

# An Algorithm for Optimal Service Provisioning using Resource Pricing

Steven Low

AT&T Bell Labs  
Murray Hill, NJ 07974  
slow@research.att.com

Pravin Varaiya \*

University of California  
Berkeley, CA 94720  
varaiya@helios.berkeley.edu

## Abstract

We propose a microeconomic approach to service provisioning in ATM networks. Our provisioning procedure consists of two algorithms, one executed by the network and the other by individual users. The network offers directly for rent its bandwidth and buffers. Users purchase freely resources to meet their desired quality based on their own traffic parameters and delay requirements. The network periodically adjusts resource prices based on user requests monitored in the previous period. We describe network's adjustment scheme and users' decision rule, and show that under the proposed price adjustment scheme, in order to minimize their own service cost, individual users will indeed request amounts of resources that optimize a measure of overall network performance. Since our approach does not require the network to know user traffic and delay parameters, it does not mandate traffic policing on the part of the network.

## 1 Introduction

We propose a microeconomic approach to provisioning services in an ATM network. A service is specified by a one-way connection (source, destination, route) and two sets of service parameters. A connection is used to transport a data stream or message from the source to the destination; the traffic parameters specify constraint on a user's traffic 'burstiness'; the quality parameters specify maximum end-to-end delay. The network offers for sale different types of services, differentiated by the triplet (connection, traffic parameters, quality parameters). There is a flow of user requests for these services depending on the ser-

vice cost. A service request is admitted if sufficient resources can be allocated along the connection's route to guarantee service quality. By resources, we mean the bandwidth and buffers in each node along the route. Before transmission, a message is segmented into small, fixed size units called cells. The bandwidth and buffers allocated to a connection can vary over links in its route.

Two related questions motivate our approach: how to price network services and what resources to allocate to each connection along its route. We will ignore the 'fixed cost' of providing and operating the network infrastructure and focus on the so-called 'congestion cost' [8]. We propose a provisioning procedure in which individual users, rather than the network, decide what resources to reserve along their routes based on their own traffic and quality parameters. They pay a price for the resources they reserve and the network periodically adjusts these prices to coordinate user requests. We derive network's adjustment scheme and users' decision rule and show their optimality.

More specifically, a network offers a set  $S$  of services. A unit of type  $s$  service is provided by a type  $s$  connection with the associated traffic and quality parameters, and is sold for a unit price of  $w_s$ . The network can produce any amount of type  $s$  service provided the required resources do not exceed capacity. Users request services to maximize their individual utilities subject to budget constraints, and the result of this maximization is summarized by an aggregate demand function  $D_s(w_s)$ . We assume that demand function can be easily measured by counting the number of service requests, both admitted and rejected. Further assume that the network charges its services so as to maximize a welfare function:

$$\max_{w,x} W'(w,x) \quad (1)$$

$$\text{subject to } f(x) \leq 0 \quad (2)$$

\*Research supported by Pacific Bell, the MICRO program, and the NSF Grant IRI-9120074.

$$x_s \leq D_s(w_s), \quad s \in S \quad (3)$$

where  $W'(w, x)$  is the sum of network revenue and user surplus, (2) is capacity constraint, and (3) says that the amount of service sold is at most the demand for that service. Standard price-adjustment schemes [9, pp. 188–189] can be used to reach an equilibrium, in which the network sets a price  $w$  and supply  $x$ , users present their demands  $D_s(w_s)$ , and the network increases or decreases price according as demand is greater or less than supply, until  $(w, x)$  converges.

