A Pricing Model for Networks with Priorities*

Hong Jiang, Ikhlaq Sidhu & Scott Jordan+
jhong@hoserve.att.com, isidhu@usr.com & scott@eecs.nwu.edu
Bell Laboratories, U. S. Robotics & Northwestern University

Abstract

In this paper, a centralized resource allocation model based on priority schemes is first proposed for best-effort networks in which traffic from all priorities are fully multiplexed. Then a priority pricing scheme is used to distribute the resource allocation process. Users and the network exchange their characterizations on QoS, resource demand and prices via contract negotiation. The contract negotiations take place among three network layers: users, priority classes and the network, in a distributed and dynamic fashion.

Key words: connection establishment, contract negotiation, pricing and resource allocation.

1 Introduction

Resource allocation and pricing for high speed networks have attracted much research interests [1] [2] [3] [5] [6] [9] [10] [11] [12] in recent years. The connection establishment process described in [5] [6] relies on resource reservation as a mechanism to guarantee desired QoS of each connection. This type of resource management demands significant effort in user traffic characterization and admission control. Most importantly, the degree of statistical multiplexing is only limited to virtual circuits within a virtual path since further multiplexing requires a more complex resource management procedure. These drawbacks of resource reservation lead to a very different type of resource management often used by networks that provide best-effort services, in which the networks resort to other mechanisms to enhance network performance and provide diverse QoS: Priority schemes are a prime example of this. The trade-offs between resource reservation and priority schemes are: Best-effort networks with priorities do not guarantee the QoS, even though a network could use priorities to guarantee the QoS by controlling the amount of traffic admitted into each priority level. The characterization or the measurement of various QoS in terms of user demand could be challenging for a priority network; however, the advantage is that traffic sources are fully multiplexed among all priority classes.

In this paper, we adopt a connection establishment process using priorities to manage network resources. The connection establishment process consists of two separate stages: user and network characterizations, and then contract negotiation. Finally, a centralized mathematical model to maximize network economic efficiency is formulated, and the methods using pricing to distribute this maximization problem are outlined.

* This work was completed at Northwestern University.

+ Corresponding author: Tel: 847-467-1243, Fax: 847-491-4455, Department of Electrical Engineering & Computer Science, Northwestern University, 2145 Sheridan Rd., Evanston, IL 60208-3118

1.1 User Characterization

A user's objective here is to maximize his consumer surplus, which is equal to the benefit minus the cost of a service by choosing the appropriate demand for each priority class.

Similarly, a user's benefit of a service is defined in terms of demand and the QoS. In existing models [5] [6], it is assumed that users do not have the flexibility to choose a circuit bundle (or QoS). For instance, video users will use the video circuit bundles and voice users will only use the voice circuit bundles. The main argument for this assumption is that a video user usually will not choose a voice circuit bundle to send his video stream because there might be a mismatch between the voice circuit bundle's QoS and the desire QoS of his video application.

However, in a best-effort network with priorities, the QoS of the priority classes are not guaranteed and are varying in time depending on user demand. A user can choose any one of the M priority classes or he can mix different priority classes to transmit his traffic stream. For instance, a video user may compress a video stream at two levels, and he might want to use two priority classes to transmit the stream. In case of network congestion, the lower priority packets will be dropped first. However, for non-real-time services, if each data packet is equally important, then the user might only choose one out of M priority classes to send his data stream.

By incorporating user demand cross elasticity[1], better economic efficiency could be achieved. For example, when the QoS of a priority class becomes unsatisfactory to a user, he may decide to send fewer packets to this priority class and more packets to others. Hence, user benefit functions are defined as $ben(QoS_1, \lambda_{i1}, QoS_2, \lambda_{i2}, \ldots QoS_M, \lambda_{iM})$ to reflect the demand cross elasticity across priority classes, where λ_{ij} is user i's demand of priority class j. The definition of the benefit function shows that a user's benefit of a service depends on both the demand and the QoS of all priorities.

Consider a network with two priority classes, a user benefit function could look like the one in Figure 1 with QoS_1 and QoS_2 fixed. Certainly such a definition of benefit function is very broad and ideal, and could be difficult for a user to obtain. In practice, a user can choose a simplified version of the benefit function, or even a discrete benefit function as an approximation.

