Minimum Cost Traffic Shaping: A User's Perspective on Connection Admission Control

Matthias Falkner, Michael Devetsikiotis, Ioannis Lambadaris
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario K1S 5B6
Canada

Abstract— **We propose a minimum cost method for traffic shaping in the context of QoS-based networks. Given the user's desired QoS and the network's resource availability, our procedure determines the least-cost parameters for a traffic shaper which still guarantees access to the network whilst satisfying the QoS constraints. We illustrate our scheme using on-off sources and formulate the QoS constraints by effective bandwidths.**

I. Introduction

Connection Admission Control in QoS-based networks typically relies on a traffic contract between the user and the network. The user is expected to characterize the traffic stream under a desired QoS level. This will give the network an indication of the anticipated load of the connection. The network then determines its resource availability. If sufficient resources are available to carry the connection, the call is admitted to the network. A specific set of user-declared traffic characterizations and the corresponding QoS are then registered in a traffic contract. The user's traffic stream is subsequently policed to ensure that the actual traffic stream does not violate the traffic contract. Furthermore, the user may shape its traffic stream to comply with the traffic contract. Finally, if insufficient resources are available, a connection is not granted by the network, thus preserving the QoS of existing calls.

In this paper, we propose to develop traffic shaping methodologies as a solution to an optimization problem, where the user establishes a connection and selects corresponding traffic shaping parameters satisfying the QoS constraints at a minimum cost. Alternatively, the traffic shaping problem can be interpreted as trying to find the minimum cost parameters of the policing or traffic shaping function which enable the user to gain access to the network whilst satisfying the QoS requirement of the connection.

The remainder of this paper is organized as follows: the next section aims to position our scheme within the context of (σ, ρ)-constrained connection admission control (CAC). To this aim, we restrict the scope of our scheme in this paper to on-off traffic and only consider traffic shaping. In section III we introduce the notion of cost. We present a general formulation of our model in section IV and show an example using on-off sources in section V. In section VI we discuss the issue of sensitivity analysis on the results and formulate possible extensions. We conclude the paper and provide suggestions for future research in section VII.

II. Network dimensioning and CAC under (σ, ρ) Constraints

The notion of a (σ, ρ)-constraint [1] provides an upper bound for the number of arrivals in a given interval of a particular arrival process. Let A_n denote the number of cell arrivals in slot n in a discrete time model. A given arrival process $\{A_n\}$ is (σ, ρ)-constrained in the time interval (k, m); $k, m \in Z^+$ if

$$\sum_{n=k}^{m} A_n \leq \rho(m-k+1) + \sigma.$$

Intuitively, ρ and σ can be interpreted as a mean rate and the maximum burst size of the given arrival process. Note that any arrival process $\{A_n\}$ is bounded above by an infinite family of (σ, ρ)-constraints. Two important properties of (σ, ρ)-constraints are worth mentioning:

Property–I If K arrival processes are multiplexed and each of the arrival processes is (σ_k, ρ_k)-constrained, $k = 1 \ldots K$, then the aggregate arrival process is (σ, ρ)-constrained, where $\sigma = \sum_{k=1}^{K} \sigma_k$ and $\rho = \sum_{k=1}^{K} \rho_k$.

Property–II If a (σ, ρ)-constrained arrival process feeds a G/D/1/σ queue with FCFS service discipline and a constant service rate $\rho, 0 < \rho < 1$, then no cells will be lost due to buffer overflows.

An arrival process $\{A_n\}$ can be shaped to conform to a $(1, \rho)$-constraint if it is fed through a G/D/1/σ queue [2], [3] with deterministic service rate ρ measured in cells per unit time. The burst size of the departure process is limited by the single server, hence the stream is $(1, \rho)$-constrained. This model represents a traffic shaper, because every cell is subject to a queuing and processing delay thus modifying the original inter-arrival process.

The above mentioned properties can be extended to form a calculus of (σ, ρ)-constraints for end-to-end virtual path (VP) connections [3]. The network consists of a series of output buffer switches connected by links. Each switch is composed of a switching fabric (SF) and one processor sharing node (PSN) per output link. Each PSN in turn is made up of a number of buffers served by a single server. For simplicity, we assume that VPs are exclusively associated with a single buffer in each PSN along the end-to-end path. Each VP carries a number of arrival streams or virtual circuits (VC) which all share a common source and

0-7803-5284-X/99/$10.00 © 1999 IEEE.

