

# A Game Theoretic Framework for Bandwidth Allocation and Pricing in Broadband Networks

Haïkel Yaïche, Ravi R. Mazumdar, *Senior Member, IEEE*, and Catherine Rosenberg, *Senior Member, IEEE*

**Abstract**—In this paper, we present a game theoretic framework for bandwidth allocation for elastic services in high-speed networks. The framework is based on the idea of the Nash bargaining solution from cooperative game theory, which not only provides the rate settings of users that are Pareto optimal from the point of view of the whole system, but are also consistent with the fairness axioms of game theory. We first consider the centralized problem and then show that this procedure can be decentralized so that greedy optimization by users yields the system optimal bandwidth allocations. We propose a distributed algorithm for implementing the optimal and fair bandwidth allocation and provide conditions for its convergence. The paper concludes with the pricing of elastic connections based on users' bandwidth requirements and users' budget. We show that the above bargaining framework can be used to characterize a rate allocation and a pricing policy which takes into account users' budget in a fair way and such that the total network revenue is maximized.

**Index Terms**—Bandwidth allocation, elastic traffic, game theory, Nash bargaining solution, pricing.

## I. INTRODUCTION

CURRENT high-speed networks have to support applications which have no way of predicting their traffic requirements in advance, but have stringent loss requirements and can tolerate variations in transfer delays. These performance characteristics mean that the sources can be made to modify their data transfer rates according to network conditions. These services are referred to as *elastic services*. Their source rates are adjusted according to the network conditions so the network can carry a variable number of bursty connections in an efficient manner. Typical services, which share these properties, are TCP/IP based services, ATM available bit rate (ABR) services, or services using bandwidth-on-demand on a multiple access system.

These applications are expected to ride “on top of” (at least partially since some minimum bandwidth may be reserved) bandwidth-guaranteed connections and utilize any residual bandwidth. Since the available bandwidth will change depending on the amount of “background” bandwidth-guaranteed

services being carried, the incoming elastic sources will have to continually change their rates based on some notification by the network on the available bandwidth. Thus the notion of *rate control* of sources arises.

Since potentially there are many sources distributed in the network which will be competing for the use of the available bandwidth, there are several issues which arise and must be dealt with. These are: 1) efficient bandwidth allocation to the different sources taking into account their different needs and performance requirements; 2) the crucial notion of fairness; 3) the ability to implement the allocation scheme in a distributed manner with minimal communication overheads; and 4) the issue of pricing the bandwidth in such a way that the network revenue will be maximized if the users are allocated bandwidth according to 1) and 2) above.

In this paper, we propose a game theoretic framework, which is very powerful, to address the above issues. In particular, by drawing upon the Nash bargaining framework from cooperative game theory [24], [25], we show that one can obtain a unified framework in which we can address issues of network efficiency, fairness, revenue maximization, and pricing. The advantage of such a framework is that we have precise mathematical characterization of the solutions and their properties, and therefore a precise framework in which different solutions can be compared.

The idea of using the Nash bargaining solution (NBS) in the context of telecommunication networks is not new. This was first presented in the context of packet-switched (data) networks by Mazumdar *et al.* [22]. The properties of Pareto optimality as well as the development of local optimization procedures which lead to Pareto-optimal solutions (the local procedures being greedy schemes) were studied in a series of papers by Douligieris and Mazumdar [10], [8], [9] in the context of data networks. This paper is thus an extension of those ideas as well as a new approach in the context of elastic services in broadband networks. Preliminary results have been presented in [29] and [30].

The issue of rate control for elastic sources has been the focus of much attention. In the ATM ABR context the primary concern has been to develop algorithms which adapt quickly to congestion while trying to be fair in a so-called max – min sense [5], [13], [16]. This notion of fairness is different from the notion of the min – max solutions in game theory. More recently, [17], [18] and [21] have considered the problem of rate allocation and charging based on knowledge of user utility functions. All consider the issue of maximizing the social benefit, which is the sum of the user utilities. In [17] it is also shown

Manuscript received January 9, 1998; revised November 26, 1998 and May 8, 2000; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor S. H. Low. This work was supported by a contract from the Centre National d'Etudes des Télécommunications (CNET), France Telecom, through the Consultations Thématiques program.

H. Yaïche is with the Department of Electrical Engineering and Computer Science, Ecole Polytechnique de Montréal, Montréal H3C 3A7, Canada (e-mail: yaïche@comm.polymtl.ca).

R. R. Mazumdar and C. Rosenberg are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907-1285 USA (e-mail: mazum@ecn.purdue.edu; cath@ecn.purdue.edu).

Publisher Item Identifier S 1063-6692(00)09116-0.

that the socially optimizing solution can be obtained as the solution to a user optimization problem. Furthermore, it is shown that the solution obtained has the property of *proportional fairness* if the utility functions are logarithmic functions of the allocated bandwidth. Allocating bandwidth based on user willingness-to-pay is considered. Both [18] and [21] provide distributed algorithms for achieving the socially optimal rate allocations. The pricing issues the authors consider are different; in [17] users state their prices and the network allocates the bandwidth accordingly, while in [21] the network charges a price based on user bandwidth demands. The combination of flow control and pricing has also been addressed in [6] and [26].

The utility function approach used in [17] and [21] suffers from the point of view that user utilities or preferences are only known in some qualitative sense. Thus, although reasonable assumptions can be made on the behavior of utility functions, such an approach cannot be used to provide concrete numerical answers. Hence, the approach we take is to consider measurable performance characteristics rather than abstract utility functions. In the context of elastic services, one important measure is allocated rate. We propose a game theoretic framework based on choosing this measure. We demonstrate that not only is it possible to address the issues of fairness and efficiency, but the framework also allows us to put the solution in proper context.

Using the Nash bargaining framework from cooperative game theory [24], [25] we show that *proportional fairness* (as introduced in [17]) is in fact an NBS. The bargaining framework allows us to address the bandwidth allocation problem with nonzero minimum bandwidth guarantees [known as minimum cell rate (MCR) in the ABR context] while also accounting for peak-rate requirements of sources [referred to as peak cell rate (PCR) in the ABR context]. We then provide a distributed algorithm implemented at network links (or nodes), which achieves the desired bandwidth allocations that are Pareto optimal and fair. This algorithm is based on the gradient of the dual of the basic optimization problem which results when computing the NBS [2]. The algorithm proposed in [21] is also based on the dual of the social optimum problem with second-order differentiability or  $C^2$  assumptions on the user utility functions. The performance functions we consider are not in  $C^2$ , and hence we provide a proof of the convergence of our algorithm to the desired allocations.

We then address the issue of pricing and its relation to bandwidth allocation. It is shown that based on a user's budget or *willingness-to-pay* and its bandwidth demands, a bargaining framework can be developed to allocate the network bandwidths to the users in a way which is optimal in the Pareto sense and is fair to the users. Furthermore, based on this, we can develop a pricing scheme based on the congestion in the network for which network revenue is maximized when the network operates at the allocations corresponding to the bargaining solution. This pricing scheme has the following property: a user is never charged more than its declared budget but could be charged less than its budget if the amount of congestion in the network links used by its connection is low.

The outline of this paper is as follows: In Section I, we present the salient facts about the NBS which is the base for our frame-

work. Section II considers the optimal and fair rate allocation problem for elastic connections which have both minimum and peak rate constraints. We discuss both the centralized (system optimality) as well as the user-based contexts. In Section III, we propose a distributed algorithm to implement the solution and analyze its behavior in terms of convergence. In Section IV, we then show how the game theoretic framework we have introduced leads to a very elegant framework for charging and allocating bandwidth resources based on user budgets or *willingness-to-pay*. Technical proofs are deferred to the Appendix.

## II. BASIC FRAMEWORK

In this section, we present the salient concepts and results from cooperative game theory and the Nash bargaining (or arbitrated) solutions (NBS) which are used in the sequel. For details, we refer the reader to the book by Muthoo [24] and the paper by Nash [25].

The basic setting of the problem is as follows: There are  $N$  users (connections) which compete for the use of a fixed resource (bandwidth). Each user  $i$  ( $i \in \{1 \dots N\}$ ) has a performance function  $f_i$  and a desired initial performance  $u_i^0$  which is the minimal performance required by the user without any cooperation in order to enter the game. Each performance function is defined on a subset of  $\mathcal{R}^N$  termed  $X$ , which is the set of game strategies of the  $N$  users. In a context of network resource allocation,  $X$  could represent the space of allocated rate vectors. The initial performance of each user represents a minimum guarantee that the network must provide the user. Therefore, we will assume throughout our framework that each user involved in the game can achieve its initial performance. In other words, there exists at least a vector in  $X$  for which the performance vector  $f = (f_1, \dots, f_N)$  is superior or equal to the initial performance vector  $u^0$ .