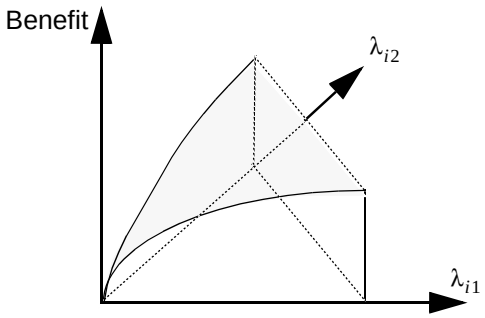

Figure 1 Benefit vs. demands

1.2 Network Characterization

The network's objective is to maximize its total user benefit by optimally allocating resources among priority classes and by choosing the best combination of QoS.

The network charges users on a per packet basis. Thus the total charge to a user for a connection would be $\sum_j P_j \lambda_{ij}$, where P_j is the price charged for priority j packets. The caveat of this pricing base has been discussed in [13] since some users with bursty traffic streams might not be charged enough to offset the negative externalities, such as prolonged delay and additional packet loss. Therefore, there might be some loss of network efficiency here. However, the benefit of complete sharing among all users might well compensate for the loss.

In high-speed networks, multiple services with diverse QoS require that the networks also provide diverse QoS. QoS in a best-effort network is often defined by mean delay and loss probability. Figure 2 shows the approximate QoS requirements of typical applications: Voice applications are very delay-sensitive but relatively loss-insensitive, and the data applications such as e-mail are loss-sensitive but relatively delay-insensitive. Hence, the networks need to decide a priority policy that can best meet users' QoS needs among many priority policies. There has been vast literature on designing priority policies for computer networks [1] [4] [7] [14].

1. Cross elasticity here means that users have the flexibility to choose among all priority classes.

After the network determines a priority policy, it needs to communicate the policy to its users so that they can take advantage of such a policy and determine their demand for each priority class. On the network's side, given user demand of each priority class, the network needs to characterize the QoS of each priority class and its sensitivity in response to user demand. If the QoS and its sensitivity can be measured by the network, then the tasks of user traffic characterization can be greatly reduced.

2 Contract Negotiation

A contract negotiation process is a distributed resource allocation and efficiency maximization process. Pricing is important in giving users incentives to choose the right QoS and demand for both himself and the network. Just imagine, if all four priority classes have the same price, then every user will choose the class with the best QoS, thus rendering the priority policy useless.

In the contract negotiation, the network must first decide a priority policy and inform its users of the policy so that users can take full advantage of the priorities by tailoring their demand accordingly. Second, the network and the users reach an agreement through an iterative process. The network first sets the prices for all priority classes, and declares the projected QoS of each class. The users respond with their demand to each priority class based on the information given by the network. The network then computes the QoS given the aggregate demand to each priority class, and announces a new set of prices and QoS. This iterative process continues until the projected QoS truly matches the QoS provided to the users, and until total user benefit is maximized. As user benefit functions and demand change over time, the negotiation takes place at a granularity comparable to the change of users' characteristics.

Before mathematical models are established for the network maximization problem, the notations and the assumptions are stated as follows. Without loss of generality, four priority classes are still used in the model.

λ_{ij} : the arrival rate or demand of user i to priority class j.

$\lambda_j = \sum_i \lambda_{ij}$: the aggregate arrival rate or demand to priority class j.

D_l : the mean delay of priority class l projected and announced by the network.

L_k : the loss probability of priority class k projected and announced by the network.

$ben_i(D_1, L_1, D_2, L_2, D_3, L_3, D_4, L_4, \lambda_{i1}, \lambda_{i2}, \lambda_{i3}, \lambda_{i4}) = ben_i(\{D_l\}, \{L_k\}, \{\lambda_{ij}\})$: the benefit function of user i in terms of mean delay, loss probability, and demand to each priority class.

$B = \sum_i ben_i$: the aggregate benefit of all users.

$FD_l(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = FD_l(\{\lambda_j\})$: the mean delay of priority class l as a function of demand to each priority class.

