Proceedings of the IEEE International Conference on Networking, 2001

Optimal Provisioning and Pricing of Internet Differentiated Services in Hierarchical Markets*

Errin W. Fulp[1] and Douglas S. Reeves[2]

[1] Department of Computer Science, Wake Forest University, P.O. Box 7311,
Winston-Salem N.C. 27109-7311 USA
fulp@wfu.edu

[2] Department of Computer Science *and* Department of Electrical and Computer Engineering, N.C. State University, Box 7534, Raleigh N.C. 27695 USA
reeves@eos.ncsu.edu

Abstract. Network service providers contract with network owners for connection rights, then offer individual users network access at a price. Within this hierarchy, the service provider must carefully provision and allocate (price) network resources (e.g. bandwidth). However, determining the appropriate amount to provision and allocate is problematic due to the unpredictable nature of users and market interactions. This paper introduces methods for optimally provisioning and pricing differentiated services. These methods maximizes profit, while maintaining a low blocking probability for each service class. The analytical results are validated using simulation under variable conditions. Furthermore, experimental results will demonstrate that higher profits can be obtained through shorter connection contracts.

1 Introduction

The Internet continues to evolve from its small and limited academic origins to a large distributed network interconnecting academic and commercial institutions. In this distributed environment, individual users rely on network service providers for network access [5]. Network service providers contract with network owners for connection rights (large amounts over long periods of time), then offer individual users network access (small amounts over short periods of time) at a price. Within this hierarchy, the service provider must carefully provision and allocate (price) network resources (e.g. bandwidth). However, determining the appropriate amount to provision and allocate is problematic due to the unpredictable nature of users and market interactions. Furthermore, provisioning and allocation is more complex with Differentiated Service (DS) enabled networks, since multiple Quality of Service (QoS) classes exist.

* 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.

It has been demonstrated that resource pricing is an efficient mechanism for resource management, optimal allocations, and revenue generation [2], [3], [4], [6], [7]. However, the majority of these methods are not based on a market hierarchy, and do not consider how to provision resources. Other work has investigated DS resource provisioning [12], but not retail pricing. In contrast, this paper addresses these questions (provisioning and pricing) together within the context of a DS enabled network consisting of hierarchical markets [1]. Goals include **maximizing profit**, as well as maintaining a **low blocking probability**.

The remainder of this paper is structured as follows. Section 2 describes the general design of the hierarchical market economy. Service provider provisioning and allocation strategies are presented in section 3, that maximize profit and reducing the blocking probability. In section 4, the economy is demonstrated under variable conditions, and the monetary advantage of shorter term service level agreements is presented. Finally, section 5 provides a summary of the hierarchical market economy and discusses some areas of future research.

2 The Hierarchical Market Model

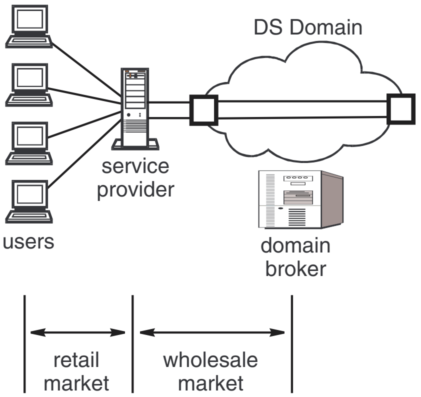

Fig. 1. Example hierarchical market economy consisting of users, service providers, and domain brokers.

As seen in figure 1, the network model is composed of three types of entities (users, domain brokers, and service providers) and two types of markets (wholesale and retail). An individual user, executing an application, requires bandwidth of a certain QoS class along a path. Users may start a session at any time, request different levels of QoS, and have varying session lengths. Furthermore, users desire immediate network access (minimal reservation delay). In contrast, the domain broker owns large amounts of bandwidth (or rights to bandwidth) and is only interested in selling large DS connections[1] [14]. The service provider plays a very important role in the network economy. Interacting with users and domain brokers, the service provider purchases bandwidth from domain brokers (provisioning), then re-sells smaller portions to individual users (allocation). Buying and selling occurs in two different types of markets: the wholesale market and the retail market.

2.1 Network Resource Markets

In our network economy, service level agreements for future DS connections are bought and sold in the wholesale market. These forward contracts represent large bandwidth amounts over long periods of time [5]. Domain brokers sell contracts for large DS connections, with an associated Service Level Agreement (SLA), across a specific network. An offer specifies the location, delivery date, class q, price g_q, and term. The market then attempts to match a buyer with the seller and a forward contract is created. This is how bandwidth is currently traded in many on-line commodity markets, such as RateXchange and Interxion. If a service provider agrees to purchase a DS connection of capacity s_q, the associated cost is $g_q \cdot s_q$ for the agreed term.

