Pricing Bandwidth for ABR Rate Control and Multimedia Traffic*

Errin W. Fulp† and Douglas S. Reeves†

Abstract

Multimedia applications are expected to play a more prevalent role in integrated service networks. One method of efficiently transmitting such traffic uses the ABR service class. However, rate control for this class becomes more difficult due to the bursty and somewhat unpredictable behavior of multimedia traffic. This paper presents a microeconomic-based ABR rate control technique that models the network as competitive markets. Prices are affixed to ABR bandwidth based upon supply and demand, and users purchase bandwidth to maximize their individual QoS. This yields a *state-less* rate control method that provides a Pareto optimal and QoS-fair bandwidth distribution, high utilization (up to 95% in simulation results), and better performance than max-min or demand-based weighted max-min. Proofs that this method achieves weighted max-min fairness, Pareto optimality, QoS-fairness and price stability are provided, as well as simulation using MPEG-compressed video traces.

1 Introduction

ATM integrated-service networks are designed to accommodate a variety of network applications. These applications range from simple file transfers to complex multimedia programs that transmit voice and video. Multimedia applications are of special interest since they are expected to play a more prevalent role in the future. However, the transmission of such applications is not trivial due to their bursty and long range dependent behavior [8].

One approach to transmitting video over ATM networks uses the Available Bit Rate (ABR) service class [16, 24]. This service class is one in which network characteristics may change during the lifetime of a connection. For this reason, the ABR service class is suitable for traffic that can adapt to changing network conditions. The ABR service class is also appropriate for bursty traffic that cannot easily specify its source characteristics a priori. When transmitting video, the video encoding must adapt to network conditions [9]. As a result, dynamically adjusting bandwidth

*F49620-96-1-0061 and F49620-97-1-0351. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the AFOSR or the U.S. Government.

†Departments of Electrical and Computer Engineering, and Computer Science, N. C. State University, ewfulp|reeves@eos.ncsu.edu

requirements in response to network conditions can yield high bandwidth utilization. In [24], Roberts demonstrated how the ABR service class can be used to transmit MPEG-compressed video resulting in utilization over 95% as compared to 30-60% using the Variable Bit Rate (VBR) service class. Therefore, the ABR service class is a viable option for transmitting compressed video.

Due to the dynamic nature of the ABR service class, proper rate control is essential and becomes more difficult when video is transmitted. Explicit rate control relies on network feedback provided by Resource Management (RM) cells that are circulated for each connection [2]. A RM-cell traveling from the source to the destination will be referred to as moving *upstream*, while a RM-cell traveling from the destination to the source will be referred to as moving *downstream*. The RM-cell consists of several fields, one of which is the Expected Rate (ER). This field indicates the maximum rate the network can support for this user. As the RM-cell travels along the path, a switch and/or destination may alter its contents. Exactly how this is done depends on the strategy. Once the cell reaches the destination it is returned to the source, which must alter transmission based on the RM-cell information.

Methods that perform rate allocation can be generally classified on whether they maintain per-connection state information [12]. Methods that maintain state information that is directly used in the calculation of the allowable rate of a user will be referred to as *state-maintaining*. Alternatively, if per-connection state information is not required for the calculation of the allowable rate, it will be referred to as *state-less*. Of these two categories, a state-less method is preferred. Such a method does not require the overhead (storage and computational) of connection tables when computing allowable rates. Also, state-less implementations are scalable to larger networks since additional data structures are not required.

When a switch becomes congested, many of these ABR rate control strategies attempt to allocate bandwidth in a fair (max-min) manner [2]. The max-min fairness criterion states that any user is entitled to as much bandwidth as any other. When a link is bottlenecked, the bandwidth is divided equally among the users of the link. If a user requires less than this amount, the difference is divided equally among the remaining users. This process is repeated until all users of the link have been allocated a maximum amount of bandwidth. A more detailed description is provided

in [3]. However, max-min does not take into account the fact that some sources (for example, compressed video) may be able to reduce their transmission rate more easily than others. When congestion occurs, allocating equal amounts of bandwidth may not be best when considering the individual QoS expected by each user [16].

Another allocation criterion is weighted max-min. This extension of max-min incorporates the use of weights when determining allocation amounts. We now present some notation, followed by a definition of weighted max-min.

Given a set of users A of a single type of resource with supply S. Each user j has a weight w^j and desires a maximum allocation of b^j. Let a^j be the allocation for user j, where $0 \leq a^j \leq b^j$ for each $j \in A$. Denote $\{a^j\}$ as the array of allocations of all the users. User j is either "completely satiated" or "non-satiated" with their allocation, a^j. Let C be the set of completely satiated users, where $a^j = b^j$ for each $j \in C$. Let N be the set of non-satiated users, where $0 \leq a^j < b^j$ for each $j \in N$. Therefore, $C \cup N = A$ and $\sum_{j \in A} a^j \leq S$ must always be true.

Definition 1.1. Weighted max-min fair: An allocation of resources $\{a^j\}$ is weighted max-min fair if,

$$a^j = b^j \quad \text{for all} \quad j \in C \tag{1}$$

$$a^j = \frac{w^j}{\sum_{l \in N} w^l} \cdot \left(S - \sum_{k \in C} a^k \right) \quad \text{for all} \quad j \in N \tag{2}$$

Weighted max-min fair achieves max-min fair when the weights are equal; however, the weights provide flexibility when determining how resources are allocated. Users can be given "priority" by assigning larger weights, thus receiving larger amounts of a resource. Therefore, it is important to assign the weights appropriately to maximize network performance.

Examples of ABR rate control methods that achieve weighted max-min fairness include [10, 16, 17] and can be differentiated based on how weights are assigned. Using the Minimum Cell Rate (MCR), that is declared in the RM cell, as the weight was done by [10] and [17]. A source with a

larger MCR would receive a larger portion of the available bandwidth; yet no evidence was provided indicating weights assigned in this manner are appropriate or provide better results than max-min. Furthermore, both implementations were state-maintaining. Lakshman, et al. introduced another state-maintaining ABR rate control method for transmitting compressed video, where weights were based on the desired bandwidth of a user (demand-based weighted max-min) [16]. Simulation results demonstrated that this method can perform better than max-min; however, it assumes traffic that requires the most bandwidth is more sensitive to bandwidth reductions, which we will demonstrate is not necessarily true.

