Congestion Pricing Flow Control for Computer Networks

Errin W. Fulp*, Maximilian Ott†, Daniel Reininger†and Douglas S. Reeves*

Abstract

Network applications require certain performance guarantees that can be provided if enough network resources are available. Consequently, contention for the limited network resources may occur. For this reason, networks use flow control to manage network resources fairly and efficiently. This paper presents a distributed microeconomic flow control technique. This method models the network as competitive markets, where pricing of bandwidth based upon supply and demand. This yields a decentralized flow control method that provides a Pareto optimal bandwidth distribution, high utilization (up to 95% in simulation results) and better QoS than max-min. Proof of stability and a Pareto optimal distribution are provided as well as results from various simulations.

1 Introduction

Current and future networks must accommodate a wide variety of network applications. These applications range from programs that transmit simple text to complex multimedia applications that require voice and video transmission. All of these applications need network resources, such as bandwidth and buffer space, to obtain a certain Quality of Service (QoS). QoS may include bounds on, the delay of packets, variation in the delay or packet loss probability. Consequently, contention may occur for the finite amount of resources. For this reason networks need a method of flow control to manage these limited resources in a fair and efficient manner.

There are two goals associated with flow control, fairness among applications and the balance between throughput and QoS [1] [6]. Defining fairness is difficult because of the various types of applications and their desired QoS. For example, some real-time applications require rigid bounds on transmission delay, while other applications may not. For this reason similar applications (or those with similar QoS requirements) can be grouped together and fairness can be easily defined. However, fairness between different groups is more complex. The balance between throughput and QoS is the concept that the network should seek high resource utilization, but not at the expense of poor QoS (and vice versa). There are several

*Departments of ECE and CSC, North Carolina State University, ewfulp|reeves@eos.ncsu.edu
†C&C Research Laboratories, NEC USA, max|djr@ccrl.nj.nec.com

different categories of flow control for computer networks. Strategies range from window flow control to microeconomic approaches and each will be briefly discussed next.

Window flow control has been implemented in many networks [1]. This flow control method places an upper bound on the number of packets that can be transmitted (from sender to receiver). The receiver acknowledges reception of a packet from the sender. The acknowledgment acts as a permit, since once it is received the sender can transmit another packet. Since transmission of a packet requires a permit, the receiver can alter the transmission speed of the sender by slowing (or stopping) the transmission of acknowledgments. This may be needed if the receiver is overwhelmed or if congestion occurs in the network. For this reason, window flow control is considered a reactive strategy. A reactive strategy adjusts data flow in the network via feedback. Window flow control is not well suited for high speed networks due to large propagation delays in the network [1]. Another disadvantage is few performance guarantees can be made.

Rate control techniques can provide some guarantees (for example, minimum bandwidth) to applications. Assuming the bandwidth desired by each application is known, a static (fixed) amount of bandwidth is allocated and the source regulated. Since each source is regulated, this type of flow control is considered preventive. A preventive flow control method avoids congestion by requiring applications to transmit no more than what is agreed to; therefore policing is required to ensure users adhere to their allocated amount. A disadvantage to preventive flow control is the possible under utilization of resources. This will occur if an application requires its allocated bandwidth infrequently.

Game theory is another method of flow control. In this approach, each application is considered a player in a cooperative or non-cooperative game. A cooperative game requires players communicate information about their strategies. Alternatively, a non-cooperative game requires players to work individually without information from others. In either type of game, the goal of each player is to optimize their performance. Non-cooperative games have the advantage of less player-to-player communication overhead [22]. Nonetheless, the use of this information, in cooperative games, can result in a Pareto optimal allocation [12] [16]. A Pareto optimal allocation is one where no player can increase their allocation without someone else decreasing theirs. While optimal allocations can be achieved, it is difficult to apply this strategy to large networks and various types of traffic sources [22].

Microeconomic flow control is the focus of this paper. This technique models the network as an economy, then applies microeconomic principles for resource allocation. A simple economy consists of two types of agents, consumers and producers. Consumers require resources to satisfy their wants. For this reason, they seek an amount of resources that maximizes their utility. Utility is a measurement of satisfaction. Producers create or own the resources sought by consumers. They seek to maximize their utility by selling or renting their resources. Economics provides many models under which resources are allocated as well as mathematical tools to demonstrate important allocation goals.

Considering a computer network as an economy, one can view applications as consumers and switches as producers. As previously mentioned, applications require resources to transmit data through the network. Applications seek to maximize their utility by obtaining these resources. The utility of an application may be measured by the QoS provided. Switches, in the network, own these resources (such as link bandwidth, buffer space and processor time) and maximize their utility by renting or selling. Using this framework, microeconomics can

be used to define how network resources are allocated.

One approach of applying microeconomics to computer networks involves a maximization of utility functions [8] [9] [13] [14] [17] [21]. A utility function maps a resource amount to a satisfaction value. Using this function, one can compare the satisfaction levels of different resource amounts. The maximization process determines the optimal resource allocation such that utility is maximized subject to budget and resource availability constraints [10]. The maximization process requires knowledge of the individual utility curves and constraints. Since the computation required for the maximization process increases as the number of users increases, these methods are not scalable to networks with a large number of users. To provide scalability, some approaches group users and use a single utility curve to represent the group. The maximization process is then performed for the smaller number of groups instead of individual users. Groups can be created based on desired QoS [8] [9] [21] or on traffic types (or service classes) [13] [14]. In either case users must fall into one of the limited number of categories and all users must be satisfied with their representative utility curve. Accurately grouping users together may be problematic due to the wide variety of applications and their diverse resource requirements. Another problem is that these approaches generally require a centralized entity to determine the optimal allocation amount. This is undesirable because the economy relies on one entity, which is not reliable or fault tolerant.

In a congestion pricing approach, users are charged for the resources they use and resources are priced to reflect supply and demand [2] [4] [11] [15]. With such a model, prices can be set to encourage high utilization of network resources as well as a fair distribution (reactive flow control approach). Users act independently, attempting to maximize their own utility and prices are set based on local resource conditions. It has been shown that pricing based on supply and demand results in higher utilization than traditional flat (single) pricing [2] [11] [15]. Ferguson, et al. proposed a flow control mechanism based on the pricing of network resources [4] [3]. Prices of links in the system were iteratively adjusted until a equilibrium of supply and demand was reached. They were able to prove that the system achieved a Nash equilibrium; yet they required demands to be constant until the equilibrium price was determined. If the demands changed, the prices were no longer valid. Our approach uses congestion pricing in a competitive market. Similar to other microeconomic flow control methods, our approach is decentralized, seeks an equilibrium price and reaches a Pareto optimal distribution. However, in addition to those goals our approach,

- Maximizes the individual QoS observed by the application.
- Allows and adapts to changes in demands from users over time.
- Minimizes signaling and reservation requirements.
- Creates a strategy that is independent of traffic types.

The remainder of this paper is structured as follows. Section 2 reviews economic market models. Section 3 describes the pricing technique in detail. Section 4 proves our pricing strategy achieves a equilibrium price and describes how a fair Pareto optimal distribution is reached. Section 5 discusses how the pricing policy contends with network dynamics such as, users entering/exiting and multimedia traffic. Section 6 describes the simulation results for

various networks and traffic types and includes comparison to a traditional Connection Admission Control (CAC) method. Finally, section 7 reviews the pricing technique, summarizes the results and discusses some open questions.

2 Market Models

A simple economic 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. Consumers require resources to satisfy wants. Producers create or own the resources sought by consumers. These 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.

We will use a competitive market model for our network economy. In this model consumers and producers are considered price takers [23]. A price taker must accept the price determined from the current supply and demand. Using the current price, each agent is assumed to have a specific behavior. Consumers seek an amount of resources so their utility is maximized. In essence, each consumer acts independently or selfishly. The amount a consumer is ultimately able to acquire depends on their budget and the current resource price. Producers sell or rent their resources to maximize their utility (for example, profit). Prices are set with respect to supply and demand. The price increases if the demand is greater than the supply and decreases when the demand is less than the supply. When they are equal, the market and price is in equilibrium. This moment is referred to as "clearing the market" and the resulting allocation is Pareto optimal [23]. This is also known as the First Welfare Theorem, which established the claim that competition is beneficial on an equal basis. This model was chosen for our computer network economy because of its ability to achieve certain desirable goals, such as Pareto optimal distribution and price stability. The competitive market also has a structure that is simple to implement, and a well founded mathematical basis for analysis.