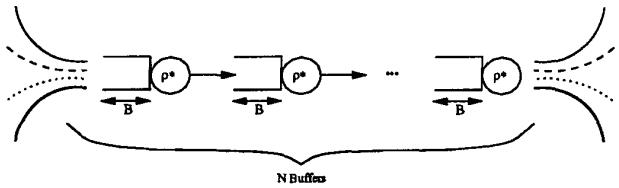

Fig. 1. Virtual Path Model

destination. A VP can thus be modeled by a series of N single server queues in tandem, as depicted in Fig. 1.

Let ρ^* denote the bandwidth allocation for the VP, i.e the bandwidth allocated at each queue. At each of the N PSN, the VP goes through a buffer of size B. Assume that the VP carries K VCs, each of which are (σ_k, ρ_k)-constrained, $k = 1, \ldots, K$, such that the aggregate arrival stream is (σ, ρ)-constrained by property I. We further assume that each of the PSNs has a work-conserving scheduling policy with minimum bandwidth parameter [3] $\mu = 0$, as is the case for PGPS (packetized general processor sharing) schedulers. It can be shown that if buffer n, $n = 1 \ldots N$, is sized according to

$$B_n \geq \lfloor \sum_{k=1}^{K} \sigma_k + n \rfloor; \quad n = 1 \ldots N,$$

then no cells are lost on the entire VP [3]. Furthermore, an upper bound for the end-to-end cell delay for cell v is given by

$$\delta_v \leq max\{d_{v-1}, a_v\} - a_v + \frac{N}{\rho^*} + \Pi, \qquad (1)$$

where a_v denotes the arrival time of cell v, d_v denotes the departure time of cell v and Π denotes the total propagation delay along the VP.

The above framework for network dimensioning under (σ, ρ)-constraints suggests the following algorithm for CAC: a newly arriving $(\tilde{\sigma}, \tilde{\rho})$-constrained VC can be admitted to the network if each buffer carrying the VC satisfies the following conditions

$$\tilde{\rho} \leq \rho^* - \sum_{k=1}^{K} \rho_k; \tilde{\sigma} \leq min\{B_n - \sum_{k=1}^{K} \sigma_k - n\}; n = 1 \ldots N, \quad (2)$$

If both conditions are satisfied, the respective amounts of buffer and bandwidth are reserved along the VP. The QoS provided after successful admission guarantees no cell losses and a maximum cell delay as in (1). The price to pay for this high level of QoS is a decrease in the statistical multiplexing gain. The network cannot fully exploit statistical multiplexing, since any VC which violates the above CAC rules is not admitted to the network, even though the network's resource utilization could be increased by statistical multiplexing more connections together.

For details on resource provisioning using (σ, ρ)-constraints, in particular proofs of the results presented in this section and extensions to cover arbitrary VP structures, the reader is referred to Kesidis [3].

III. Costing Aspects

In this section we briefly review the notion of cost in multiservice networks. Pricing and costing issues have not been the main focus of research by the engineering community. Walrand [4] claims that the main reason for this is the lack of economic background in engineering. Nevertheless, a number of researchers [5], [4], [6] and references therein have picked up on this concept and investigated the roles of tariffs in communication networks. Kelly [5], for example, uses the notion of cost to induce the user to declare the true values for the mean and the peak cell rates of a given on-off traffic source during CAC. Then the user is charged according to a linear function $f(m, M)$ of the declared mean rate m and the measured mean rate M, where $f(m, M) = a(m) + b(m)M$. The values for the fixed cost $a(m)$ and the variable cost $b(m)$ are chosen such that $f(m, M)$ is tangent to the curve

$$\alpha(M) = \frac{1}{\theta} log \left[1 + \frac{M}{h}(e^{\theta h} - 1) \right]$$

at the point $M = m$, as shown in Fig. 2. Note that $\alpha(M)$ represents the effective bandwidth for an on-off source with mean rate M and peak rate h, where h is assumed to be given. θ is a QoS parameter related to the blocking probability and will be introduced later. The concavity of the function $\alpha(M)$ with respect to M then ensures that $f(m, M) \geq \alpha(M)$, with equality holding only at the point $M = m$. The user is charged a premium for not declaring the true mean rate of the source. For further details the reader is referred to Kelly [5].