In this paper we introduce a state-less ABR rate control technique based on microeconomics and the competitive market model. We will also define a new notion of allocation fairness (*equitable allocation*), which is based on the ability of an application to reduce demands. We will prove that our method is weighted max-min fair as well as indicate how weights can be assigned to provide an equitable allocation. Using this framework, users are consumers, switches are producers and prices are affixed to link bandwidth. While link bandwidth is priced and users pay for their usage, it is important to note that this is done for **rate control** only, not revenue generation or cost recovery[1]. Advantages our ABR rate control strategy include:

- State-less implementation
- High bandwidth utilization
- Equitable allocations (QoS-fair or utility-fair) for various applications.
- Control of individual QoS
- Efficient allocations (Pareto optimal allocation)

The remainder of this paper is structured as follows. Section 2 discusses network pricing research. Section 3 reviews the competitive market model. Section 4 describes our rate control method in detail. Section 5 describes how equilibrium and Pareto optimal distribution are achieved. Section 6 describes the simulation results including comparisons to max-min and Lakshman's ABR rate

[1]Pricing ABR services for revenue, *not* rate control, is described in [5].

control method, using actual MPEG compressed traces. Finally, section 7 summarizes the results and discusses some open questions.

2 Pricing and Networks

Pricing and computer networks has gained more attention in both the research community and the private sector. A review of pricing and networks can be found in [11]. Some research deals strictly with cost recovery and/or profit maximization [5]. Alternatively, some research has applied economics to resource allocation and flow control [1, 6, 7, 13, 15, 19, 20, 26]. In general, these methods set prices to influence the bandwidth demands of the users. For example, Ferguson, et al. introduced a virtual circuit flow control method based on pricing link bandwidth [6]. Prices in the network are adjusted until an equilibrium of supply and demand is reached. They were able to prove the method reaches a Nash equilibrium, but did not permit network dynamics (such as VBR sources or dynamically varying the number of users). In contrast, Fulp, et al. used pricing to control congestion in computer networks and allowed network dynamics to occur [7]. They proved their method has price stability and reaches a Pareto optimal distribution. This method was described in general terms and was not implemented with any specific network technology in mind. In this paper we will introduce a microeconomic-based rate control method that is appropriate for the ATM ABR service class. Links will be modeled as asynchronous competitive markets that yield efficient and equitable allocations. This is done without the price distribution overhead of [7].

3 Competitive Market Model

The competitive market model consists of scarce resources and two types of agents, consumers and producers. A resource is an item (or service) which is valued by agents in the economy. Since it is scarce, there is never enough of the resource to satisfy all the agents all the time. For this reason, allocation decisions must be made. The agents come together at a market, where they buy or sell resources. Usually these exchanges are intermediated with money and the exchange rate of a resource is called its price. In the competitive market, prices are adjusted until supply equals demand. At this price the market is in *equilibrium* and the resulting allocation is Pareto

optimal [27]. For our discussion we consider a single competitive market that consists of one type of resource and m consumers, where j represents the jth consumer.

Consumer j has wealth w^j and acts independently (selfishly) purchasing resources to increase their *utility*, where utility is a measurement of overall happiness. The utility obtained from an amount of a resource is determined from a *utility function*. The utility function maps a resource amount to a real number, that corresponds to a satisfaction level. Assuming $u^j(\cdot)$ is the utility function of consumer j, if user j prefers an amount x over y then $u^j(x) > u^j(y)$. The utility function can be used to compare resource amounts based on the satisfaction the user will receive. For a single resource market, it is assumed that the utility function is monotonically increasing [27]. In addition, a user normally becomes satiated with some amount, above which the utility may decrease[2]. We will assume consumer j desires a maximum resource amount b^j; therefore $u^j(b^j)$ is the highest utility consumer j can achieve. When maximizing utility, consumer j must adhere to their budget constraint. Assuming consumer j wishes to purchase an amount a^j, where $a^j \leq b^j$, at price p the budget constraint $p \cdot a^j \leq w^j$ must be true. The wealth signifies purchasing power of each consumer, since consumers with more wealth can afford more resources. Therefore, the wealth can also be viewed as a weight when resources are allocated.

The competitive market always seeks the equilibrium price that causes supply to equal demand. It will be shown in the next section that the equilibrium price can be determined directly; however in a decentralized economy some terms are not known. For this reason, the equilibrium price is determined via a *tâtonnement process* [28][3]. The tâtonnement process iteratively adjusts the price with respect to excess demand. The excess demand is a function of the total (aggregate) demand and supply of the resource. The price increases if the demand is greater than the supply and decreases when the demand is less than the supply. It is important to note that the demand and supply at the current price must be known before an adjustment can occur. The iterative process repeats until a price is reached such that supply equals demand; at this point the market and price are in equilibrium. This is referred to as "clearing the market", where consumers maximize utility given their budget constraints and producers maximize profits. We will refer to the prices calculated

[2]Not to be confused with the indifference curve which is normally convex.

[3]Alternatively, an auction or bidding procedure can be used [22].

before the equilibrium price is reached as *intermediate prices*. Buying and selling normally do not occur with the intermediate prices [27]; however, this constraint will not apply to our ABR rate control method. Once the market is in equilibrium the resulting allocation is weighted max-min fair and Pareto optimal, which is proven in the next section.

3.1 Fairness and Optimality

The allocation provided by a competitive market in equilibrium can be described as *weighted max-min fair* and *efficient* (Pareto optimal). Furthermore, with appropriate wealth distribution the allocation is also *equitable*. In this section we formally define the terms efficient (Pareto optimal) and equitable allocations, and introduce three theorems that indicate conditions under which a competitive market can achieve these important goals.

3.1.1 Weighted Max-Min and Pareto Optimal Allocation

We repeat the notation of section 1. However, the symbols are now intepreted according to the notion of a competitive market. Given a single type of resource with supply S and a set of consumers A. Each consumer j desires a maximum allocation of b^j and has wealth w^j. Let a^j be the allocation for consumer j, where $0 \leq a^j \leq b^j$ for all $j \in A$. Denote $\{a^j\}$ as the array of allocations of all the consumers. Consumer j is either "completely satiated" or "non-satiated" with their allocation, a^j. Let C be the set of completely satiated consumers and N be the set of non-satiated consumers, where $C \cup N = A$.

