Proceedings of the INFORMATIK 2001, Workshop on Advanced Internet Charging and QoS Technology (ICQT'01)

Optimal Provisioning and Pricing of Differentiated Services Using QoS Class Promotion*

Errin W. Fulp† and Douglas S. Reeves‡

†Department of Computer Science, Wake Forest University
Winston-Salem N.C. USA, email: fulp@wfu.edu
‡Department of Computer Science *and* Department of Electrical and Computer Engineering
N.C. State University, Raleigh N.C. USA, email: reeves@eos.ncsu.edu

Abstract: *This paper introduces a new method for optimally provisioning and pricing differentiated services, that maximizes profit and maintains a small blocking probability. Resources are provisioned per Quality of Service (QoS) class over the long-term (service level agreement duration), then priced based on user demand over the short-term. Unique to this method is the ability to dynamically promote traffic from one QoS class to a higher QoS class, based on estimated demand statistics. This additional flexibility encourages better short-term utilization of the classes, resulting in higher profits while maintaining a low blocking probability. Experimental results will demonstrate QoS class promotion can obtain higher profits, as compared to other provisioning and allocation methods.*

Keywords: *Internet differentiated services, pricing, provisioning, allocation, network QoS.*

1. Introduction

Currently the Internet provides only best-effort service with no Quality of Service (QoS) guarantees of packet delay, delay variation, or loss. Yet, this best-effort service is insufficient for an increasing number of applications (e.g. multimedia oriented). Differentiated Services (DiffServ) is one proposed enhancement to the Internet to provide reliable QoS in a scalable fashion [2]. Under this mechanism, a finite set of QoS classes are available to aggregate flows that traverse a DiffServ enabled network (DiffServ domain). A connection across a DiffServ domain would have a Service Level Agreement (SLA), which details the maximum bandwidth, QoS class, location (ingress and egress routers), cost, and the term (duration) [2, 22]. Different service classes are provided through proper resource provisioning and prioritization, where higher (more stringent QoS) service classes require more resources (e.g. bandwidth). Higher QoS classes would cost more, thus demand a higher price than lower QoS classes, per packet transmitted [23]. A DiffServ connection serves multiple users requiring the same QoS, and persists over a long period of time. In contrast, users require smaller bandwidth amounts for shorter periods of time. In this framework, resources must be provisioned per connection in the long term, then portions of the connection are allocated to users in the short term.

Determining the appropriate amount to provision and allocate is problematic due to the different time scales, multiple QoS classes, and the unpredictable nature of users. In this paper, we present a method that optimally provisions and allocates differentiated services based-on microeconomic theory. Since provisioning and allocation are interdependent, it is important to address these issues simultaneously. However, previous microeconomic-based research has only investigated these issues in isolation. It has been demonstrated that pricing is an effective method for achieving fair allocations as well as revenue generation [1, 4, 6, 10, 14, 18, 23]. However, these methods do not consider how to provision resources. Other work has investigated resource provisioning [3, 5, 12, 11, 19], but not resource allocation to individual users. In [8], a hierarchical model was introduced to provision and allocate

*This work was supported by DARPA and AFOSR (grant F30602-99-1-0540). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the DARPA, AFOSR or the U.S. Government.

DiffServ bandwidth. Bandwidth was provisioned in a wholesale market, then allocated (via pricing) to individual users in a retail market. Objectives of this method included maximizing profits and maintaining a low blocking probability. However, each class was considered independent and users did not change classes based on prices. Furthermore, the amount allocated to users was constrained by the SLA contract. However, if higher class connections are available that have the same ingress-egress pair, any spare capacity could be used to transmit lower QoS class traffic, possibly yielding higher profits.

In this paper, a method for provisioning and allocating DiffServ connections, similar to [8], will be presented; however, a hierarchical model is not necessary. We will assume the network manager can either create or purchase (in a wholesale market) DiffServ connections. Furthermore, *QoS class promotion* will be used to increase utilization and profit. Class promotion can occur when demand for a lower class is greater than the amount provisioned (SLA agreement). If higher class connections have bandwidth available and have the same ingress-egress pair, then lower class traffic can be sent using the higher class connection. This additional flexibility will result in better utilization of resources and higher profits, while maintaining a low blocking probability.