Jiang [7] provides a high-level description of a unified CAC framework involving the user and the network, consisting of respective agents which interact in two stages. In the first stage, the user agents characterize the information streams and define the desired performance parameters. At the same time, the network agents determine the network resource availability. In the second stage, both network and user agents come together in a market type environment to establish which information streams are carried by the network at agreed upon prices.

In the following section we build on the concepts described above. We look at CAC from a user's point of view. We take a *microeconomic* approach and eliminate the assumption that users and the network come together in a

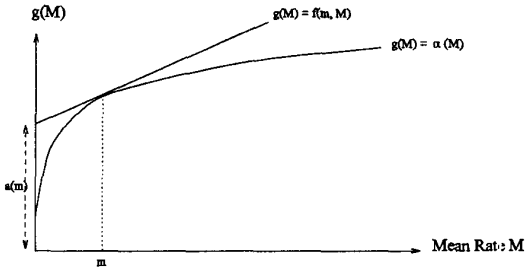

Fig. 2. Effective Bandwidth Pricing

market environment. In our view, a single user is *given* a price for theuse of network resources, which could be established using the framework described above. However, it is important to note that a single user cannot influence the price by declaring his or her demand for bandwidth. We feel that this assumption is more realistic in a networking environment, since such price determination schemes could not possibly be executed in real time every time a single user wishes to establish a connection. In what follows, we thus take a more focused approach on the operations at the user-network interface, rather than taking a global view of the entire network. We concentrate on using the concept of cost for the determination of the policing / shaping parameters.

IV. Minimum Cost Traffic Shaping

We now show that under the assumption that all streams can be shaped, the user faces the task of determining the shaping parameters such that the CAC constraints and the desired QoS constraints are met. In particular, we show that there may exist many such possibilities with which the user's traffic can be shaped. To differentiate between these multiple possibilities, we use the notion of a shaping cost. We thus formulate the shaping problem as a constrained minimization problem, where the user selects the traffic shaping settings such that the connection is admitted to the network whilst satisfying the user's QoS requirements and minimizing the economic cost of the connection.

Our modified network model is depicted in Fig. 3. We assume that VPs are pre-established with limited buffer / bandwidth resources. All of the N buffers along a VP have the same buffer capacity B, bandwidth allocation ρ^* and QoS parameter values. A newly arriving VC demands buffer and bandwidth resources according to the CAC scheme presented in section II. In return, the network imposes a cost $C(\alpha(.), \sigma, \rho)$ for the usage of these resources.

All user traffic follows an on-off model and is QoS-based and thus voluntarily shaped. Users wish to establish VCs requiring a particular QoS level, which implies that

- the user wishes to satisfy the constraints imposed by the CAC algorithm to establish the VC.
- the user is able to specify the QoS-constraints in the following form involving cell loss and worst-case delay

$$P(cell\ loss) < \epsilon_l$$
$$P(worst-case\ cell\ delay > \delta) < \epsilon_d,$$

i.e. that the user knows the values for ϵ_l, ϵ_d and δ. These constraints will be considered in more detail in section *IV.A*. Such functional forms have been found for on-off traffic using effective bandwidths [8], [9], [10], [11]. We also assume that the user has a limited budget constraint $B(\alpha(.), \sigma, \rho)$, representing the maximum amount of network resources that the user can afford. We consider the user to act rationally in an economic sense, i.e. that the user wishes to minimize cost in light of a limited budget constraint.