Observe that bandwidth and buffers are ‘substitutable resources’ to meet a service quality (see §2). In this paper, we exploit this tradeoff of different resource combinations to optimize an overall measure of network performance. For each allocation indexed by  $\mu$ , let  $W(\mu)$  denote the welfare achieved in (1) by an equilibrium price–supply vector  $(w(\mu), x(\mu))$ . We are interested in an iterative and decentralized algorithm that solves for a  $\mu^*$  that meets service quality and maximizes  $W(\mu)$ . The algorithm leads to our service provisioning procedure in which the network offers directly for rent its bandwidth and buffers, and the users purchase freely resources to meet their desired service quality. A user bases its decision on the knowledge of its *own* traffic and quality parameters, and on the resource price. The network periodically adjusts the prices based on the monitored user request for resources on the entire network. Collectively, users of type  $s$  service effectively do two optimizations in each period: one selects a resource combination along the route to minimize service cost  $w_s$ ; the other selects a demand  $D_s(w_s)$  for type  $s$  service that maximizes surplus. It is decentralized in that each user only needs to know the resource price at nodes along its route in addition to its own traffic and quality parameters. The solution makes critical use of the bandwidth–buffer tradeoff described by burstiness curve in [4, 5]: for the network, it determines the resource combination to maximize welfare; for individual users, it provides a simple rule for requesting resources (see (7)).

Our approach differs in three ways from the conventional service provisioning approach in which the network decides beforehand resources to be allocated to the users. First, under our approach, users freely rent resources and package them into services that best meet their needs. Second, since service price is the rent a user pays for the resources it reserves, our procedure ties this price to network performance measured by the welfare function. Third, since the network only guarantees the availability of purchased resources, it is the users’ responsibility to shape their traffic in order that the allocated resources can provide

the desired quality. Our approach relieves the network of the difficult task of traffic policing and enforcement and can potentially adapt to time-varying user needs expressed by the traffic and quality parameters.

Pricing has been used previously for control and optimization in communication networks, e.g. [1, 3]. As a consequence of the need to guarantee service quality in an ATM network, various bandwidth allocation schemes have recently been designed to guarantee certain amount of bandwidth to a connection despite changes in the number and burstiness of concurrent connections, e.g. [2]. This justifies our approximation of each network node as allocating a fixed bandwidth to a connection. We seek among the many possible resource combinations one that maximizes welfare by exploiting the bandwidth–buffer tradeoff.

The paper is organized as follows. In §2, we introduce our network model and service parameters. In §3, we describe our service provisioning procedure, i.e., the network’s price adjustment scheme and users’ decision rule, that does not require the network to know user traffic and quality parameters. In §4, we define optimality and show that our provisioning scheme is asymptotically optimal by formulating optimal allocation as a game problem. All proofs are omitted and can be found in [6].

## 2 Model

We consider a network with a set  $L$  of links. Link  $l \in L$  comprises a transmission capacity of  $C_l$  cells per second, or cps, and buffers for  $B_l$  cells. A set  $R$  of routes is specified;  $r \in R$  also denotes the set of links along that route. The network offers a set  $S$  of services. A unit of type  $s$  service is sold at a price of  $w_s$ <sup>1</sup> and is provided by a connection over route  $r_s$  for one unit of time. Our service provisioning scheme will relate this unit price  $w_s$  to the cost of resources needed to provide the service, as elaborated in §4. Under the scheme, the price is adjusted periodically to achieve an optimal allocation. We shall assume that user demand, or requests, for type  $s$  service is given by the aggregate demand function  $D_s(w_s) = \nu_s \exp(-w_s)$ ,  $\nu_s := \lambda_s T_s$ .  $\lambda_s \exp(-w_s)$  is the (average) rate type  $s$  requests arrive and  $T_s$  is the (average) duration of a type  $s$  connection. A type  $s$  request is admitted and assigned a connection with route  $r_s$  provided there is available spare bandwidth of  $\mu_{l_s}$  cps and spare buffers for  $b_{l_s}$  cells, in each link

<sup>1</sup>The price  $w_s$  may be some fictitious currency for control purposes that the network can adjust at will.

$l \in r_s$ . The routing  $s \rightarrow r_s$  is fixed, but the allocation

$$s \rightarrow \{\mu_{l_s}, b_{l_s}; l \in r_s\}$$

can be freely chosen provided the service quality constraint is met as explained next.