The remainder of this paper is structured as follows. Section 2 describes the general design of the DiffServ model, where interactions between individual users and the network manager are defined. Optimal strategies for bandwidth provisioning and allocation are then presented in section 3. This method uses QoS class promotion to maximize profits, while maintaining a low blocking probability. In section 4, the monetary advantage class promotion is demonstrated. Finally, section 5 provides a summary of the provisioning and allocation method and discusses some areas of future research.

2. Differentiated Service Model

An example DiffServ network that consists of users and a network manager is given in figure 1. As previously mentioned, a DiffServ connection represents a large amount of bandwidth over long periods of time [9]. The connection has an associated SLA that specifies the maximum bandwidth, QoS class, location (ingress and egress routers), cost, and term (duration) [2, 22]. While the network manager will be responsible for multiple DiffServ connections that have different ingress-egress routers, for brevity we will only consider connections that have the same

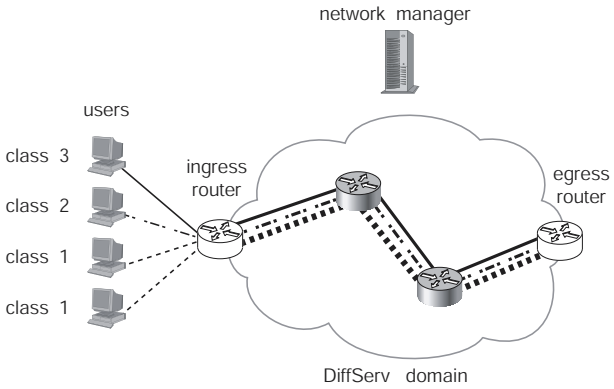


Figure 1: Example DiffServ enabled network consisting of users and a network manager. Three different DiffServ connections exist, each providing a certain QoS class.

ingress-egress routers (a requirement for QoS class promotion). Assume Q QoS service classes exist, where $i > j$ indicates i is a higher QoS class than j. A DiffServ connection is created or purchased (in a wholesale market [8]) by the network manager. The cost associated with the connection is denoted as $c_i(s_i)$, where i is the QoS class and s_i is the bandwidth. This cost function can be linear [8] or non-linear (discount per quantity bandwidth provisioned). Once the DiffServ connection is established, portions of the connection are sold (allocated) to individual users at a price.

The price of DiffServ bandwidth (charged to users) will be usage-based, where the user cost depends on the current price and the amount consumed. An important issue is the time scale associated with the price. For example, prices could remain fixed for long periods of time or continually change based on current congestion levels [7]. Spot market prices are updated over short periods of time to reflect congestion [7]. While this method does provide fair allocations under dynamic conditions, users can not accurately predict the cost of their sessions, due to possible price fluctuations. In contrast, fixed prices provide predictable costs; however, the user has no incentive to curtail consumption during peak (congested) periods. As a compromise, we will use prices based on slowly varying parameters such as Time of Day (ToD) statistics. As noted in [13, 16, 17], the aggregate demand for bandwidth changes considerably during certain periods of the day. A day will be divided into T equal length periods of time, where $t = 1, ..., T$. To provide predictability, these prices (next day) are known a pri-

ori by the users via a price-schedule $\{p_{i,t}\}$, where $p_{i,t}$ is the price of class i bandwidth during the t ToD period. The bandwidth of a DiffServ connection is sold on a first come first serve basis; no reservations are allowed. If the amount is not available at the beginning of the session, the user is considered blocked. However, users who can not afford $p_{i,t}$ are **not** considered blocked. Furthermore, users require a certain minimum QoS but can use any higher QoS class. This choice will be based on QoS class prices and the application requirements.