3 A Proposed Pricing Policy

This proposed flow control method is based on a competitive market model, where pricing is done to promote high utilization and Pareto optimal distribution. There are three entities in this network economy: users (those who execute network applications), Network Brokers (NB) and switches, as seen in figure 1. Using the competitive market nomenclature, users are consumers, switches are producers and network brokers are used to assist the exchange of resources in the market. While there are many resources in a computer network, this paper focuses on the pricing of link bandwidth.

3.1 Switch

In our competitive market, the switch owns the link bandwidth that is sought by consumers. The network consists of several switches interconnected with links. For a unidirectional link

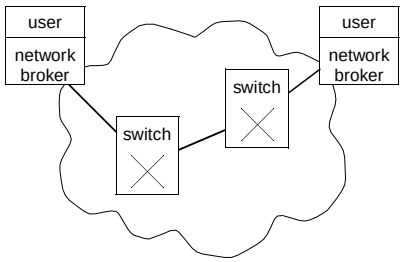


Figure 1: Example network consisting of users, network brokers and switches.

between two switches, we consider the sending switch as owner of the bandwidth of that link. Each switch prices its link bandwidth based on local supply and demand for that link. Therefore a single switch, having multiple output links, will have one price associated with each output port. For example in figure 6, switch 1 owns links 1 and 2, and will determine a price for each. The entire network can be viewed as multiple competitive markets, one market per link (similar to the New York Stock Exchange). These 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 failure of the entire economy.

The price computation for link i is performed 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 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 constant and 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 equation,

$$p_{n+1}^i = p_n^i + c \cdot \left(\frac{d_n^i - \alpha \cdot S^i}{\alpha \cdot S^i} \right) \tag{1}$$

The form of the price equation is referred to as a tâtonnement process and is used in a competitive market to set the price with respect to the current supply and demand [18] [24]. In a tâtonnement process the new price is equal to the previous price plus a correction function. The correction function provides feedback based on 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 more rapidly 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 positive ($d_n^i > \alpha \cdot S^i$). The price will decrease as the demand decreases and increase as the demand increases. An *equilibrium price* p_*^i is reached at link i when the supply equals the demand. At this point the market

clears for link i and the allocation of bandwidth is Pareto optimal [23]. The positive constant c amplifies the feedback signal and its value ultimately controls how quickly the price will increase or decrease (speed of adjustment). Note that the equation can yield negative prices. We will assume that the price will not fall below a certain non-negative minimum price (set by the switch).

After the new price, p^i_{n+1}, is calculated, a new price quote is forwarded to each NB using this link. The price quote for link i, denoted as q^i_{n+1}, consists of; p^i_{n+1}, d^i_n, S^i, c and α. The NB will use all of the information in the price quote to determine the amount of bandwidth to purchase.

3.2 User

The user, executing a network application, requires bandwidth for transmission. The amount of bandwidth desired is determined from the application and is denoted as b_m. We assume b_m is constant for the duration of the application. In a later section we will allow b_m to vary over time, which is desirable for multimedia transmission.

Based on prices and wealth, the user can afford a range of bandwidth (less than or equal to b_m), and some amounts will be preferred over others. In economics these preferences are represented with a utility function. The utility function maps a resource amount to a real number, that corresponds to a satisfaction level. Assuming $U(\cdot)$ is a utility function, if the user prefers an amount x over y (this is represented using the notation $x \succ y$) then $U(x) > U(y)$. The utility curve can be used to compare resource amounts based on the satisfaction the user will receive. For this economy we will use *QoS profiles* for utility curves. Based on psycho-visual experiments, the QoS profile is a two dimensional graph, as seen in figure 5. The profile can be approximated by a piece-wise linear curve with three different slopes. The slope of each linear segment represents the rate at which the performance of the application degrades when the network allocates a portion of the desired bandwidth (b_m). The horizontal axis measures the bandwidth ratio of allocated bandwidth to desired bandwidth (b_m). The break points associated with this axis correspond to mean opinion scores, and depends on the video material as well as the video compression quantizer. 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. As seen in the figure, if the allocated bandwidth is equal to the desired bandwidth (b_m), 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 is given in [19].

Finally, 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 the user, W (\$/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 rate are used.

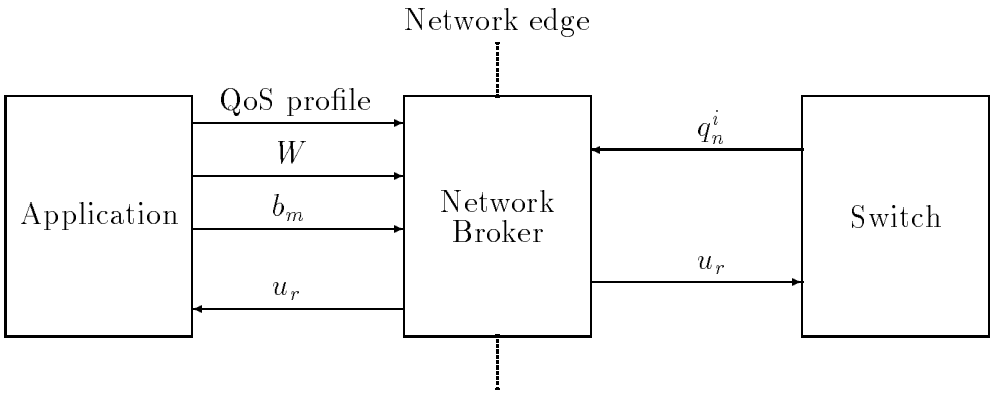


Figure 2: Information exchanged between the application, NB and switch.

3.3 Network Broker

Users can only enter the network economy through a network broker (NB). This entity is an agent for the user and is located between the user and the edge of the network. Representing the user in the economy the NB performs the following tasks: connection admission control, policing, and purchase decisions. Although the NB works as an agent for the user (making purchasing decisions), we assume that the NB operates honestly in regards to both the switches and the user.

The NB controls network admission by initially requiring the user to have enough wealth to afford at least an *acceptable* QoS; otherwise, the user is denied access. The purpose of this requirement is to be certain all users are viable consumers in the market and to prevent overloading the economy. We believe the social welfare of the economy is better when it consists of fewer users each receiving a good QoS, instead of many users each receiving a poor QoS. Hence, we are attempting to maximize the number of users in the economy, where each user can afford an acceptable QoS. If the desired bandwidth is constant, then the test is relatively simple. However, for sources where the desired bandwidth will change over time, a more complex admission test is required.

The NB monitors the user and the prices by gathering and storing information about each, as seen in figure 2. From the user, the NB collects and stores; the QoS profile, b_m and W. The NB also stores the route, R, that connects source to the destination, where R consists of v links, $\{l^i, i = 1 \ldots v\}$. For each link on R, a price quote, q^i, is collected. Where $\vec{q} = \{q^i, i = 1 \ldots v\}$ is the vector of price quotes for the route. Price quotes will change over time, since they represent link supply and demand. The NB will only store the most recent price quote from each link in the route. The NB will divide the budget rate, W, into a vector of v budget rates $\vec{w}$. Where $\vec{w} = \{w^i, i = 1 \ldots v\}$ and w^i corresponds to link i. Separate budgets are used to localize the effect of prices to each link. This prevents spending the entire budget on one expensive link. Of course depositing and withdrawing to and from these individual budgets is possible and perhaps advantageous. Using this information the NB levies the user for their consumption. Users will be charged based on usage (similar to electricity), since bandwidth is a non-storable item. Using this information the NB polices the user, ensuring only the bandwidth purchased is actually used.

Finally, the NB determines the amount of bandwidth to purchase, u_r, that will maximize the utility of the user. This value is based on the budget, current prices and QoS profile, and may change when a new price quote, q^i_{n+1} is received. We denote the rth change in the bandwidth to use as t_r, and the interval of time between the change t_r and t_{r+1} as the rth update interval, U_r. The length of U_r will vary depending on the reception of q^i_{n+1}. If $u_{r+1} \neq u_r$, the user will start sending at the u_{r+1} immediately. There is no need for direct confirmation/feedback from the switches. Exactly how the NB determines u_{r+1} is described next.

3.3.1 Determining the Bandwidth to Use

When determining u_{r+1}, the NB will first calculate the maximum and minimum bandwidth that can be used. The maximum bandwidth that can be used at link i is,

$$b^i_{max} = \frac{w^i}{p^i} \ , i = 1 \ldots v$$

therefore the maximum bandwidth the NB can afford is,

$$\widehat{b}_{max} = \min_{i=1\ldots v} \{b^i_{max}\} \ .$$

Note this equation maximizes the bandwidth at the current prices. The minimum bandwidth that can be used is determined from the QoS profile, b_m and the value that corresponds to the lowest acceptable QoS score. It is possible that $\widehat{b}_{max} < b_{min}$ (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.