Let  $U \subset \mathcal{R}^N$  be a nonempty convex closed and upper-bounded set. In our context, the set  $U$  denotes the set of achievable performance. Let  $u^0 \in \mathcal{R}^N$  such that  $U_0 = \{u \in U / u \geq u^0\} \neq \emptyset$ . Here  $u^0$  denotes the initial agreement point. Let  $\mathcal{G} = \{(U, u^0) / U \subset \mathcal{R}^N\}$  denote the set of achievable performance with respect to the initial agreement point.

We first define the notion of Pareto optimality in the context of multiple-criteria objectives which occurs in the typical game setting with multiple players.

*Definition 2.1:* The point  $u \in U$  is said to be Pareto optimal if for each  $v \in U$ ,  $v \geq u$ , then  $v = u$ .

The interpretation of a Pareto optimum is that it is impossible to find another point which leads to strictly superior performance for all the players simultaneously. In general, in a game with  $N$  players (or equivalently for a set of  $N$  objectives), the Pareto-optimal points form an  $N - 1$  dimensional hypersurface, which implies that there are an infinite number of points which are Pareto optimal. From the definition of Pareto optimality, it is clear that an optimal network operating point should be a Pareto-optimal point. The question that arises is at which of the (infinitely many) Pareto-optimal points should we operate the system?

One way in which we can define suitable Pareto-optimal points for operation is by introducing further criteria. From the

perspective of resource sharing, one of the natural criteria is the notion of fairness. This, in general, is a loose term and there are many notions of fairness. One of the commonly used notions is that of max – min fairness which penalizes large users. From the definition of max – min fairness [3], it can be seen that it corresponds to a Pareto optimum. However, it is not easy to take into account the notions that users might have different requirements within this framework. A much more satisfactory approach is to use the fairness axioms from game theory as the fairness criteria [25].

We now define the NBS, which encapsulates the above requirements of yielding Pareto optima as well as being fair in a precise sense. Except in trivial cases, it differs from the max – min solution.

*Definition 2.2:* A mapping  $S: \mathcal{G} \rightarrow \mathcal{R}^N$  is said to be an NBS if:

- 1)  $S(U, u^0) \in U_0$ .
- 2)  $S(U, u^0)$  is Pareto optimal.
- 3)  $S$  satisfies the linearity axiom if  $\phi: \mathcal{R}^N \rightarrow \mathcal{R}^N$ ,  $\phi(u) = u'$  with  $u'_j = a_j u_j + b_j$ ,  $a_j > 0$ ,  $j = 1, \dots, N$  then  $S(\phi(U), \phi(u^0)) = \phi(S(U, u^0))$ .
- 4)  $S$  satisfies the irrelevant alternatives axiom if  $V \subset U$ ,  $(V, u^0) \in \mathcal{G}$ , and  $S(U, u^0) \in V$  then  $S(U, u^0) = S(V, u^0)$ .
- 5)  $S$  satisfies the symmetry axiom if  $U$  is symmetric with respect to a subset  $J \subseteq \{1, \dots, N\}$  of indices (i.e.,  $u \in U$  and  $i, j \in J$ , then if  $u_i^0 = u_j^0$  then  $S(U, u^0)_i = S(U, u^0)_j$  for  $i, j \in J$ ).

*Remark 2.1:* The items 3, 4, and 5 above are the so-called axioms of fairness. The linearity property of the solution implies that the bargaining solution is scale invariant, i.e., the bargaining solution is unchanged if the performance objectives are affinely (i.e., of the form  $au + b$ ) scaled. The irrelevant-alternatives axiom states that the bargaining point is not affected by enlarging the domain if agreement can be found on a restricted domain. The symmetry property states that the bargaining point does not depend on the specific labels, i.e., users with the same initial points and objectives will realize the same performance.

Having defined the NBS, we define the optimal point as follows:

*Definition 2.3:* Let  $u^*$  be given by  $S(U, u^0)$ . Then  $u^*$  is the (Nash) bargaining point and  $f^{-1}(u^*)$  is called the set of the (Nash) bargaining solutions.

The following result, due to Stefanescu [27], provides for a characterization of the Nash bargaining point and will form the basis for the results in the sequel.

*Theorem 2.1:* Let  $f_i(\cdot): X \rightarrow \mathcal{R}$ ,  $i = 1, 2, \dots, N$  be concave upper-bounded functions defined on  $X$  which is a convex and compact subset of  $\mathcal{R}^N$ . Let  $f(x) = (f_1(x), \dots, f_N(x))$ .

Let  $U = \{u \in \mathcal{R}^N: \exists x \in X \text{ s.t. } f(x) \geq u\}$ . Denote by  $X(u) = \{x \in X: f(x) \geq u\}$  and  $X_0 = X(u^0)$  the subset of strategies that enable the users to achieve at least their initial performances.

Then there exists a bargaining solution and a unique bargaining point  $u^*$ . Moreover the set of the bargaining solutions ( $f^{-1}(u^*)$ ) is determined as follows:

Let  $J$  be the set of users able to achieve a performance strictly superior to their initial performance, i.e.,  $J$  is defined as  $\{j \in \{1 \dots N\}: \exists x \in X_0, f_j(x) > u_j^0\}$ . Each vector  $x$  in the bargaining solution set verifies  $f_J(x) > u_J^0$  and solves the following maximization problem ( $P_J$ ):

$$(P_J) \quad \text{Max} \prod_{j \in J} (f_j(x) - u_j^0) \quad x \in X_0.$$

Hence,  $u^*$  satisfies that  $u_j^* > u_j^0$  for  $j \in J$  and  $u_j^* = u_j^0$  otherwise.

*Remark 2.2:* From the assumption that there exists a nonempty set  $J$  of users who can achieve performance superior to their initial performance, it implies that  $P_J > 0$ . Note that for each  $j \in \bar{J}$ ,  $\forall x \in X_0$   $f_j(x) = u_j^0$ . The users in  $\bar{J}$  are not considered in the optimization above.

It can be readily shown that if each function  $f_j$  ( $j \in J$ ) is injective on  $X_0$ , then the bargaining solution set is a singleton and therefore there exists a unique NBS (in the space  $X$ ).

We now state an equivalent optimization problem, which will also result in the NBS. The proof can be found in the Appendix. We first need the following result whose proof is given in the Appendix.

*Lemma 2.1:* Let  $g: X \rightarrow \mathcal{R}_+^*$  be concave. Then  $h = \ln(g(\cdot)): \mathcal{R}_+ \rightarrow \mathcal{R}$  is concave. If  $g$  is injective, then  $h$  is strictly concave.

Using the above, we can now formulate an equivalent optimization problem, which we will consider in the sequel.

*Theorem 2.2:* In addition to the assumptions in Theorem 2.1, let  $\{f_j\}; j \in J$  be injective on  $X_0$ .

Consider the two maximization problems ( $P_J$ ) and ( $P'_J$ ):

$$(P_J) \quad \text{Max} \prod_{j \in J} (f_j(x) - u_j^0) \quad x \in X_0$$

$$(P'_J) \quad \text{Max} \sum_{j \in J} \ln(f_j(x) - u_j^0) \quad x \in X_0.$$

Then:

- 1) ( $P_J$ ) has a unique solution; the bargaining solution set is a singleton.
- 2) ( $P'_J$ ) is a convex program and has a unique solution.
- 3) ( $P_J$ ) and ( $P'_J$ ) are *equivalent*. Hence, the unique solution of ( $P'_J$ ) is the bargaining solution.

*Remark 2.3:* In [17], it has been shown that if the user utility functions are logarithmic, then the maximization of the sum of the utility functions leads to an allocation which has been termed as *proportionally fair* by Kelly. In light of Theorem 2.2, this corresponds to a NBS. However, the definition of a NBS does not require the user objectives to be logarithmic functions. In general, all that can be said in the case when the sums of user utilities are maximized as considered in [17], [21] is that the allocation will be a Pareto optimum. This optimum is referred to as a social optimum.

*Remark 2.4:* Since the NBS is Pareto optimal, it implies that there exists a set of weights  $\{w_i\}_{i \in J}$  such that  $u^* = (f_1(x^*), \dots, f_N(x^*))$  where

$$u^* = \arg \max_{x \in X_0} \sum_{i \in J} w_i f_i(x). \quad (1)$$

This follows from the fact that every Pareto point can be obtained as the solution to the maximization of the sum of the weighted objectives (see [1]).

### III. OPTIMAL AND FAIR BANDWIDTH ALLOCATION FOR ELASTIC CONNECTIONS

It is natural to adopt a game theory approach to model and address the issue of network resource allocation. In the context of flow control in packet-switched networks, many schemes were based on the use of game theory and gave a characterization for some candidate points. Some of them considered Nash equilibrium points [4], [10] and others considered Pareto-optimal points [8], [9]. In [22], the Nash bargaining point was proposed as a suitable solution for the design of an optimal and fair flow control.

As in [22], we consider the Nash bargaining point as the desired point for the operation of the network. This is due to the Pareto optimality and fairness property associated with NBSs. It is important to note that NBSs are not related to Nash equilibria which (except in the case of inessential games) are Pareto inefficient [11], [1]. Nash equilibria are important in that they arise in the context of greedy optimization.