The network manager of a DiffServ enabled network is responsible for establishing DiffServ connections and allocating portions of the connection to individual users. Within this model, acquiring resources for a DiffServ connection is provisioning, while selling portions of the connection will be referred to as allocation. Primary goals of the network manager will be profit maximization and minimizing the blocking probability experienced by users.

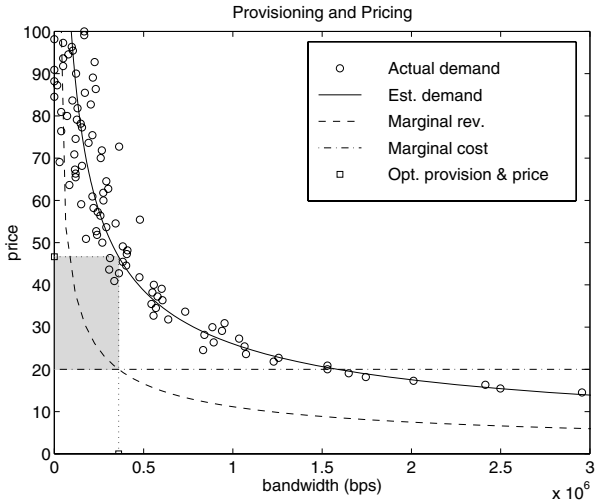


Figure 2: The network manager seeks the point where the marginal revenue equals the marginal cost. If optimal provisioning and pricing occurs, the amount of profit is given in the shaded area.

3. Optimal Resource Provisioning and Allocation

Assume multiple QoS classes belonging to the set Q are required between the same ingress-egress routers. Therefore, multiple DiffServ connections are required, each providing a different QoS class. For our discussion, i will uniquely identify a QoS class and DiffServ connection. The network manager is interested in maximizing the profit of all connections for this ingress-egress pair. This is done when the difference between the revenue generated minus the cost is maximized, as seen in the following formula,

$$\max \left\{ \sum_{i \in Q} \sum_{t=1}^{N} [r_i(x_{i,t}) - c_i(s_i)] \right\} \qquad (1)$$

The revenue generated by connection (QoS class) i during ToD period t is $r_i(x_{i,t})$ and is based on $x_{i,t}$ which is the user demand for this class. Note the profit maximization is over the SLA term (N consecutive ToD periods) and all QoS classes. Viewing this as an optimization problem, the first order conditions are

$$\sum_{i \in Q} \sum_{t=1}^{N} \frac{\partial r_i(x_{i,t})}{\partial x_{i,t}} = N \cdot \sum_{i=1}^{Q} \frac{\partial c_i(s_i)}{\partial s_i} \qquad (2)$$

Note the supply (SLA provisioning amount) for each class, s_i, is constant for each ToD period. The left-hand side of equation 2 is referred to as the marginal revenue, which is the additional revenue obtained if the network manager is able to sell one more unit of DiffServ bandwidth. The right-hand side of equation 2 is referred to as the marginal cost, which is the additional cost incurred. This relationship between revenue and cost can be depicted graphically, as seen in figure 2. A solution to the optimization problem exists, if the cost and revenue functions are continuous and convex. Therefore, to determine the appropriate provisioning amounts and prices these functions must identified.

The Cobb-Douglas demand function will be used to model aggregate user demand (multiple users seeking the same QoS class). The Cobb-Douglas demand function is commonly used in economics because it is continuous, convex, and has a constant elasticity [20]. A constant elasticity assumes users respond to proportional instead of absolute changes in price, which is more realistic. Therefore, this demand function is popular for empirical work. For example, the Cobb-Douglas demand function has been used to describe Internet demand in the INDEX Project, where user demand for different Internet access speeds was modeled [21]. Therefore, we believe this function is also appropriate for DiffServ. The Cobb-Douglas function has the following form,

$$x_{i,t} = \beta_{i,t} \cdot \prod_{j \in Q} p_{j,t}^{\alpha_{ij,t}} \qquad (3)$$