After the $\widehat{b}_{max}$ and b_{min} have been calculated, u_{r+1} can be determined. The following procedure will attempt to find the maximum bandwidth at the current prices and budgets. It also calculates the price impact of the change in consumption on itself. In microeconomics this is similar to *internalizing externality*. The initial u_{r+1} is,

$$u_{r+1} = \begin{cases} b_m & \text{if } \widehat{b}_{max} \geq b_m \\ \widehat{b}_{max} & \text{if } \widehat{b}_{max} < b_m \text{ AND } \widehat{b}_{max} \geq b_{min} \\ \emptyset & \text{otherwise, } b_{min} \text{ was not affordable} \end{cases} \tag{2}$$

Using the price quotes, the NB must determine if the u_{r+1} will cause a price change that the user cannot afford, minimizing the externality of the bandwidth used. The highest price the user can afford at link i is,

$$\frac{w^i}{u_{r+1}} \ . \tag{3}$$

The new price caused by u_{r+1} at link i is,

$$p^i + c \cdot \left(\frac{u_{r+1} + d^i - \alpha \cdot S^i}{\alpha \cdot S^i} \right) \ . \tag{4}$$

The new price given in equation 4 can not exceed the maximum price affordable, given in equation 3. Using these equations the following inequality provides a bound on feasible u values,

$$w^i \geq u_{r+1} \cdot \left[p^i + c \cdot \left(\frac{u_{r+1} + d^i - \alpha \cdot S^i}{\alpha \cdot S^i} \right) \right] \quad . \tag{5}$$

Solving (5) for u_{r+1} yields the bandwidth at link i whose price change the user can afford. The inequality (5) has a closed form or it can be solved iteratively. As described earlier, once the NB has determined its u_{r+1} it will start sending immediately at this rate. No signaling is performed. This provides a significant reduction in overhead; however an over an allocation of bandwidth may occur. Consider the following scenario. Assume many users are using one link and the price has reached an equilibrium value. Now assume one user ends their session and this reduction of bandwidth results in a lower price. If the remaining users react to this lower price, over-allocation of bandwidth may occur. One simple approach to prevent this situation is to require each user to divide their amount of bandwidth to use, u_{r+1}, by the number of users remaining at the link. This should only be done if the renegotiation was the result of a lower price. An over-allocation may still occur if many users using a link start sending at a higher rate simultaneously due to their application (not price); however this would require a correlation of these events. In general, adjusting the price based on $\alpha \cdot S^i$ and the high capacity of most links diminish the significance of this problem.

4 Optimality

As with any allocation strategy there are certain optimal allocation goals. Since pricing is used, optimality will be described in microeconomics terms. There are two important goals this technique strives for; Pareto optimal allocation and price stability.

4.1 Pareto Optimality

In this section we define the conditions, that are necessary for the competitive markets to reach a Pareto optimal distribution of bandwidth. Pareto optimality is the allocation of finite resources such that no sub-set of users can improve on their allocation without lowering the utility of another, given that supply equals demand. This is a standard goal in microeconomics for social benefit of resource distribution. The proof provided is based on one by Akira Takayama [23] and was adapted for our competitive market model. This proof was chosen because it does not require strict assumptions on the utility functions of users, as other Pareto proofs require.

Notation 4.1. The network economy consists of several individual markets. A single market is composed of n consumers (user and NB pair) and a single producer (link). Let d^i be the demand of consumer i where $i = 1 \ldots n$ and $d = \sum_{i=1}^{n} d^i$. Denote the demand set of i as D^i, where D^i is a subset of R^n. An initial allocation of resources to consumer i is denoted as $\bar{d}^i$ and $\bar{d} = \sum_{i=1}^{n} \bar{d}^i$. Let S be the resource supply of the producer. Given a price p, the profit of the producer is $p \cdot S$.

Definition 4.1. Feasibility: An array of demand vectors $\{d^i\}$ is said to be feasible if $d = S + \bar{d}$.

Definition 4.2. Pareto Optimal: A feasible $\{\hat{d}^i\}$ is said to be Pareto optimal if there does not exist a feasible $\{d^i\}$ such that $d^i \succeq \hat{d}^i$ for all $i = 1 \ldots n$ with $\succ$ for at least one i.

Definition 4.3. Competitve Equilbrium: An array of vectors $[p, \{\hat{d}^i\}, S]$ is called a competitive equilibrium, if $\hat{d}^i \in D^i,\ i = 1 \ldots n$ and

(i) $\hat{d}^i \succeq d^i$ for all $d^i \in D^i$ such that $\hat{p} \cdot d^i \leq \hat{p} \cdot \hat{d}^i, i = 1 \ldots n$

(ii) $\hat{d} = S + \bar{d}$

Definition 4.4. Local Nonsatiation Point: A point $d^i \in D^i$ is called a local nonsatiation point, if there exists a $\delta > 0$ with $B_\delta(d^i) \cap (D^i \backslash d^i) \neq \emptyset$ such that for any $\epsilon, 0 < \epsilon < \delta$, with $B_\epsilon(d^i) \cap (D^i \backslash d^i) \neq \emptyset$, we have $d_0^i \succ d^i$ for some $d_0^i \in B_\epsilon(d^i) \cap D^i$, where $B_\delta(d^i)$ and $B_\epsilon(d^i)$ are open balls with center d^i and radii δ and ϵ , respectively.

Definition 4.5. Locally Nonsaturating $\succeq$: The preference ordering $\succeq$ is called locally nonsaturating if given any local nonsatiation point d^i, $d_0^i = d^i$ implies that d_0^i is also a local nonsatiation point.

Assumption 4.1. The preference $\succeq$ ordering is locally nonsaturating for every consumer.

Lemma 4.1. *Let $\hat{d}^i$ be a locally nonsatiation point for the ith consumer when price $\hat{p}$ prevails. Then under assumption 4.1, $d^i \succ \hat{d}^i$ implies $\hat{p} \cdot d^i > \hat{p} \cdot \hat{d}^i$.*

Proof. Suppose this is not true; therefore $\hat{p} \cdot d^i \leq \hat{p} \cdot \hat{d}^i$. Since $\hat{d}^i$ is the *chosen point* at price $\hat{p}$, $\hat{d}^i \succeq d^i$, which is a contradiction. □

Theorem 4.2. *Let $[\hat{p}, \{\hat{d}^i\}, S]$ be a competitve equilibrium such that $\hat{d}^i$ is a local nonsatiation point for all $i = 1 \ldots n$. Suppose assumption 4.1 holds true for all i. Then $[\{\hat{d}^i\}, S]$ is a Pareto optimum.*

Proof. Suppose $[\{\hat{d}^i\}, S]$ is not a Pareto optimum. Then there exists $[\{d^i\}, S]$ such that $d^i \in D^i$, $i = 1 \ldots n$ and

(i) $d = S + \bar{d}$

(ii) $d^i \succeq \hat{d}^i$ for all $i = 1 \ldots n$

(iii) $d^i \succ \hat{d}^i$ for some i

Hence from lemma 4.1 we have

$$\sum_{i=1}^{n} \hat{p} \cdot d^i > \sum_{j=1}^{n} \hat{p} \cdot \hat{d}^i \quad \text{or} \quad \hat{p} \cdot d > \hat{p} \cdot \hat{d}$$

But definition 4.3, condition (ii) requires

$$\hat{p} \cdot \hat{d} = \hat{p} \cdot S + \hat{p} \cdot \bar{d}$$

Therefore we have

$$\hat{p} \cdot d > \hat{p} \cdot S + \hat{p} \cdot \bar{d}$$

Which contradicts the feasibility of $\{d^i\}$. □

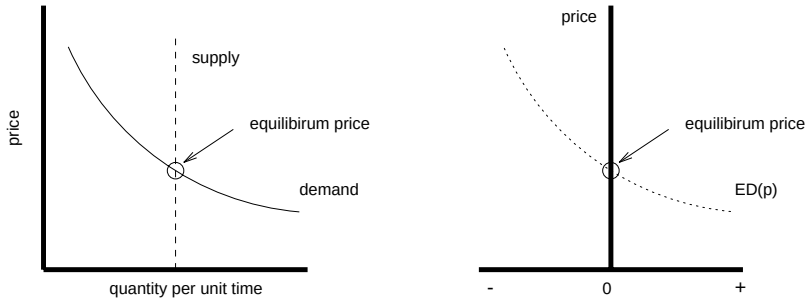


Figure 3: Example supply, demand and excess demand curves.

4.2 Price Stability

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, pricing mechanism should alter the price to return to equilibrium. A proof that the proposed pricing technique achieves this is given next.