The definition of a Nash bargaining point is highly dependent on the consideration of an initial performance point (termed  $u^0$  in the previous section). It represents a minimum performance that a user wants to achieve and the user will not enter the game if it is not possible. In the context of elastic services, for each connection (user) the initial performance can be viewed as a performance achieved by the minimum rate (MR) they want guaranteed by the network.

First, we consider a centralized (or global) model in which network resources are the available link capacities and each connection aims at maximizing its allocated rate beyond its minimum desired rate. Given that there are many users who all share the same objective, the network performs an allocation which is fair to all the users while at the same time efficient from the point of view of the network. As argued above, this corresponds to finding the NBS for the allocation problem.

Then, we show that the NBS from the point of view of the network can be achieved by solving a user-level greedy optimization problem by suitable modification of the user objectives. The required modification comes in the form of implied costs associated with the global problem and these in turn play a role in network revenue maximization.

#### A. Network Optimal Rate Allocations

We consider a static model for the centralized (network) problem in which  $N$  connections demand use of the network and are identified by the routes (or paths) they take. We assume there are  $L$  links or nodes within the network. Each connection is assumed to be elastic with a peak rate (PR) and an MR to be guaranteed by the network. Connections compete for available bandwidth resources within the network. These resources are network link available capacities and they are assumed to be fixed (nontime-varying). With respect to the abstract framework already presented, the admissible rate vector space  $X_0$  is

determined by network capacity constraints and the minimum and peak rates of the connections. It is defined as follows:

$$X_0 = \{x \in \mathcal{R}^N / x \geq \text{MR } x \leq \text{PR and } Ax \leq C\} \quad (2)$$

where  $C$  is the vector of link capacities, PR is the vector of peak rates of the connections, and  $A = (a_{lp})_{l,p}$  is an  $L \times N$  incidence matrix, i.e.,  $a_{lp}$  is equal to 1 if the link  $l$  belongs to the path  $p$  and 0 otherwise.

In the context of elastic services, it is natural to assume that each connection aims to obtain an allocated bandwidth greater than its minimum rate and as close to its peak bandwidth requirement as possible. Therefore, with respect to the framework described above, the performance function  $f_i$  for a user  $i$  is simply defined as  $x_i$ . Moreover,  $\text{MR}_i$  represents the initial (or minimum) performance desired by user  $i$ .

For simplicity and without loss of generality, we assume that on each link the spare capacity is strictly superior to the sum of the  $\text{MR}_i$ s of the connections crossing this link. If this assumption is not valid, then our model and results are still valid for the subset of connections to which we can allocate more than the corresponding minimum rate. One can show that this assumption ensures that  $X_0$  has a nonempty interior.

With respect to the framework described in Section I, the NBS of the centralized model is an optimal and fair rate allocation of network available capacities to the  $N$  connections. From Theorem 1.1, the NBS is the solution of the following convex global optimization problem ( $S$ ):

$$\begin{cases} ax_{\{x\}} \prod_{i=1}^N (x_i - \text{MR}_i) \\ x_i \geq \text{MR}_i, & i \in \{1 \dots N\} \\ x_i \leq \text{PR}_i, & i \in \{1 \dots N\} \\ (Ax)_l \leq (C)_l, & l \in \{1 \dots L\}. \end{cases}$$

*Proposition 3.1:* Under the hypothesis that  $\sum_{i=1}^N a_{li} \text{MR}_i < C_l$ ;  $l = 1, 2, \dots, L$ , there is a unique NBS for the centralized problem ( $S$ ) which is characterized as follows:

There exist  $\mu_l \geq 0$  ( $l \in \{1 \dots L\}$ ) and  $\beta_i \geq 0$  ( $i \in \{1 \dots N\}$ ) such that:

- for each  $i \in \{1 \dots N\}$

$$x_i = \text{MR}_i + \min \left\{ (\text{PR}_i - \text{MR}_i); \frac{1}{\sum_{l=1}^L \mu_l a_{li}} \right\}; \quad (3)$$

- $x_i \leq \text{PR}_i$   $i \in \{1 \dots N\}$ ;
- $Ax \leq C$ ;
- $(Ax - C)_l \mu_l = 0$   $l \in \{1 \dots L\}$ .

*Proof:* Now under the assumption that  $\sum_{i=1}^N a_{li} \text{MR}_i < C_l$ ;  $l = 1, 2, \dots, L$ , the set  $X_0$  is nonempty, convex, and compact.

Define

$$f(x) = \prod_{i=1}^N (x_i - \text{MR}_i)$$

then  $f(\cdot): X_0 \rightarrow \mathfrak{R}^+$  is strictly concave.

Noting that the constraints are linear in  $\{x_i\}$  and  $f(x)$  is  $C^1$ , it implies that the first-order Kuhn–Tucker [23] conditions are necessary and sufficient for optimality.

Let  $\mathcal{L}(x, \lambda, \beta, \mu)$  denote the Lagrangian where  $\lambda_i \geq 0$ ;  $i = 1, 2, \dots, N$ ,  $\beta_i \geq 0$ ;  $i = 1, 2, \dots, N$ , and  $\mu_l \geq 0$ ;  $l = 1, 2, \dots, L$  denote the Lagrange multipliers associated with the MR, PR, and capacity constraints respectively.

Then

$$\begin{aligned} \mathcal{L}(x, \lambda, \beta, \mu) = & f(x) - \sum_{i=1}^N \lambda_i (\text{MR}_i - x_i) \\ & - \sum_{i=1}^N \beta_i (x_i - \text{PR}_i) - \sum_{l=1}^L \mu_l ((Ax)_l - C_l). \end{aligned}$$

Then the first-order necessary and sufficient conditions are given by

$$1 + \left( \lambda_i - \beta_i - \sum_{l=1}^L \mu_l a_{li} \right) (x_i - \text{MR}_i) = 0;$$

$$i = 1, 2, \dots, N$$

and

$$\begin{aligned} (x_i - \text{MR}_i)\lambda_i &= 0; & \lambda_i &\geq 0; & i &= 1, \dots, N \\ (x_i - \text{PR}_i)\beta_i &= 0; & \beta_i &\geq 0; & i &= 1, 2, \dots, N \\ ((Ax)_l - C_l)\mu_l &= 0; & \mu_l &\geq 0; & l &= 1, 2, \dots, L. \end{aligned}$$

Under the assumption  $\sum_{i=1}^N a_{li} \text{MR}_i < C_l$ , we see that the constraints  $x_i \geq \text{MCR}_i$  are nonactive and hence  $\lambda_i = 0$  for all  $i = 1, 2, \dots, N$ . Furthermore,  $\beta_i = 0$  if  $x_i < \text{PR}_i$  and  $x_i = \text{PR}_i$  otherwise.

Hence, the result follows as stated.  $\square$

*Remark 3.1:* The Lagrange multiplier  $\mu_l$  has the interpretation as the implied cost associated with the network link  $l$ . It represents the marginal cost of a rate unit allocated for any connection crossing link  $l$ .

Having obtained the characterization of the optimal (in the Nash bargaining sense) rates allocated in the centralized or network framework we now address the issue of how we can define a local optimization problem (for each connection or user) which yields the above allocations.

### B. The User Problem

In the previous section, we formulated and solved the centralized network optimal rate allocation problem. In general, this will involve centralized coordination amongst the connections. In a network distributed over a vast geographical area, this will require much communication overheads. Thus, an important issue is whether such a computation can be decentralized at a user level in which the user tries to optimize its performance greedily. In general, greedy procedures lead to Nash equilibria [28] which, being Pareto inefficient, are not NBSs. Thus, clearly, users must use modified criteria if the greedy optimization is to lead to the NBS for the network.

The answer to the above question is in the affirmative. This is well known in the theory of nonlinear programming as the concept of tolls or penalties. This is also the approach used by

Kelly [17]. The basic idea is that if we think of the implied costs as the penalties to be paid by the users, then local optimization of the net user “goodput,” i.e., the desired performance minus the penalty to be paid, will yield a Pareto-optimal point. This will be the optimal of the weighted sum of the original objective functions, the weights being the penalties. Such an idea has also been discussed in the context of packet-switched networks in the thesis of Douligieris [7], where the decentralized procedure attempts to arrive at the centralized or Pareto-optimal flow control settings via the imposition of penalties.

In the decentralized model, each connection can optimize only its allocated rate. The rate for the connection is bounded from below by the MR and from above by the PR. It is assumed that each user optimizes its rate without regard to the other users (i.e., local optimization over  $x_i$  for user  $i$ ). However, offering unrestricted access to each user or connection is not in the network’s interest and thus the network penalizes or charges each user for use of network resources. This is reflected in a penalty in the user optimization criteria.