Where $p_{i,t}$ is the price for resource i during ToD t and the approximate aggregate wealth of users requiring class i is denoted by $\beta_{i,t}$. The cross-price elasticity during ToD t is $\alpha_{ij,t}$, if $j = i$ then $\alpha_{ij,t}$ is the own-price elasticity. Own-price elasticity represents the percent change in demand for class i in response to a percent change in the price of class i. The cross-price elasticity is the percentage change in the quantity demanded in response to a percent change in the price of another resource. If two resources are substitutes, the cross-price elasticity will be positive, since the price of one resource and the demand for another resource move in the same direction. The effect of cross-price elasticity is depicted in figure 3. If the cross-price elasticity is zero, then any change in price j will not affect the demand for resource i, as seen in figure 3(a). If the cross-price elasticity is positive, as seen in figure 3(b), the demand for i will change based on both prices. We will also assume that each user has a minimum desired QoS class i; therefore, $\alpha_{ij,t} = 0, \forall j < i$.

Given the aggregate demand function, the revenue earned is the price multiplied by the demand,

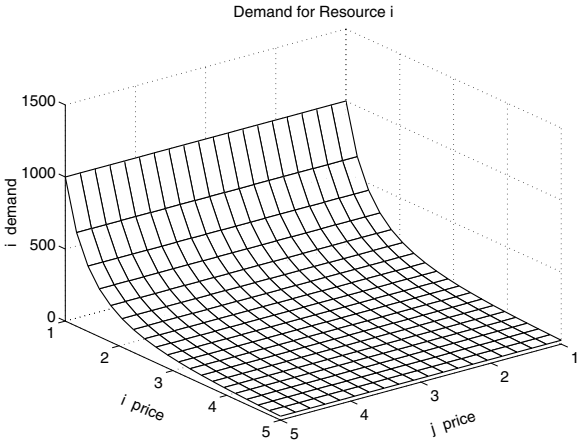
$$p_{i,t} \cdot x_{i,t} = \left(\frac{x_{i,t}}{\beta_{i,t} \cdot \prod_{j \in Q, j \neq i} p_{j,t}^{\alpha_{ij,t}}} \right)^{\frac{1}{\alpha_{ii,t}}} \cdot x_{i,t}$$

$$= x_{i,t}^{1+\frac{1}{\alpha_{ii,t}}} \cdot \beta_{i,t}^{\frac{-1}{\alpha_{ii,t}}} \cdot \left(\prod_{j \in Q, j \neq i} p_{j,t}^{\alpha_{ij,t}} \right)^{\frac{-1}{\alpha_{ii,t}}} \quad (4)$$

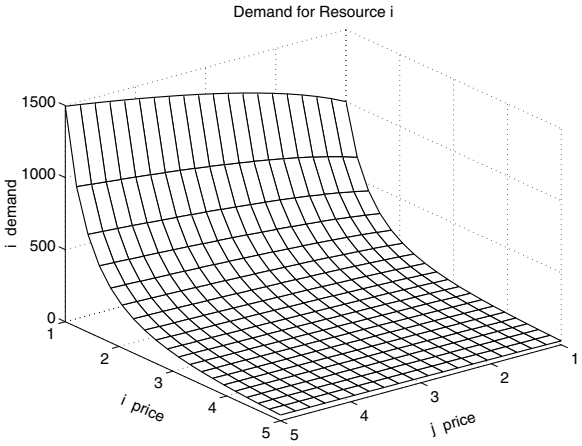
Taking the derivative of equation 4 with respect to demand yields the marginal revenue for ToD period t. Similarly, taking the derivative of the cost function yields the marginal cost. Substituting these values into equation 2 results in a system of equations can be solved for s_i (we seek the point where demand equals supply, therefore $x_i = s_i$) which is the appropriate amount to provision for QoS class i. Since the marginal equations (revenue and possibly cost) are non-linear, a direct solution can not be found. For this reason, gradient methods (e.g. Newton) can be used to determine the optimal provisioning amounts [15, 24]. Due to the time typically associated with negotiating a SLA [8], calculations can be performed off-line; therefore, convergence time is not critical.

