Computer Networks Journal, Elsevier B.V., 46(1), p41-52, Sept. 2004

Bandwidth Provisioning and Pricing for Networks with Multiple Classes of Service [⋆]

Errin W. Fulp [a] Douglas S. Reeves [b]

[a] *Department of Computer Science, Wake Forest University, Winston-Salem N.C. USA, email:* `fulp@wfu.edu`

[b] *Department of Computer Science and Department of Electrical and Computer Engineering, N.C. State University, Raleigh N.C. USA*

Abstract

Network service providers purchase large point-to-point connections from network owners, then offer individual users network access at a price. Appropriately provisioning (purchasing) and allocating (pricing) connections remains a difficult problem due to increasing demands and network dynamics. However, connection management is more complex with the deployment of Quality of Service (QoS). This paper describes a scalable connection management strategy for QoS enabled networks. The management technique maximizes profit, while reducing blocking experienced by users. Important issues regarding demand estimation, connection duration, and pricing intervals, are addressed and analyzed. Simulation results are also provided to demonstrate the viability of the proposed system.

Key words: Connection Management; SLA; DiffServ; Bandwidth Pricing; Microeconomics.

1 Introduction

The next generation Internet will provide advanced services, such as Quality of Service (QoS) guarantees, to users and their applications. As a result of these enhancements, it is expected that service providers will face an increasing number of users as well as a wide variety of applications. Under these

[⋆] 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.

demanding conditions, network service providers must carefully provision and allocate network resources (e.g. bandwidth) for their customers. Provisioning is the acquisition of large point-to-point network services (connections) over a long time scale. In contrast, allocation is the distribution of these provisioned services (via pricing) to individual users over a smaller time scale [1]. Determining the optimal amounts to provision and allocate remains a difficult problem under realistic conditions. For example, service providers must balance user needs in the short-term while provisioning connections for the long term [2]. Furthermore, this must be done in a scalable fashion to meet the growing demand for network services, while also being adaptable to future network technologies.

Microeconomics can serve as an efficient mechanism for resource management, optimal allocations, and revenue generation [2–8]. An excellent description of the issues related to pricing and managing QoS-enabled networks in a retail market is provided in [2]. Network resource provisioning has also been investigated, where bandwidth contracts are bought and sold among network brokers and service providers [9–15]. This previous research was primarily interested in the development of a wholesale market and defining general economic stability. For example, a wholesale/retail market was proposed for Internet Differentiated Services (DiffServ) networks by Semret et al. [15]. While this paper provided important insight into provisioning and peering, it did not address resource allocation (pricing). Pricing was investigated in a companion paper [16]; however, these management issues are best answered simultaneously, since provisioning and allocation are interdependent.

In [17,18] a framework was introduced for provisioning and pricing a single connection within the context of hierarchical markets. A connection was purchased in a wholesale market and access was sold to individual users in a retail market. This paper builds on this hierarchical model to provide a scalable connection management strategy for multiple connections and future network services. Unlike [17,18], important issues regarding connection balancing, demand estimation, connection duration, and pricing intervals are addressed. Objective of the connection management method is to maximize profit and bandwidth utilization, while reducing the blocking experienced by users.

The remainder of this paper is structured as follows. Section 2 describes the connection management model used, consisting of individual users and a service provider. Optimal strategies for bandwidth provisioning and allocation (pricing) are presented in section 3 with analytical results. In section 4, the performance of the connection management techniques are investigated under realistic conditions using simulation. Finally, section 5 provides a summary of connection management and discusses some areas of future research.