Walrasian price stability states that prices will adjust in response to supply and demand. 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,

$$x(p) = (d(p) - \alpha \cdot S) \ . \tag{6}$$

Example supply, demand and excess demand curves for the system are given in figure 3. As seen in this figure, the demand curve has a negative slope. This 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 (1). We will define stability as,

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

The price rule will increase the price p when it is lower than equilibrium price p_*. This is done because a positive excess demand exists. When p is greater than p_*, it is lowered towards p_* because the excess demand is negative. Therefore the price rule always moves the price towards p_*, resulting in a stable equilibrium price. 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 (1) over time can be written as,

$$\frac{dp}{dt} = c \cdot \frac{x(p)}{\alpha \cdot S} \ . \tag{7}$$

Since $\alpha \cdot S$ is constant (the switch does not produce bandwidth) define the constant a as,

$$a = \frac{c}{\alpha \cdot S} \ .$$

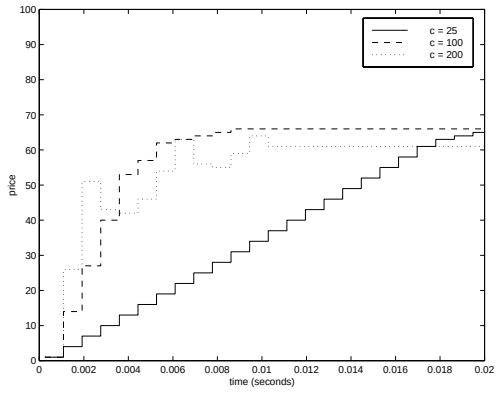


Figure 4: Price change over time using different c values.

Using the previous definitions, the price adjustment can be rewritten as,

$$\frac{dp}{dt} = a \cdot x(p) \ . \tag{8}$$

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,

$$\frac{dp}{dt} = a \cdot x'(p_*) \cdot (p - p_*) \ . \tag{9}$$

The general solution to this equation is,

$$p_n = (p_0 - p_*)\, e^{a \cdot x'(p_*) \cdot t} + p_* \tag{10}$$

where p_0 is the initial price. As seen from the solution, system is stable as time increases. However it must be the case that $x'(p_*)$ is negative, which is shown in figure 3 . The constant a is immaterial for the stability property [23]. However, the number of iterations required to reach the equilibrium price depends on the traffic, the budgets and the constant c. The switch has no control over first two items, yet some basic information can help in the selection of c. As an example, the impact of c on the number of iterations is given in figure 4. Values too low will cause the price to change slowly, resulting in more iterations. In the graph a c value of 25 requires 22 iterations, while c values of 100 to 350 require approximately 12 iterations. However values too large may raise the price too quickly, causing some users to exit before a price correction can be made.

5 Network Dynamics

Thus far, the description and 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. We allow and encourage these items into our economy; how they are handled is discussed next.

5.1 Users and Network Brokers

As described in the introduction, multimedia applications will constitute a large portion of the applications in current computer networks. The traffic generated by these applications can be described as VBR, which means the bandwidth required will change often and unexpectedly. Restricting the user to a constant desired bandwidth, as described in section 3.2, requires the user to purchase the highest amount of bandwidth expected (peak rate). For VBR sources, this approach is both difficult to implement and inefficient. Implementation is difficult since the peak rate may not be known in advance (consider live or interactive video). Purchasing only the peak rate is inefficient since the application may only require the peak rate for a short period of time. For these reasons it is advantageous to allow the user to change the desired bandwidth over time.

For a particular application, we denote the mth desired bandwidth change as t_m, and the interval of time between bandwidth changes t_m and t_{m+1} as the mth application interval, A_m. The bandwidth desired during A_m is constant and is denoted as b_m. It is important to note the length of A_m depends on the application and will vary over time. At the end of A_m the new desired bandwidth b_{m+1} is sent to the NB. Now the NB determines a new amount of bandwidth to use, u_{r+1}, when either a new price or new desired bandwidth is received. The procedure for determined u_{r+1} is described in section 3.3.1.

Once u_{r+1} is calculated the user can send at this new rate immediately. A method that requires confirmation (from the switches or a central entity) can adversely affect the performance of users and switches. If at the time u_{r+1} was determined this amount of bandwidth was available, then it is wasted until the confirmation is received. While 100% utilization can be achieved, the confirmation delay results a temporary low QoS for the user and lower profits for the switches. Our economy prevents this by, not requiring explicit confirmation and targeting a lower utilization (α).

5.2 Switches

Since the number of users and demands for bandwidth change over time, the aggregate demand, d_n, will change as well. This can be depicted in figure 3 by shifting the demand curve (for the bandwidth of a link) left or right over time. 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 4. Once this price is found, the resulting distribution is Pareto optimal as explained in the previous section. When the aggregate demand changes, the stability of the price equation ensures that the price of bandwidth always moves towards p_*. This was shown in previous section and is demonstrated experimentally in the next section.

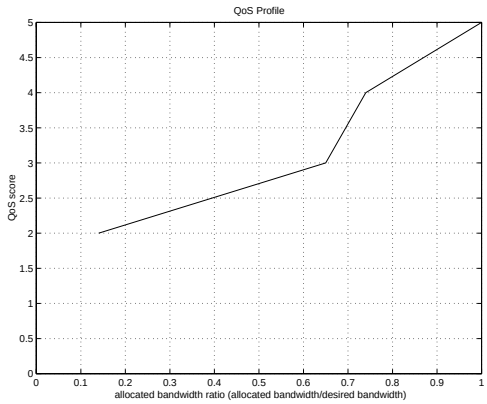


Figure 5: QoS profile used in single QoS profile experiments.

6 Experimental Results

In this section, the performance of the network economy is examined via a series of experiments. Previous microeconomic flow control techniques either do not provide experimental results or simulate limited networks (number of switches and/or traffic source types). We seek to simulate realistic network configurations, allow users to enter/exit the network, have different application types and use actual MPEG-compressed traffic. Simulation results will show that the proposed pricing technique achieves a fair Pareto distribution, equilibrium price and high network utilization.

6.1 Single QoS Profile

For these experiments each user (source) had the QoS profile given in figure 5 (single QoS profile). This profile was generated from an actual MPEG video application using the VBR traces [19]. Experiments were performed with CBR, VBR and interactive VBR traffic. The network in figure 6 was simulated for the CBR and non-interactive VBR experiments. This network configuration is appropriate for testing flow control techniques since it provides; a large number of users entering/exiting, competition among users with different route lengths and various propagation delays. The network consists of four switches and seven links. Each output port carried the traffic of 12 users and was connected to a 155 Mbps link. Links interconnecting switches were 1000 km in length. Links connecting sources to their first switch were 25 km in length. Routes were such that, users originating at switch 0 had three hop routes, users originating at switch 1 had two hop routes and users originating at switches 2 and 3 had one hop routes. Users entered the network in one second intervals (starting with user 0 and ending with user 48).

For each experiment, the pricing strategy had the following initial values. Each user had budget rates, w^i, of 3×10^8/sec [1]. Since all users have the same budget rate, they are

[1]The denomination is based on bps, where the transmission of one bit is equal to one unit of currency. If based on Mbps the budget would be 300/sec, where the transmission of one Mb is equal to one unit of currency.