We introduce  $N$  positive network parameters, denoted by  $\{\alpha_i$ ;  $i = 1, 2, \dots, N\}$ , which represent the penalty or cost incurred per unit of bandwidth or capacity by the  $N$  users, given that they share the resources. The  $\alpha_i$ s also can be interpreted as the penalty per bandwidth unit that the network imposes on user  $i$  for consuming bandwidth within the network. We show how the  $\alpha_i$ s are determined such that the corresponding rate allocations lead to the centralized NBS rate allocations.

The objective of each user is to maximize its net utility which is, for a particular rate, the difference between the utility obtained from the allocated bandwidth  $x_i$  and the cost of accessing the network given by  $\alpha_i x_i$ .

Hence, let  $(U_i)$  denote the following convex problem associated with user  $i$ :

$$\begin{cases} \max_{\{x_i\}} \ln(x_i - \text{MR}_i) - \alpha_i x_i \\ x_i > \text{MR}_i \\ x_i \leq \text{PR}_i. \end{cases}$$

The network aims to determine the optimal rate allocation to users that maximizes its total “revenue” based upon “charging”  $\alpha_i$  per unit of bandwidth to user  $i$ . Hence, the network has to solve the following convex problem  $(N)$ :

$$\begin{cases} \max_{\{x\}} \sum_{i=1}^N \alpha_i x_i \\ x_i \geq \text{MR}_i, & i \in \{1 \dots N\} \\ x_i \leq \text{PR}_i, & i \in \{1 \dots N\} \\ (Ax)_l \leq (C)_l, & l \in \{1 \dots L\}. \end{cases}$$

The following proposition shows that by appropriate choice of network costs, the  $\alpha_i$ s, the NBS of the centralized model maximizes each user’s net utility and the network total revenue.

*Proposition 3.2:* Let  $\mathbf{x} = (x_1, x_2, \dots, x_N)$  be the unique NBS of the centralized problem  $(S)$ . Let  $\alpha_i = \sum_l a_{il} \mu_l$ , where  $\mu_l$  denotes the implied cost associated with link  $l$ ;  $l = 1, 2, \dots, L$  obtained from the solution to  $(S)$ .

Then  $x_i$  is the solution to the user optimization problem

$$U_i(x_i): \max_{x_i \in X_i} \{\log(x_i - \text{MR}_i) - \alpha_i x_i\} \quad (4)$$

and  $\mathbf{x}$  is the solution to

$$\max_{\mathbf{x} \in X_0} \sum_{i=1}^N \alpha_i x_i \quad (5)$$

where

$$X_i = \{x_i: \text{MR}_i \leq x_i \leq \text{PR}_i\} \quad (6)$$

and  $X_0$  is the admissible bandwidth space as defined earlier.

*Proof:* From proposition 2.1, the solution to (5) is given by

$$x_i = \text{MR}_i + \frac{1}{\alpha_i + \beta_i}$$

where

$$\begin{aligned} \alpha_i &= \sum_{l=1}^L a_{i,l} \mu_l \\ Ax &\leq C \\ (Ax - C)_l \mu_l &= 0 \quad l \in \{1 \dots L\} \\ (x_i - \text{PR}_i) \beta_i &= 0 \quad i \in \{1 \dots N\}. \end{aligned}$$

Note by definition  $\alpha_i$  is just the implied cost for the  $i$ th connection which uses the route specified by the links  $l$  along its path, where the  $\mu_l$  are the implied costs associated with link  $l$ .

In order to show that for each  $i \in \{1 \dots N\}$ ,  $x_i$  is the unique solution of the strictly convex problem ( $U_i$ ) we introduce the Lagrangian for the optimization problem  $U_i(x_i)$ .

Let

$$\begin{aligned} \mathcal{L}(x_i, \gamma_i, \delta_i) &= \log(x_i - \text{MR}_i) - \alpha_i x_i - \gamma_i (\text{MR}_i - x_i) \\ &\quad - \delta_i (x_i - \text{PR}_i) \end{aligned}$$

denote the Lagrangian.

Then the necessary and sufficient Kuhn–Tucker conditions of ( $U_i$ ) are given by

$$\begin{aligned} -\frac{1}{x_i - \text{MR}_i} + \alpha_i + \delta_i - \gamma_i &= 0 \\ (x_i - \text{MR}_i) \gamma_i &= 0 \\ x_i &\leq \text{PR}_i \\ (x_i - \text{PR}_i) \delta_i &= 0. \end{aligned}$$

The above holds for every  $i = 1, 2, \dots, N$ . Once again noting that  $x_i > \text{MR}_i$ , we obtain  $\gamma_i = 0$ .

In a similar way, by considering the Lagrangian for the problem ( $N$ ) we obtain that the necessary and sufficient Kuhn–Tucker conditions are given by

$$\begin{aligned} \text{for each } i \in \{1 \dots N\} \quad -\alpha_i + \sum_{l=1}^L \kappa_l a_{l,i} &= 0 \\ x_i &> \text{MR}_i \quad i \in \{1 \dots N\} \\ x_i &\leq \text{PR}_i \quad i \in \{1 \dots N\} \\ Ax &\leq C \\ (Ax - C)_l \kappa_l &= 0 \quad l \in \{1 \dots L\}. \end{aligned}$$

From above, by taking  $\delta_i = \beta_i$  and  $\kappa_l = \mu_l$ , we see that the solution  $\mathbf{x}$  in terms of  $\alpha_i$  satisfies the Kuhn–Tucker nec-

essary conditions for ( $U_i$ ) and ( $N$ ). Sufficiency follows from uniqueness.  $\square$

*Remark 3.2:* The optimization problem (5) is just an instance of the characterization of the Pareto points given in Remark 2.4 with weights  $\{\alpha_i\}$ .

To summarize the results so far, we have shown how the notion of fairness can be used to obtain the optimal (in the Pareto sense) network rate allocations and then shown how they could be realized using a local procedure in which the implied costs associated with the path taken by the connection play an important role.

One problem in implementing the decentralized optimization problem is that we need knowledge of the link implied costs  $\mu_l$ , which are only obtained from the solution to the global network optimization problem. It can be argued that if that is the case there is no benefit in considering the decentralized optimization problem, since if we solve the global problem, then we can directly obtain the optimal and fair bandwidth allocations.

In the following section we show that the bandwidth allocation problem can indeed be implemented in a distributed manner, without solving the global problem directly.

#### IV. DISTRIBUTED ALGORITHM FOR FAIR BANDWIDTH ALLOCATION

As mentioned above, the algorithm could easily be implemented as a local procedure [optimization of  $U_i(\cdot)$ ] once we know the user implied costs  $\alpha_i$ . This is, however, only determined by solving the global problem, and hence there is no gain in using the local interpretation unless we can devise a local way of obtaining this solution.

In this section, we present a distributed algorithm for obtaining the optimal fair bandwidth allocations based on local algorithms and measurements. The approach we use is drawn from the so-called gradient projection methods [2] in optimization theory. The algorithms proposed in [21] are also based on the gradient projection method. However, in [21] they assume that the utility functions are concave with bounded second-order derivatives. In our context, such an assumption does not hold, and hence we provide a proof of the convergence.

We present an algorithm based on the gradient projection of a dual problem associated with the original problem ( $S$ ).<sup>1</sup> The advantage of this is that in any realistic network the number of links  $L$  is usually much smaller than  $N$  the number of users. The dual problem is based on the  $L$  local algorithms run at the different nodes which are ingress nodes for a given link. The important feature of such an algorithm is that the link updates only require information on users who use that link, and hence global information is not required. The complexity of such a procedure is much less than if it is done at the user level in terms of communication overheads. This is because at a given link only the information about connections using that link is needed, while at the user level, information about all other connections which affect the given connection is needed.

<sup>1</sup>In [30], we also propose an algorithm which is based on the primal problem as in [18], but which requires much more information exchange than the one we present here.

We now discuss the framework and algorithm below. The proofs of the technical results are given in the Appendix.