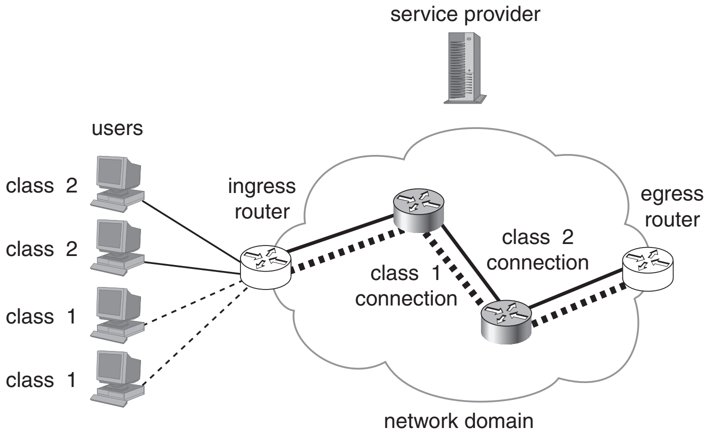


Fig. 1. Network supporting QoS, where multiple users share a single QoS connection.

2 Connection Management Model

The network model consists of users and a service provider, as seen in figure 1. Users require a certain QoS and bandwidth amount, for example effective bandwidth [19], for their network applications. Applications may start a session at any time, request different levels of QoS, and have varying session lengths. In addition, users desire immediate network access (minimal reservation delay). In contrast, service providers own large amounts of bandwidth (or rights to bandwidth) and point-to-point connections across networks [20].

For this paper, link bandwidth will be the primary resource that is priced, while QoS metrics will be used as constraints. For example a customer may request a bit rate with a certain minimum point-to-point delay or type of service (e.g., best effort), which is consistent with how service (in its limited form) is traded today [21]. Furthermore, this paper defines a *service* as a low-level network service that a network service provider offers to its customers. Services could include the Internet Integrated Services (IntServ) guaranteed service [22] or DiffServ assured forwarding [20]. A QoS connection is the actual invocation and use of a service (QoS class). How the QoS is achieved depends on the underlying network protocols; it is expected in the future that the customer will be shielded from such specifics via middleware [2,23]. Once the QoS connection is established (provisioned), portions of the connection are sold (allocated) to individual users at a price; therefore, multiple users may share a single QoS connection, as seen in figure 1.

The price of bandwidth (charged to users) will be based on use. Similar to residential electricity, bandwidth will be considered a nonstorable commodity. The time scale associated with the price is another important issue. Bandwidth prices could remain fixed for long periods of time or continually change based on current congestion levels [5]. For example, spot market prices are updated over short periods of time to reflect congestion [5]. While this method does

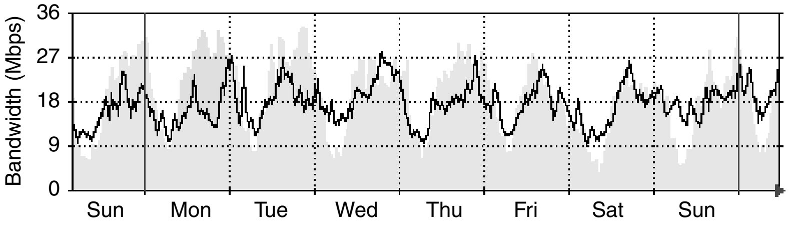


Fig. 2. Aggregate bandwidth usage for Wake Forest University. The dark line represents the incoming traffic, while the shaded area is traffic departing campus.

provide fair allocations under dynamic conditions, users cannot 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 or congested periods. Therefore, the service provider must be able to update prices in response to user demand yet provide some predictability.

As a compromise, the proposed management technique will use prices based on slowly varying parameters, such as Time of Day (ToD) statistics. As noted in [24–26] and seen in figure 2, the aggregate demand for bandwidth changes considerably during certain periods of the day. A day will be divided into equal periods called a ToD. During each ToD, every QoS class will have a fixed bandwidth price. To provide predictability, prices are known a priori 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. As described in [27], the duration of the ToD will effect performance. Users prefer a simple price structure with few price intervals. In contrast, service providers prefer more price intervals, yielding more control over user demand. Therefore, the service provider must carefully balance the preferences of the users against benefits of more control. Given the price schedule, the interval durations, and the required bandwidth, the user can predict the session cost. The bandwidth of a QoS connection is sold on a first-come, first-serve basis; advanced reservations are not allowed. Furthermore, a new user is admitted only if the QoS of existing users will be maintained. If the amount is affordable but not available in any of the acceptable QoS classes, the user is considered blocked. However, users who cannot afford $p_{i,t}$ are **not** considered blocked.