Definition 3.1. Feasibility: A price and an allocation array, $[p, \{a^j\}]$, are said to be feasible if,

(i) $S = \sum_{j \in A} a^j$

(ii) $p \cdot a^j \leq w^j$ for all $j \in A$

Definition 3.2. Competitive equilibrium: At price p_* and allocation array $\{a^j\}$, a competitive market is in equilibrium if,

(i) $[p_*, \{a^j\}]$ is feasible

(ii) $u^j(a^j) \geq u^j(\hat{a}^j)$ for all $\hat{a}^j$, where $\hat{a}^j \leq b^j$ and $p_* \cdot \hat{a}^j \leq w^j$, such that $p_* \cdot a^j \geq p_* \cdot \hat{a}^j$ for all $j \in A$

Definition 3.3. Completely Satiated: At price p consumer j is completely satiated with a^j if the amount of resources affordable is greater than what is desired, b^j.

$$\text{if} \quad \frac{w^j}{p} \geq b^j \quad \text{then} \quad a^j = b^j \tag{3}$$

Definition 3.4. Non-Satiated: At price p consumer j is non-satiated with a^j if the amount of resources affordable is less than or equal to what is desired, b^j.

$$\text{if} \quad \frac{w^j}{p} < b^j \quad \text{then} \quad a^j = \frac{w^j}{p} \tag{4}$$

Theorem 3.1. *The allocation achieved by the competitive market in equilibrium, $[p_*, \{a^j\}]$, is weighted max-min fair if the wealth of each user is used as their weight.*

Proof. From definition 3.2 (i),

$$S = \sum_{j \in A} a^j = \sum_{k \in C} a^k + \sum_{l \in N} a^l \tag{5}$$

From the definition 3.4, this can be rewritten as,

$$S = \sum_{k \in C} a^k + \sum_{l \in N} \frac{w^l}{p_*} \tag{6}$$

Solving for p_*

$$p_* = \frac{\sum_{l \in N} w^l}{S - \sum_{k \in C} a^k} \tag{7}$$

The allocation for non-satiated consumers is given from substituting equation (7) into (4); therefore the allocation by a competitive market in equilibrium is,

$$a^j = b^j \quad \text{for all} \quad j \in C \tag{8}$$

$$a^j = \frac{w^j}{\sum_{l \in N} w^l} \cdot \left(S - \sum_{k \in C} a^k \right) \quad \text{for all} \quad j \in N \tag{9}$$

□

Definition 3.5. Pareto Optimality: The allocation array $\{a^j\}$, where $\sum_{j \in A} a^j = S$, is said to be Pareto optimal if there does not exist another allocation array $\{\hat{a}^j\}$, where $\sum_{j \in A} \hat{a}^j = S$, such that $u^j(\hat{a}^j) \geq u^j(a^j)$ for all $j \in A$ with a strict inequality for at least one j.

Theorem 3.2. *The allocation of a competitive market in equilibrium, $[p_*, \{a^j\}]$, is Pareto optimal.*

Proof. Suppose $\{a^j\}$ is not Pareto optimal. Then there exists $\{\hat{a}^j\}$ where

(i) $[p_*, \{\hat{a}^j\}]$ is feasible

(ii) $u^j(\hat{a}^j) \geq u^j(a^j)$ for all $j \in A$

(iii) $u^j(\hat{a}^j) > u^j(a^j)$ for at least one j

From definition 3.2 (ii) we have

$$p_* \cdot \sum_{j \in A} \hat{a}^j > p_* \cdot \sum_{j \in A} a^j \tag{10}$$

However, definition 3.1, condition (i) requires

$$p_* \cdot \sum_{j \in A} a^j = p_* \cdot S \tag{11}$$

Therefore we have

$$p_* \cdot \sum_{j \in A} \hat{a^j} > p_* \cdot S \tag{12}$$

which contradicts the feasibility of $\{\hat{a}^j\}$. □

3.1.2 Equitable Allocation

A Pareto optimal resource allocation in microeconomics is called *efficient*, and many different efficient allocation exist for a competitive market in equilibrium (consider the different possible allocations of wealth) [22]. For this reason, we employ a social welfare criterion, the *equitable* criterion, to compare and rank efficient allocations. The equitable criterion states that each user in the economy should enjoy approximately the same level of utility [22].

Definition 3.6. Equitable Allocation: An allocation array $\{a^j\}$ is an equitable allocation if,

$$u^j(a^j) = u^k(a^k), \quad \forall j,k \in A \tag{13}$$

This is the measure of **fairness** we use for our rate control method. It is important to note this does not necessarily correspond to equal amounts of a resource (the goal of max-min). This can also been referred to as a "QoS-fair" or "utility-fair" allocation. An equitable allocation can be achieved by a competitive market in equilibrium when the wealth of each consumer is correctly assigned. This is described next[4].

User j has utility function $u^j(a^j)$ that indicates a utility value q^j for an allocation amount a^j. The inverse of the utility function, denoted as $\bar{u}^j(q^j)$, indicates an allocation amount a^j that achieves a utility value of q^j. Define the aggregate inverse utility function for all consumers as,

$$\bar{u}(\cdot) = \sum_{j \in A} \bar{u}^j(\cdot) \tag{14}$$

Since $\bar{u}^j(\cdot)$ is monotonic, $\bar{u}(\cdot)$ is monotonic and has a unique solution for any feasible utility value. At equilibrium the supply equals the demand; let q^* be the utility value for all users at which this occurs, i.e.,

$$S = \bar{u}(q^*) = \sum_{j \in A} \bar{u}^j(q^*) \tag{15}$$

[4] A similar method for determining weights in a "Fair Queueing" wireless scheduler is presented in [4]; however the method described here was done independently.

q^* can be found quite easily, since $\bar{u}(\cdot)$ is monotonic. To provide each consumer the same utility level q^* when the market is in equilibrium, the wealth of consumer j is set as follows:

$$w^j = \bar{u}^j(q^*) \tag{16}$$

Theorem 3.3. *Allocating wealth using equations (14), (15) and (16) yields an equitable allocation for a competitive market in equilibrium.*

Proof. From theorem 3.1 the allocation for consumer j in a competitive market in equilibrium is,

$$a^j = \frac{w^j}{\sum_{k \in A} w^k} \cdot S \tag{17}$$

substituting equation (16), then (15)

$$a^j = \frac{\bar{u}^j(q^*)}{\sum_{k \in A} \bar{u}^k(q^*)} \cdot S = \bar{u}^j(q^*) \tag{18}$$

The utility for consumer j is,

$$u^j(a^j) = u^j(\bar{u}^j(q^*)) = q^* \quad \text{for all} \quad j \in A \tag{19}$$

Since all users achieve a utility value of q^* this is the definition of an equitable allocation. □

4 A Proposed ABR Rate Control Strategy

This proposed ABR rate control strategy is based on a competitive market model, where pricing is done to promote high utilization as well as efficient and equitable allocations. Using the competitive market nomenclature, users are consumers and switches are producers. The resource priced is ABR bandwidth. Bandwidth will be considered a non-storable resource (similar to electricity). For this reason, users cannot purchase bandwidth with the intent to use it at a later date. This microeconomic model is incorporated into the ABR service class defined by the ATM Forum. This yields a decentralized state-less rate control method that achieves all the goals associated with a competitive market. We again emphasize that our goal is rate control with QoS. Users are not billed, nor is there any element of cost recovery or profit generation. Next we define the actions associated with switches and users in the economy.

