5}$$

where, $\rho_i$ is the token generation rate and $\sigma_i$ is the token pool size. Note that, unlike $A_i(t)$, the upper envelope $a_i(t)$ is a deterministic function.

An input traffic is called *greedy* at time origin if $A_i(t) = a_i(t)$ for all $t \geq 0$. Define

$$A(t) \triangleq \sum_{i=1}^{N} A_i(t). \tag{6}$$

The system is called *all-greedy* if

$$A(t) = a(t) \triangleq \sum_{i=1}^{N} a_i(t). \tag{7}$$

Throughout, we assume that the input traffic satisfies

$$\limsup_{t\to\infty} \frac{a(t)}{t} < C. \tag{8}$$

Our focus in the present paper is to devise an algorithm for call admission control so that the QoS parameters are guaranteed for all connections when the input traffics fluctuate in a region bounded by their upper envelopes. We take a conservative approach and propose our algorithm for an extreme situation in which all sources transmit with their maximum allowable rate (greedy sources). Similar assumptions have been used in the literature of worst-case network dimensioning [3] [4].

Assume that the input $A_i(t)$ is handled by a node which assigns to this traffic, the service curve $S_i(t; \underline{A}, \Phi) \in \mathcal{J}$, where $\underline{A} \triangleq (A_1, \ldots, A_N)$ is the vector of all traffics, $N$ is the total number of connections, and $\Phi$ is the scheduler used in the node. The service curve, $S_i(t; \underline{A}, \Phi)$, represents the minimum service guaranteed for user $i$. Fig. 2 illustrates a typical input traffic, $A_i(t)$, and the corresponding service curve, $S_i(t; \underline{A}, \Phi)$. The vertical distance at each time $t$ represents the instantaneous backlog. The horizontal distance represents the total delay incurred inside the

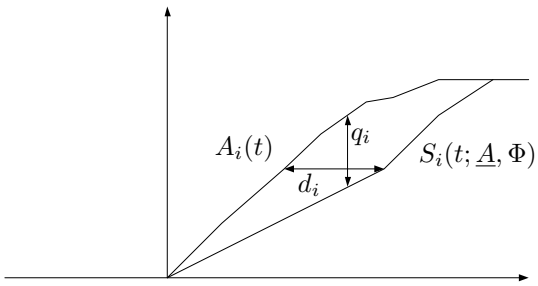


Figure 2. Input traffic of user $i$ and the corresponding service curve. The vertical distance at each time $t$ represents the instantaneous backlog and the horizontal distance represents the delay.

node for the traffic arriving at time $t$. The service given to a user depends on the traffic activity of all users that share the same bandwidth.

Throughout, we assume that the node uses a *generalized processor sharing* (GPS) scheduler, [4], with the parameters $\phi_i, i = 1, \ldots, N$, satisfying $\sum_{i=1}^{N} \phi_i < 1$. In GPS, the service provided for a session $i$, which is continuously backloged over the interval $[0, t]$, satisfies

$$\frac{S_i(t;\underline{A},\Phi)}{\phi_i} \geq \frac{S_j(t;\underline{A},\Phi)}{\phi_j}. \tag{9}$$

Equality holds in (9), if both sessions $i$ and $j$ are backloged in $[0, t]$. We represent the set of all backloged sessions at time $t$ by $\mathcal{B}(t)$. The complementary set of $\mathcal{B}(t)$ — the set of unbackloged sessions — is represented by $\mathcal{B}^c(t)$. In a GPS scheduler, the service provided for a backloged session $i$ at time $t$ is given by

$$S_i(t;\underline{A},\Phi) = \phi_i \frac{t - \sum_{j\in\mathcal{B}^c(t)} A_j(t)}{\sum_{j\in\mathcal{B}(t)} \phi_j}. \tag{10}$$

The normalized service curve will be represented by $\frac{S_i(t;\underline{A},\Phi)}{\phi_i}$. Note that the normalized service curve will be identical for all backloged traffics.

**Definition 1** For functions $A(t), B(t) \in \mathcal{J}$ define the maximum vertical distance (MVD) operator $\langle A(t), B(t)\rangle$ as

$$\langle A(t), B(t)\rangle \triangleq \max\Big\{ \sup_{t\geq 0}[A(t) - B(t)], 0\Big\}. \tag{11}$$

In [10], the MVD operation has been defined as the *scalar projection* of $A(t)$ onto $B(t)$ in the *min-plus algebra*. The operator $\langle .,. \rangle$ computes the maximum vertical distance between the first element and the second element in the brackets. From Fig. 2, we have that the maximum backlog for an input traffic with the aggregate traffic function $A_i(t)$ and the service curve $S_i(t;\underline{A},\Phi)$ is given by $\langle A_i(t), S_i(t;\underline{A},\Phi)\rangle$.