The retail market consists of a service provider selling to individual users, portions of the DS connections purchased in the wholesale market. The price of DS bandwidth will be usage-based, where the user cost depends on the current price and the amount consumed. We will use prices based on slowly varying parameters such as Time of Day (ToD) statistics [9], [10], as seen in figure 2. 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 priori by the users via a price-schedule $\{p_{q,t}\}$, where $p_{q,t}$ is the price of class q bandwidth during the t ToD period. The bandwidth of DS connection q is sold on a first come first serve basis; no reservations are allowed. Assume a user requires an amount of bandwidth b_q of service class q for the duration of their session. If the amount is not available at the beginning of the session, the user is considered blocked. However, users who can not afford b_q are **not** considered blocked. Therefore, it is important to price bandwidth to maximize profit as well as maintain a low blocking probability.

3 Optimally Provisioning and Allocating Network Resources

In this section optimal provisioning and retail pricing methods are developed for the service provider, that will maximize profit and reduce the blocking probability. The profit maximization behavior of the service provider is constrained by both markets. To maximize profits, the service provider will seek to make the

[1] Therefore, a session is a small amount of bandwidth (appropriate for a single user or application) compared to a connection.

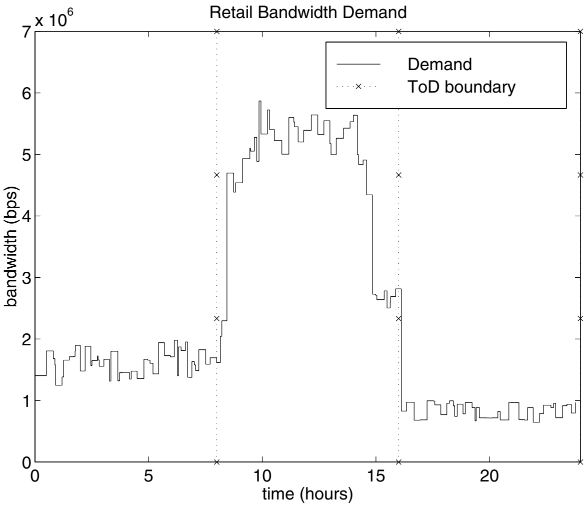

Fig. 2. Example Time of Day (ToD) changes in retail demand.

difference between the total revenue and the total costs as large as possible. The revenue from the retail market for class q during ToD t is $r_{q,t} = p_{q,t} \cdot d_{q,t}(p_{q,t})$. Where $d_{q,t}(p_{q,t})$ is a convex function representing the aggregate retail market demand for service class q during ToD period t at price $p_{q,t}$. As described in section 2.1, the cost of service class q is given from the wholesale market as $c_q = g_q \cdot s_q$. Assume the SLA term for each connection q is N consecutive ToD periods; therefore the supply during ToD $t = 1, ..., N$ is constant. From the revenue and cost, the profit maximization problem can be written as

$$\max \left\{ \sum_{q=1}^{Q} \sum_{t=1}^{N} (r_{q,t} - c_q) \right\} \tag{1}$$

Where profit maximization is over the SLA term. Viewing this as an optimization problem, the first order conditions are

$$\sum_{q=1}^{Q} \sum_{t=1}^{N} \frac{\partial r_{q,t}}{\partial s_q} = N \cdot \sum_{q=1}^{Q} \frac{\partial c_q}{\partial s_q} \tag{2}$$

The left-hand side of equation 2 is also referred to as the marginal revenue, which is the additional revenue obtained if the service provider is able to sell one more unit of DS bandwidth. The right-hand side of equation 2 is referred to as the marginal cost. This is the additional cost incurred by purchasing one more unit of DS bandwidth from the wholesale market. The service provider must purchase (provision) bandwidth from the wholesale market and price bandwidth in the retail market so the marginal revenue equals the marginal cost, as seen in figure 3. If this is done, the profit is maximized and the blocking probability is zero.