Once a type  $s$  request is admitted, a connection is set up, and the user sends a message. A message is a 'fluid flow',  $m(t), 0 \leq t \leq T$ , where  $m(t)$  is the instantaneous rate in cps, and  $T$  is the duration. A type  $s$  message must satisfy two constraints denoted by  $(b_s(\mu), \underline{\mu}_s)$ . The parameter  $\underline{\mu}_s$  is a positive real number that bounds the average message rate. It is also the minimum bandwidth required for a type  $s$  message at each node along its route. The parameter  $b_s(\mu), \mu \geq 0$ , is a non-negative, decreasing, convex function that bounds the message 'burstiness'. A type  $s$  message  $m$  is said to be *compliant* if

$$\eta := \frac{1}{T} \int_0^T m(\tau) d\tau \leq \underline{\mu}_s \quad (4)$$

and

$$\max_{0 \leq t \leq T} \int_s^t [m(\tau) - \mu] d\tau \leq b_s(\mu), \quad \mu \geq \eta \quad (5)$$

Inequality (4) says that the average message rate  $\eta$  cannot exceed  $\underline{\mu}_s$ . The left-hand side of (5) is the maximum backlog if  $m$  is transmitted over a link at a constant speed  $\mu \geq \eta$ . Hence, inequality (5) says that if  $m$  is allocated a bandwidth of  $\mu$ , then a buffer of size  $b_s(\mu)$  is sufficient to prevent cell loss. Note that the larger is  $\mu$  the smaller is  $b_s(\mu)$ . Thus the function  $b_s$ , called burstiness curve, gives the bandwidth-buffer tradeoff for zero cell loss. To incorporate cell loss, we may relax (5) and let  $b_s(\mu)$  be the buffer required to have no more than certain number of lost cells if  $m$  is transmitted over a link at a constant speed  $\mu$  [10].

A word is in order on how this traffic characterization may be used in practice. First, if the duration of a message is very long, e.g. a video program, we divide its duration into disjoint periods  $T_i, i = 1, \dots, k$ . A type  $s$  message is compliant if the portion of message on every period  $T_i$  satisfies (4-5) with  $T$  replaced by  $T_i$ . (4) then guarantees that no cell backlog carries over to the next period. Second, we do not assume that a user of type  $s$  service knows its own message  $m(t)$  or the burstiness of  $m(t)$ . We only assume that they know a bound  $b_s(\mu), \mu \geq \eta$ , on the burstiness. In fact, condition (5) can be easily enforced by passing an arbitrary user message through a leaky bucket policing device before being admitted into the network. The two parameters of a leaky bucket (and the bound on

peak message rate in each period) define a piecewise linear burstiness curve that bounds the burstiness of the output message from the leaky bucket [5, Proposition 3].

For the rest of this paper, we will use the following vector notation.  $\mu_s$  denotes the vector  $\{\mu_{l_s}, l \in r_s\}$  of bandwidth allocation for a type  $s$  connection, and  $\mu$  denotes the vector  $\{\mu_s, s \in S\}$ . Similarly,  $b_s(\mu_s)$  denotes the vector  $\{b_s(\mu_{l_s}), l \in r_s\}$  of buffers required for a type  $s$  connection, and  $b(\mu)$  denotes the vector  $\{b_s(\mu_s), s \in S\}$ .  $\underline{\mu}$  denotes the vector  $\{\underline{\mu}_s, s \in S\}$ . We may abuse notation and use ' $\mu \geq \underline{\mu}$ ' to mean ' $\{\mu_{l_s} \geq \underline{\mu}_s; l \in r_s, s \in S\}$ '. Finally,  $\langle x, y \rangle = \sum x_i y_i$  denotes the inner product of vectors  $x$  and  $y$ .