We consider the convex problem (P) (primal problem) equivalent to (S) since they have the same optimal solution (cf. Theorem 2.2):

$$\begin{cases} \min_{\{x\}} GL(x) = -\sum_{i=1}^N \ln(x_i - MR_i) \\ x_i > MR_i, & i \in \{1, \dots, N\} \\ x_i \leq PR_i, & i \in \{1, \dots, N\} \\ (Ax)_l \leq (C)_l, & l \in \{1, \dots, L\}. \end{cases}$$

Let  $X$  be a subset of  $\mathcal{R}^N$  defined by connection bandwidth constraints and let  $\mathcal{L}$  be the Lagrangian associated with (P) and defined over  $X \times \mathcal{R}^L$ .  $X$  and  $\mathcal{L}$  are defined as follows:

$$\begin{aligned} \mathcal{L}(x, \mu) \\ = -\sum_{i=1}^N \ln(x_i - MR_i) + \sum_{i=1}^N \left( \sum_{l=1}^L \mu_l a_{li} \right) x_i - \sum_{l=1}^L C_l \mu_l. \end{aligned}$$

The dual function  $d(\cdot): \mathcal{R}^L$  to  $\mathcal{R}$  corresponding to (P) is then defined as follows:

$$d(\mu) = \text{Min}_{x \in X} \mathcal{L}(x, \mu). \quad (7)$$

Since the primal is separable and has a unique solution,  $d$  can be computed explicitly. Indeed, for each  $\mu \in \mathcal{R}^L$

$$\begin{aligned} d(\mu) = \sum_{i=1}^N \left[ -\ln \left( g_i \left( \sum_{l=1, a_{li}=1}^L \mu_l \right) - MR_i \right) \right. \\ \left. + \left( \sum_{l=1, a_{li}=1}^L \mu_l \right) g_i \left( \sum_{l=1, a_{li}=1}^L \mu_l \right) \right] \\ - \sum_{l=1}^L C_l \mu_l \end{aligned} \quad (8)$$

where for each  $i \in \{1, \dots, N\}$ ,  $g_i(\cdot)$  is defined on  $\mathcal{R}$  as follows:

$$g_i(p) = \begin{cases} PR_i, & \text{if } p \leq \frac{1}{PR_i - MR_i} \\ MR_i + \frac{1}{p}, & \text{if } p \geq \frac{1}{PR_i - MR_i}. \end{cases} \quad (9)$$

The dual problem (D) is the following:

$$\text{Max}_{\mu \in \mathcal{R}_+^L} d(\mu).$$

Since  $X$  is convex,  $GL(\cdot)$  is convex over  $X$ , and there exists  $x \in X$  such that  $(Ax)_l < C_l$  for each  $l \in \{1, \dots, L\}$ . It implies that there exists a Lagrange multiplier and therefore there is no duality gap (see [2, Ch. 5]). Hence, (D) has at least one optimal solution.

Let  $\bar{U}$  be the set of solutions of the dual problem. This set is also the set of Lagrange multipliers.  $\bar{U}$  is nonempty and can be characterized in many ways ([2, Ch. 5]). The saddle point characterization allows us to show that  $\bar{U}$  is compact (see the Appendix). From duality,  $d$  is concave on  $\mathcal{R}^L$ . One can show

readily (Danskin's theorem in [2]) that  $d$  is also  $C^1$  and the partial derivatives are determined as follows:

$$\frac{\partial d}{\partial \mu_l}(\mu) = \sum_{i=1, a_{li}=1}^N g_i \left( \sum_{l=1, a_{li}=1}^L \mu_l \right) - C_l \quad l \in \{1, \dots, L\}.$$

### A. Dual-Based Algorithm

We propose an algorithm that solves the dual problem (D) and which is based on a simple gradient-projection method. The algorithm uses a constant step-size. We will show that by a suitable choice of the step-size the algorithm converges to  $\bar{U}$ , the set of the solutions of the dual problem. Moreover, since the solution to the primal problem is unique, the corresponding primal solutions converge to the unique Nash bargaining vector.

Let  $\gamma > 0$  denote the step-size (or gain) associated with the following recursion scheme of dimension  $L$ .

For each  $l \in \{1, \dots, L\}$  and  $k \geq 0$

$$\mu_l^{(k+1)} = \text{Max} \left( 0, \mu_l^{(k)} + \gamma \left( \sum_{i=1, a_{li}=1}^N \mathbf{x}_i(\mu^{(k)}) - C_l \right) \right) \quad (10)$$

where for  $g_i(\cdot)$  defined in (9)

$$x_i(\mu^{(k)}) = g_i \left( \sum_{l=1}^L a_{li} \mu_l^{(k)} \right) \quad (11)$$

denotes the allocated bandwidth at iteration  $k$  for user  $i$ , and  $\mu^{(k)} = (u_1^{(k)}, \dots, u_L^{(k)})$  denotes the implied-cost vector at iteration  $k$  with  $\mu^{(0)} \in \mathcal{R}_+^L$  arbitrary. It can be taken to be 0.

Let  $N(i)$  denote the number of links crossed by user  $i$  through the network and define

$$K = \sqrt{L} \left( \sum_{i=1}^N (PR_i - MR_i)^2 N(i) \right). \quad (12)$$

Then  $K$  defines bounds on how large we can choose the gain  $\gamma$ .

We now state the main result on the convergence of the algorithm defined above.

**Proposition 4.1:** Let  $\{\mu^{(k)}\}$  be a sequence generated by (10) such that  $\mu^{(0)} \in \mathcal{R}_+^L$  and  $\gamma \in (0, (2/K))$ .

Let  $\bar{x}$  be the Nash bargaining allocation vector [solution of (S)].

$$\lim_{k \rightarrow \infty} x(\mu^{(k)}) = \bar{x}. \quad (13)$$

The proof is given in the Appendix.

### B. Network Implementation

In the following, we present an asynchronous distributed implementation of the algorithm. The iterations can be run at each network node using local information ( $\mu_l$  for link  $l$ ) and information sent by relevant users which use link  $l$ . A user updates its local variable ( $x_i$  for user  $i$ ) using information received from the links that this user crosses. After each update, a user sends the new value to these links.

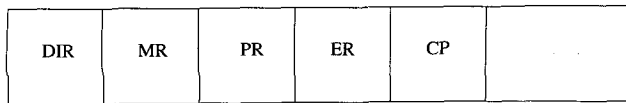


Fig. 1. RM packet structure.

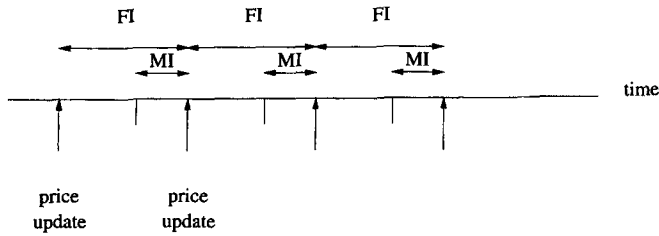


Fig. 2. Link updating and measurement process.

We propose a scheme using explicit-rate type of notification modeled after ABR schemes.

We assume that elastic sources regularly send forward resource management (RM) packets in order to get feedback from the network about the congestion state or resource availability.

The information necessary for the operation of the control scheme is conveyed by the RM packets, which are of two kinds: forward RM packets which are created by sources and conveyed along their corresponding paths, and backward ones which are created by destinations that turn around the forward RM packets. The fields of an RM packet (Fig. 1) relevant to the description of the control scheme are DIR (direction: forward or backward), MR (connection minimum rate), PR (connection peak rate), CP (congestion price), and ER (explicit rate). CP is used by the network nodes to communicate the value of the price variables ( $\mu_l$  for link  $l$ ) they control. ER stands for the maximal rate at which a given connection can transmit data.

There is a set of parameters associated with the control scheme: a constant step-size  $\gamma$  used to update the price variables, feedback intervals (FI in Fig. 2), and some measurement intervals (MI in Fig. 2). Each network link has its own feedback interval and measurement interval. A link price is updated at the beginning of each feedback interval and the total link input rate is measured during the measurement interval, as shown in Fig. 2.

If we interpret  $x_i(\mu^k)$  as the current data rate of connection and  $i$  as a function of the current network link price vector, then in (10) the sum  $\sum_{i=1, a_{li}=1}^N x_i(\mu^k)$  can be interpreted as the current total input rate at link  $l$ . It is important to note that the new price for a link  $l$  is computed when the information about current total input rate (the above sum) is available at the link. This helps determine the right values for the feedback and measurement intervals associated with network links.

In the following, we describe the local procedures associated with the allocation scheme.

#### Source Procedure:

- A source sends a forward RM packet and inserts the MR and the PR in the corresponding fields. Then, it sends the packet to the destination.
- At the reception of a backward RM packet, a source adjusts its transmission data rate according to the explicit rate notification (ER) contained in the RM packet. We consider

that a source has a variable called allowed rate (AR) which is updated as follows:  $AR \leftarrow ER$ . AR is the maximal rate at which a source is allowed to transmit.

#### Destination Procedure:

- Upon the reception of a forward RM packet, a destination creates a backward RM packet, puts zero in the CP field, and sends it back to its corresponding source.

#### Network Node Procedure:

 For a particular output link:

- At the beginning of each feedback interval (Fig. 2), the node updates the link price using the input rate measured during the previous measurement interval, a constant step-size  $\gamma$ , and the link available capacity  $C$ . The following illustrates the price updating:

$$\text{price} := \text{Max}(0, \text{price} + \gamma(\text{Input} - C))$$

- Upon the reception of a backward RM packet, ER and CP are modified using the current link price, the MR, and the PR. The modifications are done as follows (ER is modified using the new value of CP):

$$\begin{aligned} \text{CP} &:= \text{price} + \text{CP} \\ \text{ER} &:= \begin{cases} \text{PR}, & \text{if } \text{CP} \leq \frac{1}{\text{PR} - \text{MR}} \\ \text{MR} + \frac{1}{\text{CP}}, & \text{if } \text{CP} > \frac{1}{\text{PR} - \text{MR}} \end{cases} \end{aligned}$$

Once the modifications are completed, the backward RM packet is relayed back to the source.

