Optimal Provisioning and Pricing of Internet Differentiated Services in Hierarchical Markets

Errin W. Fulp and Douglas S. Reeves

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.**

Keywords— **resource pricing, differentiated services, resource allocation, capacity provisioning, profit maximization, hierarchical markets**

I. 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 [1], [2]. The service provider contracts for large connection rights (large amounts over long periods of time) from network owners, then sells access to individual users [1], [2], [3]. Users can then start sessions at any time and have varying length sessions. Within this hierarchy, the service provider seeks to maximize profit, while providing users a guarantee of network availability (a small blocking probability) [2]. To achieve these goals, the service provider must carefully provision and allocate network resources (e.g. bandwidth). Provisioning refers to obtaining resources from network owners, while allocation is the distribution of resources to users (typically achieved through pricing [4]). Furthermore, these issues are interdependent. The amount to provision depends on the previous aggregate user demand, while the allocation is restricted by the amount provisioned. For these reasons, determining the appropriate amounts to provision and allocate is difficult due to the unpredictable nature of users [2]. Furthermore, provisioning and allocation is more complex with the advent of Internet Quality of Service (QoS).

Currently the Internet provides only best-effort service with no 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 (DS) is one proposed enhancement to the Internet to provide reliable QoS in a scalable fashion [5]. Under this mechanism, a finite set of QoS classes are available to aggregate flows that traverse a DS enabled network (DS domain). The DS domain is controlled by a domain broker, that is responsible for ensuring adequate resources are reserved for the aggregate traffic flows. A connection across a DS domain would have a Service Level Agreement (SLA) and Service Level Specification (SLS), which details the maximum bandwidth, QoS class, location (ingress and egress routers), cost, and the term (duration) [5], [6]. Given this framework, a service provider could purchase a DS connection from a domain broker, then offer portions of the DS connection to users. Users could then initiate sessions that differ not only in the session duration, but also in the amount of resources requested and the QoS class. However, the additional capabilities of DS also increases the complexity of provisioning and allocating resources. The service provider must provision and allocate for multiple QoS classes. For each class, the service provider will seek to maintain a small blocking probability, while maximizing profit.

In this paper, we investigate and present methods that optimally provision and allocate differentiated services.

Errin W. Fulp is with the Computer Science Department, N. C. State University, Box 7534, Raleigh N.C. 27695 USA, email: ewfulp@eos.ncsu.edu

Douglas S. Reeves is with the Computer Science Department, and Electrical and Computer Engineering Department, N. C. State University, Box 7534, Raleigh N.C. 27695 USA, email: reeves@eos.ncsu.edu

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.

This will be done within the context of a market hierarchy, where DS connections are provisioned in one market and sold to users in another market. 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 [7], [8], [9], [10], [11], [12], [13]. However, these methods are not based on a market hierarchy, and do not consider how to provision resources. Other work has investigated resource provisioning [14], [15], [3], [16], where bandwidth contracts (SLA) are bought and sold among bandwidth brokers or Internet service providers. This research is primarily interested in the development of the wholesale market and general economic stability. For example, a wholesale/retail market was proposed for DS networks by Semret et al. [16]. In this method, game theory (and simulation) was used to study the stability of the economy for provisioning DS bandwidth. While this paper provided important insight into provisioning and peering, resource allocation (pricing) in the retail market was not addressed. Retail pricing was investigated in a companion paper [17]; however, users were modeled as large aggregate subscribers, which is inconsistent with our model. In contrast, this paper will consider the provisioning and allocation of DS connections in a hierarchical economy. Microeconomic theory will be used, which provides a means for evaluating the fairness as well as the profit of different allocations and prices. Provisioning and allocation will be done to accomplish two major goals,

- Maximize profit of the service provider
- Minimize the blocking probability in the retail market

Furthermore, the monetary advantage of short term contracts (SLA) will be examined and demonstrated experimentally.

The remainder of this paper is structured as follows. Section II describes the general design of the hierarchical market economy. The interactions between individual users, domain brokers and the service provider are defined within the wholesale and retail markets. Optimal service provider strategies for bandwidth provisioning and allocation are then presented in section III. These methods seek to maximize profits, while maintaining a low blocking probability. In section IV, the economy is demonstrated under variable conditions, and the monetary advantage of shorter term service level agreements is presented. Finally, section V provides a summary of the hierarchical market economy and discusses some areas of future research.