To specify the quality of service, the maximum end-to-end delay, we use two results from [4, 5]. Suppose a type  $s$  compliant message is transmitted over a connection with route  $r_s$ . Suppose that a bandwidth of  $\mu_{l_s}$  cps and buffers of  $b_{l_s}$  cells are allocated to that connection at each link  $l \in r_s$ . Suppose that the allocation  $\{\mu_{l_s}, b_{l_s}, l \in r_s\}$  satisfies

$$\mu_{l_s} \geq \underline{\mu}_s, \quad b_{l_s} \geq b_s(\mu_{l_s}), \quad l \in r_s \quad (6)$$

i.e. at each link, the allocated bandwidth exceeds the minimum bandwidth  $\underline{\mu}_s$ , and the allocated buffer exceeds the burstiness constraint. Then, (i) no cells will be lost at any link  $l \in r_s$  [4, Proposition 4], and (ii) the end-to-end delay is at most  $\frac{b_s(\underline{\mu}_s)}{\underline{\mu}_s}$  plus a constant 'propagation and processing delay' [4, Theorem 3]. Consequently, a maximum end-to-end delay translates into the minimum bandwidth  $\underline{\mu}_s$  required at each link along the route.

In summary, there are two service parameters for type  $s$  service: the burstiness curve  $b_s$ , and the minimum bandwidth  $\underline{\mu}_s$  required at each link along the route. A user message is compliant if it satisfies (4-5). An allocation  $\{\mu_{l_s}, b_{l_s}; l \in r_s, s \in S\}$  is compliant if it satisfies (6). Given service parameters,  $\{b_s, \underline{\mu}_s; s \in S\}$ , we want to find a compliant allocation  $(\mu, b) = \{\mu_{l_s}, b_{l_s}; l \in r_s, s \in S\}$  that is 'optimal'. We will restrict ourselves to allocations with  $b_{l_s} = b_s(\mu_{l_s})$ , since this is sufficient to prevent cell loss. We henceforth represent an allocation by a vector  $\mu = \{\mu_{l_s}, b_s(\mu_{l_s}); l \in r_s, s \in S\}$ .

### 3 Service provisioning procedure

In this section, we describe our service provisioning procedure. The specification consists of two algorithms, one executed by the network and the other executed by individual users.

**Network algorithm:**

1. Network initializes the update period  $n = 0$ .
2. It posts a rent  $(\alpha^n, \beta^n)$  in period  $n$  for resources at each link.
3. It monitors the vector of bandwidth and buffers  $(\mu^n, b^n)$  reserved in period  $n$  by users of all service types, and uses this *observed*  $(\mu^n, b^n)$  to minimize the following function over the set  $\{(\alpha, \beta) \geq 0\}$ :

$$\sum_s \nu_s \exp\left[-\frac{1}{T_s} (\langle \alpha, \mu_s^n \rangle - \langle \beta, b_s^n \rangle)\right] + \langle \alpha, C \rangle + \langle \beta, B \rangle$$

4. It uses any minimizer as rent  $(\alpha^{n+1}, \beta^{n+1})$  in the next period. It increments  $n$ , and goto step 2.

We make two remarks about the network algorithm. First, the resource price  $(\alpha_l^n, \beta_l^n)$  at link  $l$  depends only on  $l$  and is the same for all service types. Furthermore, its computation only uses information that can be monitored in the network; in particular, it does not need user traffic or quality parameter. Second, every user of the same type of service will make the same resource reservation  $(\mu_s^n, b_s^n)$  if they execute the same user algorithm described next.

**User algorithm:**

In period  $n = 0, 1, 2, \dots$ ,

1. A user of type  $s$  service requests bandwidth  $\mu_{l_s}^n$  and buffer  $b_{l_s}^n$  at link  $l$  along its route according to

$$\begin{aligned} \mu_{l_s}^n &= \underline{\mu}_s & \text{if } -\frac{\alpha_l^n}{\beta_l^n} < \frac{d}{d\mu} b_s(\underline{\mu}_s) \\ \mu_{l_s}^n &= M_s & \text{if } -\frac{\alpha_l^n}{\beta_l^n} > \frac{d}{d\mu} b_s(M_s) \\ -\frac{\alpha_l^n}{\beta_l^n} &= \frac{d}{d\mu} b_s(\mu_{l_s}^n) & \text{otherwise} \end{aligned} \quad (7)$$

and

$$b_{l_s}^n = b_s(\mu_{l_s}^n)$$

Here,  $M_s$  is the peak message rate at which  $b_s(M_s) = 0$ ,  $\underline{\mu}_s$  is the minimum bandwidth required for a type- $s$  connection to satisfy its delay requirement, and  $b_s$  is the (bound on the) burstiness curve of a conformant type  $s$  message.