$FL_k(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = FL_k(\{\lambda_j\})$: the mean loss probability of priority class k as a function of demand to each priority class.

α_l, β_k : the Lagrangian multipliers of constraints (2) and (3).

P_j : the price per packet charged by priority class j to its users.

Assume that the user benefit functions are differentiable and jointly concave in λ_{ij}, D_l and L_k, and that they are increasing in λ_{ij} and decreasing in D_l and L_k. Assume also that the mean delay functions FD_l and the mean loss probability functions FL_k are differentiable, jointly convex and increasing in λ_j for all priority class j.

First, a centralized mathematical model is formulated, where only a single node is considered. Pricing is introduced to decompose the centralized model into a distributed model. Finally, the negotiation process based on the distributed model is outlined.

2.1 A Centralized Model

The network's objective is to maximize its total user benefit subject to network constraints.

$$\max_{D_l, L_k, \lambda_{ij}} \quad ben_i(\{D_l\}, \{L_k\}, \{\lambda_{ij}\}) \tag{1}$$

subject to constraints:

$$D_l = FD_l(\{\lambda_j\}) \qquad \forall l \tag{2}$$

$$L_k = FL_k(\{\lambda_j\}) \qquad \forall k \tag{3}$$

$$D_l, L_k, \lambda_{ij} \geq 0 \qquad \forall (l, k, i, j) \tag{4}$$

The Lagrangian function of the maximization problem is:

$$G = \sum_i ben_i(\{D_l\}, \{L_k\}, \{\lambda_{ij}\}) + \sum_l \alpha_l (D_l - FD_l(\{\lambda_j\})) + \sum_k \beta_k (L_k - FL_k(\{\lambda_j\})) \tag{5}$$

Since the objective function is concave, and the delay and loss functions are convex, the Lagrangian function is jointly concave in D_l, L_k and λ_j. Again, this is a concave program, Kuhn-Tucker conditions [8] ((6)-(14)) are sufficient and necessary for the global optimal solution.

$$\frac{\partial G}{\partial \lambda_{ij}} = \frac{\partial}{\partial \lambda_{ij}} ben_i - \sum_j \alpha_l \frac{\partial FD_l}{\partial \lambda_{ij}} - \sum_k \beta_k \frac{\partial FL_k}{\partial \lambda_{ij}} \le 0$$

$$= \frac{\partial}{\partial \lambda_{ij}} ben_i - \sum_j \alpha_l \frac{\partial FD_l}{\partial \lambda_j} - \sum_k \beta_k \frac{\partial FL_k}{\partial \lambda_j} \qquad \forall (i,j) \tag{6}$$

$$\lambda_{ij}\left(\frac{\partial}{\partial \lambda_{ij}} ben_i - \sum_j \alpha_l \frac{\partial FD_l}{\partial \lambda_j} - \sum_k \beta_k \frac{\partial FL_k}{\partial \lambda_j} \right) = 0 \qquad \forall (i,j) \tag{7}$$

$$D_l - FD_l(\{\lambda_j\}) = 0 \qquad \forall l \tag{8}$$

$$L_k - FL_k(\{\lambda_j\}) = 0 \qquad \forall k \tag{9}$$

$$\frac{\partial G}{\partial D_l} = \sum_i \frac{\partial}{\partial D_l} ben_i + \alpha_l \le 0 \qquad \forall l \tag{10}$$

$$D_l \left(\sum_i \frac{\partial}{\partial D_l} ben_i + \alpha_l \right) = 0 \qquad \forall l \tag{11}$$

$$\frac{\partial G}{\partial L_k} = \sum_i \frac{\partial}{\partial L_k} ben_i + \beta_k \le 0 \qquad \forall k \tag{12}$$

$$L_k \left(\sum_i \frac{\partial}{\partial L_k} ben_i + \beta_k \right) = 0 \qquad \forall k \tag{13}$$

$$\lambda_{ij}, D_l, L_k \ge 0 \qquad \forall (l,k,i,j) \tag{14}$$

Note that out of Kuhn-Tucker conditions, there are $4I + 8 + 8$ number of equations (not including the complimentary slackness conditions and (14)), where I represents the total number of users using the network. There are also the same number of variables in these equations: D_l, L_k, λ_{ij}, α_l and β_k. Therefore, if a global optimal solution exists, then it can be found by simultaneously solving these equations. Again, solving such a problem could be computationally prohibitive for a moderately-sized network. Hence, we investigate how to decentralize the model in an attempt to obtain the same optimal total user benefit achieved by the centralized model.

