Feedback And Efficiency In ATM Networks

Liam Murphy* John Murphy† Jeffrey K. MacKie-Mason‡

Abstract

Admission control and congestion control can provide performance guarantees in ATM networks. However some users may not be able to describe their traffic accurately enough for the network to provide these guarantees. By sending a dynamic feedback signal about the current utilization of network resources, the network could provide some guarantees to adaptive users who respond appropriately: this is the basis of ABR service. We outline a user-oriented framework for network operation and control, explicitly defining how such feedback is generated by the network and what form it takes. We show through simulations that it is possible to simultaneously gain both network and economic efficiency by using a form of feedback we call responsive pricing, which is compatible with current ATM Forum UNI specifications.

1 Introduction

Asynchronous Transfer Mode (ATM) has been adopted as the transfer mode for the Broadband Integrated Services Digital Network (BISDN) [1], a service-independent network capable of supporting all existing and future communication services. ATM is also emerging as a local area networking technology, since it provides flexible bandwidth-on-demand capabilities for conventional data communications. ATM-based networks are therefore expected to accommodate a wide range of users, including some whose applications require *guarantees* on cell loss and/or delay. (Some users may be satisfied with best-effort service, for which the network offers no guarantees on loss or delay.)

In order to obtain these guarantees from the network, users have to describe their traffic inputs by specifying values for network-defined *traffic descriptors* such as peak cell rate (PCR) or sustainable cell rate (SCR). However some users may not be able to describe their traffic accurately, because

- their applications cannot be sufficiently well-characterized by the given traffic descriptors; or
- their actual traffic inputs depend on factors outside user control (such as the number of applications competing for a shared resource).

A common assumption in many proposed admission control schemes is that traffic which is not well-described cannot get specific performance guarantees.

The ATM Forum has recognized the problem of providing guarantees to users whose traffic cannot be well-described, and in response has developed a specification for Available Bit Rate (ABR) service [11]. Users who choose ABR service receive feedback from the network about the current level of network resource utilization, and can get cell loss guarantees if they respond appropriately — by reducing their input rates in times of congestion, for example.

Most suggestions for supporting ABR service assume that well-described traffic which requires performance guarantees gets priority in the use of network resources such as bandwidth or buffer space, and that the remaining resources are fairly shared among all ABR users. Two issues not explicitly addressed are why more demanding traffic should get priority over ABR traffic, and what constitutes fair sharing. A common interpretation is that the available bandwidth is shared equally among all ABR users.

It is important to realize that these proposals make *implicit* assumptions about user requirements and preferences: giving a CBR/VBR user priority over an ABR user assumes that the CBR/VBR user is more "valuable" than the ABR user, but the reverse will sometimes be true; sharing the available bandwidth equally among ABR users values all such traffic equally, but the users themselves may put widely differing values on network access and Quality of Service (QoS).

A network is only as valuable as its users perceive it to be. Therefore, we advocate that the users themselves determine relative traffic priorities. This requires a network control framework which is **user-oriented** rather than application-oriented or traffic-type-oriented [5]. Our aim in this paper is to describe some feedback schemes which could be part of such a user-oriented framework for network operation and control.

*Department of Computer Science and Engineering, Auburn University, AL 36849, USA; lmurphy@eng.auburn.edu.

†School of Electronic Engineering, Dublin City University, Glasnevin, Dublin 9, Ireland; murphyj@eeng.dcu.ie.

‡Dept. of Economics and School of Information, University of Michigan, Ann Arbor, MI 48109-1220, USA; jmm@umich.edu.

2 User-oriented network operation and control

Both TCP and proposed ABR congestion control schemes are *application-oriented*: the focus is on how a particular application type responds to feedback from the network, with the overall aim of modifying network traffic levels. A user-oriented scheme would also take user valuations into account: the network could serve higher-value users even under congestion if lower-value users could be induced to withhold their traffic until congestion lessens. By essentially swapping high-value for low-value traffic, network performance can be maintained or improved while the total value of network service increases from the users' point of view.