2. It is admitted if resources are available, and rejected otherwise.

3. If admitted, a user of type  $s$  service pays the rent

$$\sum_{l \in r_s} (\alpha_l^n \mu_{l_s}^n + \beta_l^n b_{l_s}^n)$$

and occupies the reserved resources for a (average) duration of  $T_s$ .

The decentralized nature of the user algorithm is striking: given the price  $(\alpha^n, \beta^n)$ , the bandwidth and buffer request  $(\mu_{l_s}^n, b_{l_s}^n)$  at link  $l$  depends only on the price at link  $l$ , and on the user's own burstiness curve  $b_s$ . Since  $b_s(\mu)$  is strictly decreasing and convex for  $\mu < M[4, 5]$ , the optimal  $\mu_{l_s}^n$  is unique for each  $(\alpha^n, \beta^n)$ .

## 4 Optimality

In this section, we establish the optimality of the service provisioning procedure described in §3. We first formulate the problem of optimally allocating resources among competing services. We then show that the service provisioning procedure is an distributed implementation of a solution to the allocation problem.

The network can produce any amount  $x_s$  of type  $s$  service, provided that sufficient resources are available, i.e.

$$x_s \leq D_s(w), \quad s \in S \quad (8)$$

$$\sum_s x_s \frac{\mu_s}{T_s} \leq C_l, \quad \sum_s x_s \frac{b_s(\mu_s)}{T_s} \leq B_l, \quad l \in L \quad (9)$$

and expects a revenue of  $\sum x_s w_s$ . The aggregate demand function summarizes the users' utility such that  $\sum \int_{w_s}^{\infty} D_s(v) dv$  is the user surplus [9]. Take as social welfare

$$W'(w, x, \mu) := \sum \int_{w_s}^{\infty} D_s(v) dv + \sum x_s w_s$$

so the problem is to maximize  $W'(w, x, \mu)$  subject to (8-9). Consider initially a fixed allocation  $\mu \geq \underline{\mu}$ .

**Definition 1** A set of prices and amounts of service produced  $\{w_s(\mu), x_s(\mu); s \in S\}$  form an equilibrium if, for all  $\{x_s\}$  satisfying (8-9),

$$\begin{aligned} x_s(\mu) &= \nu_s \exp[-w_s(\mu)] \text{ and} \\ \sum x_s(\mu) w_s(\mu) &\geq \sum x_s w_s(\mu) \end{aligned}$$

The conditions say that, in equilibrium, user demand is met and the network maximizes revenue.

**Proposition 1**  $(w_s(\mu), x_s(\mu))$  is an equilibrium if and only if there exist  $(\alpha(\mu), \beta(\mu)) \geq 0$  such that

$$\begin{aligned} x_s(\mu) &= \nu_s \exp[-w_s(\mu)] & (10) \\ \sum x_s(\mu) \frac{\mu_{ls}}{T_s} &\leq C_l, \quad \sum x_s(\mu) \frac{b_s(\mu_{ls})}{T_s} \leq B_l & (11) \\ w_s(\mu) &= \frac{1}{T_s} (\langle \alpha(\mu), \mu_s \rangle + \langle \beta(\mu), b_s(\mu_s) \rangle) & (12) \\ \sum x_s(\mu) w_s(\mu) &= \langle \alpha(\mu), C \rangle + \langle \beta(\mu), B \rangle \end{aligned}$$

Note that (12) says that the equilibrium price equals the resource cost for providing that service. The cost is estimated by taking as the 'shadow' price or rent of  $\alpha_l(\mu)$  per cps of bandwidth and  $\beta_l(\mu)$  per cell of buffer, in link  $l$ .

It can be verified that there is a unique equilibrium, that the equilibrium maximizes  $W(w, x, \mu)$  over  $w \geq 0, x \geq 0$ , subject to (8-9), and that the maximum welfare is

$$W(\mu) = \sum x_s(\mu) + \langle \alpha(\mu), C \rangle + \langle \beta(\mu), B \rangle \quad (13)$$

Now suppose  $\mu \geq \underline{\mu}$  can be freely chosen. We can now formally define an optimal allocation.