4.1 Switch

Using competitive market model, the switch owns the ABR link bandwidth that is sought by users. The network consists of several switches interconnected with links. For a unidirectional link between two switches, we consider the sending switch as owner of the bandwidth of that link. Each switch prices its ABR link bandwidth[5] based on local supply and demand. Therefore a single switch, having multiple output ports, will have one price associated with each output port, where i represents the ith link (competitive market) of the economy. The entire network can be viewed as multiple competitive markets, one market per link (similar to the New York Stock Exchange). The markets operate independently and asynchronously since there is no need for market communication (for example, price comparisons) or synchronization from switch to switch. Consequently, this results in a decentralized economy, where the physical failure of one switch/link does not necessarily cause the failure of the entire economy.

As described in the introduction, a switch will periodically receive RM-cells. The RM-cell provides the user feedback about the links (or destination) in their route. We propose using the link price as feedback, since users must scale bandwidth consumption due to budget constraints. At the switch, the *current* price for link i is inserted into the RM-cell traversing link i if it is greater than the price already stored in the RM-cell; the price in the RM-cell is initialized to 0 by the source node. The switch can update a RM-cell traveling upstream (source to destination) or downstream (destination to source). We assume that the price is stored in the ER field of the RM-cell. Therefore no additional field is required and no other information is placed/altered in the RM-cell. Note that the current price can be an intermediate or equilibrium price. Buying and selling can still occur with intermediate prices, since bandwidth is a non-storable resource[6].

The price for link i is calculated at the switch, at discrete intervals of time. We denote the nth calculation instant as t_n^i and the interval of time between the calculation points t_n^i and t_{n+1}^i as the nth price interval, P_n^i. The price during P_n^i is constant and is denoted as p_n^i. The demand for bandwidth at link i is measured as the total (aggregate) traffic received at its associated output

[5]Since we are pricing only ABR link bandwidth, all references to bandwidth will refer to ABR bandwidth.

[6]In other words, the purchasing decisions at the intermediate prices will not affect the purchasing decisions at the equilibrium price. See [27] for more information.

port. During the nth price interval, P_n^i, the total demand is expected to change; even so, the calculation of p_{n+1}^i will only use the demand measured at the end of the interval. For this reason, let the demand for bandwidth at link i, at the end of the nth price interval, be denoted as d_n^i. The supply of bandwidth at link i is denoted as S^i. At the end of the price interval, P_n^i, the switch updates the price of link i using the following tâtonnement process,

$$p_{n+1}^i = p_n^i \cdot \frac{d_n^i}{\alpha \cdot S^i} \tag{20}$$

As described in section 3, the tâtonnement process is used to determine the equilibrium price, where supply equals demand. This is achieved by setting the new price (p_{n+1}) with respect to the excess demand at the current price (p_n). In equation (20) excess demand is the ratio of the demand (received traffic) and the supply (bandwidth available). The bandwidth available is the total bandwidth times a constant α, where $0 < \alpha \leq 1$. This causes the price to increase after some percentage (α) of the total bandwidth has been reached. This is evident from the equation, since the price will only increase if the numerator is greater than the denominator. Since the price is calculated using the aggregate (not individual or group) demand, supply and current price, this rate control method is state-less.

As described in [16], queueing delays experienced by each user can be kept minimal since explicit rate methods can ensure that the aggregate rate of all users sharing the link remains below the total capacity ($\alpha \cdot S^i$). Users are also required to maintain a smooth transmission at their allowed rates (also required by [20]). Upon receiving a new price (via a returning RM-cell), users determine their allowable transmission rate. Exactly how the allowable rate is determined by the user is given in the next section.

4.2 User

User j, executing a network application, desires a maximum amount of bandwidth for transmission, b_m^j. This amount of bandwidth maximizes the utility of the user and is expected to change over time (for example compressed video). Denote the time between maximum desired bandwidth

change t_m^j and t_{m+1}^j as the mth maximum desired bandwidth interval B_m^j. The length of B_m^j will vary over time depending on the application of user j. The amount b_m^j is not necessarily the allowed transmission rate for the user, since the user is a consumer in the economy. The allowable transmission rate for user j will depend on the price associated with the route, which is provided via the RM-cell.

As defined by the ABR service class, user j periodically generates RM-cells that circulate through the route to obtain feedback about network conditions. When the RM-cell reaches the destination it is returned (via the same route) to the user. Since switches in the route can update the price contained in RM-cell, the user knows the highest price of the route. This price corresponds to the bottlenecked link in the route. Using this price, determining the allowable transmission rate, charging, CAC and source policing can be done. These functions can be implemented as a separate entity (for example a network broker [7]) located at the edge of the network. The user is charged continuously for the duration of the session (analogous to a meter). To pay for the expenses, we will assume the user provides an equal amount of money over regular periods of time. We will refer to this as the budget rate of user j, w^j (\$/sec). A single initial endowment could have been used, but would necessitate defining how it is spent during the session. To simplify simulation and analysis, budget rates are used.

Based on prices and wealth, user j can afford a range of bandwidth, less than or equal to b_m^j. Preferences in the amount of bandwidth to use is provided with a utility function (individually defined for each user). For this economy we will use *QoS profiles* [23] for the utility curves. The QoS profile is a function relating satisfaction to resource allocation, and is determined from psycho-visual experiments. The profile can be approximated by a piece-wise linear curve with three different slopes (examples are shown in figure 3). The slope of each linear segment represents the rate at which the performance of the application degrades when the network allocates a percentage of the maximum desired bandwidth (b_m^j). A steeper slope indicates the inability of the application to easily scale bandwidth (for example, high quality video), while a flatter slope signifies the application can more readily scale bandwidth requirements (for example, teleconferencing or data transmission). The horizontal axis measures the bandwidth ratio of allocated bandwidth to maximum desired

bandwidth (b_m^j). The vertical axis measures the satisfaction and is referred to as a QoS score. Our QoS scores range from one to five, with five representing an excellent perceived quality and one representing very poor quality. We will refer to an *acceptable* QoS score as any value greater than or equal to 3. As seen in the figure, if the allocated bandwidth is equal to the maximum desired bandwidth (b_m^j), the ratio is one and the corresponding QoS score is 5 (excellent quality). As this ratio becomes smaller the QoS score reduces as well. Profiles can be created for a variety of applications and redefined as users gain more experience. New and updated profiles can be easily incorporated within the economy as they become available. More information about QoS profiles and the relationship between bit-rate and quality can be found in [18, 21, 23].