2.2 A Distributed Model

In the distributed model in [5] [6], since the network is structured according to four layers (users, circuit bundles, virtual paths, and physical trunks), the negotiations naturally occur among these layers. Since currently only a single node is considered, the natural structure of network has only three layers: users, priority classes and the node. The analogy between the two models is that priority classes here are similar to the circuit bundles that differentiate themselves from others with QoS. The difference between priority classes and circuit bundles resides in the fact that the QoS is guaranteed and fixed for each circuit bundle, while it is varying in time for a priority class.

A distributed contract negotiation could proceed in the fashion illustrated by Figure 2. Each user negotiates with each priority class for the amount of traffic he is transmitting to this priority class given the price of the priority class and the priority-level parameters. Each priority class negotiates with the node for the amount of traffic allocated by the node given the prices. These levels of negotiations could take place at different time scales, and the contract parameters such as QoS, resource demand and prices vary dynamically.

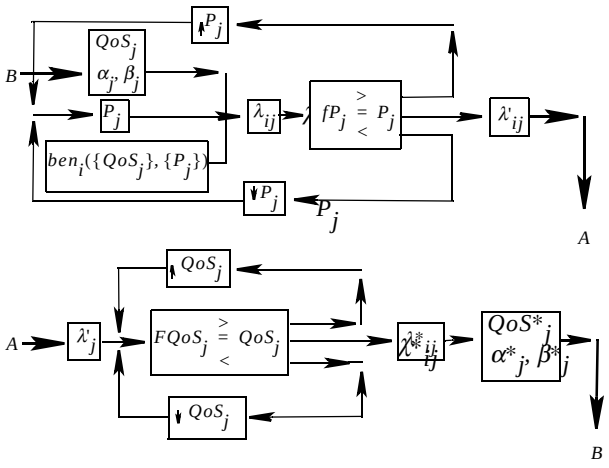

Figure 2 Distributed contract negotiation for priority networks

2.3 User-priority class negotiation:

For this level of negotiation, the parameters at a priority class level, such as the QoS (D_l, L_k) and Lagrangian multipliers (α_l, β_k), are considered fixed. Users determine their demand to priority class j, and the priority class in turn sets the price given user demand.

User i maximizes his consumer surplus:

$$\max_{\lambda_{ij}} \; ben_i(\{D_l\}, \{L_k\}, \{\lambda_{ij}\}) - \sum_j P_j \lambda_{ij} \tag{15}$$

subject to constraint: $\lambda_{ij} \ge 0$.

The optimal solution for maximal consumer surplus satisfies the following conditions:

$$\frac{\partial}{\partial \lambda_{ij}} ben_i - P_j \le 0 \qquad \forall j \tag{16}$$

$$\lambda_{ij}\left(\frac{\partial}{\partial \lambda_{ij}} ben_i - P_j\right) = 0 \qquad \forall j \tag{17}$$

Equations (16) and (17) together indicate that if there exists a demand λ_{ij}^* for priority class j such that $\frac{\partial}{\partial \lambda_{ij}^*} ben_i - P_j = 0$, then λ_{ij}^* is the optimal demand for user i. If no such positive demand can be found to satisfy the equation, then the optimal demand $\lambda_{ij}^* = 0$.

Priority class j's strategy could be derived from equations (6) and (7). The equations suggest that if priority class j sets the price to be $P_j = \sum_j \alpha_l \frac{\partial FD_l}{\partial \lambda_j} + \sum_k \beta_k \frac{\partial FL_k}{\partial \lambda_j}$ given the fixed QoS (D_l, L_k) and the Lagrangian multipliers (α_l, β_k), then it can induce the users to behave in a socially optimal way. $\frac{\partial FD_l}{\partial \lambda_j}$ and $\frac{\partial FL_k}{\partial \lambda_j}$ are the system-dependent sensitivity characterization of the QoS in terms of the aggregate demand for priority class j. An equilibrium point is found when (6) and (7) are satisfied.