Given this framework, service providers are responsible for establishing QoS connections and allocating portions of the connection to individual users. Their primary goal will be to maximize profit and bandwidth utilization.

3 Optimal Connection Provisioning and Allocation

Assume the service provider needs to establish multiple connections to the same destination (network or domain), each providing a different QoS. Each point-to-point connection has an associated Service Level Agreement (SLA) that specifies the maximum bandwidth (provisioned amount), QoS class, location (ingress and egress routers), cost, and the term (connection duration) [21,28]. For this paper, i will uniquely identify a QoS class and connection. Assume Q represents a set of ordered QoS classes, where $i > j$ indicates i is a higher (more stringent) QoS class than j. In addition, let each day be divided into N equal-length ToD periods, where $t = 1...N$. Therefore, an SLA would span several consecutive ToD periods. The service provider is interested in maximizing the profit of all connections. This is achieved when the difference between the revenue generated minus the cost is maximized, as seen in the following formula.

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

The revenue generated by connection (QoS class) i during ToD period t is $r_i(x_{i,t})$, where $x_{i,t}$ is the user demand for this class. The cost of connection i is $c_i(s_i)$, where s_i is the bandwidth capacity specified in the SLA. The profit maximization is over the SLA term (N consecutive ToDs) and all QoS classes. The first-order conditions of the optimization problem (equation 1) are

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

Note that the supply (SLA provisioning amount) for each class, s_i, is constant for each ToD. The left-hand side of equation 2 is referred to as the marginal revenue, which is the additional revenue obtained if the service provider is able to sell one more unit of bandwidth. The right side of equation 2 is referred to as the marginal cost, which is the additional cost incurred. If the cost and revenue functions are continuous and convex, the optimization problem can be solved. Therefore, to determine the appropriate provisioning amounts and prices, these functions must be identified.

The Cobb-Douglas demand function will be used to model and predict aggregate user demand (multiple users seeking the same QoS class) [29]. The Cobb-Douglas demand function is commonly used in economics, because it is continuous, convex, and has a constant elasticity [29]. A constant elasticity assumes users respond to proportional instead of absolute changes in price, which is more realistic. The INDEX Project used the Cobb-Douglas demand function to describe user demand for different Internet access speeds [30]. For this reason, this function is also appropriate for Internet QoS demand.

The Cobb-Douglas function has the following form,

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

where $p_{i,t}$ is the price for resource i during ToD t; and the approximate aggregate wealth (per unit bandwidth) 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 its price. Own-price elasticity is typically negative, since price and demand move in opposite directions [30]. 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. Therefore the model captures the preferences for different QoS classes as well as the relative amounts desired per class. The values for aggregate wealth and the elasticities are estimated from previous ToD periods, which is described further in section 4.1.

The optimization problem given in equation 1 must be solved to determine the appropriate amount to provision for each class. This amount will be constant for the duration of the connection. Given the aggregate demand function, the revenue earned is the price multiplied by the demand,

$$\begin{aligned} p_{i,t} \cdot x_{i,t} &= \left(\frac{x_{i,t}}{\beta_{i,t} \cdot \prod_{\forall j \in Q, j \neq i} p_{j,t}^{\alpha_{ij,t}}} \right)^{\frac{1}{\alpha_{ii,t}}} \cdot x_{i,t} = \\ & x_{i,t}^{1+\frac{1}{\alpha_{ii,t}}} \cdot \beta_{i,t}^{\frac{-1}{\alpha_{ii,t}}} \cdot \left(\prod\nolimits_{\forall j \in Q, j \neq i} p_{j,t}^{\alpha_{ij,t}} \right)^{\frac{-1}{\alpha_{ii,t}}} \end{aligned} \tag{4}$$

Taking the derivative of equation 4 with respect to demand yields the marginal revenue for ToD 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 that can be solved for s_i. Remember, we seek the point where demand equals supply; therefore, $x_{i,t} = s_i$, which is the appropriate amount to provision for QoS class i during ToD t. Since the marginal equations (revenue and possibly cost) are non-linear, a direct solution cannot be found; however, gradient methods (e.g., Newton) can be used to determine the optimal provisioning amounts [31,32]. Due to the time typically associated with negotiating an SLA [18], calculations can be performed off-line, since convergence time is not critical.