The following monotonicity property holds for the MVD operator [10].

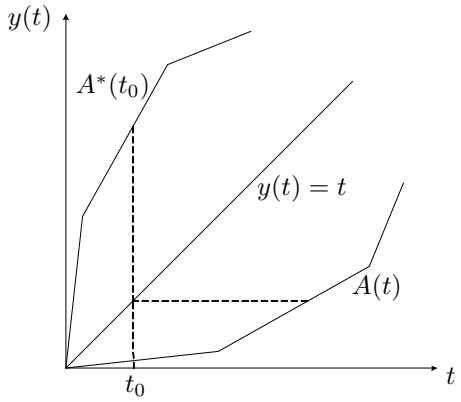


Figure 3. An illustration of the adjoint mapping.

**Proposition 1** *For $A_1(t) \leq A_2(t)$ and $B_1(t) \geq B_2(t)$, we have*

$$\langle A_1(t), B_1(t)\rangle \leq \langle A_2(t), B_2(t)\rangle. \tag{12}$$

**Definition 2** For $A(t) \in \mathcal{J}$, the *adjoint mapping* is defined as $A^*(t) \triangleq t + \inf\{d \,|\, t \leq A(t+d)\}$. The reverse mapping is defined as $A(t) = t - \inf\{d \,|\, A^*(t-d) \leq t\}$.

It is straightforward to show that there is a one-to-one relationship between $A(t)$ and $A^*(t)$ and that $A^*(A(t)) = A(A^*(t)) = t$ and $\langle t, A(t)\rangle = \langle A^*(t), t\rangle$. See Fig. 3 for an illustration of the adjoint mapping.

## 2.2 QoS Parameters

In this paper, maximum delay and maximum backlog will be used as the QoS parameters. The delay, $d_i(t;\underline{A},\Phi)$, and the backlog, $q_i(t;\underline{A},\Phi)$, for session $i$ with the input traffic $A_i(t)$ and the service curve $S_i(t;\underline{A},\Phi)$, at time $t$, are defined as

$$\begin{aligned} d_i(t;\underline{A},\Phi) &\triangleq \inf\{d \geq 0 \,|\, A_i(t) \leq S_i(t+d;\underline{A},\Phi)\} \\ q_i(t;\underline{A},\Phi) &\triangleq A_i(t) - S_i(t;\underline{A},\Phi). \end{aligned}$$

The objective is to give a set of constraints over $\underline{A}$ and $\Phi$ so that

$$\sup_{t\geq 0} d_i(t;\underline{A},\Phi) \leq D_i, \tag{13}$$

$$\sup_{t\geq 0} q_i(t;\underline{A},\Phi) \leq Q_i. \tag{14}$$

Using (11) and Definition 2, (13) and (14), are given by

$$\langle S_i^*(t;\underline{A},\Phi)\,,\, A_i^*(t)\rangle \leq D_i, \tag{15}$$

$$\langle\, A_i(t)\,,\, S_i(t;\underline{A},\Phi)\rangle \leq Q_i. \tag{16}$$

It is quite obvious that in order to have (15) and (16), the service curve should be bounded as

$$\max\{A_i(t+D_i), A_i(t) - Q_i\} \leq S_i(t;\underline{A},\Phi) \leq A_i(t).$$

Let $\bar{S}_i(t) \in \mathcal{J}$ be a deterministic function independent of the traffic vector, $\underline{A}$, and the scheduler $\Phi$ and assume $\bar{S}_i(t) \leq S_i(t;\underline{A},\Phi)$, for all $t$. If such a bound could

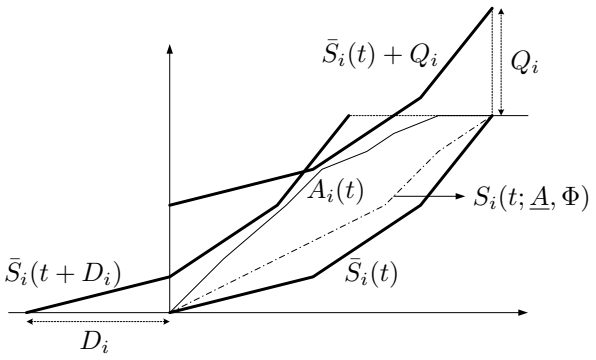


Figure 4. The deterministic lower boundary of the service curve.