To find such an equilibrium, first consider an iterative negotiation process between users and priority class j with the assumption that the prices of other priority classes except for priority class j have already reached the optimal points.

1. The priority class first announces a price $P_j^{(1)}$.
2. Users respond with their demand. The priority class then obtains the aggregate demand $\lambda_j^{(1)}$.
3. The priority class computes a new price $fP_j^{(1)} = \sum_j \alpha_l \frac{\partial FD_l}{\partial \lambda_j^{(1)}} + \sum_k \beta_k \frac{\partial FL_k}{\partial \lambda_j^{(1)}}$. If $fP_j^{(1)} = P_j^{(1)}$, then the equilibrium price P_j^* is found.
4. If $fP_j^{(1)} \neq P_j^{(1)}$, the priority class would choose a new price $P_j^{(2)}$ between the values of $P_j^{(1)}$ and $fP_j^{(1)}$. Then repeat steps 1, 2 and 3.

Theorem 1:

1) There exists an equilibrium price P_j^* for priority class j such that equations (6) and (7) are satisfied.

2) In the above iterative process, P_j^* is between the values of $P_j^{(1)}$ and $fP_j^{(1)}$.

Proof:

1) Due to the concavity and convexity of benefit function and QoS functions, $\frac{\partial}{\partial \lambda_{ij}} ben_i$ is decreasing and $\sum_j \alpha_l \frac{\partial FD_l}{\partial \lambda_j} + \sum_k \beta_k \frac{\partial FL_k}{\partial \lambda_j}$ is increasing since α_l and β_k are non-negative. Only two types of users are possible. Type n users can not afford the service, thus their demand $\lambda_j^* = 0$. Type m users choose their demand so that equation $\frac{\partial}{\partial \lambda_{ij}} ben_i - \sum_j \alpha_l \frac{\partial FD_l}{\partial \lambda_j} - \sum_k \beta_k \frac{\partial FL_k}{\partial \lambda_j} = 0$ can be satisfied (Figure 3). If the aggregate demand $\lambda_j = 0$, then $P_j^* = 0$ by assumption. Therefore, there exists an equilibrium price point.

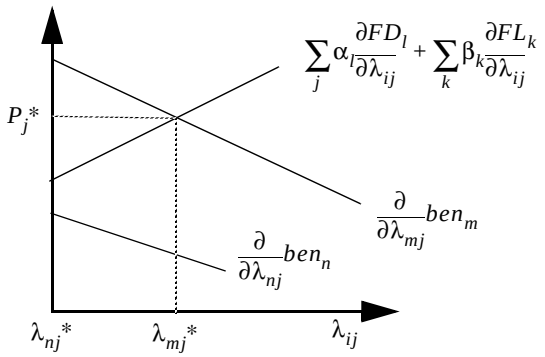

Figure 3 An equilibrium point

2) If $P_j^{(1)} \ge P_j^*$, then $\lambda_j^{(1)} \le P_j^*$ due to the concavity of benefit functions. Thus $fP_j^{(1)} \ge P_j^*$ due to the convexity of $P_j = \sum_j \alpha_l \frac{\partial FD_l}{\partial \lambda_j} + \sum_k \beta_k \frac{\partial FL_k}{\partial \lambda_j}$. Hence $fP^{(1)} \le P_j^* \le P_j^{(1)}$. Similarly, if $P_j^{(1)} \le P_j^*$, then $fP^{(1)} \ge P_j^* \ge P_j^{(1)}$. Q.E.D

Even though Theorem 1 does not guarantee that the iterative process will converge to the equilibrium price, it points out the right direction for the price setting.

Complexity arises when all four priorities are considered simultaneously during an iterative process. It is not clear how the iterative processes of

all four classes should proceed. More research needs to be done to study the interactions and convergence among these priority classes.