3.1. Allocation per Time of Day

The previous section described a method for determining the appropriate amount of bandwidth to provision per QoS class. In this section, the appropriate amount to charge per class during a ToD period will be determined. These prices will form a price schedule, given to the users.



(a) Cross-price elasticity $\alpha_{ij} = 0$.



(b) Cross-price elasticity $\alpha_{ij} = 0.25$.

Figure 3: Demand for resource i as prices for resources i and j vary.

In [8], a method was presented to set the price for a QoS class and ToD period. The price was set based on the supply and the estimated demand function; however, each class was considered independent. Prices of different classes did not effect demand; although, such behavior is realistic. Furthermore, the amount allocated to users was constrained by the SLA contract. In contrast, we propose *QoS class promotion*, which is the adjustment of QoS class prices and allocations based on ToD demand. Class promotion can occur when demand for a lower class is greater than the amount provisioned (SLA agreement). If higher class connections have bandwidth available and have the same ingress-egress pair, then lower class traffic can be sent using the higher class connection. This additional flexibility will result in higher utilization of resources and higher profits, while maintaining a low blocking probability. Therefore, for each ToD period the network manager will maximize revenue across all QoS classes,

$$\max \left\{ \sum_{i \in Q} r_i(x_{i,t}) \right\} \tag{5}$$

$$\text{subject to:} \quad x_{i,t} \leq s_i + \sum_{\forall j > i} (s_j - x_{j,t})$$

$$\sum_{i \in Q} x_{i,t} \leq \sum_{i \in Q} s_i$$

The first constraint concerns the amount of bandwidth available to service class i, which is less than or equal to the provisioned amount plus any bandwidth available in any higher classes. The second constraint ensures the total amount allocated is no more than the total amount provisioned. This can be viewed as a constrained optimization problem. Again, the resulting system of equations are nonlinear and require gradient methods to find the appropriate allocation amounts and prices [15, 24]. The resulting values for $x_{i,t}$ are the optimal bandwidth amounts for each class per ToD. Based on these values, the optimal price for each class and ToD is given using equation 3.

The other goal for the network manager is to maintain a low blocking probability. Based on the optimal provisioning and pricing equations given in the previous sections, these values will result in supply equaling demand (as seen in figure 2). For that reason, the predicted blocking probability is zero. If the estimated demand is greater than the actual demand, the blocking probability is zero. However, if the estimated demand is less than the actual demand, then the blocking probability will be greater than zero. Therefore, a zero blocking probability depends on accurate demand estimation [8].

User Type	Wealth				Elasticity	
	$w_{i,1}$	$w_{i,2}$	$w_{i,3}$	$w_{i,4}$	α_{ii}	$\alpha_{\mathrm{AF,EF}}$
EF	500	5000	3000	5000	1.75	0
AF	500	10	3000	10	1.5	0.2

Table 1: User variable values used in the numerical example. Note α_{ii} is the own-price elasticity, while $\alpha_{\mathrm{AF,EF}}$ is the cross-price elasticity.

4. A Numerical Example

This section provides an example of the allocations and profit achieved using QoS class promotion. A comparison is made with allocating only the SLA provisioned amounts during each ToD period, as done in [8], which will be referred to as SLA-based allocation.