We suggest bringing users back into the loop and thereby ensuring that performance measures are user-oriented (see Figure 1). We propose a form of feedback which we call **responsive pricing** and argue that it represents a particularly useful mechanism for maximizing network value. Users would gain by obtaining service more closely matched to their needs; network operators would gain through improved network performance and increased user satisfaction with the service they receive.

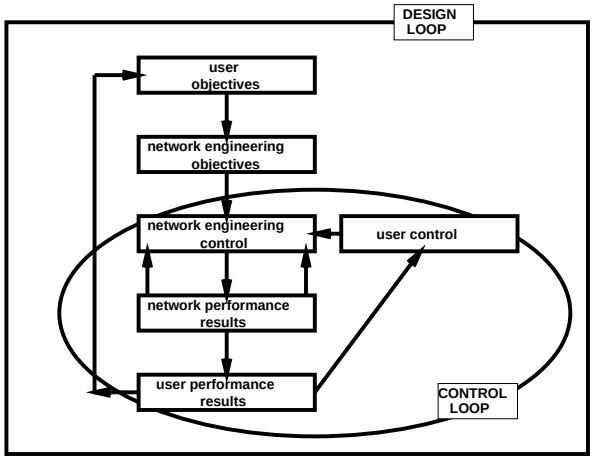

Figure 1: Network design and control loops

3 Feedback and adaptive users

Users of current data networks respond to multiple forms of feedback, on various time scales. On the longest time scale users decide whether or not to use a particular network, perhaps based on the network's charging structure, or their previous experience with it. At the connection setup level, if a user observes that the network is usually heavily loaded at certain times of the day and lightly loaded at others, she may schedule her network usage accordingly. Most people are familiar with the decision that it is not worthwhile to use a network during busy periods, but instead to do something else and defer their network usage, without necessarily recognizing this process as economic decision-making.

During a connection, **adaptive** users can adjust their traffic inputs or QOS demands to respond to feedback signals from the network about the current state of network resources. TCP applications use various congestion control algorithms such as slow-start [2] to adjust their input rates to the currently available bandwidth. The ATM Forum's ABR service allows flexible users to get loss guarantees if they respond to network feedback signals. Proposed layered MPEG coding schemes [12] would allow real-time VBR video users to obtain a QOS which would vary with the degree of network resource utilization.

Since it is already accepted that user responses can be automated using pre-programmed network interfaces, fairly sophisticated user behavior can be envisaged, and feedback strategies need not be limited by human user response times. The issue becomes one of choosing which objective function the feedback scheme should try to optimize. This objective is usually some form of efficiency.

We distinguish two different notions of efficiency: **network efficiency** refers to the utilization of network resources, such as bandwidth and buffer space; **economic efficiency** refers to the relative valuations users attach to their network service. If a network can maintain a target level of service while minimizing the resources needed to provide this service, we say that its operation is network efficient. If no user currently receiving a particular QOS values it less than another user who is being denied that QOS, we say the operation is economically efficient.

3.1 Price as a form of feedback

Adaptive users can help to increase network efficiency if they are given appropriate feedback signals. When the network load is high, the feedback should discourage adaptive users from inputting traffic; when the load is low, the feedback should encourage these users to send any traffic they have ready to transmit. In this way many of the congestion problems that occur if the offered load is fixed can be avoided. One possible feedback signal is a price based on the level of network load: when the load is high, the price is high, and when the load is low, the price is low (or zero).

Similarly, by associating a cost measure with network loading, all users can be signaled with the prices necessary to recover the cost of the current network load. Price-sensitive users — those willing and able to respond to dynamic prices — increase the overall economic efficiency by choosing whether or not to input traffic ac-

cording to their individual willingness to pay the current price. Users who value immediate network service more will choose to transmit, while those who value it less will wait for a lower price.