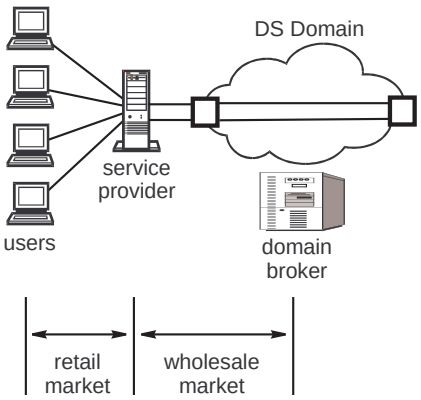


Fig. 1. Example hierarchical market economy consisting of users, service providers, and domain brokers.

II. The Hierarchical Market Model

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. In contrast, the domain broker owns large amounts of bandwidth (or rights to bandwidth) and is only interested in selling large DS connections (duration and term). These connections are much larger than an individual user would require; therefore, DS connections are not directly traded between users and domain brokers [6]. For this reason, the service provider plays a very important role in the 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. These markets as well as their interactions are discussed in the following sections.

A. Wholesale Market

Historically, network owners (traditional telephone carriers) maintained peering (bilateral) agreements to exchange traffic without settlements [2]. Yet these agreements are not possible if the cost is not perfectly balanced (one party benefits more than another). This is the case of many service providers, where the traffic transmitted into a network is greater than the traffic transmitted to the service provider. For this reason, wholesale markets are becoming increasingly popular for buying and selling bandwidth.

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 [2]. For example, in current on-line bandwidth markets the typical minimum bandwidth amount is a DS1 (2.048 Mbps in Europe and Asia), with a minimum term of one year [18]. The long term is a result of the non-commoditized market; however, as liquidity increases shorter terms are expected (for example one month). Due to the large amounts and long terms, only domain brokers and service providers participate in the wholesale market. Domain brokers sell contracts for large DS connections, with an associated SLA, across a specific network. The SLA can be static or dynamic in nature, and specifies the QoS provided (e.g. bounds on packet loss and latency) [5], [6]. Currently data carriers offer only static QoS agreements, which typically take months to negotiate; however, on-line bandwidth markets have reduced the negotiation time considerably [6]. For our wholesale market, assume Q QoS service classes exist, where $q = 1...Q$ and Q represent the highest (best) class. In the wholesale market, DS connections will be sold in an auction format similar to many current on-line bandwidth markets (e.g. RateXchange, Interxion, and Bandwidth.com). In these markets, domain brokers offer a DS connection for sale. The offer specifies the location, delivery date, class q, the price g_q, and the term. The market then attempts to match a buyer with the seller and a forward contract is created. 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. Once the connection is sold and delivered, the service provider will re-sell portions of the DS connection to individual users in the retail market.

B. Retail Market

The retail market consists of a service provider selling to individual users, the bandwidth of each DS connection purchased in the wholesale market. These DS connections represent different QoS classes and different routes; however for brevity, assume each DS connection can be uniquely identified by the associated QoS class q. The price of DS bandwidth will be usage-based, where the user cost depends on the current price and the amount consumed. An important issue in the retail market 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 [10]. Spot market prices are updated over short periods of time to reflect congestion [10]. 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, there is no incentive for the user 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 [19], [20], [21], the aggregate demand for bandwidth changes considerably during certain periods of the day. For example, demand for bandwidth during business hours may be greater than during the early morning, 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.