Given the amount provisioned and the demand functions (equation 3), the price per ToD can be determined using the following equation.

$$p_{i,t} = \left(\frac{x_{i,t}}{\beta_{i,t} \cdot \prod_{\forall j \in Q, j \neq i} p_{j,t}^{\alpha_{ij,t}}} \right)^{\frac{1}{\alpha_{ii,t}}} \tag{5}$$

Substituting s_i for $x_{i,t}, t = 1 \ldots N$ gives the price per ToD such that $x_{i,t} = s_i$. Of course, the actual demand may be less than this amount (actual demand is finite), which will result in an under allocation during the ToD.

As previously described, the prices will form a price schedule, which is given to the users in advance. Using the price schedule, users can determine the cost of their session and purchase the bandwidth that maximizes their QoS. A user is accepted into a class only if the available bandwidth can satisfy the demand of the new user while not degrading the QoS of existing users. If the amount of bandwidth desired is affordable but not available, then the user is blocked. While the system does not target a specific blocking probability, demand estimation will be performed to minimize the blocking experienced by users. Since the system is based on the competitive market, the resulting allocations (at the equilibrium prices) are fair and efficient [33].

3.1 Provisioning and Allocation Analysis

The validity of the provisioning and allocation equations can be verified by examining the cases of infinite and unity elasticity. Assume users are only interested in one type of service i; as a result, the cross-price elasticities are zero $\alpha_{ij,t} = 0, \forall i \neq j$, while the own-price elasticity is non-zero. Given this simplified Cobb-Douglas demand equation, the revenue earned by the service provider during the interval t for connection i is,

$$p_{i,t} \cdot x_{i,t} = p_{i,t} \cdot \beta_{i,t} \cdot p_{i,t}^{\alpha_{ii,t}} = \beta_{i,t} \cdot p_{i,t}^{1+\alpha_{ii,t}} \tag{6}$$

Alternatively, the revenue earned can be written as,

$$p_{i,t} \cdot x_{i,t} = \left(\frac{x_{i,t}}{\beta_{i,t}} \right)^{\frac{1}{\alpha_{ii,t}}} \cdot x_{i,t} = \beta_{i,t}^{-\frac{1}{\alpha_{ii,t}}} \cdot [x_{i,t}]^{1+\frac{1}{\alpha_{ii,t}}} \tag{7}$$

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

$$\beta_{i,t}^{-\frac{1}{\alpha_{ii,t}}} \cdot \left(1 + \frac{1}{\alpha_{ii,t}} \right) \cdot [x_{i,t}]^{\frac{1}{\alpha_{ii,t}}} \tag{8}$$

Assume, cost of connection i paid by the service provider is $c(s_i) = g_i \cdot s_i$ and the marginal cost is g_i, where g_i per unit bandwidth. From equation 2, the service provider maximizes profit when marginal revenue equals the marginal cost.

$$\beta_{i,t}^{-\frac{1}{\alpha_{ii,t}}} \cdot \left(1 + \frac{1}{\alpha_{ii,t}}\right) \cdot [x_{i,t}]^{\frac{1}{\alpha_{ii,t}}} = g_i \tag{9}$$

Solving for $x_{i,t}$, the optimal amount to provision for connection i is

$$s_i = \left[\frac{g_i}{\beta_{i,t}^{-\frac{1}{\alpha_{ii,t}}} \cdot \left(1 + \frac{1}{\alpha_{ii,t}}\right)}\right]^{\alpha_{ii,t}} = \frac{\beta_{i,t} \cdot g_i^{\alpha_{ii,t}}}{\left(1 + \frac{1}{\alpha_{ii,t}}\right)^{\alpha_{ii,t}}} \tag{10}$$