Using the price returned from the last RM-cell, the wealth, the QoS profile and the maximum desired bandwidth (b_m^j), user j must determine their allowable transmission rate. Denote the rth allowable transmission rate of user j as, a_r^j. A new allowable transmission rate, a_{r+1}^j, will be determined in response to a new price, or a change in application demand. Exactly how a_{r+1}^j is calculated is given in the next section.

4.2.1 Determining the Allowable Transmission Rate

When determining a_{r+1}^j, the user will always attempt to maximize their utility (QoS score). User j will purchase no more than b_m^j and must stay within their budget constraint. The minimum bandwidth, $\check{b}^j$, the user will accept is determined from the QoS profile, b_m^j and the value that corresponds to the lowest acceptable QoS score. Using this information a_{r+1}^j is,

$$a_{r+1}^j = \begin{cases} \min\{\frac{w^j}{p_n}, b_m^j\} & \text{if } \check{b}^j \leq \frac{w^j}{p_n} \\ \emptyset & \text{otherwise, } \check{b}^j \text{ was not affordable} \end{cases} \tag{21}$$

Where p_n is the price from the most recently received RM-cell and $\frac{w^j}{p_n}$ is the maximum amount of bandwidth affordable. As noted in the equation it is possible that the minimum is not affordable, due to the QoS constraint, prices and budgets. If this case arises, the user must either; increase the budget rate, accept a lower QoS, or drop the connection. Properly managing such a situation

is an area for future work.

5 Price Stability

In section 3.1 it was proven that a competitive market can yield a Pareto optimal and weighted max-min fair allocation; however, this occurs only when the market is in equilibrium (supply equals demand). For this reason, we must also prove that our tâtonnement process, equation (20), will reach the equilibrium price p_*. We assume for this section that the aggregate demand, d_n^i, is constant (as done in [6]). This assumption is removed in section 5.1, where the effects of network dynamics (users entering/exiting and variable user demand) are properly addressed.

The equilibrium price (p_*) occurs when a price is reached such that the demand equals the supply. At this point, the resources are fully utilized. If the demand changes, the pricing mechanism should alter the price to return to equilibrium. For that reason, adjustments in the price are driven by knowledge from the market concerning the *excess demand* at a specific price. Denote the demand for bandwidth at price p as $d(p)$. For a link in the network the excess demand at price p is,

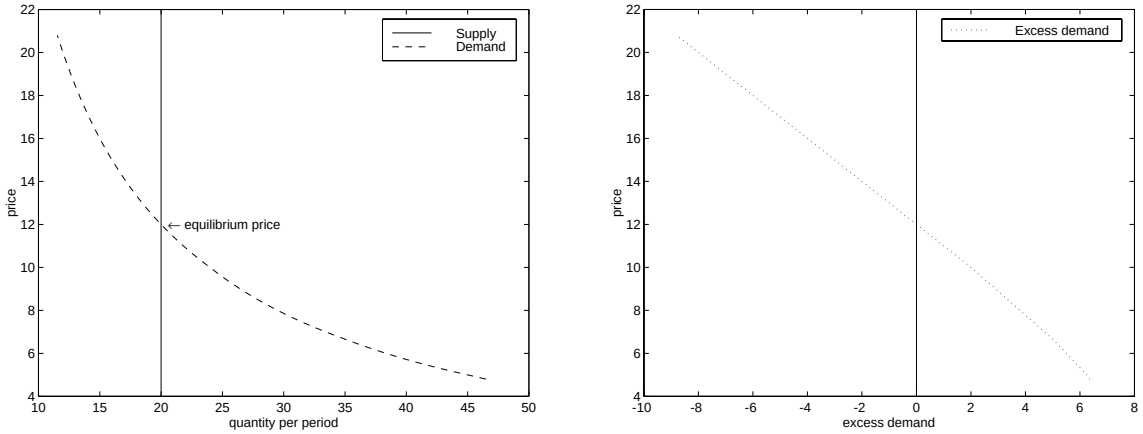
$$x(p) = p \cdot \left(\frac{d(p)}{\alpha \cdot S} - 1 \right) \tag{22}$$

Example supply, demand and excess demand curves for the system are given in figure 1. As seen in this figure, the demand curve has a negative slope which represents that an increase in price will reduce demand. The supply curve is a vertical line, because the supply of bandwidth is constant (the link does not produce bandwidth). From the supply and demand curves the *excess demand* curve can be derived.

Using these graphs we can predict the behavior of the price rule (20). We will define stability as,

$$\lim_{t \to \infty} p_t \to p_* \ .$$

The price rule will increase the price p when it is lower than equilibrium price p_*. This is done because the excess demand is greater than one. When p is greater than p_*, it is lowered towards



(a) Supply and demand curves

(b) Excess demand curve.

Figure 1: Example supply, demand and excess demand curves.

p_* because the excess demand is less than one. Therefore the price rule always moves the price towards p_*, resulting in price stability. It should be noted that the slope of the supply curve must be positive for this to be true.

The equilibrium price can be proven stable mathematically as well. Using the excess demand equation, the price adjustment from the price equation (20) over time can be written as,

$$\frac{dp}{dt} = x(p) \ . \tag{23}$$

The price adjustment can be viewed as a first-order differential equation. The local response of the equation can be analyzed in the region of an equilibrium price using the Taylor approximation,

$$\frac{dp}{dt} = x(p_*) + x'(p_*) \cdot (p - p_*)$$

$$\frac{dp}{dt} = x'(p_*) \cdot (p - p_*) \tag{24}$$

The general solution to this equation is,

$$p = (p_0 - p_*)\, e^{x'(p_*) \cdot t} + p_* \tag{25}$$

where p_0 is the initial price. As seen from the solution, p approaches p_* as time increases. However it must be the case that $x'(p_*)$ is negative, which is illustrated in figure 1(b). To provide some insight into the speed of convergence (number of iterations required by the equation), 30000 independent simulations were performed. Each consisted of a competitive market (output link) with 10, 30 or 50 users. For each simulation user j was assigned a random demand b^j and a random wealth w^j; at each price iteration, the relative difference from the equilibrium price was recorded. On average the relative difference was 0.26 after one iteration, 0.02 after two iterations and 0.002 after three iterations.