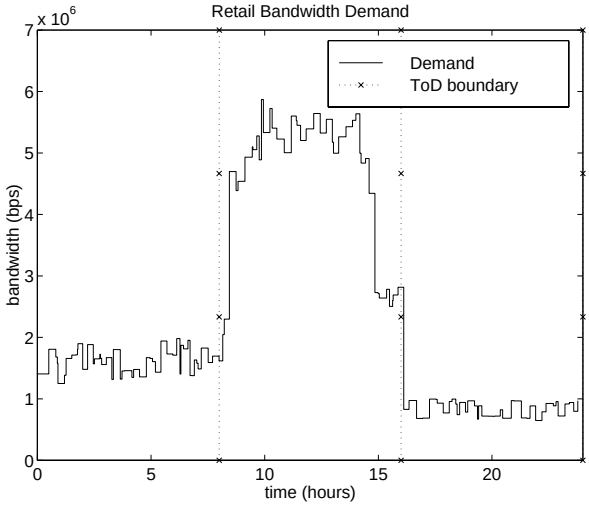


Fig. 2. Example Time of Day (ToD) changes in retail demand.

Users desire immediate network access (minimal reservation delay) and may start their sessions at any time. Assume a user requires an amount of bandwidth b_q of service class q for the duration of the session. This amount could represent the effective bandwidth of the application [22], or some other minimum. If the prices are acceptable, the user sends a request to the service provider for the desired amount. If bandwidth is available, the user is charged for the desired amount, using the price schedule, for the duration of the 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.

III. Service Provider

In this section optimal provisioning and retail pricing methods are developed for the service provider. As described in the previous section, service providers (e.g. Internet Service Providers (ISP), businesses, and university campuses) purchase bandwidth from the domain brokers, then re-sell portions of the connection to individual users. Acting as a buyer and a seller, the service provider participates in the wholesale and retail markets attempting to maximize profits and maintain a low blocking probability. To achieve these goals, the service provider must address two important questions. First, the service provider must determine how much bandwidth to provision from the wholesale market. Second, the service provider must calculate the appropriate retail market price. These problems are interdependent. The amount to provision will depend on the previous retail demand, while the allocation is restricted by the amount provisioned. Therefore, both problems are difficult to solve due to market interactions and variable user demand.

The profit maximization behavior of the service provider is constrained by both markets. To maximize profits, the service provider will seek to make the 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 the aggregate retail market demand for service class q during ToD period t at price $p_{q,t}$. As described in section II-A, 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. This relationship between revenue and cost can be depicted graphically, as seen in figure 3. If the service provider purchased an amount of bandwidth where the marginal revenue exceeded the marginal cost, profits would not be maximized since the addition of one more unit of bandwidth more add more revenue than the cost. If the marginal revenue were less than the marginal cost, reducing the bandwidth by one unit would lower costs more than revenue, which would increase profits.

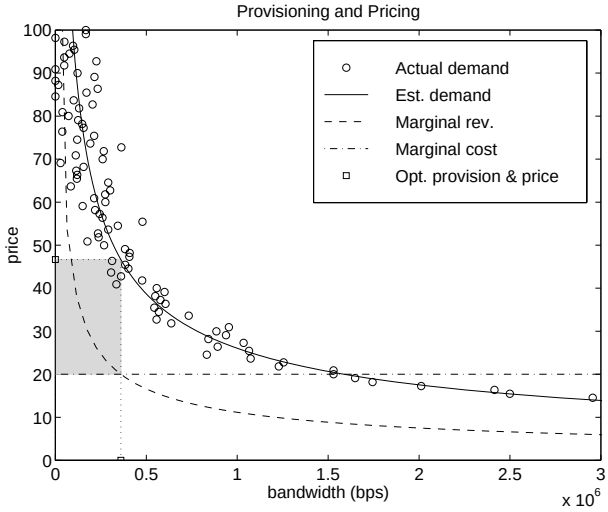


Fig. 3. The service provider seeks the point where the marginal revenue equals the marginal cost. If the optimal provisioning and retail pricing occurs, the amount of profit is given in the shaded area.

Applying this rule to the hierarchical market economy, 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. How the service provider achieves this in a hierarchical market economy is described in the following sections.

A. Single ToD Wholesale Provisioning and Retail Pricing

In this section, we will determine the optimal amount of bandwidth to provision for a single class q for one ToD period t (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 [23],

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

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. For example, an elasticity of 2 would mean a 1 percent increase in price would result in a 2 percent decrease in demand[1]. The

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

larger the price-demand elasticity value, the more elastic the demand [23]. The elasticity value will depend on individual tastes as well as the level of service provider competition. The Cobb-Douglas demand curve is commonly used in economics because the elasticity is constant, unlike linear demand curves [24]. 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 used for describing Internet demand in the INDEX Project [25]; therefore, we believe this curve is also appropriate for the retail market. Given the aggregate demand function, the revenue earned is,

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

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 $g \cdot s$; therefore, 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 (this is discussed in more detail in section III-C).

incorporated in the demand equation.