Price signals thus have the potential to increase both network and economic efficiency, though whether a particular pricing scheme increases either notion of efficiency depends on the implementation. The charge to an ATM network user could have several components, such as a connection fee, a charge per unit time or per unit of bandwidth, premium charges for certain services, and so on. In this paper, we focus attention on only one type of pricing: a responsive component which varies with the state of network congestion. *By responsive pricing we do not mean a charge which counts the number of bytes or cells regardless of the network conditions.* On the contrary, we propose charging only when network congestion indicates that some users may be experiencing QOS degradation, with the size of the charges related to the degree of congestion. If the network is lightly loaded and all users are getting acceptable QOS, the responsive prices would be zero.

4 Modeling user adaptation to feedback

We propose taking advantage of the adaptability of certain user types [4], [9] to improve efficiency over a short time horizon. We have modeled two types of adaptive user:

- *Inelastic.* An inelastic application requires a delay guarantee, but can tolerate loss and is adaptive. For example, this might be the second level of a two-level video codec.
- *Elastic.* This type of user waits until feedback from the network indicates that they can input traffic, then transmits and requires that their cells are not lost in the network. A possible example of an elastic user type would be a non-real-time data transfer with no ARQ capability, where already-transmitted cells are not buffered at the sender.

With these two types we have heterogeneity across applications. We are also able to model within-type heterogeneity by specifying users of a given type who value their QOS differently.

4.1 Responsive pricing schemes

In our simulations we have been comparing three different schemes for allocating a simple network's resources. The first is a conventional approach that makes no use of feedback and user adaptation. The second is a closed-loop form of feedback and adaptation; the third is a closed-loop variation we call "tight loop" because it shortens the delay in the control loop.

4.1.1 No feedback

Our proposal to improve efficiency through involving the users in network control is somewhat controversial. Most in the network engineering community seem to assume that a network will be tuned for efficiency given a set of admitted user connections. The only room for interaction with the users in such a setting is through the connection setup negotiation. Therefore, as a baseline, we simulate a network that does not provide feedback.

4.1.2 Closed-loop feedback

This feedback network uses a simple scheme [7], [10]. The network state is measured by buffer occupancy at the network access points. This occupancy is converted into a price per cell, which is then transmitted back to each active adaptive application. The applications then decide on how many cells to send during the next period.

In this network, users send some cells in period t, and network performance is affected by the aggregate number of cells received during the period. At the end of period t the network sends a signal back to users based on the network utilization in period t. Users then decide how many cells to send in period $t+1$, based on their observed period t performance and their application requirements.

4.1.3 Smart market pricing

A "smart market" approach to adaptation has been proposed in [3]. A user sends cells to the network interface which include in each header an indication of how much the user is willing to pay to get that packet into the network in the current period. Then, during the pricing period, the network interface sorts the "bids" on the incoming packets, and admits to the network only as many as it can accommodate without degrading performance below some bound.

The network interface admits packets in descending order of their bids. Users are charged not the amount that they declared they were willing to pay, but the value of the *maximum* bid on a cell that is *not admitted* to the network. Thus, users pay just the congestion cost (the amount that the highest-value denied packet would have paid for immediate transport) and they get to keep all of the excess value that they attribute to delivery. This form of pricing by auction has several nice properties, described in [3].

4.2 Simulation details

4.2.1 Inelastic user

This user has a delay bound on the traffic, but can tolerate only sending a fraction of the cells that are ready to go in the period in question. We assume that cells not sent in the period are useless to the user and are discarded.

We assume that the user has a concave benefit function, $benefit_t(X)$, on the number of cells sent, X. This is fixed for each period but is allowed to vary from one period to the next. If the user sends a number of cells X_t in period t, then the benefit to the user is $benefit_t(X_t)$. How the user decides on the number of cells to send in period t depends on the price given to the user, P_t:

$$\frac{\partial benefit_t(X)}{\partial X} = P_t \qquad (1)$$

The user solves this equation for X_t, the number of cells sent in period t.