5.1 Network Dynamics

Thus far, the analysis of the network economy has not considered the dynamic nature of an actual computer network. The dynamics we are interested in include; users entering/exiting the network, and allowing Variable Bit Rate (VBR) sources. Although prevalent in actual networks, these dynamics have been either or both excluded in other microeconomic flow control methods. If the number of users and/or the demands for bandwidth change over time, then the aggregate demand, d_n, for a link will vary as well. As a result, there is not a single equilibrium price, p_*, for all time. However, the market can be viewed as having multiple equilibrium prices, each for some segment of time. During a segment the pricing technique will seek the equilibrium price as described in section 5. Once this price is found, the resulting distribution is Pareto optimal. When the aggregate demand changes, the stability of the price equation ensures that the price always moves towards p_*.

6 Experimental Results

In this section the performance of the price-based rate control method is investigated via simulation. Experiments performed will consist of a realistic network configuration, allow users to enter/exit the network, have different application types and use actual MPEG-compressed traffic. A comparison is made with two other ABR rate allocation methods, max-min and weighted max-min. The max-min fairness criterion was chosen since it is sought by many current ABR rate control methods [2]. The

max-min implementation was centralized and no communication overhead was included; therefore the max-min results presented here should be considered better than what is possible in practice. The weighted max-min rate control algorithm by Lakshman, et al. [16] was selected because it is described as an ABR rate allocation method for transmitting compressed video. Weights are equal to the desired bandwidth of each application; therefore this method will be referred to as "demand-based weighted max-min". This method requires frame prediction to allocate bandwidth before it is required; however a look-ahead buffer was used instead. For this reason, the performance of this method should be considered best possible[7]. Experimental results will show that the proposed price-based rate control technique achieves high link utilization and equitable (QoS-fair) allocations, as well as better QoS control than than max-min or demand-based weighted max-min.

Similar to [5], a rate based simulator was used that propagated rate changes and RM-cells through the network. This resulted reduced simulation times, considering the number of users, traffic type and network modeled. The network simulated consisted of 152 users, four switches and seven links, as seen in figure 2. Each output port carried traffic from 38 users and connected to a 55 Mbps link. Links interconnecting switches were 1000 km in length, while links connecting sources to their first switch were 25 km in length. Users had routes ranging from one to four hops, and entered the network at random times uniformly distributed between 0 and 120 seconds. The network can be described as a"parking lot" configuration, where multiple sources use a primary path. This configuration was agreed upon by members of the ATM Forum [14] as a suitable benchmark for allocation methods; it models substantial competition between users with differing routes and widely-varying propagation delays.

Since there is a variety of applications that transmit compressed video, user applications were considered one of two types: Multimedia on Demand (MoD) or teleconferencing. We are interested in determining if the rate allocation methods are able to provide equivalent QoS scores (utility) regardless of application type (equitable allocation). MoD applications require the transmission of high quality voice and video. These applications can scale bandwidth requirements only within a limited range, since bandwidth control is achieved through quantizer control [23]. The QoS profile

[7] A correction was made to the algorithm presented in [16] and was confirmed by the author.

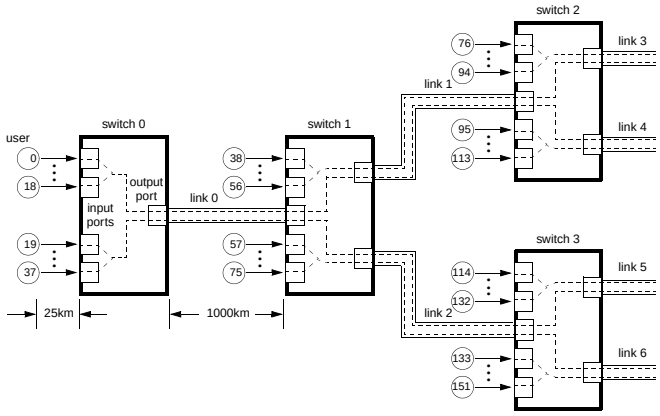


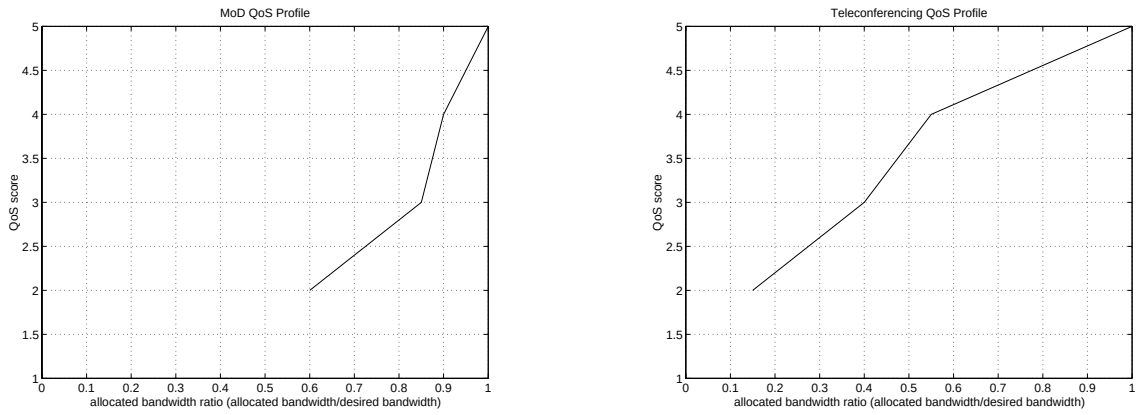
Figure 2: Network configuration used in simulations.

associated with MoD applications is given in figure 3(a). As seen in the profile, the acceptable bandwidth ratio range (i.e., resulting in a QoS score greater than or equal to 3) is relatively small, 0.85 to 1.0. Teleconferencing applications, in contrast, transmit a lower quality voice and video and can scale bandwidth requirements within a larger range. This is primarily due to quantizer control as well as the ability to transmit below the standard 24 or 30 frames-per-second. The QoS profile associated with teleconferencing applications is given in figure 3(b); the acceptable bandwidth ratio range is 0.4 to 1.0. User j, where $j = 0 \ldots 151$, was considered MoD if $\mod(j,5) < 3$, otherwise the user was considered teleconferencing. Regardless of the type of application, the source for each user was one of 15 MPEG-compressed traces obtained from Oliver Rose at the University of Würzburg, Germany [25][8]. Identifying each trace with a unique number (0 - 14), user j transmitted video trace $mod(j, 15)$.