As mentioned in the beginning of this section, the value of the demand-price elasticity can reflect level of competition among service providers. If the number of service providers is very large, then the retail market is a competitive market [23]. In this market model, free entry exists, so new service providers will continue to enter the economy as long as profits are positive. Users are very elastic (α approaches ∞), because there is a large selection of service providers. If the service provider has a monopoly, the elasticity approaches 1 and profits increase [23]. This is because users have little or no choice of service providers. From equations 8 and 9, the optimal revenue under these two extreme cases is as predicted.

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

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

B. Multiple ToD Wholesale Provisioning and Retail Pricing

The previous section described optimal provisioning and pricing methods for a single ToD. However, the SLA duration is much larger than a single ToD period. 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 III-A, 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. Therefore, the aggregate wealth and elasticity can change from one ToD period to the next. The revenue earned during ToD period t is,

$$p \cdot d_t(p) = p \cdot \beta_t \cdot p^{-\alpha_t} = \beta_t \cdot p^{1-\alpha_t} \tag{13}$$

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

$$p \cdot d_t(p) = \left(\frac{\beta_t}{d_t(p)}\right)^{\frac{1}{\alpha_t}} \cdot d_t(p) = \beta_t^{\frac{1}{\alpha_t}} \cdot [d_t(p)]^{1-\frac{1}{\alpha_t}} \tag{14}$$

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

$$\beta_t^{\frac{1}{\alpha_t}} \cdot \left(1 - \frac{1}{\alpha_t}\right) \cdot [d_t(p)]^{-\frac{1}{\alpha_t}} \tag{15}$$

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

$$\sum_{t=1}^{N} \left(\beta_t^{\frac{1}{\alpha_t}} \cdot \left(1 - \frac{1}{\alpha_t}\right) \cdot [d_t(p)]^{-\frac{1}{\alpha_t}} \right) = N \cdot g \qquad (16)$$

To determine the optimal supply s_*, we must solve the previous equation 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 [26]. 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,

$$p_{*,t} = \left(\frac{\beta}{s_{*,t}} \right)^{\frac{1}{\alpha_t}} \qquad (17)$$

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

C. Retail Market Demand Estimation and Blocking Probabilities

As described in sections III-A and III-B, 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 [25].

Estimating the demand requires some underlying structure (model) that appropriately describes the actions of the consumers. The retail demand for QoS class q at ToD period t was described by the Cobb-Douglas equation

$$d_{q,t}(p) = \beta_{q,t} \cdot p^{-\alpha_{q,t}} \qquad (18)$$

Equation 18 has two unknown parameters $\alpha_{q,t}$ and $\beta_{q,t}$, which are estimations of the aggregate demand-price elasticity and wealth respectively. However, the non-linear relationship must be transformed to estimate these values using standard regression techniques. For this reason, a log-transformation is used, where the log of each side of equation 18 is taken. This yields the equation

$$\ln d_{q,t}(p) = \ln \beta_{q,t} - \alpha_{q,t} \cdot \ln p \qquad (19)$$

Equation 19 is linear in terms of the logs of the variables. Now the parameters can be determined using standard linear regression techniques [27]. The estimated value of $\alpha_{q,t}$ is not an unbiased estimate; however, it is consistent [27]. For that reason, some precision in the estimation will be lost. Determining the regression parameters $\alpha_{q,t}$ and $\beta_{q,t}$ does require previous demand and price data [28], [24], [25]. Therefore, retail market time-series data must be collected during previous ToD periods (e.g. same ToD period of the previous day). The resulting estimated demand function is used for provisioning and retail market pricing, as seen in figure 3. For this reason, correct estimation is essential for profit maximization.

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). 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, which will be demonstrated in the next section.