be devised, the delay and backlog at time $t$ will be bounded above by

$$\begin{aligned} d_i(t;\underline{A},\Phi) &\leq \inf\{d \geq 0 \,|\, A_i(t) \leq \bar{S}_i(t+d)\}, \\ q_i(t;\underline{A},\Phi) &\leq A_i(t) - \bar{S}_i(t). \end{aligned}$$

Then, the QoS parameters, $D_i$ and $Q_i$, are guaranteed for the input traffic $A_i(t)$, if the following constraints hold

$$\langle \bar{S}_i^*(t), A_i^*(t) \rangle \leq D_i, \quad (17)$$

$$\langle A_i(t), \bar{S}_i(t) \rangle \leq Q_i, \quad (18)$$

$$S_i(t;\underline{A},\Phi) \geq \bar{S}_i(t). \quad (19)$$

A call request can be accepted into the network if (17)-(19) hold. This requires that $A_i(t)$ be constrained to the region

$$\bar{S}_i(t) \leq A_i(t) \leq \min\{\bar{S}_i(t+D_i), (\bar{S}_i(t)+Q_i)\} \quad (20)$$

as illustrated in Fig. 4.

## 3 Universal Service Curve

Assume a scenario with $N$ sources using a node with capacity, $C$, and a work-conserving GPS scheduler. The GPS weights are assigned during the call admission process. The traffic of each source $i$ is guaranteed by a service curve $S_i(t;\underline{A},\Phi)$. Since the total service given in a time unit cannot exceed the link bandwidth, the variation of the traffic descriptors and/or the scheduling parameters of a single source might affect the performance of all sources. In fact, admission of a new traffic $A_i(t)$ into a network, using a work-conserving GPS scheduler, will reduce the service given to all ongoing backloged sessions.

The maximum delay and the maximum backlog of session $i$, under an all-greedy regime, are defined as

$$d_i(\underline{a},\Phi) \triangleq \langle S_i^*(t;\underline{a},\Phi), a_i^*(t) \rangle, \quad (21)$$

$$q_i(\underline{a},\Phi) \triangleq \langle a_i(t), S_i(t;\underline{a},\Phi) \rangle \quad (22)$$

where $\underline{a} \triangleq (a_1(t),\ldots,a_N(t))$ is the set of traffic upper bounds, and $S_i(t,\underline{a},\Phi)$ is the service curve for user $i$ under an all-greedy regime. It can be shown that for each session $i$, the worst-case delay and backlog are obtained under an all-greedy regime, that is

$$\sup_{t\geq 0} d_i(t;\underline{A},\Phi) \leq d_i(\underline{a},\Phi), \quad (23)$$

$$\sup_{t\geq 0} q_i(t;\underline{A},\Phi) \leq q_i(\underline{a},\Phi). \quad (24)$$

**Theorem 1** *If $A(t) = a(t)$, and $a_j(t)$ is concave and upper semi-continuous for all $j \in \{1,\ldots,N\}$, the service curve $S_i(t;\underline{a},\Phi), i = 1,\ldots,N$, will be convex and lower semi-continuous in the interval $[0,T_i]$.*

Theorem 1 is not new. Similar results for affine upper boundary processes have been reported in [4]. Theorem 1 can be considered as the generalization of the results of [4].

The concave upper boundary processes, $a_i(t)$, used in Theorem 1, are of practical importance; for instance, leaky bucket regulators defined in (4) and (5) have concave upper-boundary processes. Leaky bucket regulators are used in ATM Forum [11] and IntServ [2] standards.

The objective of this paper is to find the conditions under which

$$d_i(\underline{a},\Phi) \leq D_i, \quad (25)$$

$$q_i(\underline{a},\Phi) \leq Q_i. \quad (26)$$

Define the *depletion time*, $T_i$, for each session $i$ in an all-greedy regime, as

$$T_i \triangleq \inf\{t > 0 \,|\, a_i(t) \leq S_i(t,\underline{a},\Phi)\}. \quad (27)$$

Without loss of generality, we assume that the sessions are numbered in the increasing order of depletion times, $T_1 \leq T_2 \leq \ldots \leq T_N$. The service curve $S_i(t;\underline{a},\Phi)$, $i = 1,\ldots,N$, for a GPS scheduler is then given by

$$S_i(t;\underline{a},\Phi) = \begin{cases} \phi_i \dfrac{Ct - \sum_{j=1}^{k} a_j(t)}{\sum_{j=k+1}^{N} \phi_j} & \text{for } t \in [T_k, T_{k+1}], \\ & k = 0,\ldots,i-1, \\ \infty & \text{for } t > T_i. \end{cases}$$