The pricing strategy had the following initial values. MoD users had a budget rate[9], w, of 3×10^7/sec, while teleconferencing users had a budget rate of 1.5×10^7/sec. Teleconferencing users are given a lower budget because they are able to scale bandwidth requirements more readily. This was done to achieve a more equitable allocation. Although a method for determining the wealth of each user was presented in section 3.1.2, for this simulation wealth was assigned based on the bandwidth ratio required to achieve a QoS score of 3. While the wealth assignment method is less

[8]Traces can be obtained from the ftp site `ftp-info3.informatik.uni-wuerzburg.de` in the directory `/pub/MPEG`

[9]The denomination is based on bps; if based on Mbps, the budget would be 300/sec.



(a) QoS profile for MoD users.

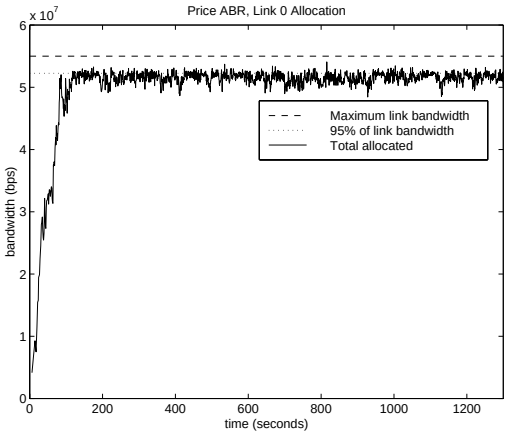
(b) QoS profile for teleconferencing users.

Figure 3: QoS profiles used in simulations.

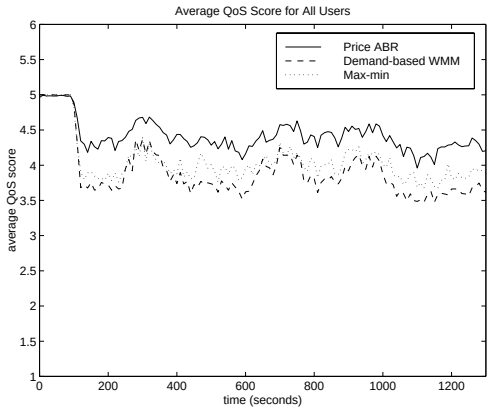
complex, it will result in an allocation slightly less equitable than possible. Switches initialized their prices to 50 and α (the target utilization) to 95%. This utilization target is extremely aggressive when coupled with QoS requirements. Switches updated their link prices at 10 msec intervals, a compromise between the desire for responsiveness, and the need for stability.

For comparisons, we are interested in the link bandwidth utilization and the QoS provided to each user. Allocation graphs are provided to measure the utilization of link bandwidth. To measure the QoS observed, average QoS graphs, percent Good or Better (GoB) measurements and average QoS scores are provided. Average QoS graphs measure the average QoS score observed over time and are based on all users or on individual type. The percent Good or Better (GoB) measurement is the average percentage of time a user had a quality score of at least 3.

Results from the simulation are given in figures 4 and 5, and table 1. As seen in figure 4(a), the allocation provided by the price method for link 0 indicates the total allocation stayed in the vicinity of 95% (α, the target utilization), yet never crossed 100%. Therefore, pricing was able to properly manage bandwidth demand (allocation results for the other links are very similar). The average QoS and percent GoB values provide a clear distinction between the rate control methods. For all users, the max-min and demand-based weighted max-min methods yielded lower average QoS and percent GoB values. This indicates, on average, users experienced lower QoS scores and enjoyed an acceptable QoS for shorter durations than the pricing method. More importantly, the

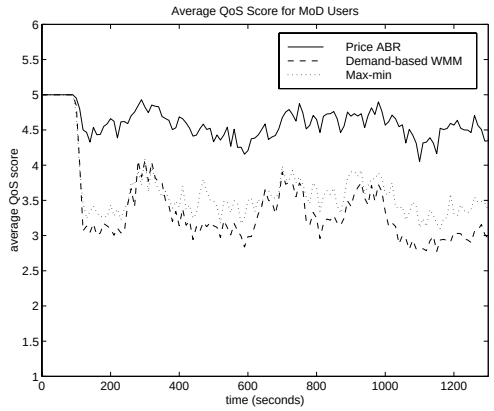


(a) Price ABR link 0 allocation.

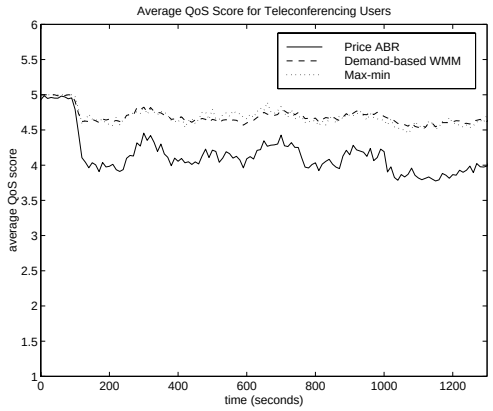


(b) Average QoS score for all users.

Figure 4: Allocation and average QoS score values.



(a) Average QoS score for MoD users.



(b) Average QoS score for teleconferencing users.

Figure 5: Average QoS score values for MoD and teleconferencing users.

	%GoB			Average QoS Score		
	All	MoD	Teleconf.	All	MoD	Teleconf.
Price ABR	90	90	90	4.43	4.63	4.14
Demand-based WMM	72	54	99	3.88	3.36	4.68
Max-min	80	66	99	4.25	3.92	4.76

Table 1: Percent GoB and average QoS scores.

pricing method provided both application types similar QoS scores and percent GoB values. This represents a more *equitable* (QoS-fair) allocation by the price method than max-min or demand-based weighted max-min. This is due to the inability of max-min or demand-based weighted max-min to differentiate between MoD users and teleconferencing users. However when equitable allocations are desired, allocation decisions must consider the fact that a reduction in bandwidth reduces the QoS for MoD users more quickly than teleconferencing users (as defined by their profiles). This was accomplished by the pricing method via wealth distribution, which resulted in an equitable allocation for all users regardless of application type.

7 Conclusions

This paper introduced a state-less ABR rate control method based on microeconomics. The computer network is modeled as multiple competitive markets. Switches own the resources sought by users, and price their resources based on local supply and demand. A user requires these resources to maximize their individual QoS. This competitive market structure encourages high utilization, with equilibrium pricing and Pareto optimal resource distribution. Proofs that the competitive market can achieve weighted max-min fair, efficient (Pareto optimal) and equitable (QoS-fair) allocations were provided as well. There are fewer restrictions on the network than required by other methods based on microeconomics, and behavior during the convergence period is described, as well as illustrated experimentally. This paper also discussed how this economy properly handles network dynamics, such as users entering/exiting, and VBR traffic sources. Simulation results demonstrate the ability of the economy to successfully allocate bandwidth of a network to a large number of diverse users, each transmitting an actual MPEG-compressed video trace. Utilization for this network was over 95% and the allocation of link bandwidth provided substantially better control of QoS than max-min or demand-based weighted max-min [16]. Finally, we believe the implementation cost will be very reasonable, since most of the functionality is in the host systems (network edge) rather than in the switches or routers, and the method can be incorporated into the existing ABR service class (no additional fields in the RM-cell are required and connection tables are not needed when determining allowable rates). While this paper has advocated microeconomics

theory solely for ABR rate control, our approach can potentially be applied to usage-based billing and cost recovery.