IV. Experimental Results

In this section, the optimal provisioning and pricing techniques described in the section III are investigated under variable conditions using simulation. Results will demonstrate the ability of the service provider to correctly manage the retail market under dynamic conditions. The effects of correct demand estimation and the blocking probability will also be shown, as well as the increase in profits resulting from shorter SLA terms.

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. The service provider purchased a single DS connection from the domain broker in the wholesale market. Users then purchased small portions of this DS connection from the service provider in the retail market. Each user 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). The session du-

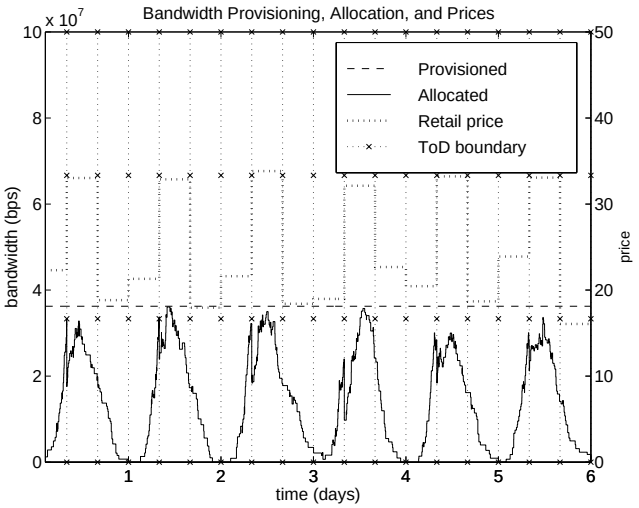
ration was a uniformly distributed between 0 to 8 hours. If the user could afford the service, but the desired amount was not available, the user was blocked until the next day. This repeats until the end of the simulations.

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. In both experiments, we were interested in measuring the profit earned over the six days to determine the advantage of shorter SLA terms. In addition, the blocking probabilities were determined to measure the effectiveness of the demand estimation. To provide data for retail demand estimation, price and demand information was gathered from additional simulations, where the price was iteratively changed and the resulting demand per ToD was collected. Once the data was collected, appropriate demand curves were generated for each experiment.

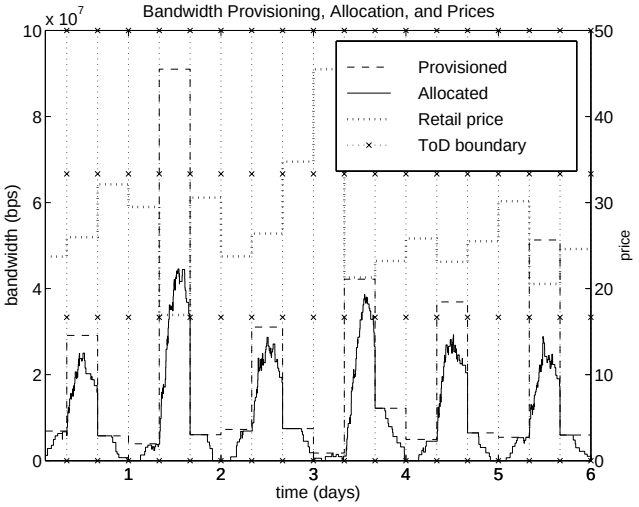
Figure 4(a) shows the provisioning, allocation, and pricing results, when the SLA term was 6 days. As seen in this figure, the provisioned amount was 36.2 Mbps for the duration of the simulation, while the price per ToD varied from 16.0 to 33.8. Prices during the second ToD of each day were high, since the demand was higher. 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 only three ToD periods. The blocking probabilities for ToD periods 5, 8 and 11 were: 0.125, 0.03, and 0.11 respectively. During these peak ToD periods, the estimated demand was less than the actual demand. Therefore, the estimated parameters were incorrect.

Figure 4(b) 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. The blocking probabilities for periods 4, 7, 10, and 13 were: 0.22, 0.35, 0.2, and 0.2 respectively. Again, this indicates the predicted demand was too small for these periods. However, these were the first ToD periods of the day (non-peak).

These simulations provide a demonstration of the provisioning and retail pricing methods presented in this paper. In each case 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. The blocking probability was zero for most ToD periods, except when the estimated demand was less than the actual demand. Therefore, better demand estimation techniques are needed.