The optimal retail price $p_{i,t}$ during interval t is,

$$p_{i,t} = \left(\frac{s_i}{\beta_{i,t}}\right)^{\frac{1}{\alpha_{ii,t}}} \tag{11}$$

The value of the own-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 [34]. In this market model free entry exists, so new service providers will continue to enter the economy as long as profits are positive. User demand is very elastic (elasticity approaches $-\infty$) in this case given the large selection of service providers. In contrast, if the service provider has a monopoly the elasticity approaches -1 and profits increase, since users have little or no choice [34]. From equations 10 and 11, the optimal revenue under these two extreme cases is as predicted.

$$\lim_{\alpha_{ii,t} \to -\infty} \frac{\beta_{i,t} \cdot g_i^{1+\alpha_{ii,t}}}{(1 + \frac{1}{\alpha_{ii,t}})^{1+\alpha_{ii,t}}} = 0 \tag{12}$$

$$\lim_{\alpha_{ii,t} \to -1} \frac{\beta_{i,t} \cdot g_i^{1+\alpha_{ii,t}}}{(1 + \frac{1}{\alpha_{ii,t}})^{1+\alpha_{ii,t}}} = \beta_{i,t} \tag{13}$$

3.2 Increasing Performance via QoS Class Promotion

The preceding developed equations for optimal provisioning and allocation, where a fixed amount of bandwidth is provisioned for the duration of the connection (called SLA-based provisioning). The service provider may change prices per ToD; yet, profits could increase if the amount of bandwidth available (supply) per ToD was also adjustable. Unfortunately, the SLA specifies a strict connection capacity.

QoS class promotion refines the QoS class prices (given by equation 5) and provisioned amounts, while maintaining the SLA capacity constraints associated with each class ($s_i, \forall i \in Q$). Class promotion can occur when the predicted demand for a lower class is greater than the amount provisioned (SLA agreement) and higher class bandwidth can be made available through higher prices. QoS promotion is based on estimated demand and is performed before the final price schedule is determined. As a result, prices will not dynamically change during a ToD if QoS promotion is performed.

The connection supply (s_i) and prices ($p_{i,t}$) are determined first using the equations and methods given in the previous section. Then for each ToD, the service provider will maximize revenue again across all QoS classes with the same destination,

$$\begin{gathered}\max\left\{\sum_{\forall i \in Q} r_i(x_{i,t})\right\}\\ \text{subject to: } x_{i,t} \le s_i + \sum_{\forall j > i}(s_j - x_{j,t}),\ \sum_{\forall i \in Q} x_{i,t} \le \sum_{\forall i \in Q} s_i, \text{ and}\\ p_{i,t} < p_{j,t} \quad \forall j > i\end{gathered} \tag{14}$$

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 that the total amount allocated is no more than the total amount provisioned. The last constraint prevents price inversion, which is a higher class having a lower price than a lower class. The resulting system of equations requires non-linear methods to find the appropriate allocation amounts and prices [31]. As previously described, the preferences for different QoS classes and amounts is captured by the Cobb-Douglas demand equation via the elasticity and aggregate wealth parameters. The resulting values for $x_{i,t}$ are the optimal bandwidth provisioning amounts for each class per ToD. Based on these values, the optimal price for each class and ToD is determined using equation 5. These prices will form the final price schedule given to users.

Note that QoS class promotion may restrict the supply available to higher class traffic (via increasing prices) in lieu of a larger bandwidth supply for lower class traffic. QoS-promotion results in a larger total group of users (over all QoS classes) and higher profits, which is beyond the ATM concept of allowing lower class (ABR) traffic the use of any unused *higher class* (e.g., VBR or CBR) bandwidth [35].