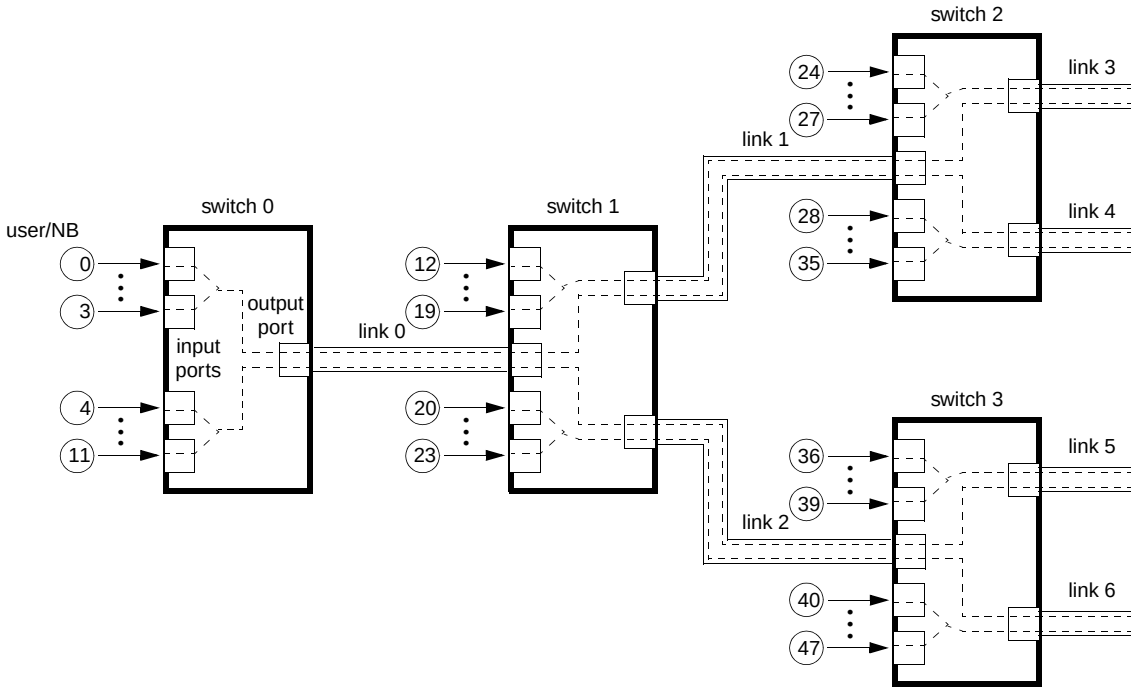


Figure 6: Network used for single QoS profile experiments.

considered equal (purchasing power). This should cause all users to be treated fairly, with no disproportionate allocation if all require the same amount. Switches initialized their prices to 1. The price equation c constant was set to 50 and α (the target utilization) was 90%. We also assumed that there was no propagation delay between the user and their NB. Switches updated their link prices at an interval equal to 20 times the longest propagation delay of any user connected to it.

We are interested in the resource utilization and the QoS provided to the users in the network. The number of users and the utilization provides insight into the efficiency of the allocation. To measure the QoS observed we calculated the percent Good or Better (GoB) quality of service. The percent GoB is the average percentage of time a user had a quality score of at least 3. The performance can also be seen with the allocation and QoS score graphs accompanying each experiment. The allocation graphs show the individual usages as well as the total usage at a particular link. The QoS score graph shows the number of users receiving; excellent QoS (a score of 4.95 or greater), a good QoS (a score between 4.94 and 3) or poor QoS (a score less than 3), over time. We are interested in comparing this pricing technique to a traditional CAC method, which is briefly described in the next section.

6.1.1 Dynamic CAC

Similar to the traffic management of Available Bit Rate (ABR) service offered by ATM networks, this technique requires signaling to allocate resources. Once a user wishes to (re)negotiate for more bandwidth they must send a request (similar to a RM-cell in ATM ABR service) to the switches in the route. Each switch reads the request and determines the amount of link bandwidth it can allocate. Allocation is performed on a first come first

served basis. If the switch is unable to allocate the requested bandwidth, it changes the request to its available amount; therefore a switch is only able to reduce the rate. The new request is then forwarded to the next switch, where the process is repeated. The last switch then forwards the request back to the user. The result is an allocation which is the minimum amount available in the route. Once the message has reached the NB, it must decide if the allocated amount is sufficient using the QoS profile given in figure 5. If so, it will start sending at the allocated value. Although this method can allocate all of available the link bandwidth, it requires a round trip propagation delay before a higher rate can be used. If the user wishes to use less resources, it can do so immediately but must also notify the switches in the route.

6.1.2 CBR Results

CBR experiments allow better understanding of the algorithm performance. It is easy to determine if the resulting allocations are Pareto optimal and if the price reaches equilibrium. All previous pricing methods have used CBR sources, but few with complex network configurations. For example Ferguson et al. used microeconomics to allocate bandwidth among Virtual Circuits in a simple network [3]. In addition, our experiment allows sources to enter and exit the network over time. For this experiment each user required a bandwidth of 20 Mbps for a duration of 60 seconds.

The allocation and QoS graphs for this experiment are given in figures 7 and 8. Using the dynamic CAC method, each output link could only support eight users. The remaining users of this link were denied access. For example at link zero, the first seven users were allocated their peak amount and the eighth user was allocated the remaining 15 Mbps. The result of this allocation was, the first seven users received an excellent score while the eighth user received a good QoS score. If the first seven users were forced to reduce their bandwidth, more users and fairer (equal) QoS scores could result. This occurred when the pricing method was simulated, where each link supported ten users. As users initially entered the network, they used their desired bandwidth (20 Mbps) as seen in figure 8. At link zero, the seventh user entered increasing the price and causing others currently using the link to scale back. This continued until the ninth user accessed the link. The price then reached equilibrium and the remaining users (10 - 12) could not afford to start their session. This provides a fair Pareto optimal allocation, where each user had approximately the same allocation as well as the same QoS score. This was expected since all users have the same purchasing power. The other links behaved in a similar fashion. The pricing technique allowed 40 users (ten per link) in the network as compared to 32 for the CAC method, a 25% increase. As users exited the link, the reduction in price caused others to increase their consumption, as seen in figure 8. Therefore, the pricing mechanism also attempts to maximize utilization, by lowering the price and encouraging usage.

6.1.3 VBR Results

For these experiments the traffic of each user was one of three MPEG-compressed traces. To date no other microeconomic flow control method has provided experimental results with actual MPEG sources. Statistics about the traces are given in table 1 and their duration was

Name	Average Rate (Mbps)	Minimum Rate (Mbps)	Peak Rate (Mbps)
Video 1	6.78	1.49	25.3
Video 2	5.98	1.76	17.8
Video 3	4.13	1.48	17.3

Table 1: Statistics about MPEG videos.

approximately 20 minutes. For VBR sources a single equilibrium price will not exist. This is due to the changing demands of the sources over time, as described in section 5. However, at any point in time there exists such a price and the pricing rule always adjusts towards it.

The utilization and the percent GoB values were similar for both techniques. For example, the average utilization of link 0 was 70% and percent GoB was 99.9% for the dynamic CAC and the price method. The difference in performance can be noted in QoS graphs, figures 9 and 10. The QoS values provided by the CAC method ranged from 2.2 to 5. This was the result of signaling delay. Consider a point where a user needs more bandwidth. They must wait twice the propagation delay before sending at the new rate (if it was available). During this delay the user can only send at the rate currently allocated, resulting in a very low QoS score (2.2 in some cases). This was especially true for sources with longer routes. The result is a higher variation in the observed QoS scores. The pricing technique does not suffer from signaling delays since users immediately send at the new rate (if they can afford it), as seen in figure 10. During times where the total bandwidth approaches the 90% level (recall that α was set to 90%) users must reduce their demands due to an increase in price. Otherwise all users operated at an excellent QoS level.

6.1.4 Interactive VBR Results

This experiment consisted of 32 users and a single switch. Each user was connected to the switch via a 25 km link and the switch had one 155 Mbps output link (all traffic was routed to the output link). The traffic of each user was one of the three MPEG-compressed videos described in table 1. The interactivity of each video included altering the frame rate and the resolution. The frame rate could be either 30, 20 or 10 fps. The resolution was either 480×640 or 240×320. The frame rate (or resolution) was constant for a period of time uniformly distributed between 30 and 850 seconds. At the end of a period, a new frame rate (or resolution) was selected, where each frame rate (or resolution) was equally probable. Sources started their sessions in one second intervals.

Results for the interactive videos are presented in figures 11 and 12. The dynamic CAC method allocated up to 100% of the link bandwidth, while the price method allocation was in the targeted range of 90% (α), yet never over allocated the link. The variation around 90% is due to price changes caused by users entering/exiting and changing demands. The percent GoB was similar for the two methods, 97.35% for the dynamic CAC method and 96.50% for the price method. However, as observed in the VBR experiments the CAC method suffered from signaling delays and rarely had all 32 users experiencing an excellent QoS score. In contrast, the price method was able to achieve a higher number of excellent scores.

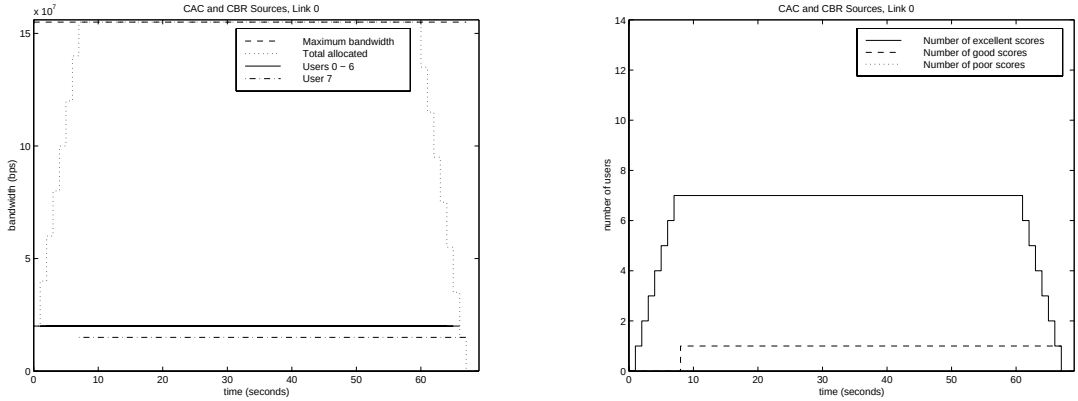


Figure 7: Link 0 allocation and QoS score graphs for the CAC method and CBR sources.