4.2.2 Elastic user

At the start of the connection, the elastic user guesses the average price and the average PCR over the lifetime of the connection; these are denoted by $P^\star$ and $PCR^\star$ respectively. These could be based on past experience or given by the network. We assume that $P^\star > 0$ and $PCR^\star > 0$. An elastic user cares about the ultimate delay in sending a completed file. This user is flexible, so their send decision depends on the expected effect of sending some cells now on overall delay. We assume that the user thinks that the price in the next period will be the average price $P^\star$, regardless of the actual price now. We further assume that their forecast future PCR is the average $PCR^\star$. This is the simplest possible user forecasting model.

The benefit that the user places on a connection used to transmit x cells is as follows:

$$ben(x, v) = v - h(d(x)) \qquad (2)$$

where $d(x)$ is the delay for x cells, h is the cost function for the delay, and v is the value the users would get if all their cells were delivered immediately (i.e. $h(0) = 0$).

The model for the delay cost for the user, $h(d)$, has the following attributes: $h(0) = 0; \quad h(\infty) = v$. An example of this could be a function that approaches v asymptotically like $h(d) = v(1 - e^{-d})$ which is shown in Figure 2.

By comparing the expected benefits of sending now versus later using these simplistic forecasting models, the send decisions for an elastic user can be derived. The details are lengthy and are contained in [6].

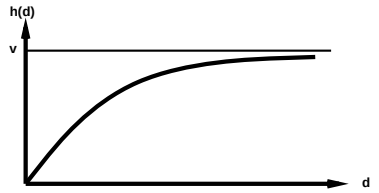

Figure 2: Possible Delay Cost Function

4.2.3 Preliminary simulation results

To give a sense of the gains that are possible with responsive pricing, we offer some preliminary results that compare no feedback to the closed-loop pricing scheme.

The simulated network is a high-speed ATM 155 Mbps link shared by inelastic and elastic users. Video sources are modeled as inelastic users; data sources are modeled as elastic users. The network and source models were simulated using SES/*workbench* [13], a discrete-event simulator that allows hardware and software simulation.

The model takes in cells over a pricing period and gives a price to all the sources sharing the link. The price reflects the congestion (if any) in the buffer and hence on the virtual path. The pricing period is short compared to the video frame time: a value of about 0.05 of a frame time was chosen. To achieve feasible run times we neglect cell scale effects.

Source Type		% Loss	User Value
Unpriced	Inelastic	0	240
	Elastic	30.4	146
	Combined	**19.1**	**386**
Priced	Inelastic	4.4	239
	Elastic	0.1	204
	Combined	**1.7**	**443**
Priced vs. Unpriced		**-91.0%**	**+14.8%**

Figure 3: Performance and Economic Gains from User Feedback (preliminary)

A fixed number of inelastic sources are always admitted and active. In addition, a number of elastic applications are active at any given time. To simplify, a random number of new elastic connections are initiated each period, with a random number of cells to be delivered. Thus, the number of continuing connections in any period after the first will be random (as varying numbers of connections are completed). The distribution of elastic message sizes is chosen so that the average load added to the network each period is within the tolerance for a reasonable call admission algorithm. However, sometimes the amount of active elastic traffic will be large, and the network will suffer some performance difficulties (cell delays and losses).

In the simulations we have generated 20 video sources with random frame sizes to represent the inelastic traffic, and a random number of between 1 and 39 elas-

tic data sources with random frame sizes. Our network experiences 80% utilization on average. With closed-loop pricing, packet loss drops from 19% to under 2%, while the net benefits perceived by the users increase by nearly 15%.

We are also developing simulations for smart-market pricing feedback. The user send decision is more complex to model, since the cutoff price is not known before the user attaches a bid to a cell. Therefore the bid is set to tradeoff the value of getting a cell delivered immediately against the value of getting it delivered at a later, uncertain time. Thus we model the user's probability distribution over future cutoff prices in a way that is consistent with the system's actual behavior. We also model how the user expects her future bidding behavior to change if her current bid is too low to obtain service.