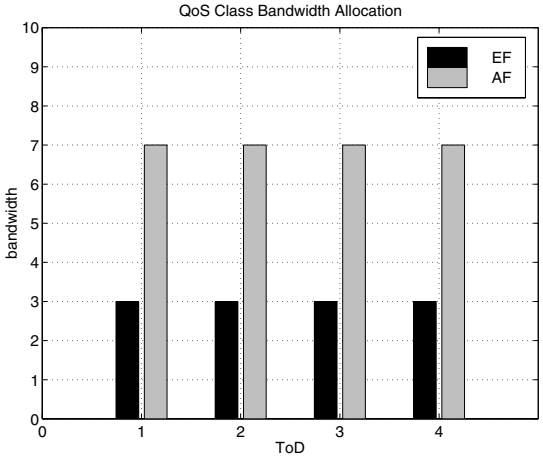
Assume two different DiffServ classes are required between a pair of ingress-egress routers. The classes are Expedited Forwarding (EF) and Assured Forwarding (AF), where EF is considered a higher QoS class than AF. Furthermore, assume the SLA term is four consecutive ToD periods[1]. The cost for each class was linear with respect to the amount provisioned. The EF class cost was 10 tokens per unit bandwidth, while the AF class had a cost of 5 tokens per unit bandwidth[2]. Users were distinguished based on the minimum QoS desired. One set of users required EF, while the other required at least AF. Note, AF traffic can be promoted to the EF class. Values for the wealth and elasticity for each set of users are given in table 1. As seen in the table, the aggregate wealth of each group changed per ToD, while the elasticities remained constant. Given these parameters, the optimal provisioning and allocation amounts were solved numerically for the SLA-based and QoS class promotion techniques.

Results are given in figure 4, where allocation amounts for the QoS classes are given for each ToD period. As seen in figure 4(a), the SLA provisioning amounts were 9 units for EF and 10 units for AF. Allocating bandwidth based on these values as done in [8] (no promotion), resulted in a total profit

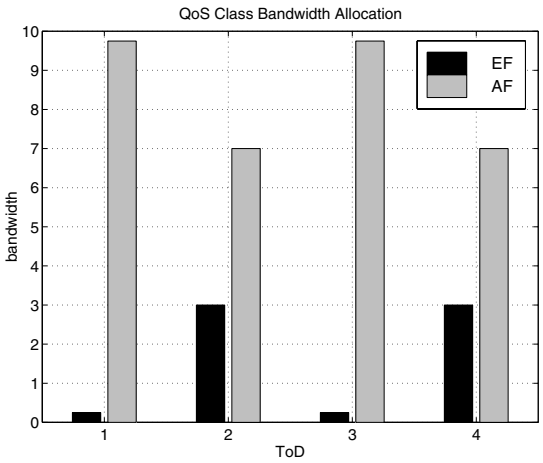
[1]The term of an actual SLA would be much longer, however length will not impact the results presented.

[2]Tokens were used as a generic currency, where one token had the value of one unit of bandwidth.

of 1358.86 tokens. Figure 4(b) shows the allocation of bandwidth using QoS class promotion. In the first ToD period, AF traffic was promoted (allocated amount increased) since these users can accept any QoS class and the wealth of both sets of users was the same. Since EF users required the highest QoS class, they had to accept higher prices. During the second ToD period, no AF traffic was promoted since these users had a smaller aggregate wealth. Profits could have increased if the EF allocation was increased (promoted); however, the extra capacity is not available (bound by the SLA). During the third ToD, the AF allocation was promoted since the aggregate wealth of AF users was higher than the EF users. Finally, during the last ToD period, the AF allocation was not promoted since the aggregate wealth of the EF users was higher. Note the allocation for any class never exceeded the total provisioned for the class plus any amount provisioned for higher classes. The resulting profit using QoS class promotion was 1640.49 tokens, a 20.73% increase.



(a) SLA-based bandwidth allocation with no QoS class promotion.



(b) Bandwidth allocation with QoS class promotion. AF may be promoted to EF.

Figure 4: QoS class provisioning and allocation amounts for two traffic classes, Expedited Forwarding (EF) and Assured Forwarding (AF). EF is considered a higher QoS class than AF.