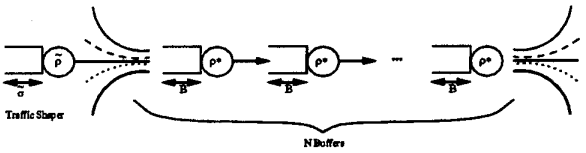

Fig. 3. Modified Virtual Path Model using Traffic Shaping

A. CAC and QoS Constraints

In order to gain access to the network, the VC has to satisfy (2). Since the traffic shaper is represented by a G/D/1/ $\tilde{\sigma}$ queue with service rate $\tilde{\rho}$ its output will be $(1, \tilde{\rho})$-constrained. A closer look at (2) reveals that any two-tuple $(\tilde{\sigma}, \tilde{\rho})$ of the set

$$A = \left\{ (\tilde{\sigma}, \tilde{\rho}) : 1 \leq \tilde{\sigma} \leq B - N - \sum_{k=1}^{K} \sigma_k ; \tilde{\rho} \leq \rho^* - \sum_{k=1}^{K} \rho_k \right\}$$

serves as a possible candidate to determine the shaper's parameters, since the output of a shaper with parameters buffer size $\tilde{\sigma}$ and rate $\tilde{\rho}$ is $(1, \tilde{\rho})$-constrained. We can thus formulate the following CAC constraints:

$$1 \leq B - N - \sum_{k=1}^{K} \sigma_k ; \quad \tilde{\rho} \leq \rho^* - \sum_{k=1}^{K} \rho_k . \tag{3}$$

Under (σ, ρ)-constrained CAC, the QoS provided by a VP guarantees no cell loss and a maximum cell delay as in (1) if the CAC conditions (3) are met. Cell losses can thus only occur at the shaper, and hence the cell loss probability (CLP) of the traffic shaper also determines the end-to-end CLP for the VP. A constraint for the buffer size can generally be formulated as follows, given the user's desired end-to-end CLPϵ_l

$$P(Q > \tilde{\sigma}) \triangleq f(\tilde{\alpha}(.), \tilde{\sigma}, \tilde{\rho}) < \epsilon_l . \tag{4}$$

In the case of on-off sources, the effective bandwidth approximation for the queue length [8], [9], [10], [11] can be used to obtain a particular constraint for the buffer size $\tilde{\sigma}$ as

$$\tilde{\sigma} \geq -\frac{log \epsilon_l}{\tilde{\theta}} .$$

The QoS parameter $\tilde{\theta}$ of the traffic shaper can now be introduced as promised in section III. $\tilde{\theta}$ can be determined as the solution to the equation $\tilde{\alpha}(\theta) = \tilde{\rho}$, where $\tilde{\alpha}(\theta)$ is the effective bandwith of the newly arriving stream.

Similarly, a QoS-constraint for the worst-case end-to-end delay generally takes the form

$$P(D_{WC} > \delta) \triangleq g(\tilde{\alpha}(.), \tilde{\sigma}, \tilde{\rho}, \delta) < \epsilon_d , \tag{5}$$

given the user's desired worst-case cell delay δ and the desired probability of achieving this delay ϵ_d.

Let us split the worst-case end-to-end cell delay D_{WC} into the worst-case delay across the VP D_{VP} plus the delay

introduced by the traffic shaper D_{TS}. Note that the upper bound for the worst-case delay across the VP is $d_{UB} = B(N+1)/\rho^* + \Pi$. In the case of on-off sources we can again use the effective bandwidth approximation for the queue length to specify the following delay constraint

$$\begin{aligned} P(D_{WC} > \delta) &= P(D_{TS} + D_{VP} > \delta) \\ &\leq P(D_{TS} + d_{UB} > \delta) < \epsilon_d, \end{aligned}$$

giving

$$P(D_{TS}+d_{UB} < \delta) = P(\tilde{\rho} D_{TS} < \tilde{\rho}(\delta - d_{UB})) \approx e^{-\tilde{\theta}\tilde{\rho}(\delta - d_{UB})}$$

or

$$\tilde{\rho}(\delta - d_{UB}) \geq -\frac{log \epsilon_d}{\tilde{\theta}}.$$

If the parameters for the traffic shaper satisfy constraints (4) and (5), the user's desired QoS requirements for the end-to-end VP are met.