**Definition 2** An allocation  $\mu$  is optimal, or welfare-maximizing, if it maximizes the welfare  $W(\mu)$  in (13).

To directly maximizing (13) the network needs to know user traffic and quality parameters  $(b_s, \underline{\mu}_s)$ . The solution proposed below does not require such knowledge, and hence does not require any traffic policing and enforcement on the part of the network, though users may still want to shape their messages to comply with  $(b_s, \underline{\mu}_s)$  so that the end-to-end delay is met. From (10-11) and the convexity of  $G(\mu, \alpha, \beta)$  in  $\alpha, \beta$ , we obtain an alternative expression for the maximum welfare,

$$W(\mu) = \min_{(\alpha, \beta) \geq 0} G(\mu, \alpha, \beta)$$

where

$$\begin{aligned} G(\mu, \alpha, \beta) &= \\ &\sum_s \nu_s \exp[-\frac{1}{T_s} (\langle \alpha, \mu_s \rangle - \langle \beta, b_s(\mu_s) \rangle)] \\ &+ \langle \alpha, C \rangle + \langle \beta, B \rangle \end{aligned}$$

Hence,  $\mu^*$  is a welfare-maximizing allocation if

$$W(\mu^*) = \max_{\mu \geq \underline{\mu}} W(\mu) = \max_{\mu \geq \underline{\mu}} \min_{(\alpha, \beta) \geq 0} G(\mu, \alpha, \beta) \quad (14)$$

The following result is key to our solution. Note that  $G$  is convex in  $(\alpha, \beta)$  but not generally concave in  $\mu$ .

**Proposition 2** There exists a saddle-point  $(\mu^*, \alpha^*, \beta^*)$  to the max-min problem (14) that is welfare-maximizing, i.e.

$$\begin{aligned} &G(\mu^*, \alpha^*, \beta^*) \\ &= \max_{\mu \geq \underline{\mu}} \min_{(\alpha, \beta) \geq 0} G(\mu, \alpha, \beta) \\ &= \min_{(\alpha, \beta) \geq 0} \max_{\mu \geq \underline{\mu}} G(\mu, \alpha, \beta) \end{aligned}$$

Note that the max-min problem in the proposition is equivalent to the following game played by  $U_s, s \in S$ , against  $N$ :

$$U_s : \min_{\mu_s \geq \underline{\mu}_s} \sum_{l \in r_s} (\alpha_l \mu_{ls} + \beta_l b_s(\mu_{ls})) \quad (15)$$

$$N : \min_{(\alpha, \beta) \geq 0} G(\mu, \alpha, \beta) \quad (16)$$

where we recall that  $\mu_s = \{\mu_{ls}, l \in r_s\}$ . The proposition says that if player  $N$  chooses the minimizer  $(\alpha^*, \beta^*)$ , then player  $U_s$  will choose the optimal  $\mu_s^*$  since  $(\mu^*, \alpha^*, \beta^*)$  is a saddle point. Note again that  $\alpha$  and  $\beta$  in (15) can be conveniently interpreted as the rent for one unit of bandwidth and buffer, respectively.

This interpretation suggests the following algorithm to reach  $(\mu^*, \alpha^*, \beta^*)$ . Suppose the network charges each user during connection setup a rent of  $\alpha_l$  per cps of bandwidth and  $\beta_l$  per cell of buffer in link  $l$ . The expected cost to a user per request of type  $s$  is then  $\sum_{l \in r_s} (\alpha_l \mu_{ls} + \beta_l b_s(\mu_{ls}))$ . Users may rent any amounts of bandwidth  $\mu_{ls} \geq \underline{\mu}_s$  and buffers  $b_s(\mu_{ls})$  at each link  $l \in r_s$ . We assume that, given prices  $(\alpha, \beta)$ , users will try to minimize their expected service cost subject to their own quality requirement  $\mu_s \geq \underline{\mu}_s$ . That is, they will take on the role of  $U_s$  in (15). Since the objective function (15) is separable in  $\mu_{ls}$ , (15) is equivalent to

$$\min_{\mu_{ls} \geq \underline{\mu}_s} \alpha_l \mu_{ls} + \beta_l b_s(\mu_{ls}), \quad l \in r_s$$

Observe then that decision (7) in user algorithm of §3 is simply the Kuhn-Tacker condition [7] which, by strict convexity of burstiness curve  $b_s$ , is sufficient for the minimization.