- A node, at regular intervals, measures (Fig. 2) the total input rate at the link.

It can be readily see that the ER contained in a backward RM packet does not increase when going through network nodes in the backward direction. In addition, the implementation of the scheme does not differentiate between network access nodes and the other nodes as far as the update of ER is concerned.

For the good operation of the control scheme, it is important to dimension for each link the feedback and measurement intervals. Indeed, the feedback interval should be large enough to allow the sources traversing a particular link to react to the new price (after update) conveyed by the backward RM packets and for a link to experience the result of the sources' reaction. The total input rate at a link should be measured during that period, i.e., when the response of sources to the new price has reached the link.

The rate of convergence is governed by  $\gamma$  which depends on the knowledge of  $K$  defined earlier. This is the only quantity which needs to be broadcast to all nodes.

## V. PRICING FRAMEWORK FOR ELASTIC SERVICES

We now address the issue of rate allocation together with the pricing issue in the context of elastic-rate connections considering users' bandwidth requirements and users' budgets (willingness-to-pay) for bandwidth above their guaranteed minimum cell rates. As already shown in Proposition 2.2, if the network charges according to the user-implied costs, the network revenue is maximized when allocated rates are according to the NBS. This key property will allow us to formulate a pricing framework for the network to charge the users.

The scenario we consider is the following.

Each user informs the network of its budget (or maximum amount the user is willing to pay for the required bandwidth beyond the guaranteed rate) in a simple way. At connection setup, a user communicates its budget for bandwidth allocated beyond its MR. Users expect that the network will take their budgets into consideration when allocating the available bandwidth to all competing users. The budget may be declared by the user or chosen from a given set of values provided by the network operator. For example, the network can provide different tariffs with the proviso that the user will not pay more than the selected amount which determines the type of service the network may guarantee. The choice of a user reflects the valuation based upon the budget for an amount of bandwidth beyond guaranteed rate. What the network operator undertakes is to provide a fair share of the bandwidth based on the user budgets and bandwidth requirements.

From the network viewpoint, it must set up a charging scheme such that the user budget is never exceeded, and yet the network maximizes its revenue when it allocates the bandwidth according to the requirements above.

#### A. Model and Assumptions

We consider the model of  $N$  users similar to the one described in Section III-A. Each user (connection)  $i$  has a minimum rate  $MR_i$  and a peak rate  $PR_i$ . Each user  $i$  chooses a parameter  $B_i$  which represents the total cost (budget) it is willing to pay for the excess bandwidth beyond the minimum cell rate. As mentioned above, these could be from tariffs published by the network (for example Gold, Silver, or Bronze services, which can guarantee different levels of bandwidth such as providing PR or some amount between MR and the PR depending on the network conditions). One desirable property is that depending on network conditions, a user must not be penalized for choosing a larger budget if resources are not congested.

As in the centralized model (Section III-A), we adopt the following simplifying assumption (without loss of generality): on each link, the spare capacity is assumed to be strictly superior to the sum of the  $MR_i$ s of the connections crossing this link.

Bargaining theory also provides us with the necessary framework to address fair and efficient bandwidth allocation subject to both user bandwidth requirements as well as budgets. The basic framework is one of asymmetric NBSs [24]. The idea is that given that users have different budgets, they desire a corresponding proportional share of the bandwidth. At the same time, from the point of efficiency, it is desired to operate the network at allocations which correspond to a Pareto-optimal point. Asymmetric NBSs are solutions which satisfy all the assumptions of the usual NBS except the property of symmetry. This is because their allocated bandwidth must reflect their different preferences in terms of their budgets.

Based on this we can now state the main result.

*Proposition 5.1:* Let  $B_i \geq 0$  denote the budget (or willingness to pay) of user  $i$ . Then the optimal and fair asymmetric NBS is given by the unique solution to

$$\max_{\mathbf{x} \in X_0} \prod_{i=1}^N (x_i - MR_i)^{B_i} \quad (14)$$

and, in particular, the solution is given by the following:

- If  $B_i = 0$  then  $x_i = MR_i$ .
- If  $B_i > 0$  and  $\sum_{l=1}^L \mu_l a_{li} > 0$  then

$$x_i = MR_i + \min \left( PR_i - MR_i, \frac{B_i}{\sum_{l=1}^L \mu_l a_{li}} \right) \quad (15)$$

where for each  $l \in \{1 \dots L\}$ ,  $(Ax - C)_l \mu_l = 0$ .

- If  $\sum_{l=1}^L \mu_l a_{li} = 0$ , then  $x_i = PR_i$ .

*Proof:* The proof readily follows, as in Proposition 2.1, by noting that the solution to  $\prod_{i=1}^N (x_i - MR_i)^{B_i}$  is the same as that of  $\sum_{i=1}^N B_i \ln(x_i - MR_i)$ .

The rest of the details can be worked out as in Proposition 2.1 and are hence omitted.  $\square$

Based on the above solution, there is a natural pricing structure that we propose. Let  $p_i(x_i)$  denote the price charged to user  $i$  ( $i = 1, 2, \dots, N$ ). Let

$$p_i(x_i) = T(B_i) + (x_i - MR_i) \sum_{l=1}^L a_{li} \mu_l \quad (16)$$

where  $T(\cdot)$  is some tariff function based on the willingness-to-pay  $B_i$ . It could be viewed as a fixed price for access with an  $MR_i$  guarantee (as, for example, in Gold, Silver, or Bronze services, or other differentiated services [15]).

The main property of such a price structure is that it contains two components, the first being a fixed tariff associated with the minimum guarantee (and budget) and the second being a congestion-based price (which can be viewed as an ‘‘elastic price’’) on the actual bandwidth allocation costs. Exactly as in Proposition 2.2, if the network charges user  $i$  according to  $p_i(x_i)$ , then the network revenue is maximized. This is stated in the following proposition.

*Proposition 5.2:* Let  $x$  be the solution to the asymmetric Nash bargaining problem. Then  $x$  solves

$$\begin{cases} \max_{\{x\}} \sum_{i=1}^N p_i(x_i) \\ x_i \geq MR_i, & i \in \{1 \dots N\} \\ x_i \leq PR_i, & i \in \{1 \dots N\} \\ (Ax)_l \leq (C)_l, & l \in \{1 \dots L\}. \end{cases}$$

We conclude by discussing some properties of the above solution.

- If a user  $i$  has no budget for the share of bandwidth beyond  $MR_i$ , then the allocated rate is  $MR_i$ .
- If a user  $i$  has a budget for the share of bandwidth beyond the minimum rate, then the allocated rate is greater than the minimum rate.
- If the network resources along a user’s path are free ( $\alpha_i = \sum_{l=1}^L a_{li} \mu_l = 0$ , i.e., the links used are not congested), then the allocated rate is the peak rate.
- If the network resources along a user’s path are not free and the user’s budget exceeds the network path cost per unit of bandwidth by more than a factor of  $(PR - MR)$ , then the user is allocated its peak rate.
- If the network resources along a user’s path are not free and the user’s budget is less than the path cost per band-

width unit, then the user is allocated a rate between the minimum and peak rate proportional to the budget of the user. As a result, if two users share the same resources and one of them is willing to pay double of the other, then that user receives double the share of bandwidth beyond minimum rate.

- If two users share the same resources, have the same maximum excess bandwidth (difference between peak and minimum rates), and are willing to pay the same price, then they get the same share of excess bandwidth.
- In all scenarios, the user is never charged more than  $B_i$ , which is its budget or willingness-to-pay. The user will be charged less than  $B_i$  if the users budget is very high in comparison to the path costs and so will not be overly penalized.
- Finally, the pricing structure we propose provides the necessary regulation to prevent users from inflating their budgets.

## VI. CONCLUSION

In this paper we have presented a game theoretic framework for the allocation of optimal rates to elastic connections which share common bandwidth. This framework puts into focus the recent work of Kelly [17], [21] and allows us to go further in showing how we can come up with a charging scheme and a joint allocation and pricing policy which is efficient and which presents nice fair properties. We have also provided a distributed asynchronous algorithm to implement the solution using an RM-based scheme found in the ATM ABR context.

Future work will address the issues of the algorithmic implementation in the context when randomness is introduced due to measurements, as well as the fact that real situations involve nonstatic scenarios.

## APPENDIX

In this section, we present the relevant proofs of the results in Section I and the proofs on the convergence of the distributed algorithm presented in Section III.

We first state the following well-known result on the relation between arithmetic and geometric means which is needed to prove Lemma 1.1.