**Proposition 2** *Under an all-greedy regime, the maximum delay and the maximum backlog of session $i$ are bounded, respectively, by $d_i(\underline{a},\Phi) \leq T_i$ and $q_i(\underline{a},\Phi) \leq S_i(T_i;\underline{a},\Phi)$.*

The bounds in Proposition 2 are subject to change under the arrival of new calls and depletion of ongoing backloged sessions. In the sequel, we will find the bounds with respect to a universal service curve, $\bar{S}(t)$.

Define the *minimum service curve*, $S(t;\underline{a},\Phi)$, as

$$S(t;\underline{a},\Phi) \triangleq \min_{1\leq i\leq N} \left\{ \frac{S_i(t;\underline{a},\Phi)}{\phi_i} \right\}. \quad (28)$$

**Lemma 1** *For $A(t) = a(t)$, and $a_j(t)$ concave and upper semi-continuous for all $j \in \{1,\ldots,N\}$, the minimum service curve $S(t;\underline{a},\Phi)$ is convex and lower semi-continuous.*

**Definition 3** The *universal service curve*, $\bar{S}(t)$, is a positive, increasing, lower semi-continuous, convex function, independent of the traffic bounds $\underline{a}$ and the scheduler $\Phi$, and satisfies

$$\bar{S}(t) < S(t; \underline{a}, \Phi). \tag{29}$$

In the following theorem, we present sufficient conditions under which (29) is satisfied.

**Theorem 2** *Let $S(t) \in \mathcal{J}$ and $\bar{S}(t) \in \mathcal{J}$ be two positive, increasing, lower semi-continuous, convex functions, and assume $\dot{S}(0^+) > \dot{\bar{S}}(0^+)$. If there exist time indices $0 = t_0 < t_1 < \ldots < t_L = T$, such that*

$$S(t_j) + \dot{S}(t_j^+)(t_{j+1} - t_j) > \bar{S}(t_{j+1}), \tag{30}$$

*for all $j = 0, 1, 2, \ldots$ and $\lim_{j\to\infty} t_j \geq T$, then $\bar{S}(t) < S(t)$ for $0 < t < T$.*

Theorem 2 will be used in Section 4 to instrument a call admission controller for IntServ/DiffServ networks.

## 4 Call Admission Control

A call set-up is initiated by a source requesting a certain allotment of bandwidth and buffer size, along with a specified QoS (maximum delay and maximum backlog). A call admission controller is then activated to handle the application of the requested call. The call is accepted if (i) enough bandwidth and buffers are available (ii) the requested QoS can be guaranteed (iii) admission of the new call will not drive the QoS for the ongoing sessions below their prescribed thresholds. In fact, the call admission controller should verify that (17)-(19) are satisfied for all $i$, once the new call is accepted. The call admission controller might optimize a suitable cost function by shaping the source traffics, $A_i(t)$ — provided that it is allowed by the user — and selecting the scheduler, $\Phi$. We devise such an algorithm here.

Theorem 1 and Theorem 2 are used to find sufficient conditions for $S(t; \underline{a}, \Phi) \geq \bar{S}(t)$. We can use $\dot{S}(\bar{T}_{\ell-1}^+; \underline{a}, \Phi) \geq \dfrac{S(\bar{T}_{\ell-1}; \underline{a}, \Phi) - S(\bar{T}_{\ell-2}; \underline{a}, \Phi)}{\bar{T}_{\ell-1} - \bar{T}_{\ell-2}}$, to prove the following corollary. Define

$$\bar{I}_\ell \triangleq \bar{T}_\ell - \bar{T}_{\ell-1}. \tag{31}$$

**Corollary 1** *For a given universal service curve $\bar{S}(t)$ with $\bar{S}_\ell \triangleq \bar{S}(\bar{T}_\ell)$, $\ell = 1, \ldots, L$, and assuming that all traffics are backloged at time $0^+$, if the regulator parameters of the input traffics are so that*

$$\bar{S}_\ell \bar{I}_{\ell-1} \leq S(\bar{T}_{\ell-1}; \underline{a}, \Phi)(\bar{I}_\ell + \bar{I}_{\ell-1}) - S(\bar{T}_{\ell-2}; \underline{a}, \Phi)\bar{I}_\ell \tag{32}$$