B. Proposed Traffic Shaping Scheme

During the CAC procedure, the user has to declare values for $\tilde{\sigma}$ and $\tilde{\rho}$ which describe the traffic stream. These values are subsequently used for traffic shaping. The user is thus faced with declaring the appropriate values for $\tilde{\sigma}$ and $\tilde{\rho}$ so as to satisfy all of the above constraints. The user can select the particular values for $(\tilde{\sigma}, \tilde{\rho})$ as the solution to the following optimization problem:

$$\begin{aligned} &\text{Min } C(\tilde{\alpha}(.), \tilde{\sigma}, \tilde{\rho}) && \text{(OBJ)} \\ &\text{s.t. } P(Q > \tilde{\sigma}) \triangleq f(\tilde{\alpha}(.), \tilde{\sigma}, \tilde{\rho}) < \epsilon_l && \text{(C1)} \\ &\quad P(D_{WC} > \delta) \triangleq g(\tilde{\alpha}(.), \tilde{\sigma}, \tilde{\rho}, \delta) < \epsilon_d && \text{(C2)} \\ &\quad 0 < \tilde{\sigma} \leq \sigma_{avail} && \text{(C3)} \\ &\quad 0 < \tilde{\rho} \leq \rho_{avail} && \text{(C4)} \\ &\quad C(\tilde{\alpha}(.), \tilde{\sigma}, \tilde{\rho}) < B(\tilde{\alpha}(.), \tilde{\sigma}, \tilde{\rho}). && \text{(C5)} \end{aligned}$$

The objective function represents the user's rational behaviour in wanting to minimize the economic cost of the connection. Constraints (C1) and (C2) represent the QoS-constraints described in section *IV.A*. Constraints (C3) and (C4) represent the CAC-constraints that have to be met in order to gain admission. Here, σ_{avail} and ρ_{avail} represent the available buffer and bandwidth capacities , thus taking traffic streams from other users into account. Constraint (C5) ensures that the user's budget limit is not exceeded.

This non-linear optimization problem can be solved by the Lagrangean method under mild conditions for the constraints and the objective function. Arrow and Enthoven [12] have shown that if $C(\tilde{\alpha}(.), \tilde{\sigma}, \tilde{\rho})$ is differentiable and quasi-concave [1] and all of the constraints are differentiable and quasi-convex [2] and assuming that a solution for $(\tilde{\sigma}, \tilde{\rho})$ exists, then the Lagrangean method will find it. Furthermore, if the conditions satisfy the Kuhn-Tucker conditions and the quasi-concave programming conditions, then the solution is unique. We will illustrate the method on an example in the following section.

The optimization problem can be illustrated best using the contours of the constraints as shown in Fig. 4. The horizontal and vertical lines represent the physical constraints (C3) and (C4). The additional vertical line represents constraint (C2), showing the value of ρ for which $P(D_{TS} > \delta - d_{UB}) = \epsilon_d$ for given ϵ_d, δ (it can be shown that this line is vertical for on-off sources). Similarly, the curved line shows the values for (σ, ρ) for which the $P(Q > \sigma) = \epsilon_l$, where again ϵ_l is given, thus representing constraint (C1). The dashed lines indicate the values for (σ, ρ) for which the total cost takes on a certain value. The optimization problem tries to find the smallest such dashed line such that the constraints are still satisfied.

We will now illustrate our minimum cost traffic shaping scheme under (σ, ρ)-constrained CAC using on-off sources as an example.

V. Example: On-Off Sources

The illustration of our scheme in this example assumes that the user is able to completely specify the on-off traffic model operating in discrete time and modeled by a two state markov modulated arrival process with transition probability matrix R. A single arrival occurs when the model is in its on-state. No arrivals occur during the off-state. The model jumps from the on-state to the off-state in any slot according to the matrix R. The effective bandwidth of the traffic is thus given by

$$\alpha(\tilde{\theta}) = \frac{1}{\tilde{\theta}} log \left[\frac{r_{11} + r_{22}e^{\tilde{\theta}} + \sqrt{(r_{11} + r_{22}e^{\tilde{\theta}})^2 + 4r_{12}r_{21}e^{\tilde{\theta}}}}{2} \right],$$

where $\tilde{\theta} = sup_\theta(\alpha(\theta) \leq \tilde{\rho})$. We also assume that the user wishes to gain access to a particular VP, and that the network has communicated the upper bound for the delay across this VP, d_{UB}, to the user. Thus the user does