Acknowledgements The authors wish to thank Maximilian Ott and Daniel Reininger of C & C Research Laboratories, NEC USA for their significant contributions to this research.

References

[1] N. Anerousis and A. A. Lazar. A Framework for Pricing Virtual Circuit and Virtual Path Services in ATM Networks. *ITC-15*, pages 791 – 802, 1997.

[2] ATM Forum Technical Committee. Traffic Management Specification. Available through `ftp://ftp.atmforum.com/pub/approved-specs/af-tm-0056.000.ps`, 1996.

[3] D. Bertsekas and R. Gallager. *Data Networks*. Prentice Hall, second edition, 1992.

[4] G. Bianchi, A. T. Campbell, and R. R.-F. Liao. On Utility-Fair Adaptive Services in Wireless Networks. In *Proceedings of the IEEE Sixth International Workshop on Quality of Service*, pages 256 – 267, 1998.

[5] C. Courcoubetis, V. A. Siris, and G. D. Stamoulis. Integration of Pricing and Flow Control for Available Bit Rate Services in ATM Networks. In *Proceedings of the IEEE GLOBECOM*, pages 644 – 648, 1996.

[6] D. F. Ferguson, C. Nikolaou, J. Sairamesh, and Y. Yemini. Economic Models for Allocating Resources in Computer Systems. In S. Clearwater, editor, *Market Based Control of Distributed Systems*. World Scientific Press, 1996.

[7] E. W. Fulp, M. Ott, D. Reininger, and D. S. Reeves. Paying for QoS: An Optimal Distributed Algorithm for Pricing Network Resources. In *Proceedings of the IEEE Sixth International Workshop on Quality of Service*, pages 75 – 84, 1998.

[8] M. W. Garret and W. Willinger. Analysis, Modeling and Generation of Self-Similar VBR Video Traffic. In *SIGCOMM*, pages 269–280, London, 1994.

[9] S. Gupta and C. L. Williamson. A Performance Study of Adaptive Video Coding Algorithms for High Speed Networks. In *Proceedings of the IEEE 20th Conference on Local Computer Networks*, pages 317 – 325, 1995.

[10] Y. T. Hou, H. H.-Y. Tzeng, and S. S. Panwar. A Weighted Max-Min Fair Rate Allocation for Available Bit Rate Service. In *Proceedings of the IEEE GLOBECOM*, pages 492 – 497, 1997.

[11] S. Jordan and H. Jiang. Connection Establishment in High Speed Networks. *IEEE Journal on Selected Areas in Communications*, 13(7):1150 – 1161, Sept 1995.

[12] L. Kalampoukas. *Congestion Management in High Speed Networks*. PhD thesis, University of California Santa Cruz, 1997.

[13] F. Kelly, A. K. Maulloo, and D. K. H. Tan. Rate Control for Communication Networks: Shadow Prices, Proportional Fairness and Stability. *Journal of the Operational Research Society*, 49:237 – 252, 1998.

[14] A. Kolarov and G. Ramamurthy. Comparison of Congestion Control Schemes for ABR Service in ATM Local Area Networks. In *Proceedings of the IEEE GLOBECOM*, pages 913 – 918, 1994.

[15] Y. A. Korilis, T. A. Varvarigou, and S. R. Ahuja. Incentive-Compatible Pricing Strategies in Noncooperative Networks. In *Proceedings of the IEEE INFOCOM*, 1998.

[16] T. V. Lakshman, P. P. Mishra, and K. K. Ramakrishnan. Transporting Compressed Video over ATM Networks with Explicit Rate Feedback Control. In *Proceedings of the IEEE INFOCOM*, pages 38 – 47, 1997.

[17] C. S. C. Lee, K. F. Cheung, and D. H. K. Tsang. Generalized Weighted Fairness Criterion: Formulation and Application on Prioritized ABR Serivce. In *Proceedings of the IEEE Symposium on Computers and Communications*, pages 512 – 516, 1997.

[18] J. G. Lourens, H. H. Malleson, and C. C. Theron. Optimization of Bit-Rates for Digitally Compressed Television Services as a Function of Acceptable Picture Quality and Picture Complexity. In *Proceedings of the IEE Colloquium on Digitally Compressed Television by Satellite*, 1995.

[19] S. H. Low. Equilibrium Allocation of Variable Resources for Elastic Traffics. In *Proceedings of the IEEE INFOCOM*, 1998.

[20] J. Murphy, L. Murphy, and E. C. Posner. Distributed Pricing for ATM Networks. *ITC-14*, pages 1053 – 1063, 1994.

[21] E. Nakasu, K. Aoi, R. Yajima, K. Kanatsugu, and K. Kubota. Statistical Analysis of MPEG-2 Picture Quality for Television Broadcasting. In *Proceedings of the 7th International Workshop on Packet Video*, volume 11, pages 702 – 711, Nov. 1996.

[22] W. Nicholson. *Microeconomic Theory, Basic Principles and Extensions.* The Dryden Press, 1989.

[23] D. Reininger and R. Izmailov. Soft Quality-of-Service for VBR+ Video. In *Proceedings of the International Workshop on Audio-Visual Services over Packet Networks, AVSPN'97*, Sept. 1997.

[24] L. G. Roberts. Can ABR Service Replace VBR Service in ATM Networks. In *Proceedings of the IEEE COMPCON*, pages 346 – 348, 1995.

[25] O. Rose. Statistical Properties of MPEG Video Traffic and Their Impact on Traffic Modeling in ATM Systems. Technical Report 101, University of Würzburg Institute of Computer Science, Feb. 1995.

[26] J. Sairamesh, D. F. Ferguson, and Y. Yemini. An Approach to Pricing, Optimal Allocation and Quality of Service Provisioning in High-speed Packet Networks. In *Proceedings of the IEEE INFOCOM*, pages 1111 – 1119, 1995.

[27] A. Takayama. *Mathematical Economics.* Cambridge University Press, 1985.

[28] L. Walras. *Elements of Pure Economics.* Richard D. Irwin, 1954. trans. W. Jaffé.