*for all $\ell = 2, \ldots, L$, and $\sum_{i=1}^N \phi_i \leq 1$, then*

$$S(t; \underline{a}, \Phi) \geq \bar{S}(t), \tag{33}$$

*for $0 < t < T_L$.*

Corollary 1 gives a set of sufficient conditions for $S(t; \underline{a}, \Phi) \geq \bar{S}(t)$. These conditions depend on the service curve $S(t; \underline{a}, \Phi)$. It is important to note that the true value of the service curve is simply required on $L$ distinct points. Once a new call is accepted into the network or an existing call is terminated, the true service curve should be updated. These values depend on the characteristics of upper envelope functions. In the following algorithm, we propose a recursive procedure that can be used to evaluate $S(\bar{T}_\ell; \underline{a}, \Phi)$ for $\ell = 1, \ldots, L$.

**Algorithm 1** *The following recursion can be used to obtain the true service curve on the set $\{\bar{T}_1, \ldots, \bar{T}_L\}$:*

*(1) $k = 0$ and $S_\ell^0 = S(\bar{T}_\ell; \underline{a}, \Phi)$, $\ell = 1, \ldots, L$;*

*(2) Let $k = k + 1$, and define for $\ell = 1, \ldots, L$,*

$$S_\ell^k = \frac{C\bar{T}_\ell - \sum_{i=1}^N a_i(\bar{T}_\ell)\, 1\{a_i(\bar{T}_\ell) \leq \phi_i S_\ell^{k-1}\}}{\sum_{i=1}^N \phi_i\, 1\{a_i(\bar{T}_\ell) > \phi_i S_\ell^{k-1}\}} \tag{34}$$

*where $1\{x\}$ is the identifier function — $1\{x\} = 1$ if the predicate $x$ is "true" and $1\{x\} = 0$ if $x$ is "false";*

*(3) If $S_\ell^k = S_\ell^{k-1}$ for all $\ell = 1, \ldots, L$, then $S(\bar{T}_\ell; \underline{a}, \Phi) = S_\ell^k$, $\ell = 1, \ldots, L$, stop;*
*else go to Step 2.*

### 4.1 DiffServ Networks

In a backbone network, the traffic flows are usually aggregated into certain hierarchical classes. In such a network, per-user quality of service guarantee is not practiced. For instance, the DiffServ standard, proposed by the Internet Engineering Task Force (IETF), provides a framework for assignment of an aggregation of flows into certain classes of service, called the per-hop behaviors (PHB) [1]. For DiffServ, a direct application of a GPS scheduling for a single traffic flow may not be feasible. For such cases, we assume that the traffic of all connections belonging to a certain class, are stored in a single FIFO queue. The scheduler handles the traffic of each queue with a prescribed GPS weight.

Assume there exist $L$ queues represented by $\Pi_\ell$, $\ell = 1, \ldots, L$. The traffic in queue $\Pi_\ell$ is served by a GPS scheduler with parameter $|\Pi_\ell|\phi_\ell$ where $|\Pi_\ell|$ is the number of connections in the $\ell$th queue — all traffics in $\Pi_\ell$ have the same GPS weight. The GPS parameters $\phi_\ell$ are ordered as $\phi_1 > \ldots > \phi_L$. Hence, the traffic in $\Pi_1$ uses the highest hierarchy and receives the best treatment. The objective is to assign the traffics into queues such that the QoS constraints for all sessions are satisfied.

Define $\hat{a}_\ell(t) \triangleq \sum_{i\in\Pi_\ell} a_i(t)$. The service given to the aggregate traffic $\hat{a}_\ell(t)$ is controlled by the fairness coefficient $|\Pi_\ell|\phi_\ell$. In fact, the GPS parameter of each queue is adapted to the number of connections assigned to that queue.

Let $\hat{d}_\ell$ and $\hat{q}_\ell$ be, respectively, the delay and the backlog of the aggregate traffic $\hat{a}_\ell(t)$.