In our service provisioning procedure, the network takes on the role of player  $N$  in (16) to calculate the price  $(\alpha, \beta)$ . As noted above, if the network charges according to  $(\alpha^*, \beta^*)$ , then users will indeed request the optimal allocation  $\mu^*$ .

The question remains whether the provisioning procedure is optimal, i.e. whether the allocation and price vectors  $(\mu^n, \alpha^n, \beta^n)$ ,  $n = 0, 1, \dots$ , produced when the network plays against the users under the provisioning procedure approach optimal allocations. This is assured by the following theorem.

### 3c.2.5

**Theorem 3** For any initial price  $(\alpha^0, \beta^0)$ , any convergent subsequence of  $(\mu^n, \alpha^{n+1}, \beta^{n+1})$ ,  $n \geq 0$ , produced under the service provisioning procedure converges to a saddle point of  $G$  (which is optimal).

Hence, the service provisioning procedure eventually achieves a welfare-maximizing allocation. The optimal price  $(\alpha^*, \beta^*)$  determines a minimum-cost allocation  $\mu^*$ .  $(\mu^*, \alpha^*, \beta^*)$  yields the service price  $w^*$  and the amount  $x_s^* = D_s(w^*)$  of service produced that form an equilibrium.

## 5 Conclusion

We have proposed an alternative approach to service provisioning in an ATM network, which does not require the network to know user needs. In this approach, the network offers directly for rent its bandwidth and buffers and the users purchase them freely to meet their desired quality. The service provisioning procedure is based on a solution of the problem of allocating bandwidth and buffers to meet several types of service requests, differentiated by bounds on the average rate and burstiness of the message and on the end-to-end delay. While our argument is preliminary, it does suggest an alternative to the popular approach in which the network decides which services will best meet user needs. Here, the users decide the resources they need and the network coordinates their choices via resource pricing in order to optimize an overall measure of network performance.

## References

- [1] D. Ferguson, Y. Yemini, and C. Nikolaou. Microeconomic algorithms for load balancing in distributed computer systems. *Proceedings of the 8th International Conference on Distributed Computing Systems*, June 1988.
- [2] C. Kalmanek, H. Kanakia, and S. Keshav. Rate controlled servers for very high speed networks. *Proceedings of Globecom'90*, pages 12–20, December 1990.
- [3] J. F. Kurose and R. Simha. A microeconomic approach to optimal resource allocation in distributed computer systems. *IEEE Transactions on Computers*, 38(5), May 1989.
- [4] S. Low and P. Varaiya. A simple theory of traffic and resource allocation in ATM. *Proceedings of Globecom'91*, pages 1633–1637, December 1991.
- [5] S. Low and P. Varaiya. Burstiness bounds for some burst reducing servers. *Proceedings of Infocom'93*, pages 2–9, March 1993.
- [6] S. Low and P. Varaiya. A new approach to service provisioning in ATM networks. *IEEE/ACM Transactions on Networking*, 1(5):547–553, October 1993.
- [7] David G. Luenberger. *Linear and Nonlinear Programming, 2nd Ed.* Addison-Wesley Publishing Company, 1984.
- [8] Jeffrey K. MacKie-Mason and Hal R. Varian. Pricing the Internet. presented at Public Access to the Internet, JFK School of Government, Harvard University, 1993.
- [9] Hal R. Varian. *Microeconomic Analysis*. W. W. Norton & Company Inc., 1978.
- [10] Michael Wong and Pravin Varaiya. A deterministic fluid model for cell loss in ATM networks. *Proceedings of Infocom'93*, March 1993.