5. Conclusions

An integral part of the DiffServ framework is network resource provisioning and allocation (pricing), which occurs over different time scales. Network managers must provision resources (large amounts over long periods of time) then allocate these resources to individual users (smaller amounts over short periods of time). Determining the appropriate amount to provision and allocate is problematic due to the different time scales, multiple Quality of Service (QoS) classes, and the unpredictable nature of users. This paper introduced a method for optimally provisioning and pricing differentiated services. Resources were provisioned per QoS class over the long-term, then priced based on user demand over the short-term. Unique to this method is QoS class promotion. Class promotion can occur when demand for a lower class is greater than the amount provisioned (SLA agreement). If higher class connections have bandwidth available and have the same ingress-egress pair, then lower class traffic can be sent using the higher class connection. This additional flexibility results in better utilization of resources and higher profits, while maintaining a low blocking probability. This was demonstrated numerically, where QoS class promotion increased profits over 20% as compared to not allowing QoS class promotion.

Future work includes investigating sampling procedures and DiffServ connection selection. Correct

estimation of the aggregate demand is essential for the provisioning and pricing method presented in this paper. While route selection was not the focus of this paper, the profit maximization techniques could be used to determine which DiffServ connections to purchase.

References

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

[2] Y. Bernet, J. Binder, S. Blake, M. Carlson, B. E. Carpenter, S. Keshav, E. Davies, B. Ohlman, D. Verma, Z. Wang, and W. Weiss. A Framework for Differentiated Services. IETF Internet Draft, February 1999.

[3] C. Courcoubetis and V. A. Siris. Managing and Pricing Service Level Agreements for Differentiated Services. In *Proceedings of the IEEE Seventh International Workshop on Quality of Service*, June 1999.

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

[5] G. Fankhauser, D. Schweikert, and B. Plattner. Service Level Agreement Trading for the Differentiated Services Architecture. Technical Report 59, TIK, 1999.

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

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

[8] E. W. Fulp and D. S. Reeves. Optimal Provisioning and Pricing of Internet Differentiated Services in Hierarchical Markets. In *Proceedings of the IEEE International Conference on Networking*, 2001.

[9] G. Huston. *ISP Survival Guide: Strategies for Running a Competitive ISP*. John Wiley & Sons, 1999.

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

[11] R. R.-F. Liao and A. T. Campbell. Dynamic Core Provisioning for Quantitative Differentiated Service. In *Proceedings of the International Workshop on Quality of Service*, 2001.

[12] Øystein Foros and B. Hansen. Competition and Compatibility among Internet Service Providers. Presented at the Second Berlin Internet Economics Workshop, 1999.

[13] R. Morris and D. Lin. Variance of Aggregated Web Traffic. In *Proceedings of the IEEE INFOCOM*, 2000.

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

[15] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer-Verlag, 1999.

[16] A. Odlyzko. The Economics of the Internet: Utility, Utilization, Pricing, and Quality of Service. Technical Report 99-08, DIMACS, Feb. 1999.

[17] I. C. Paschalidis and J. N. Tsitsiklis. Congestion-Dependent Pricing of Network Services. *IEEE/ACM Transactions on Networking*, 8(2):171–184, April 2000.

[18] D. Reininger, D. Raychaudhuri, and M. Ott. Market Based Bandwidth Allocation Policies for QoS Control in Broadband Networks. In *The First International Conference on Information and Computational Economics*, pages 101 – 110, 1998.

[19] N. Semret, R. R.-F. Liao, A. T. Campbell, and A. A. Lazar. Peering and Provisioning of Differentiated Internet Services. In *Proceedings of the IEEE INFOCOM*, 2000.

[20] H. R. Varian. *Microeconomic Analysis*. W. W. Norton & Co., 1992.

[21] H. R. Varian. Estimating the Demand for Bandwidth. Available at `http://www.INDEX.Berkeley.EDU/public/index.phtml`, 1999.

[22] D. Verma. *Supporting Service Level Agreements on IP Networks*. Macmillan Technical Publishing, 1999.

[23] X. Wang and H. Schulzrinne. Pricing Network Resources for Adaptive Applications in a Differentiated Services Network. In *Proceedings of the IEEE INFOCOM*, 2001.

[24] S. Yakowitz and F. Szidarovszky. *An Introduction to Numerical Computations*. Macmillan, second edition, 1989.