5 Discussion and conclusions

Many proposals have been made to incorporate feedback into network control and resource allocation schemes, such as TCP congestion control or ABR service in ATM networks. We suggest taking these proposals one step further by including the user in the feedback control loop. Responsive pricing gives users an incentive to consider the effects of their usage on other users. Price-sensitive users adjust their traffic inputs by balancing the price and their own valuation of network service.

Responsive pricing only charges users per cell when congestion is experienced. Users unwilling to pay the current price can simply wait for it to drop to an acceptable level (if the price never goes to zero, the network is always somewhat congested and capacity expansion is indicated).

Responsive pricing directly links network control to user valuation of the service, and accommodates users who value some traffic more than others regardless of the application types. Simulations show that it is possible for the network to carry more traffic, and for users to value the traffic more highly. Many objections to responsive pricing have been raised. For a discussion on questions of feasibility, overhead, profit, fairness and other issues we refer to [5] and [8].

Acknowledgments

The second author was visiting the California Institute of Technology when this paper was written and thanks Prof. R. J. McEliece and Pacific Bell for their support. The third author acknowledges financial support from NSF grant SES-93-20481.

References

[1] M. de Prycker, *Asynchronous Transfer Mode : Solution for Broadband ISDN*, 2nd Ed., Ellis Horwood, 1993.

[2] V. Jacobson, 'Congestion Avoidance and Control', *Proc. ACM SIGCOMM '88 Symp.*, Sept. 1988.

[3] J. MacKie-Mason and H. Varian, 'Pricing the Internet,' in *Public Access to the Internet*, B. Kahin and J. Keller, eds., Prentice-Hall, Englewood Cliffs, NJ, 1995. Available from URL `http://www.spp.umich.edu/spp/papers/jmm/Pricing_the_Internet.pdf`

[4] J. MacKie-Mason, J. Murphy and L. Murphy, 'ATM Efficiency Under Various Pricing Schemes,' presented at *3rd International Conference on Telecommunications Systems Modelling and Analysis*, Nashville, TN, March 1995.

[5] J. MacKie-Mason, L. Murphy and J. Murphy, 'The Role of Responsive Pricing in the Internet', *Journal of Electronic Publishing : Special Issue on Internet Economics*, University of Michigan Press (to appear).

[6] J. Murphy, *Resource Allocation In ATM Networks*, Ph.D. thesis, School of Electronic Engineering, Dublin City University, Ireland, 1995.

[7] J. Murphy and L. Murphy, 'Bandwidth Allocation By Pricing In ATM Networks', IFIP Transactions C : Communication Systems, No. C-24, 1994, p. 333-351. Available from URL `http://www.eeng.dcu.ie/~murphyj/band-price/band-price.html`

[8] L. Murphy and J. Murphy, 'Feedback and Pricing in ATM Networks', *Proc. IFIP TC6 Third Workshop on Performance Modelling and Evaluation of ATM Networks*, Ilkley, England, July 1995, p. 68/1-68/12. Available from URL `http://www.eeng.dcu.ie/~murphyj/brad-price/brad-price.html`

[9] L. Murphy and J. Murphy, 'Pricing for ATM Network Efficiency', *Proc. 3rd International Conference on Telecommunication Systems Modelling and Analysis*, Nashville, TN, March 1995, p. 349-356. Available from URL `http://www.eeng.dcu.ie/~murphyj/atm-price/atm-price.html`

[10] J. Murphy, L. Murphy and E. C. Posner, 'Distributed Pricing For Embedded ATM Networks', *Proc. International Teletraffic Congress ITC-14*, Antibes, France, June 1994, p. 1053-1063. Available from URL `http://www.eeng.dcu.ie/~murphyj/dist-price/dist-price.html`

[11] P. Newman, 'Traffic Management for ATM Local Area Networks', IEEE Communications Magazine, pp 44-50, August 1994.

[12] P. Pancha and M. El Zarki, 'MPEG Coding For Variable Bit Rate Video Transmission', IEEE Communications Magazine, pp 54-66, May 1994.

[13] SES, *SES/workbench Reference Manual*, Release 2.1, Feb. 1992.