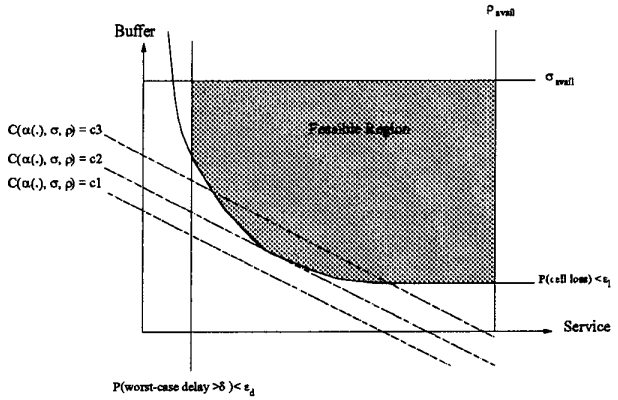

Fig. 4. Graphical representation of the optimization problem

[1] $Y = f[x]$ is said to be quasi-concave iff, whenever $f[x_1] \geq c$ and $f[x_2] \geq c$ then also $f[kx_1 + (1-k)x_2] \geq c$, for all $0 < k < 1$.

[2] $Y = f[x]$ is said to be quasi-convex iff, whenever $f[x_1] \leq c$ and $f[x_2] \leq c$ then also $f[kx_1 + (1-k)x_2] \leq c$, for all $0 < k < 1$.

not need to know exactly the number of hops on the VP to discount the QoS-requirements. Under the above assumptions we can use the effective bandwidth approximation for markovian traffic to formulate the QoS-constraints as

$$\epsilon_l \geq e^{-\tilde{\theta}\tilde{\sigma}}; \quad \epsilon_d \geq e^{-\tilde{\theta}\tilde{\rho}(\delta - d_{UB})}.$$

As far as the CAC-constraints are concerned, we assume that the network informs the user about the maximum values for $(\tilde{\sigma}, \tilde{\rho})$ which guarantee admission, i.e. we assume that

$$\sigma_{avail} = B - N - \sum_{k=1}^{K} \sigma_k \geq 1; \quad \rho_{avail} = \rho^* - \sum_{k=1}^{K} \rho_k$$

are given to the user. Since the users traffic is voluntarily shaped with a buffer of size $\tilde{\sigma}$ and service rate $\tilde{\rho}$ the output stream is always $(1, \tilde{\rho})$-constrained irrespective of the value of $\tilde{\sigma}$ we select. In this example we also assume that the user has no budget constraint. This will reduce the complexity of the optimization problem without loss of generality. Finally, we assume that the user is charged for the buffer and bandwidth resources according to a linear function at a cost of C_σ and C_ρ per unit of time respectively. The optimization problem can then be formulated as follows:

$$\begin{aligned}
&\text{Min } C_\sigma\tilde{\sigma} + C_\rho\tilde{\rho} && \text{(OBJ)}\\
&\text{s.t. } \tilde{\theta}\tilde{\sigma} \geq -ln(\epsilon_l) && \text{(C1)}\\
&\quad \tilde{\theta}\tilde{\rho}(\delta - d_{UB}) \geq -ln(\epsilon_d) && \text{(C2)}\\
&\quad 1 \leq \tilde{\sigma} \leq \sigma_{avail} && \text{(C3)}\\
&\quad 0 < \tilde{\rho} \leq \rho_{avail}, && \text{(C4)}
\end{aligned}$$

bearing in mind that $\tilde{\alpha}(\theta) = \tilde{\rho}$ holds and therefore $\tilde{\theta}$ is a function of $\tilde{\rho}$. The contour plot of the feasible region is shown in Fig. 4.

The form of the Lagrangean is as follows:

$$\begin{aligned}
\mathcal{L} \;=\; & C_\sigma\tilde{\sigma} + C_\rho\tilde{\rho}\\
& - \mu_1(\tilde{\theta}\tilde{\sigma} + ln(\epsilon_l))\\
& - \mu_2(\tilde{\theta}\tilde{\rho}(\delta - d_{UB}) + ln(\epsilon_d))\\
& - \mu_3(\sigma_{avail} - \tilde{\sigma})\\
& - \mu_4(\rho_{avail} - \tilde{\rho}),
\end{aligned}$$

and the solution to this equation is the set of values for $\tilde{\theta}, \tilde{\sigma}, \tilde{\rho}, \mu_1, \mu_2, \mu_3$ and μ_4 for which the following Kuhn-Tucker conditions hold: $\tilde{\sigma} > 0, \partial\mathcal{L}/\partial\tilde{\sigma} = 0, \tilde{\rho} > 0, \partial\mathcal{L}/\partial\tilde{\rho} = 0, \mu_i \geq 0, \partial\mathcal{L}/\partial\mu_i \leq 0, \mu_i(\partial\mathcal{L}/\partial\mu_i) = 0$ for $i = 1 \ldots 4$.