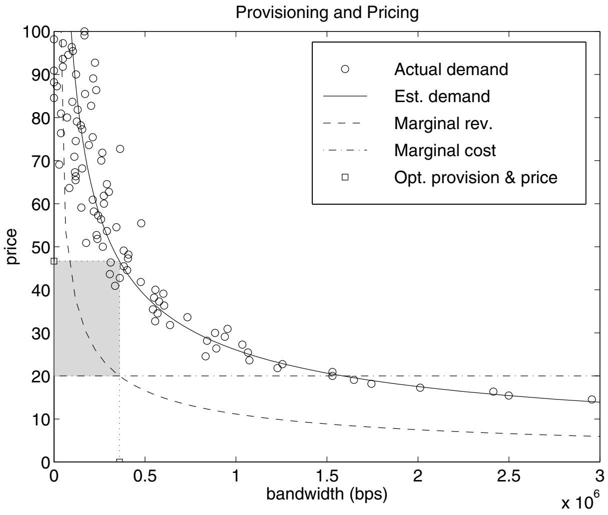

Fig. 3. The service provider seeks the point where the marginal revenue equals the marginal cost. If the optimal provisioning and retail pricing occurs, the amount of profit is given in the shaded area. (Demand data taken from the experimental section)

3.1 Single ToD Wholesale Provisioning and Retail Pricing

In this section, the optimal amount of bandwidth to provision for a single class q for one ToD period t will be determined (the q and t subscripts will be dropped for brevity). Assume the aggregate retail market demand at the retail price p has a Cobb-Douglas form [8],

$$d(p) = \beta \cdot p^{-\alpha} \tag{3}$$

Where β and α are constants describing the aggregate wealth and price-demand elasticity respectively. Price-demand elasticity represents the percent change in demand, in response to a percent change in price[2]. The larger the price-demand elasticity value, the more elastic the demand [8]. The Cobb-Douglas demand curve is commonly used in economics because the elasticity is constant, unlike linear demand curves [11]. This 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 successfully used for describing Internet demand in the INDEX Project [13]; therefore, we believe this curve is also appropriate for the retail market. Given the aggregate demand function, the revenue earned is,

$$p \cdot d(p) = p \cdot \beta \cdot p^{-\alpha} = \beta \cdot p^{1-\alpha} \tag{4}$$

[2] Typically elasticity is represented as a negative value, since demand and price move in opposite directions. However, the sign is already incorporated in the demand equation.

Alternatively, the revenue earned by the service provider can be written as,

$$p \cdot d(p) = \left(\frac{\beta}{d(p)}\right)^{\frac{1}{\alpha}} \cdot d(p) = \beta^{\frac{1}{\alpha}} \cdot [d(p)]^{1-\frac{1}{\alpha}} \tag{5}$$

As previously described, the marginal revenue is the first derivative of the revenue equation with respect to the demand; therefore, the marginal revenue is,

$$\beta^{\frac{1}{\alpha}} \cdot \left(1 - \frac{1}{\alpha}\right) \cdot [d(p)]^{-\frac{1}{\alpha}} \tag{6}$$

The cost function for bandwidth is $g \cdot s$ and the marginal cost is g. From equation 2, the service provider maximizes profit when marginal revenue equals the marginal cost.

$$\beta^{\frac{1}{\alpha}} \cdot \left(1 - \frac{1}{\alpha}\right) \cdot [d(p)]^{-\frac{1}{\alpha}} = g \tag{7}$$

Solving for $d(p)$, the optimal amount to provision s_* is

$$s_* = \left[\frac{g}{\beta^{\frac{1}{\alpha}} \cdot \left(1 - \frac{1}{\alpha}\right)}\right]^{-\alpha} = \frac{\beta \cdot \left(1 - \frac{1}{\alpha}\right)^{\alpha}}{g^{\alpha}} \tag{8}$$

During the wholesale market auction, the service provider can use equation 8 to determine the bid amount at the offered price g. Once the auction has closed, the service provider must price bandwidth for the retail market. The optimal retail price p_* is,

$$p_* = \left(\frac{\beta}{s_*}\right)^{\frac{1}{\alpha}} \tag{9}$$

This price causes the demand (equation 3) to equal the supply (equation 8); therefore, the predicted blocking probability is zero (discussed in section 3.3).

The validity of the derived equations can be examined at infinite and unity elasticity. User demand will become very elastic (α approaches ∞), if there is a large selection of service providers (large service provider competition drives profits to zero). In contrast, if the service provider has a monopoly, the elasticity approaches 1 and profits increase [8]. From equations 8 and 9, the optimal revenue under these two extreme cases is as predicted.

$$\lim_{\alpha \to \infty} \frac{\beta \cdot (1 - \frac{1}{\alpha})^{\alpha-1}}{g^{\alpha-1}} = 0 \tag{10}$$

$$\lim_{\alpha \to +1} \frac{\beta \cdot (1 - \frac{1}{\alpha})^{\alpha-1}}{g^{\alpha-1}} = \beta \tag{11}$$