*Lemma 6.1:* Let  $q_i$  ( $i \in \{1 \dots N\}$ ) be strictly positive real numbers satisfying  $\sum_{i=1}^N q_i = 1$ . Let  $a_i$  ( $i \in \{1 \dots N\}$ ) be positive real numbers. Then, (unless the  $a_i$  are all equal)

$$\prod_{i=1}^N a_i^{q_i} < \sum_{i=1}^N q_i a_i.$$

*Proof of Lemma 1.1:*

- 1) Let  $x$  and  $y \in X$ , and  $\lambda \in [0, 1]$ . Since  $g$  is concave and  $\ln$  is an increasing function, the following inequalities hold:

$$\begin{aligned} h(\lambda x + (1 - \lambda)y) &= \ln(g(\lambda x + (1 - \lambda)y)) \\ g(\lambda x + (1 - \lambda)y) &\geq \lambda g(x) + (1 - \lambda)g(y) \\ h(\lambda x + (1 - \lambda)y) &\geq \ln(\lambda g(x) + (1 - \lambda)g(y)). \end{aligned}$$

Let

$$A = \ln(\lambda g(x) + (1 - \lambda)g(y)) - (\lambda \ln(g(x)) + (1 - \lambda) \ln(g(y))).$$

It is easy to see that

$$A = \ln \left( \frac{\lambda g(x) + (1 - \lambda)g(y)}{g(x)^\lambda g(y)^{1-\lambda}} \right).$$

Using the result of Lemma 6.1 above, it readily follows that

$$g(x)^\lambda g(y)^{1-\lambda} \leq \lambda g(x) + (1 - \lambda)g(y).$$

Hence,  $A$  is positive and  $h$  is concave.

- 2) Let  $x$  and  $y \in X$  such that  $x \neq y$ . Let  $\lambda \in [0, 1]$  and  $\lambda \notin \{0, 1\}$ . Since  $g$  is injective,  $g(x) \neq g(y)$ . Using the same arguments as above, invoking Lemma 5.1

$$g(x)^\lambda g(y)^{1-\lambda} < \lambda g(x) + (1 - \lambda)g(y).$$

Hence,  $A$  is strictly positive and  $h$  is strictly concave.  $\square$

Based on Lemma 1.1, we can readily prove Theorem 1.2.

*Proof of Theorem 1.2:*

- 1) This follows from the fact that  $f_j$  is injective for each  $j \in J$ .
- 2) One can show that  $X_0$  is nonempty, convex, and closed. From Lemma 1.1, it follows that the objective function is strictly concave. The objective function represents a lower barrier for the  $x \in X_0$  such that  $f(x) \not\geq u^0$ . It is upper bounded in  $X_0$  excluding the elements satisfying  $f(x) \not\geq u^0$ .
- 3) It can easily be shown that a solution of  $(P_J)$  is a solution of  $(P'_J)$ , and conversely, by writing down the Kuhn–Tucker conditions.  $\square$

We now prove the convergence of the algorithm presented in Section III. We first show that the set  $\bar{U}$  of solutions to the dual problem is compact.

*Proposition 6.1:*  $\bar{U}$ , the set of dual solutions, is a compact set of  $\mathcal{R}_+^L$ .

*Proof:* Let  $\bar{x}$  be the unique solution of problem  $(P)$ . Then,  $\bar{U}$  is the set of all  $\bar{\mu} \in \mathcal{R}_+^L$  such that  $(\bar{x}, \bar{\mu})$  is a saddle point of  $\mathcal{L}$  [2, ch. 5].

Let  $\bar{\mu}$  be an element of  $\bar{U}$ . Then,  $\forall x \in X$ ,  $\mathcal{L}(\bar{x}, \bar{\mu}) \leq \mathcal{L}(x, \bar{\mu})$ . We know that there exists  $y \in X$  such that for each  $i \in \{1, \dots, N\}$ ,  $MR_i < y_i < PR_i$  and for each  $l \in \{1, \dots, L\}$ ,  $(Ay - C)_l < 0$ . One can show that  $y \neq \bar{x}$  and therefore  $GL(y) > GL(\bar{x})$  and  $\mathcal{L}(\bar{x}, \bar{\mu}) \leq \mathcal{L}(y, \bar{\mu})$ . Let  $m(y) = \min\{-(Ay - C)_1, \dots, -(Ay - C)_L\}$ . Then

$$\sum_{l=1}^L \bar{\mu}_l \leq \frac{GL(y) - GL(\bar{x})}{m(y)}.$$

As a result,  $\forall \bar{\mu} \in \bar{U}$ ,  $\|\bar{\mu}\|_1 \leq ((GL(y) - GL(\bar{x}))/m(y))$ . So,  $\bar{U}$  is a bounded set. It is closed because it is equal to  $d^{-1}(GL(\bar{x}))$ . Hence,  $\bar{U}$  is a compact of  $\mathcal{R}_+^L$ .  $\square$

We now evaluate the gradient of  $d(\cdot)$ .

*Proposition 6.2:*  $d(\mu)$  is continuously differentiable on  $\mathcal{R}^L$  and for each  $l \in \{1, \dots, L\}$

$$\frac{\partial d}{\partial \mu_l}(\mu) = \sum_{i=1, a_i=1}^N g_i \left( \sum_{l=1, a_i=1}^L \mu_l \right) - C_l, \quad \mu \in \mathcal{R}^L$$

*Proof:*  $\mathcal{L}(x, \cdot)$  is concave and differentiable for each  $x \in X$ .  $\mathcal{L}(\cdot, \cdot)$  is continuous. For each  $\mu \in \mathcal{R}^L$ ,  $\mathbf{x}(\mu)$  is the unique vector minimizing  $\mathcal{L}(\cdot, \mu)$  over  $X$ . Hence, the conclusion of the proposition follows (Danskin's theorem in [2]).  $\square$

The following lemma shows that the gradient of the dual function  $d$  is Lipschitz.

**Lemma 6.2:** The gradient of  $d(\cdot)$ , denoted by  $\nabla d(\cdot)$ , is Lipschitz on  $\mathcal{R}^L$ . Let  $K$  denote the Lipschitz constant. For each  $i \in \{1, \dots, N\}$ , let  $N(i)$  be the number of links utilized by user  $i$ . Then

$$K = \sqrt{L} \left( \sum_{i=1}^N (\text{PR}_i - \text{MR}_i)^2 N(i) \right).$$

*Proof:* It can readily be seen from the definition of  $g_i(p)$  for each  $i \in \{1, \dots, N\}$ , the Lipschitz constant of  $g_i(p)$  is  $(\text{PR}_i - \text{MR}_i)^2$ .

Let  $\mu$  and  $\mu'$  be two elements of  $\mathcal{R}^L$ , and let  $\|\mu\|_1$  denote the  $L_1$  norm of  $\mu \in \mathcal{R}^L$ .

For each  $l \in \{1, \dots, L\}$  define

$$f_l(\mu) = \sum_{i=1}^N a_{li} x_i(\mu).$$

Then

$$|f_l(\mu) - f_l(\mu')| \leq \sum_{i=1, a_{li}=1}^N |x_i(\mu) - x_i(\mu')|.$$

Noting that  $x_i(\mu) = g_i(\sum_{l=1}^L a_{li} \mu_l)$ , for each  $l \in \{1, \dots, L\}$

$$|f_l(\mu) - f_l(\mu')| \leq \sum_{i=1, a_{li}=1}^N (\text{PR}_i - \text{MR}_i)^2 \|\mu - \mu'\|_1.$$

Hence

$$\|f(\mu) - f(\mu')\|_1 \leq \sum_{l=1}^L \sum_{i=1, a_{li}=1}^N (\text{PR}_i - \text{MR}_i)^2 \|\mu - \mu'\|_1.$$

Let  $N(i)$  be the number of links crossed by user  $i$ . Since  $\forall \mu \in \mathcal{R}^L$ ,  $\|\mu\| \leq \|\mu\|_1 \leq \sqrt{L} \|\mu\|$ , the result follows. Note  $\|\cdot\|$  denotes the Euclidean norm.  $\square$

The next result shows that the sequence  $\mu^{(k)}$  converges to a point in  $\bar{U}$ .

**Proposition 6.3:** Let  $\{\mu^{(k)}\}$  be a sequence generated by (10) such that  $\mu^{(0)} \in \mathcal{R}_+^L$  and  $\gamma \in ]0, (2/K_2)[$ . Then

$$\mu^{(k)} \rightarrow \bar{U} \quad \text{as } k \rightarrow \infty.$$

*Proof:* Let  $\gamma \in (0, (2/K_2)]$ . Since  $d(\mu)$  is  $C^1$  over the closed and convex set  $\mathcal{R}_+^L$  and the gradient of  $d$  is  $K$ -Lipschitz, via [2, prop. 2.3.2, Ch. 2], every limit point of  $\{\mu^{(k)}\}$  is an element of  $\bar{U}$ . We now show that the sequence  $\{\mu^{(k)}\}$  is bounded. Since,  $d$  is concave on  $\mathcal{R}^L$ ,  $\mathcal{R}_+^L$  is a convex and closed set, and  $\bar{U}$  nonempty and bounded, the following level set is compact:  $\{\mu \in \mathcal{R}_+^L | -d(\mu) \leq -d(\mu^{(0)})\}$  (see [2, Appendix B, prop. B.9]). Using the descent Lemma and the projection characterization, one can show that for each  $k \geq 0$

$$-d(\mu^{(k+1)}) \leq -d(\mu^{(k)}).$$

Hence,  $\{\mu^{(k)}\}$  is bounded. This implies that there exists a convergent subsequence. Since the set of its limit points is included in  $\bar{U}$ , the result of the proposition follows.  $\square$

**Remark 6.1:** Note the above result only states that the set of all limit points of  $\mu^{(k)}$  are dual optimal. In general, there is no unique limit. Also, a stepsize in the interval  $(0, (2/K))$  guarantees an increase of the dual function  $d$  at each iteration.