2.4 Priority classes-the node negotiation:

The negotiation between the priority classes and the node results in the optimal solutions for class-level parameters, such as QoS and the prices charged to the priority classes by the node. The node's strategy could be derived from (8)-(9). The node maximizes its total user benefit subject to the delay and loss constraints ((8)-(9)). Priority classes maximize their consumer surplus. Note that in equations (10)-(13), Lagrangian multipliers (α_l, β_k) are positive. Hence, priority class j is compensated the amount equal to $\alpha_j D_j + \beta_j L_j$ for the delay and loss. Thus the consumer surplus here is the sum of the benefit of the class plus the negative cost (compensation) provided by the node. Each priority class first obtains its benefit sensitivity with respect to its loss probability and mean delay. Equations (10)-(13) are the strategy for the priority classes to maximize their consumer surplus.

The distributed process above is only an outline. Many issues need to be resolved, especially the design of iterative processes deserves further study.

3 Conclusion

A connection establishment process is outlined for best-effort networks with priorities, where priority pricing is introduced to enhance network economic efficiency. Since the QoS of best-effort networks cannot be guaranteed, both the optimal QoS of and user demand for each priority class need to be determined by the network. It is shown that the optimal price charged by a priority class to its users equals the weighted sum of delay and loss sensitivities with respect to its demand. The optimal prices for delay and loss charged by the node to a priority class are negative and are equal to the sensitivity of the aggregate benefit of the priority class with respect to delay and loss respectively. Based on these optimal conditions for maximal network efficiency, a framework of a distributed negotiation process is outlined.

References

[1] A. Y. Lin and J. A. Silveste. "Priority queueing strategies and buffer allocation protocols for traffic control at an ATM integrated broadband switching system." *IEEE Journal on Selected Areas in Communications*, 9(9): 1524-1536, December, 1991.

[2] F. P. Kelly. "Tariffs, policing and admission control for multiservice networks." *10th UK Teletraffic Symposium*, BT Laboratories, Martlesham Heath, April 1993.

[3] F. P. Kelly. "Routing in circuit-switched networks: Optimization, shadow prices and decentralization." *Advanced Applied Probability*, 20: 112-144, 1988.

[4] G. Chen and I. Stavrakakis. "ATM traffic management with diversified loss and delay requirements." *Proc. Infocom'96*, Vol. 3, pp1037-1044, San Francisco, CA, March 24-28, 1996.

[5] H. Jiang and S. Jordan. "The role of price in the connection establishment process." *European Transactions on Telecommunications*, 6(4):421-429, July-August 1995.

[6] H. Jiang and S. Jordan. "A pricing model for high speed networks with guaranteed quality of service," *Proc. Infocom'96*, Vol. 2, pp888-895, San Francisco, CA, March 24-28, 1996.

[7] H. J. Chao and N. Uzun. "An ATM queue manager handling multiple delay and loss priorities." *IEEE/ACM Transactions on Networking*, 3(6):652-659, December 1995.

[8] J. Franklin. "Methods of mathematical economics: linear and nonlinear programming: fixed-point theorems." *New York: Springer-Verlag*, 1980.

[9] J. K. Mackie-Mason and H. R. Varian. "Pricing the Internet", *Second International Conference on Telecommunication Systems Modeling and Analysis*, pp378-393, Nashville, Tennessee, March 24-27, 1994.

[10] J. Murphy and L. Murphy. "Bandwidth allocation by pricing in ATM networks." preprint.

[11] R. Cocchi, D. Estrin, S. Shenker and L. Zhang. "Pricing in computer networks: motivation, formulation, and example." *IEEE/ACM Transactions on Networking*, 1(6):614-627, December 1993.

[12] S. H. Low and P. P. Varaiya. "A new approach to service provisioning in ATM networks." *IEEE Transactions on Networking*, 1(3):547-553, 1993.

[13] S. Jordan and H. Jiang. "Connection establishment in high speed networks." *IEEE Journal on Selected Areas in Communications*, 13(7): 1150-1161, September, 1995.

[14] Y. Takagi, S. Kino, and T. Takahashi. "Priority assignment control of ATM line buffers with multiple QoS classes." *IEEE Journal on Selected Areas in Communications*, 9(7): 1078-1091, September, 1991.