4 Management Performance and Experimental Results

Section 2 described a framework for optimally managing network connections that was evaluated analytically; however, performance under realistic conditions is equally important. For this reason, this section provides examples of system performance via simulation. The use of actual data in simulation experiments is often preferred. For the system described in this paper, data is needed that captures the effect of price on demand. Unfortunately such information is not easily available. Therefore, the simulations performed in this section modeled individual behavior instead of using actual demand-price data. However, the modeled behavior is in accordance with the findings of the INDEX Project [30].

Experiments simulated users interacting with a service provider for network access. Users started their sessions at random times using a Poisson distribution with mean equal to 9:00 am each day, simulating peak hours. The duration of a session was uniformly distributed between 0 and 12 hours. Users had an own-price elasticity α_{ii} uniformly distributed between 1.75 and 2.5 (consistent with the INDEX Project) and a wealth β_i uniformly distributed between 1×10^8 and 2×10^8 tokens[1] . The minimum demand of each user was uniformly distributed between 0.25 Mbps and 2 Mbps to represent a variety of traffic. Although experiments will differ on the issue investigated, the common objectives are maximizing bandwidth utilization and profit, while minimizing the blocking probability experienced by users.

4.1 User-Demand Estimation and Blocking

Determining the optimal amount of bandwidth to provision and the associated prices requires knowledge of the aggregate user demand. Due to the dynamic nature of networks, demand can change over time; therefore, demand prediction and estimation must be employed. Future demand is estimated using demand-price data observed during previous ToD periods. During a ToD the price is fixed and the demands are recorded, where the demand during a ToD will vary based on the number of users and their applications. This is repeated for different prices resulting a set of demand-price data. Given the demand-price data, the demand curve parameters (α and β) can be estimated by transforming equation 3 into a linear form (taking the logarithm of both sides). Using the linear form of equation 3, ordinary least squares techniques can estimate the parameters. However, the shape of the curve (and thereby the performance of the management technique) will depend on which data points are used for parameter estimation.

[1] Tokens were used as a generic currency.

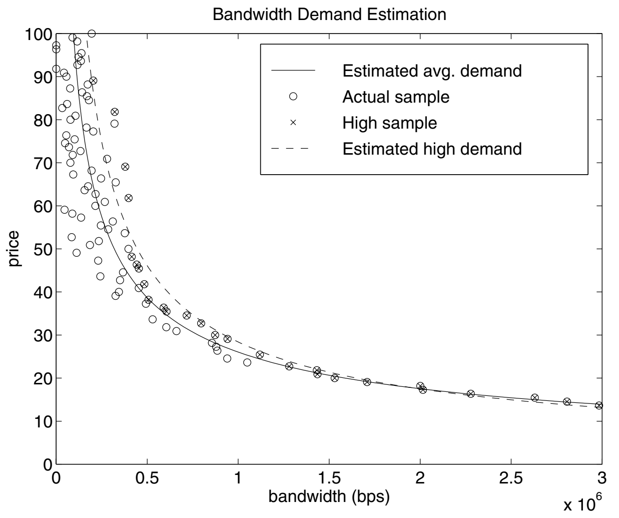
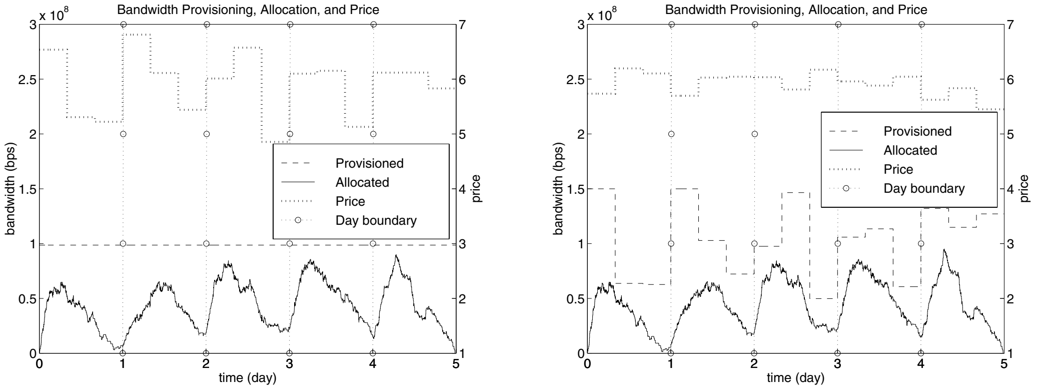