3.2 Multiple ToD Wholesale Provisioning and Retail Pricing

This section considers provisioning for a single class q (the q subscript will be dropped for brevity) over N consecutive ToD periods. These consecutive ToD periods represent the agreed SLA term from the wholesale market. As described in section 3.1, assume the aggregate retail market demand, during ToD period t at the retail price p, has a Cobb-Douglas form,

$$d_t(p) = \beta_t \cdot p^{-\alpha_t} \tag{12}$$

Where β_t and α_t are constants describing the aggregate wealth and elasticity respectively for ToD period t. The aggregate wealth and elasticity can change from one ToD period to the next. As described in section 3.1 the service provider maximizes profit when marginal revenue equals the marginal cost. Over multiple ToD periods this is

$$\sum_{t=1}^{N} \left(\beta_t^{\frac{1}{\alpha_t}} \cdot \left(1 - \frac{1}{\alpha_t} \right) \cdot [d_t(p)]^{-\frac{1}{\alpha_t}} \right) = N \cdot g \tag{13}$$

To determine the optimal supply s_*, we must solve equation 13 for $d_t(p)$. However, since the equation is non-linear, a direct solution can not be found. For this reason, gradient methods (e.g. Newton-Raphson) can be used to determine the optimal provisioning amount [15]. Due to the wholesale market auction negotiation time, this calculation can be performed off-line; therefore, convergence time is not critical. Once the auction has closed, the optimal price for ToD period t is,

$$p_{*,t} = \left(\frac{\beta}{s_{*,t}} \right)^{\frac{1}{\alpha_t}} \tag{14}$$

Therefore, in the multiple ToD case, the supply for each ToD period is constant, while the price may vary, as seen in figure 4.

3.3 Retail Market Demand Estimation and Blocking Probabilities

As described in sections 3.1 and 3.2, determining the optimal amount of bandwidth to provision and the retail price requires knowledge of the retail demand curve. However, due to the dynamic nature of the retail market demand can change over time. Such changes may reflect ToD trends, pricing, or the introduction of new technology. For this reason, demand prediction and estimation will be employed [13], where the demand curve parameters (α and β) are estimated using previous ToD measurements. The other goal for the service provider is to maintain a low blocking probability. Based on the optimal provisioning and pricing equations given in the previous two sections, these values will result in supply equaling demand (as seen in figure 3) yielding a zero blocking probability. 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, which will be demonstrated in the next section.

4 Experimental Results

In this section, the optimal provisioning and pricing techniques described in the section 3 are investigated under variable conditions using simulation. The experiments simulated 6 days, where each ToD was 8 hours in duration (3 ToD per day). The model consisted of 200 users, a domain broker, and a service provider. Users had an elasticity α uniformly distributed between 1.1 and 2.75, and a wealth β uniformly distributed between 1×10^8 and 3.5×10^8. Furthermore, the demand of each user b_t was uniformly distributed between 0.5 Mbps and 2 Mbps (consistent with multimedia traffic). Each day, users started their sessions at random times using a Poisson distribution with mean equal to the first ToD of that day. This distribution caused the second ToD period of each day to have a high utilization (simulating peak hours). Two separate experiments were performed. The first experiment assumed the SLA term was equal to 6 days, while the second experiment assumed the SLA term was equal to one ToD.

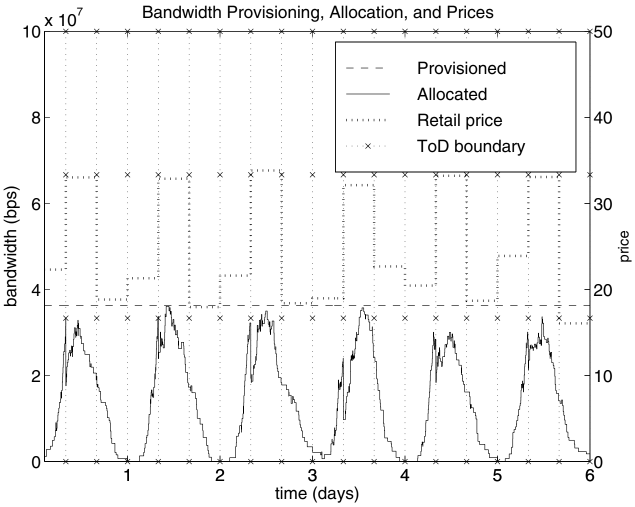

Fig. 4. Retail provisioning, allocation, and pricing simulation results for a six day SLA.