(a) One SLA (term equaled six days).



(b) Eighteen SLA's (each term equaled one ToD).

Fig. 4. Retail provisioning, allocation, and pricing simulation results.

V. Conclusions

Network services are typically provided through a hierarchical market economy. In many cases, a buyer of a resource (or service) in one market will provide some addi-

tional value, then sell the resource in another market. 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). A service provider (e.g. Internet service provider) purchases Differentiated Services (DS) connections from a domain broker in a wholesale market, then sells access to individual users in a retail market. Bandwidth bought and sold in wholesale market is for large amounts over long periods of time. In contrast, the retail market offers bandwidth for shorter periods of time and smaller amounts, which is more accessible for individual users. Within this market hierarchy, the service provider must carefully provision resources from the wholesale market and allocate resources in the retail market. Accepting the risk of long and large bandwidth commitments in the wholesale 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. The blocking probability was zero for most ToD periods, except when the estimated demand was less than the actual demand. Therefore, better demand estimation techniques are needed.

Future work includes investigating sampling procedures and DS connection selection. Correct estimation of the retail demand is essential for the provisioning and pricing methods presented in this paper. In the experiments, only the previous ToD period was used for generating the demand curves; however, other sampling periods (more and/or longer duration) may result in better estimated demand functions. While route selection was not the focus of this paper, the profit maximization techniques could be used to determine which DS connections to purchase. Similarly, long-term profits and costs (instead of short-term) could be used to determine which routes should be purchased (investments), instead of using only immediate profit measurements.

REFERENCES

[1] D. W. Crawford, "Internet Services: A Market for Bandwidth or Communication?," in *Internet Economics* (L. W. McKnight and J. P. Bailey, eds.), pp. 379 – 400, The MIT Press, 1997.

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

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

[4] E. W. Fulp, *Resource Allocation and Pricing for QoS Management in Computer Networks*. PhD thesis, North Carolina State University, 1999.

[5] 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.

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

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

[8] 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*, pp. 644 – 648, 1996.

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

[10] 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*, pp. 75 – 84, 1998.

[11] 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*, vol. 49, pp. 237 – 252, 1998.

[12] J. F. Kurose and R. Simha, "A Microeconomic Approach to Optimal Resource Allocation in Distributed Computer Systems," *IEEE Transactions on Computers*, vol. 38, pp. 705 – 717, May 1989.

[13] J. Murphy, L. Murphy, and E. C. Posner, "Distributed Pricing for ATM Networks," *ITC-14*, pp. 1053 – 1063, 1994.

[14] 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.

[15] G. Fankhauser, D. Schweikert, and B. Plattner, "Service Level Agreement Trading for the Differentiated Services Architecture," Tech. Rep. 59, TIK, 1999.

[16] 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.

[17] N. Semret, R. R.-F. Liao, A. T. Campbell, and A. A. Lazar, "Market Pricing of Differentiated Internet Services," in *Proceedings of the 7th International Workshop on Quality of Service*, 1999.

[18] G. Cheliotis, "Bandwidth Trading in the Real World: Findings and Implications for Commodities Brokerage." Presented at the Third Berlin Internet Economics Workshop, 2000.

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

[20] A. Odlyzko, "The Economics of the Internet: Utility, Utilization,

Pricing, and Quality of Service," Tech. Rep. 99-08, DIMACS, Feb. 1999.

[21] I. C. Paschalidis and J. N. Tsitsiklis, "Congestion-Dependent Pricing of Network Services," *IEEE/ACM Transactions on Networking*, vol. 8, pp. 171–184, April 2000.

[22] F. Kelly, "Notes on Effective Bandwidths," *Stochastic Networks: Theory and Applications*, vol. 4, pp. 141 – 168, 1996.

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

[24] 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*, pp. 101 – 110, 1998.

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

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

[27] H. H. Kelejian and W. E. Oates, *Introduction to Econometrics: Principles and Applications*. Harper & Row, second ed., 1981.

[28] S. Chang, *Practitioner's Guide to Econometrics*. University Press of America, 1984.