Fig. 3. Estimating demand curve parameters using all the demand data (estimated average demand) or using only the highest observed demand per price (estimated high demand).

If all the demand-price data is used to estimate the parameters, then the resulting curve is the average of the data, as seen in figure 3. Unfortunately, many actual demand data points are above the estimated demand, as seen in the figure, which results in blocking. For example, using the average-demand to estimate the demand of 200 users during one day yielded 24 blocked users and a 1.25×10^{13} profit. In contrast, if only the highest demand measured for a ToD price is considered, fewer actual data points are above the estimated curve which implies lower blocking. Using highest-demand estimation with the same group of 200 users resulted in no blocking and a profit of 1.66×10^{13}, which is a 33% profit increase over the average-demand estimation. Therefore, the more conservative highest-demand estimation yields lower blocking and higher profits as compared to the average-demand estimation. The remaining experiments used highest-demand estimation for demand prediction, and no blocking occurred.

4.2 SLA Term

When purchasing connections, service providers are interested in SLAs that specify a certain QoS, delivery points, bandwidth amount, and price. Given a group of connections that satisfies these requirements, one remaining parameter is the SLA term (contract duration). For example on-line bandwidth markets (bandwidth.com, Band-X, and InterXion) offer connections with durations ranging from a few weeks to several years. Given the variety of SLA lengths, it is important to determine the effect of contract duration on service provider performance.



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

(b) 15 SLAs (term equaled a ToD).

Fig. 4. Provisioning and allocation simulation results for a single domain and QoS class. Results indicate shorter SLA terms (Figure b) yield higher profits, which is due to an increase in control given to the service provider.

Two separate experiments were performed, each simulated 200 users over 5 days where each ToD was 8 hours in duration (3 ToD per day). The first experiment assumed the SLA term was equal to 5 days, while the second experiment assumed the SLA term was equal to one ToD [2] . Results of the two experiments are given in figure 4. In each case, the service provider was able to maximize profit, given the estimated user demand and the SLA term. The five day SLA provisioned 99 Mbps for the five days, as seen in figure 4(a). Prices fluctuated between 5 and 7 in accordance to user demand, resulting in a 1.12×10^{14} profit. The 15 SLA experiment provisioned between 50 and 150 Mbps, closely following the demand of the users as seen in figure 4(b). Prices were from 5.5 to 6.25 per SLA which is a smaller range than the single 5 day SLA. In addition, the 15 SLA experiment yielded a profit of 1.94×10^{14}, which is a 73% increase over the single 5 day SLA. The performance increase of smaller SLA durations is due to the additional control given to the service provider. The service provider was able to provision and price bandwidth to meet the changing demands, instead of relying solely on pricing. Therefore smaller SLA's can increase performance due to more control.

4.3 QoS Class Promotion

QoS promotion is the adjustment of allocations and prices based on the actual demand during a ToD. Assume two different QoS classes are required to the

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

User Type	**Wealth**					**Elasticity**	
	$\beta_{i,1}$	$\beta_{i,2}$	$\beta_{i,3}$	$\beta_{i,4}$	$\beta_{i,5}$	α_{ii}	$\alpha_{\mathrm{AF,EF}}$
EF	2×10^8	2×10^8	1×10^8	2×10^8	2×10^{12}	2.25	0
AF	1×10^5	1×10^5	1×10^9	1×10^5	1×10^5	1.75	0.41