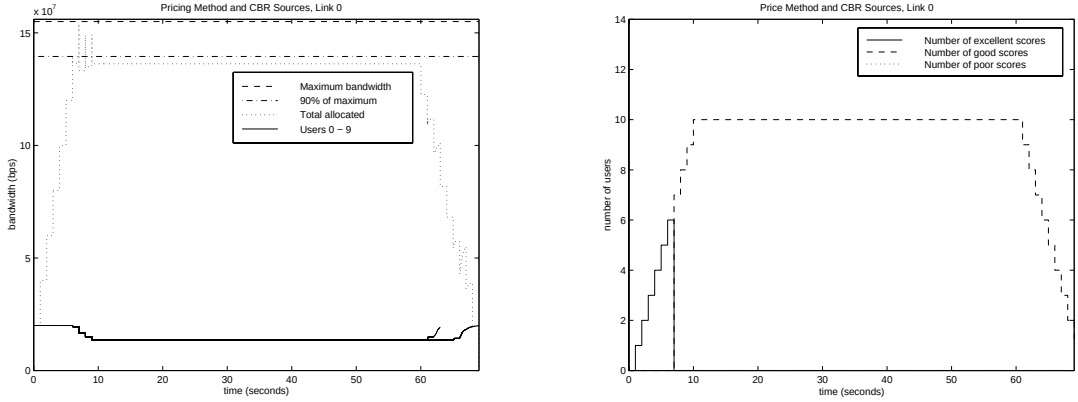


Figure 8: Link 0 allocation and QoS score graphs for the price method and CBR sources.

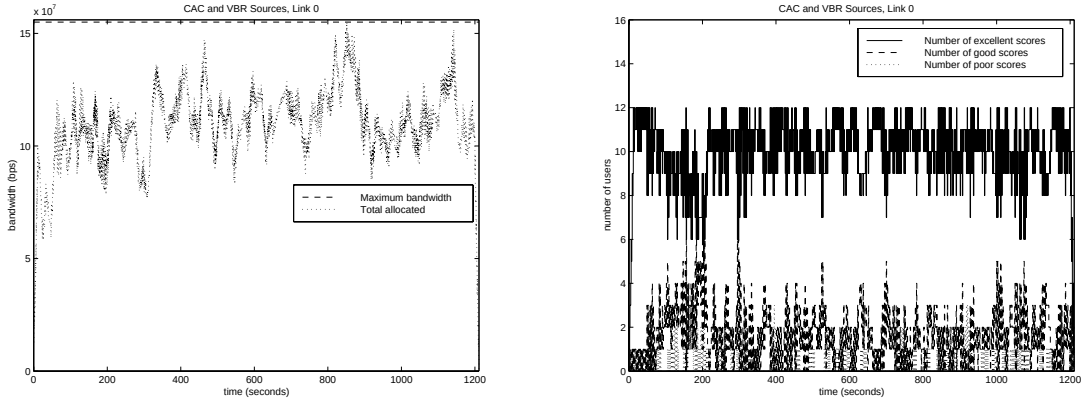


Figure 9: Link 0 allocation and QoS score graphs for the CAC method and VBR sources.

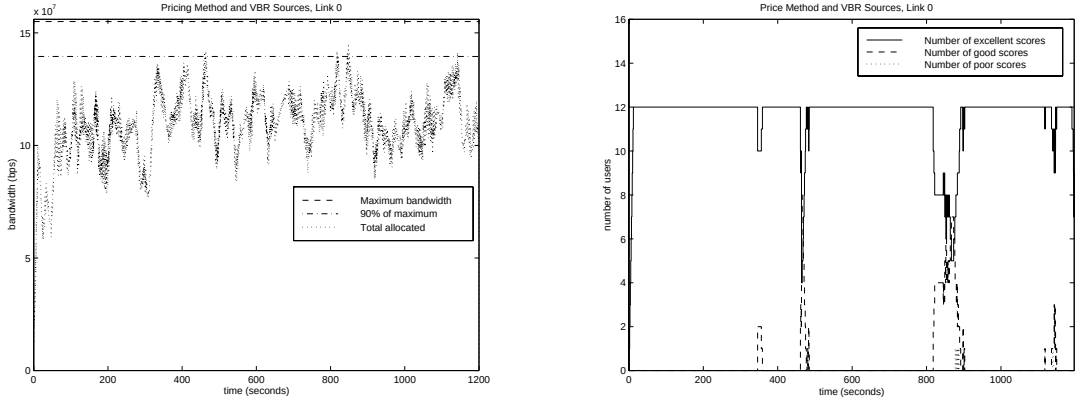


Figure 10: Link 0 allocation and QoS score graphs for the price method and VBR sources.

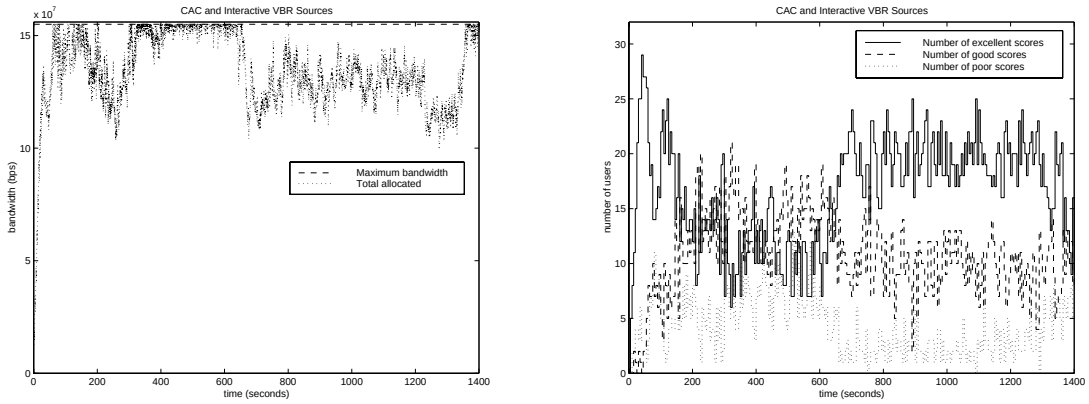


Figure 11: Link bandwidth allocation and QoS score graphs for the CAC method (32 interactive VBR sources).

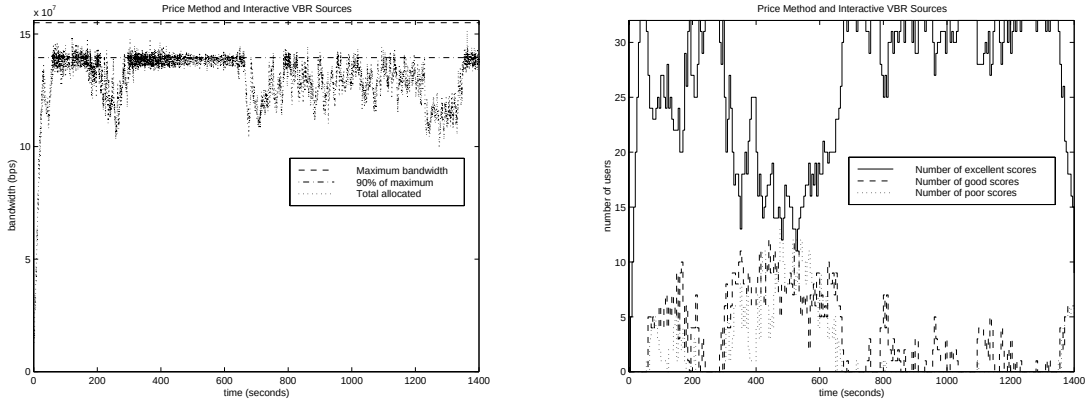


Figure 12: Link bandwidth allocation and QoS score graphs for the price method (32 interactive VBR sources).