We assume that constraints (C3) and (C4) are not binding, i.e. that sufficient resources are available to find a solution for $(\tilde{\sigma}, \tilde{\rho})$. In the calculation of $\partial\mathcal{L}/\partial\tilde{\rho}$ we need to find $\partial\theta/\partial\tilde{\rho}$. Using $\tilde{\alpha}(\theta) = \tilde{\rho}$ we get $\partial\theta/\partial\tilde{\rho} = \partial\theta/\partial\tilde{\alpha}(\theta) = 1/(\partial\tilde{\alpha}(\theta)/\partial\theta)$.

After some algebra we can find $\tilde{\theta}$ as the solution to either $\theta\tilde{\alpha}(\theta) = -ln(\epsilon_d)/(\delta - d_{UB})$ or $\theta^2(\partial\tilde{\alpha}(\theta)/\partial\theta) = -C_\sigma ln(\epsilon_l)/C_\rho$, depending on whether (C2) is binding or not. Given $\tilde{\theta}$ the remaining values can be found using $\tilde{\rho} = \tilde{\alpha}(\tilde{\theta}), \tilde{\sigma} = -ln(\epsilon_l)/\tilde{\theta}, \mu_1 = C_\sigma/\tilde{\theta}, \mu_{3,4} = 0$ and

$$\mu_2 = \begin{cases} C_\rho + \frac{\frac{C_\sigma}{\tilde{\theta}}\frac{ln\epsilon_l}{\tilde{\theta}}\frac{\partial\theta}{\partial\tilde{\rho}}}{(\delta - d_{UB})(\tilde{\theta} + \tilde{\alpha}(\tilde{\theta})\frac{\partial\theta}{\partial\tilde{\rho}})} \\ 0 \end{cases},$$

again, depending on whether (C2) is binding or not. If such a solution can be found it has the desirable property of guaranteeing access to the network, provided the traffic stream is shaped by a $G/D/1/\tilde{\sigma}$ queue with service rate $\tilde{\rho}$. The users QoS will be met at a minimum cost per time unit of the connection.

VI. Discussion and Generalizations

The above formulation of the CAC problem has several desirable properties. First of all, formulating the problem as above provides a procedure that is easily implemented in real-time, given that the arrival stream is approximated by on-off sources. The user and the network only have to exchange minimal information required to determine the CAC and QoS constraints. Examining the region defined by the constraints, as depicted in Fig. 4, immediately reveals the feasibility of the solution. If this region defines an empty set, the user can re-examine the demanded QoS and repeat the process with alternative parameters. If the region defines a singleton, the solution follows immediately. If the region defines a set of values for σ and ρ, the solution follows from the mathematical procedure outlined above. This mathematical procedure guarantees a solution provided the network determines an appropriate cost function. If the constraints are quasi-convex, a quasi-concave cost function guarantees a solution. Similarly, if the constraints are quasi-concave, a quasi-convex cost function guarantees a solution. We thus see it as the network's responsibility to determine the appropriate form of the cost function to generate a solution.

The values taken by the Lagrangean multipliers μ_i also reveal important information on the sensitivity of the result. If $\mu_i > 0$, then constraint i is binding and furthermore indicates how the value of the objective function would change in response to a small relaxation of constraint i. Thus we get from the Lagrangean multipliers an indication on how much the cost changes in response to a small change in the desired QoS values of ϵ_l or ϵ_d respectively.

Furthermore, our model can easily be extended. The cost function we have used in section V seems reasonable if the network charges the user simply for the quantity of buffer and bandwidth resources consumed. This particularly applies if the resource utilization is low. However, in case of high resource utilization, the cost function could incorporate an element of a prohibitive cost [4] by letting the cost function for the resources increase as utilization rises. As resource utilization rises, more and more users would not be able to satisfy their budget constraints and the resources would be allocated to those users who were willing and able to pay the higher price. User with 'urgent' connections can still access the network. For the network, such a cost function implies higher revenues. The above scheme is then still valid, provided the new cost function is still

differentiable and quasi-concave. Although closed-form solutions may not be feasible, real-time numerical algorithms can also be implemented.