Using the above result, the continuity of  $x(\mu^{(k)})$  and the uniqueness of the solution  $\bar{x}$  to the primal, the main result readily follows.

## REFERENCES

- [1] J.-P. Aubin, *Optima and Equilibria, GTM 140*. New York, NY: Springer-Verlag, 1998.
- [2] D. P. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1995.
- [3] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
- [4] K. Bharathkumar and J. M. Jaffe, "A new approach to performance oriented flow control," *IEEE Trans. Commun.*, vol. COM-29, pp. 427–435, April 1981.
- [5] F. Bonomi and K. W. Fendick, "The rate-based flow control framework for available bit-rate ATM services," *IEEE Network*, pp. 25–39, March/April 1995.
- [6] C. Courcoubetis, V. A. Siris, and G. D. Stamoulis, "Integration of pricing and flow control for available bit-rate services in ATM networks," in *Proc. IEEE Globecom'96*, London, U.K., pp. 644–648.
- [7] C. Douligeris, "Optimal flow control and fairness in communication networks: A game theoretic perspective," Ph.D. dissertation, Columbia Univ., New York, 1989.
- [8] C. Douligeris and R. R. Mazumdar, "On Pareto-optimal flow control in an integrated environment," in *Proc. 25th Allerton Conf.*, Urbana, IL, Oct. 1986.
- [9] —, "More on Pareto-optimal flow control," in *Proc. 26th Allerton Conf.*, Urbana, IL, Oct. 1987.
- [10] —, "User optimal flow control in an integrated environment," in *Proc. Indo-U.S. Workshop Syst. Signals*, Bangalore, India, Jan. 1988.
- [11] P. Dubey, "Inefficiency of Nash equilibria," *Math. Oper. Res.*, vol. 11, no. 1, pp. 1–8, 1986.
- [12] R. R. Mazumdar and B. T. Doshi, Eds., *Eur. Trans. Telecommun.—Focus on Elastic Services Over ATM Networks*, 1997, vol. 8, pp. 5–63.
- [13] E. J. Hernandez-Valencia, L. Benmohammed, S. Chong, and R. Nagarajan, "Rate-control algorithms for ATM ABR service," *Eur. Trans. Telecommun.—Focus on Elastic Services Over ATM Networks*, vol. 8, no. 1, pp. 7–20, 1997.
- [14] "Traffic control and congestion control in B-ISDN," ITU-T, Geneva, Recommendation I.371, June 1996.
- [15] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski. (1999, June) Assured forwarding PHB group. Internet RFC 2597. [Online] Available: <ftp://ftp.isi.edu/innotes/rfc2597.txt>.
- [16] R. Jain, "Congestion control and traffic management in ATM networks: Recent advances and a survey," *Comput. Networks ISDN Syst.*, vol. 28, pp. 1723–1730, 1996.
- [17] F. Kelly, "Charging and rate control for elastic traffic," *Eur. Trans. Telecommun.—Focus on Elastic Services Over ATM Networks*, vol. 8, no. 1, pp. 33–37, 1997.
- [18] F. Kelly, A. Maulloo, and D. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," Statistical Lab., Cambridge, U.K., preprint, 1997.
- [19] P. Key, "Fixed-point models and congestion pricing for TCP and related schemes," presented at the Workshop Mathematical Modeling of TCP, Paris, France, Dec. 7–8, 1998, [Online] Available: <http://www.dmi.ens.fr/mistral/tcpworkshop.html>.
- [20] H. Khalil, *Nonlinear Systems*. Englewood Cliffs, NJ: Prentice Hall, 1996.
- [21] S. H. Low and D. E. Lapsley, "Optimization flow control, I: Basic algorithm and convergence," *IEEE/ACM Trans. Networking*, vol. 7, pp. 861–874, Dec. 1999.
- [22] R. R. Mazumdar, L. Mason, and C. Douligeris, "Fairness in network optimal flow control: Optimality of product forms," *IEEE Trans. Commun.*, vol. 39, pp. 775–782, May 1991.
- [23] M. Minoux, *Mathematical Programming: Theory and Algorithms*. Chichester, U.K.: Wiley, 1986.

- [24] A. Muthoo, *Bargaining Theory with Applications*. Cambridge, U.K.: Cambridge Univ. Press, 1999.
- [25] J. Nash, "The bargaining problem," *Econometrica*, vol. 18, pp. 155–162, 1950.
- [26] S. Shenker, "Fundamental design issues for the future Internet," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1176–1188, Sept. 1995.
- [27] A. Stefanescu and M. W. Stefanescu, "The arbitrated solution for multi-objective convex programming," *Rev. Roum. Math. Pure Appl.*, vol. 29, pp. 593–598, 1984.
- [28] E. Van Damme, *Stability and Perfection of Nash Equilibria*. New York, NY: Springer Verlag, 1991.
- [29] H. Yaïche, R. R. Mazumdar, and C. Rosenberg, "A game theoretic framework for rate allocation and charging of available bit rate (ABR) connections in ATM networks," in *Broadband Communications'98*, P. Kuehn and R. Ulrich, Eds, pp. 222–233.
- [30] H. Yaïche, R. R. Mazumdar, and C. Rosenberg, "Distributed algorithms for fair bandwidth allocation to elastic services," in *Proc. INFOCOM*, Tel Aviv, Israel, Mar. 2000.

**Haïkel Yaïche** was born in Sfax, Tunisia. He received the Engineering degree in computer science from the Ecole Nationale Supérieure d'Electro-technique, d'Electronique, d'Informatique et d'Hydraulique de Toulouse, France. He is currently working toward the Ph.D. degree at the Ecole Polytechnique de Montréal, Canada.

His research interests include congestion control in broadband networks.



**Ravi R. Mazumdar** (SM'94) was born in Bangalore, India. He received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Bombay, India, in 1977, the M.Sc. DIC degree in control systems from Imperial College, London, U.K., in 1978, and the Ph.D. degree in systems science from the University of California, Los Angeles (UCLA), in 1983.

He is currently Professor of electrical and computer engineering at Purdue University, Lafayette, IN. Prior to joining Purdue, he was Professor of mathematics at the University of Essex, Colchester, U.K., from 1996 to 1999. From 1988 to 1996, he was Professor at INRS-Télécommunications, a graduate research institute of the Université du Québec, Canada, and an Invited Professor of electrical engineering at McGill University, Montreal, Canada. During 1985–1988, he was an Assistant Professor of electrical engineering at Columbia University, New York, NY. He has held visiting positions at UCLA, the University of Twente, The Netherlands, the Indian Institute of Science, Bangalore, India, and the Ecole Nationale Supérieure des Télécommunications, Paris, France. His research interests are in game theory, applied probability and stochastic analysis focusing on applications in telecommunication networks, statistical signal processing, and mathematical finance.

Dr. Mazumdar is a Fellow of the Royal Statistical Society.



**Catherine Rosenberg** (SM'95) received the Dipl.Ing. from the Ecole Nationale Supérieure des Télécommunications de Bretagne, Brest, France, in 1983, the M.S. degree in computer science from the University of California, Los Angeles (UCLA), in 1984, and the Ph.D. degree in computer science from Université de Paris, Orsay, France, in 1986.

She is currently an Associate Professor of electrical and computer engineering at Purdue University, Lafayette, IN. Prior to joining Purdue, she was the Head of the Department of Broadband Satellite

Networking at Nortel Networks, Harlow, U.K., and a Visiting Professor in the Department of Electrical Engineering at Imperial College, London, U.K., from 1996 to 1999. From 1988 to 1996, she was on the faculty of the Ecole Polytechnique de Montréal, Canada. She was a Member of Technical Staff at AT&T Bell Laboratories, Holmdel, NJ, from 1987 to 1988, and was with Alcatel, Lannion, France, from 1984 to 1987. She has held visiting appointments at the Université de Paris, Jussieu, France, and the Indian Institute of Science.

Dr. Rosenberg is an Associate Editor for IEEE Communications Magazine, IEEE Communications Surveys and Telecommunications Systems. Her research interests are in broadband networks, IP, ATM, broadband satellite networks (GEO or LEO based), traffic engineering, and wireless networks.