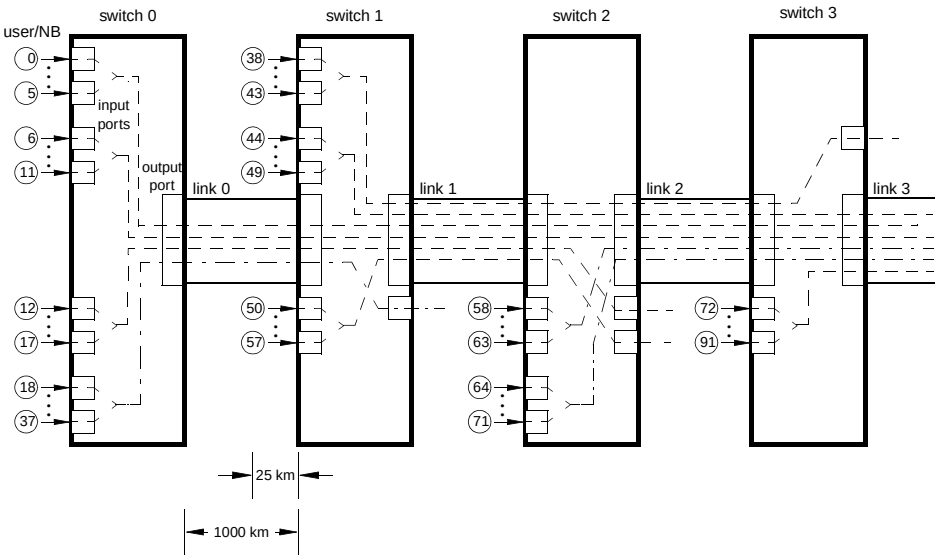


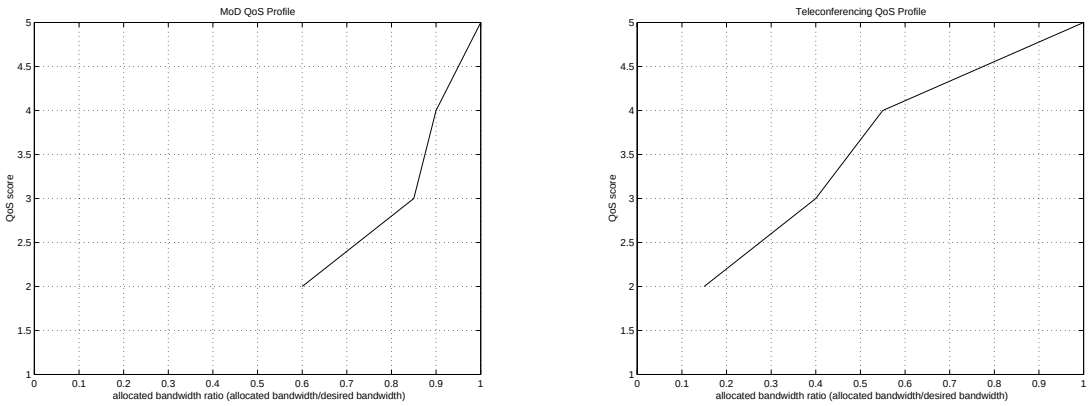
Figure 13: Network configuration used in multiple QoS profile simulations.

6.2 Multiple QoS Profiles

In this section, the performance of the economy is investigated when users have different QoS profiles. A comparison to max-min is also provided, since max-min fairness is a goal of many flow control techniques [1]. Experimental results will show that the proposed pricing technique achieves a fair Pareto distribution, provides higher QoS scores than max-min, and high network utilization.

The network simulated consisted of 92 users/NB, four switches and four primary links, as seen in figure 13. 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 one primary path. This configuration was agreed upon by members of the ATM Forum for allocation comparisons since it provides competition among users with different routes and various propagation delays [7].

For this simulation applications were one of two types, Multimedia on Demand (MoD) or teleconferencing. 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 [19]. The QoS profile associated with MoD applications is given in figure 14(a). As seen in the profile, the acceptable bandwidth ratio range (a QoS score greater than or equal to 3) is relatively small, 0.85 to 1.0. Teleconferencing applications 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 14(b), and the acceptable bandwidth ratio range is 0.4 to 1.0. The total number of MoD applications was 56 (3/5) and



(a) (b)

Figure 14: (a) MoD QoS profile. (b) Teleconferencing QoS profile.

the total number of teleconferencing applications was 36 (2/5). 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 [20][2]. Each trace is a thirty minute segment of the original video and each was encoded with constant quality using the same MPEG-1 encoder card. Relevant statistics of each video are presented in [5] and [20]. As reported in [20], the Hurst parameters indicate all videos exhibit long-range dependency, and significant peak-to-mean ratios ranging from 18.4 to 4.63 based on average frames; therefore it is evident that these are very difficult sources to regulate. To date no other microeconomic flow control method has provided experimental results with actual MPEG sources or various application types.

The pricing strategy had the following initial values. MoD users had budget rates, w^i, of 3×10^8/sec, while teleconferencing users had budget rates of 1.5×10^8/sec. Teleconferencing users have a lower budget since they are able to scale bandwidth requirements more readily. Switches initialized their prices to 1, their price equation c constant to 50 and α (the target utilization) to 95%. We also assumed no propagation delay between the user and their NB. Switches updated their link prices at an interval equal to 20 times the shortest propagation delay of any user connected to it.

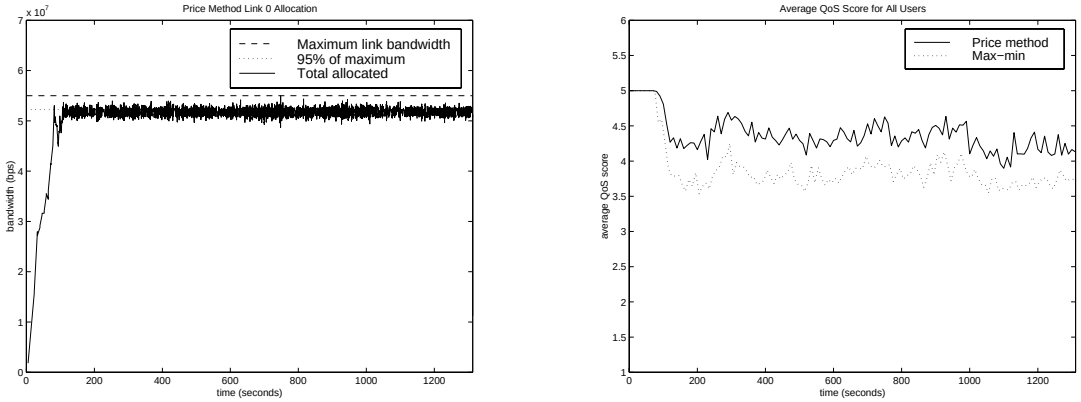
The max-min fairness criterion states that any user is entitled to as much bandwidth as any other. When a link is bottlenecked, the bandwidth is divided equally among the users of the link. If a user requires less than this amount, the difference is divided equally among the remaining users. This is process is repeated until all users of the link have been allocated a maximum amount of bandwidth. There is no distinction between application types. A more detailed description for networks is provided in [1]. In this simulation, the max-min allocation for the entire network was calculated after each source renegotiated and the resulting QoS scores were then recorded. The exact max-min algorithm was implemented. No implementation overhead or propagation delays were included. Consequently, the max-

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

min results presented here are better than or equal to what is actually possible. Regardless, we are only interested in comparing to the best performance max-min can provide.

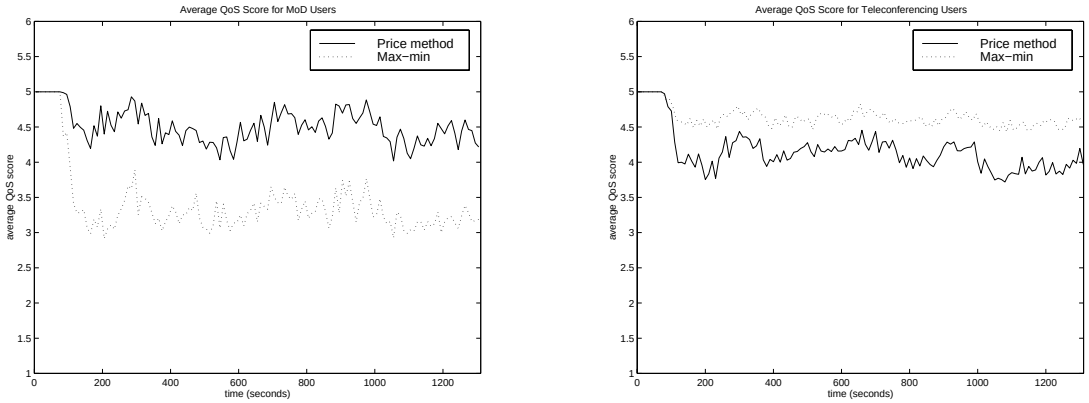
For comparisons, we are interested in the link bandwidth utilization and the QoS provided to each user. Allocation graphs are provided to measure the utilization of link bandwidth. To measure the QoS observed, average QoS graphs, percent Good or Better (GoB) measurements and average QoS scores are provided. 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.