**Theorem 3** *The delay, $\hat{d}_\ell$, and the backlog, $\hat{q}_\ell$, are bounded above by*

$$\hat{d}_\ell \leq \langle \bar{S}^*(t), \hat{a}^*_\ell(|\Pi_\ell|\phi_\ell t)\rangle, \quad (35)$$

$$\hat{q}_\ell \leq \langle \hat{a}_\ell(t), \bar{S}(t)|\Pi_\ell|\phi_\ell\rangle. \quad (36)$$

A call admission controller should be such that for all $\ell \in \{1, \ldots, L\}$,

$$\langle \bar{S}^*(t), \hat{a}^*_\ell(|\Pi_\ell|\phi_\ell t)\rangle \leq \min_{i\in\Pi_\ell} D_i \quad (37)$$

$$\langle \hat{a}_\ell(t), \bar{S}(t)|\Pi_\ell|\phi_\ell\rangle \leq \sum_{i\in\Pi_\ell} Q_i, \quad (38)$$

where $D_i$ and $Q_i$ are the QoS parameters of session $i$.

Using Theorem 3, one can propose a call admission controller for a border gateway of a DiffServ domain that utilizes a GPS scheduler with dynamic weight assignments. Assume a call request is initiated by the QoS parameters, $D_i$ and $Q_i$.

**Theorem 4** *The call could be accepted into $\Pi_{\hat{\ell}}$ in DiffServ network if*

$$\hat{\ell} = \max \ \ell \quad (39)$$

$$\text{s.t.} \ \ \langle \bar{S}^*(t), \hat{a}^*_\ell(|\Pi_\ell|\phi_\ell t)\rangle \leq \min_{i\in\Pi_\ell} D_i, \quad (40)$$

$$\langle \hat{a}_\ell(t), \bar{S}(t)|\Pi_\ell|\phi_\ell\rangle \leq \sum_{i\in\Pi_\ell} Q_i, \quad (41)$$

$$S(t; \underline{a}, \Phi) \geq \bar{S}(t) \quad (42)$$

*where $S(t; \underline{a}, \Phi)$ is defined as*

$$S(t; \underline{a}, \Phi) = \frac{Ct - \sum_{\ell\in\mathcal{G}(t)} \hat{a}_\ell(t)}{\sum_{\ell\in\mathcal{G}^c(t)} |\Pi_\ell|\phi_\ell} \quad (43)$$

*with $\mathcal{G}(t)$ being the set of all subnetworks satisfying*

$$\mathcal{G}(t) \triangleq \left\{ \ell \,\middle|\, \frac{\hat{a}_\ell(t)}{|\Pi_\ell|\phi_\ell} \leq S(t; \underline{a}, \Phi) \right\} \quad (44)$$

*and $\mathcal{G}^c(t)$ the complementary set of $\mathcal{G}(t)$.*

## 5 Conclusion

In this paper, we have introduced a methodology for call admission control in a hybrid IntServ and DiffServ network. The approach has been based on the worst-case service curve provisioning. In general, the service curve is a function of the scheduler, the parameters of a regulator for the input flow, and the volume of the crossing traffic. We have shown that, assuming a concave upper envelope for input traffic and using a GPS scheduler in the node, an elaborate call admission controller can be found for which the true service curve will be always lower bounded by a deterministic universal service curve. The universal service curve is independent of the traffic fluctuations and can be used as a vehicle to measure the maximum guaranteed delay and the maximum guaranteed backlog of the users.

## References


[1] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," *Internet Engineering Task Force, RFC 2475*, 1998.

[2] R. Braden, D. Clark, and S. Shenker, *Integrated services in the internet architecture: An overview*. IETF RFC 1633, July 1994.

[3] R. L. Cruz, "A calculus for network delay, part I and part II," *IEEE Trans. Inform. Theory*, vol. 37, pp. 114–141, Jan. 1991.

[4] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services network: the single node case," *IEEE/ACM Trans. Networking*, vol. 1, pp. 334–357, June 1993.

[5] C. S. Chang, "Stability, queue length and delay of deterministic and stochastic queueing networks," *IEEE Trans. Automatic Control*, vol. 39, pp. 913–931, 1994.

[6] R. L. Cruz, "Quality of service gurantees in virtual circuit switched networks," *IEEE J., Select., Areas Commun.*, vol. 13, pp. 1048–1057, August 1995.

[7] H. Sariowan, *A service curve approach to performance guarantees in integrated-services networks*. PhD thesis, University of California, San Diego, 1996.

[8] S. Valaee, "A methodology for virtual network partitioning: The deterministic approach," *Preprint*, 2002.

[9] C. S. Chang, "On deterministic traffic regulation and service guarantees: a systematic approach by filtering," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1097–1110, May 1998.

[10] C. S. Chang, "Deterministic traffic specification via projections under the min-plus algebra," *Proc. IEEE INFOCOMM*, pp. 43–50, 1999.

[11] ATM Forum, *Traffic management specification*. Version 4.0, April 1996.