Finally, we comment on some of the assumptions that we have made in this paper. Our model currently hinges on finding a function form for the QoS constraints. In the example we have used effective bandwidth functions to express the QoS constraints. Such functions have been determined for many of the prevalent traffic models, including also long range dependent traffic models [13]. We do realize that the user may not know the effective bandwidth function a priori, in particular if the underlying traffic model is unknown. However, the on-off model or aggregations thereof can be used to approximate a large number of traffic patterns. To find effective bandwidths for a larger set of traffic models or empirically is currently an active area of research.

VII. Conclusion

In this paper we have addressed the issue on how to determine the parameters for a traffic shaper or what parameters to declare during CAC by identifying a minimization problem. In our approach, traffic shaping involves the minimization of the cost of the connection, in terms of the network resources bandwidth and buffer, such that access to the network is obtained and the desired QoS is provided by the connection. We have discussed the general forms of such QoS and CAC constraints for CAC procedures based on (σ, ρ)-constraints and effective bandwidths.

Of course our scheme is also applicable with QoS constraints not based on effective bandwidths, as long as these remain functions of the buffer and bandwidth resources and satisfy the quasi-convexity conditions. Research is underway on extending our model to cover empirical traffic, such as pre-recorded video traces or real-time traffic. Moreover, the QoS constraints we have used in this paper can be easily extended to include further QoS measures, such as delay jitter or message delay. In such cases, our scheme would have to be extended by expressing the functional relationship of the additional or alternative QoS measures by quasi-convex functions of buffer and bandwidth. Similar extensions need to be made for alternative CAC schemes or traffic shaping/policing models. Some of these extension are presented in [14].

Acknowledgments

The authors would like to acknowledge the Natural Sciences and Engineering Research Council of Canada.

References

[1] R. C. Cruz, "A calculus of network delay, Part 1: Network Elements in isolation," *IEEE Transactions on Information Theory*, vol. 37, pp. 114–131, 1991.

[2] M. Schwartz, *Broadband Integrated Networks*, Prentice Hall, New Jersey, USA, 1st edition, 1996.

[3] G. Kesidis, *ATM Network Performance*, Kluwer Academic Publishers, Boston, MA, USA, 1st edition, 1996.

[4] J. Walrand and P. Varaiya, *High-performance Communication Networks*, Morgan Kaufmann, New Jersey, USA, 1st edition, 1996.

[5] F. Kelly, "On Tariffs, policing and admission control for multiservice networks," *Operations Research Letters 15*, pp. 1–9, 1994.

[6] J. K. Mackie-Mason and H. R. Varian, "Pricing the Internet," in *International Conference on Telecommunication Systems Modelling*, Nashville, TN, USA, March 1994, pp. 378–393.

[7] H. Jiang and S. Jordan, "Connection Establishment in High-Speed Networks," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 7, pp. 1150–1161, September 1995.

[8] C. S. Chang, "Stability, Queue Length, and Delay of Deterministic and Stochastic Queueing Networks," *IEEE Transactions on Automatic Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[9] H. Ahmadi R. Guerin and M. Naghshineh, "Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 968–981, September 1991.

[10] F. Kelly, "Notes on Effective Bandwidths," in *Royal Statistical Society Lecture Notes Series*, Oxford, 1996, vol. 4, pp. 141–168, Oxford University Press.

[11] G. Kesidis, J. Walrand, and C. S. Chang, "Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources," *IEEE ACM Transactions on Networking*, pp. 424–428, August 1993.

[12] K. J. Arrow and A. C. Enthoven, "Quasi-concave programming," *Econometrica*, vol. 29, pp. 779–800, 1961.

[13] N. G. Duffield and N. O'Connell, "Large Deviations and overflow probabilities for the general single-server queue, with applications," *Math. Proc. Cambridge Phil. Soc.*, pp. 363–375, 1995.

[14] M. Falkner, M. Devetsikiotis, and I. Lambadaris, "A Framework for Cost-based Connection Admission Control," *Carleton University Technical Report*, 1998.