For this simulation, the price method bandwidth allocation for link 0 (representative of the other links) is given in figure 15(a). The allocation graph indicates that the total allocation of bandwidth stayed in the vicinity of 95% (α, the target utilization), yet never crossed 100%. The fluctuation around 95% is the result of users entering/exiting and changing demands. The average QoS score graph, figure 15(b), shows that the price method always provided a higher average QoS score. This is also indicated in table 2, where the price method average QoS score was 4.37 as compared to 3.88 for max-min. The percent GoB for the price method was also 20% higher than max-min. This indicates that users, under the price method, enjoyed an acceptable QoS for a longer duration. The difference between the price method and max-min is more distinct when considering the QoS provided to the two types of applications individually. In figure 16(a), the price method provides a higher QoS score for MoD applications than max-min. This is also indicated in the MoD values in table 2, where the average QoS score is 24% higher and the percent GoB was 40% greater. This is due to the inability of max-min to differentiate between MoD users and teleconferencing users. When a link becomes congested, the max-min distributes bandwidth equally among bottlenecked users. However, a reduction in bandwidth reduces the QoS for MoD users more quickly than teleconferencing users (as defined by their profiles). This is also evident in the average QoS graph for teleconferencing users, figure 16(b) and the average QoS scores in table 2. In contrast, the pricing method provides more bandwidth to MoD users than teleconferencing users. As a result the average QoS score for either type is almost equal. We believe it is more desirable to allocate so as to provide comparable QoS rather than equal bandwidth amounts. In table 2 all percent GoB and the average QoS scores differ by no more than 8% for the price method. In contrast, the MoD and teleconferencing percent GoB values differ by more than 46% for max-min. For this simulation, the price method was able to price link bandwidth in such a manner that lead to high utilization and better QoS performance than tradition max-min. Users were able to purchase link bandwidth, maximizing their QoS score individually and yielding a high percent GoB.



(a) (b)

Figure 15: (a) Price method link 0 allocation. (b) Average QoS score for all users.



(a) (b)

Figure 16: (a) Average QoS score for MoD users. (b) Average QoS score for teleconferencing users.

	%GoB			Average QoS Score		
	All	MoD	Teleconf.	All	MoD	Teleconf.
Price method	90	88	91	4.37	4.51	4.15
Max-min	72	53	99	3.88	3.41	4.63

Table 2: Percent GoB and average QoS scores.

7 Conclusions

This paper introduced a decentralized flow control method based on microeconomics. A computer network was viewed as multiple competitive markets consisting of three entities; users (those who execute network applications), Network Brokers (NB) and switches. Using competitive market nomenclature, users were consumers, switches were producers and network brokers were used to assist in the exchange of network resources. For this paper we considered link bandwidth as the resource exchanged in these markets. Each switch owns the bandwidth of any output link connected to it and prices bandwidth based on local supply and demand. Pricing is done local and asynchronously, resulting in a decentralized economy. Pricing based on local conditions also encourages high utilization of the bandwidth. The user, executing a network application, requires link bandwidth for transmission and is represented in the economy with a NB. The network broker collects prices and determines usage levels that maximize the user's utilization. Once a new amount is determined, it is immediately used (no signaling is required). An equilibrium price is reached when the demand equals the supply.

We proved that this method can reach a Pareto optimal distribution of bandwidth. Here the users can not increase their utility without lowering the utility of someone else. This is an important goal of resource distribution in microeconomics. We were also able to prove that the pricing equation reaches an equilibrium. At this point the supply equals the demand and the price is stable. If there is a shift in the demand, the price is adjusted to move back to equilibrium.

Simulations were performed over various networks using CBR, VBR and interactive VBR sources. The results indicated high resource utilization is achieved as well as an increase in the number of users. A comparison to a traditional CAC method demonstrated how the pricing method provides fairer distributions without the overhead of signaling. The utilization was slightly lower than a traditional CAC, because the price method targets a percentage of the total resource. However this was offset by gains in the number of users and the fair distribution. Results also show that the price method achieves better QoS than max-min, a goal of other flow control techniques.

This paper provided a preliminary outline and some promising experimental results. We were able to demonstrate a microeconomic flow control method that: is decentralized, yields a Pareto optimal distribution, has price stability, and can adequately manage network dynamics such as users entering/exiting and multimedia traffic. Future work is needed to provide answers for open questions such as,

- What type of connection admission should be performed for VBR sources, especially live or interactive sources where a priori information is limited?
- What method of traffic policing should be performed?
- Can the minimum price be set such that is provides excellent QoS and maximizes the monetary gain of the switch?
- How should the budget be distributed? Will budget shifting result in bidding wars among users?

- How are routes selected? Should they be based on price only or other performance measures? Should alternative routes be sought if the price becomes too expensive?
- What impact will propagation long delays have? Does it raise unfair situations?
- Can this system be implemented recursively to reduce the propagation delay of prices? For example divide networks into smaller regions each with their own set of edge nodes (NB's). Instead of querying prices for each switch along the route, a user would query only the edges of each network.

References

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

[2] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang. Pricing in Computer Networks: Motivation, Formulation, and Example. *IEEE/ACM Transactions on Networking*, 1(6):614 – 627, Dec 1993.

[3] 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.

[4] D. F. Ferguson, C. Nikolaou, and Y. Yemini. An Economy for Flow Control in Computer Networks. In *IEEE INFOCOM'89*, pages 110 – 118, 1989.

[5] E. W. Fulp and D. S. Reeves. Dynamic Bandwidth Allocation Techniques. Technical Report TR-97/08, Center for Advanced Computing and Communications, Aug. 1997.

[6] M. Gerla and L. Kleinrock. Flow Control: A Comparative Survey. *IEEE Transactions on Communications*, 28(4):553 – 574, April 1980.

[7] R. Jain. Congestion Control and Traffic Management in ATM Networks: Recent Advances and A Survey. *Computer Networks and ISDN Systems*, Feb. 1995.

[8] H. Ji, J. Y. Hui, and E. Karasan. GoS-Based Pricing and Resource Allocation for Multimedia Broadband Networks. In *IEEE INFOCOM'96*, pages 1020 – 1027, 1996.

[9] H. Jiang and S. Jordan. A Pricing Model for High Speed Networks with Guaranteed Quality of Service. In *IEEE INFOCOM'96*, pages 888 – 895, 1996.

[10] D. M. Kreps. *A Course in Microeconomic Theory.* Princeton University Press, 1990.

[11] A. Krishnamurthy, T. D. C. Little, and D. Castañon. A Pricing Mechanism for Scalable Video Delivery. *Multimedia Systems*, 4:328 – 337, 1996.

[12] L. N. Kumar, C. Douligeris, and G. Develekos. Implementation of a Decentralized Pareto Optimal Algorithm. *Computer Communications*, 17(8):600 – 610, August 1994.

[13] J. F. Kurose and R. Simha. A Microeconomic Approach to Optimal Resource Allocation in Distributed Computer Systems. *IEEE Transactions on Computers*, 38(5):705 – 717, May 1989.

[14] S. Low and P. Varaiya. An Algorithm for Optimal Service Provisioning using Resource Pricing. In *IEEE INFOCOM'94*, pages 368 – 373, 1994.

[15] J. K. MacKie-Mason and H. R. Varian. Pricing Congestible Network Resources. *IEEE Journal on Selected Areas in Communications*, 13(7):1141 – 1149, Sept 1995.

[16] R. Mazumdar, L. G. Mason, and C. Douligeris. Fairness in Network Optimal Flow Control: Optimality of Product Forms. *IEEE Transactions on Communications*, 39(5):775 – 782, May 1991.

[17] J. Murphy and L. Murphy. Bandwidth Allocation by Pricing in ATM Networks. In *ITC*, June 1995.

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

[19] 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.

[20] 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.

[21] 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 *IEEE INFOCOM'95*, pages 1111 – 1119, 1995.

[22] S. J. Shenker. Making Greed Work in Networks: A Game-Theoretic Analysis of Switch Service Disciplines. *IEEE Transactions on Networking*, 3(6):819 – 831, December 1995.

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

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