Figure 4 shows the provisioning, allocation, and pricing results, when the SLA term was 6 days. As seen in this figure, the provisioned amount was 36.2 Mbps for the duration of the simulation, while the price per ToD varied from 16.0 to 33.8. Prices during the second ToD of each day were high, since the demand was higher (peak demand). In contrast, the prices for the other ToD periods were low to encourage consumption. The total profit for the simulation was 1.54×10^{13}. The blocking probability was nonzero for ToD periods 5, 8 and

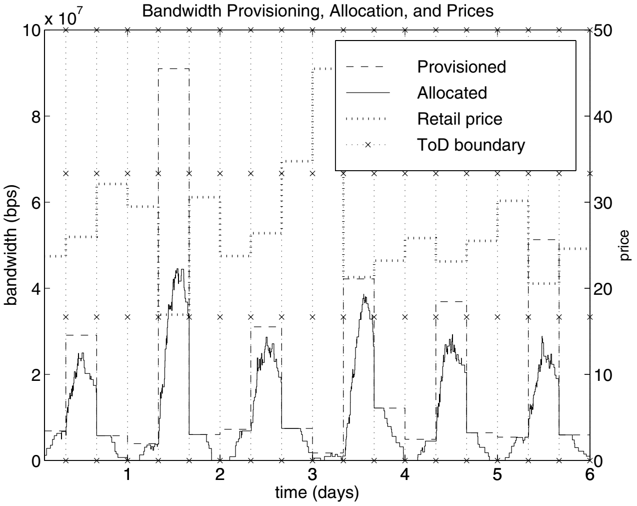

Fig. 5. Retail provisioning, allocation, and pricing simulation results for eighteen SLA's (each term equaled one ToD).

11. During these peak ToD periods, the predicted demand was less than the actual demand.

Figure 5 shows the provisioning, allocation and pricing results when the SLA term was one ToD (18 consecutive SLA's were contracted). In this simulation, the service provider could provision and price bandwidth for each ToD period. The bandwidth provisioned range from 1.7 Mbps to 90.0 Mbps, while the retail price ranged from 16 to 45.5. The total profit was 3.39×10^{13}, over twice as high as the 6 day SLA. Therefore, smaller SLA terms gave the service provider more control (provisioning and pricing), which increased profits. Similar to the other experiment, the blocking probability was nonzero for four ToD periods 4, 7, 10, and 13. Again, this indicates the predicted demand was too small for these periods. However, these were the first ToD periods of the day (non-peak).

5 Conclusions

Network services are typically provided through a hierarchical market economy. This paper introduced a hierarchical economy consisting of two types of markets (retail and wholesale) and three types of entities (service provider, domain broker, and users). Within this market hierarchy, the service provider must carefully provision resources from the wholesale market and allocate resources in the retail market. The service provider seeks to maximize profit and maintain a low blocking probability. However, achieving these objectives is problematic due to the unpredictable nature of the markets. This paper defined optimal buying/selling strategies that maximizes profit while maintaining low blocking probability per

DS connection. These methods rely on retail market estimation to determine the appropriate retail market supply and the retail market price. Simulation results were provided to demonstrate the optimal provisioning and retail pricing methods presented in this paper. The service provider was able to maximize profit given the estimated user demand and the SLA term. Shorter SLA terms were shown to yield higher profits, since the service provider is able to precisely provision based on the ToD statistics. Future work includes investigating sampling procedures and providing retail bandwidth guarantees.

References

1. 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.
2. 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.
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. 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.
5. G. Huston. *ISP Survival Guide: Strategies for Running a Competitive ISP*. John Wiley & Sons, 1999.
6. 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.
7. J. Murphy, L. Murphy, and E. C. Posner. Distributed Pricing for ATM Networks. *ITC-14*, pages 1053 – 1063, 1994.
8. W. Nicholson. *Microeconomic Theory, Basic Principles and Extensions*. The Dryden Press, 1989.
9. A. Odlyzko. The Economics of the Internet: Utility, Utilization, Pricing, and Quality of Service. Technical Report 99-08, DIMACS, Feb. 1999.
10. I. C. Paschalidis and J. N. Tsitsiklis. Congestion-Dependent Pricing of Network Services. *IEEE/ACM Transactions on Networking*, 8(2):171–184, April 2000.
11. 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.
12. 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.
13. H. R. Varian. Estimating the Demand for Bandwidth. Available through `http://www.INDEX.Berkeley.EDU/public/index.phtml`, 1999.
14. D. Verma. *Supporting Service Level Agreements on IP Networks*. Macmillan Technical Publishing, 1999.
15. S. Yakowitz and F. Szidarovszky. *An Introduction to Numerical Computations*. Macmillan, second edition, 1989.