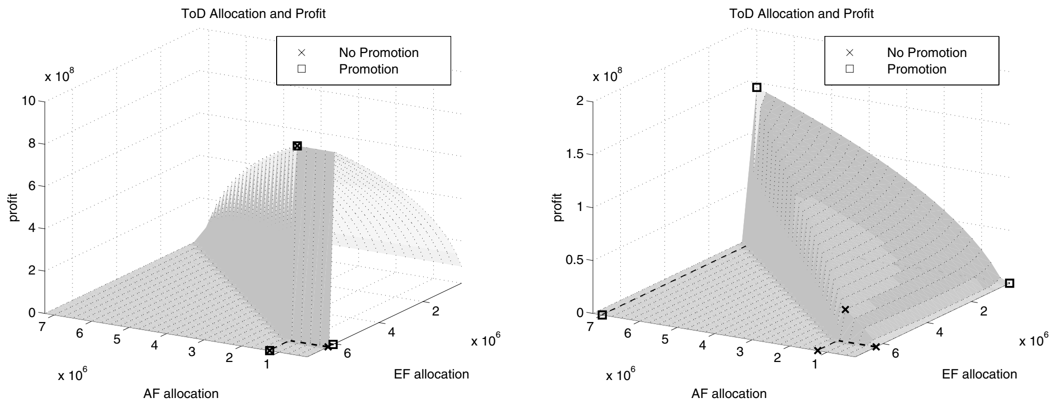
Fig. 5. Demand curve parameters used in the QoS promotion example. Note α_{ii} is the own-price elasticity, while $\alpha_{\mathrm{AF,EF}}$ is the cross-price elasticity.

same destination. The QoS classes are DiffServ Expedited Forwarding (EF) and Assured Forwarding (AF), where EF is considered a higher QoS class than AF. The SLA term was five consecutive ToD periods and 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. 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 figure 5. The aggregate wealth of each group changed per ToD, while the elasticities remained constant.

The results are given in figure 6. The SLA provisioning amounts were 6 Mbps for EF and 1.25 Mbps for AF. During ToD 1, 2, 4, and 5 QoS promotion has no advantage since the demand for EF is higher than AF. This is depicted in figure 6(a) which shows the allocations for ToD 5. Prices during these ToD intervals are 4 tokens for EF and 0.17 for AF. In contrast, during ToD 3 it is advantageous to sell EF bandwidth as AF since demand for EF is less than AF. The price for EF was 12 tokens while AF was 9 tokens. During this ToD, EF is allocated 0.25 Mbps and AF is allocated 7 Mbps, which results in higher profits as seen in figure 6(b). EF users were forced to pay higher prices, while more AF bandwidth was sold to more users yielding higher profits. Allocating bandwidth without QoS promotion resulted in a total profit of 8.1×10^8. Using QoS promotion (during ToD 3) yielded a total profit of 9.3×10^8, which is 15% higher.

4.4 Price Interval Duration

As described in the introduction, the duration of a price interval can range from extremely large to very small. Smaller intervals provide greater congestion control since prices can quickly adjust based on user demand [36,37]. This flexibility also increases profits, since prices can be set to encourage usage. Ideally, a service provider wants to update the price whenever the aggregate user demand changes. Changing the price at a rate faster than the change in demand would not increase profits. Although smaller price intervals are advantageous to the service provider, users prefer a simpler price structure [38]. Generally, users are adverse to the possibility of multiple price changes



(a) During ToD 5 allocations and profits are equal for both strategies, QoS promotion has no advantage. Similar results observed for ToD 1, 2, and 4.

(b) During ToD 3 QoS promotion yields higher profits by promoting AF traffic to EF service.

Fig. 6. QoS class allocation amounts and profits for different ToD periods. Two classes exist, Expedited Forwarding (EF) and Assured Forwarding (AF), where EF is considered a higher QoS class than AF.

during a session, even if it could result in a lower cost [37]. As a result, a service provider could lose customers, which would result in lower profits, if rival service providers offer a simpler pricing structure (fewer intervals). Therefore, interval duration is an important question to address.

The effect of price interval duration on service provider profit was investigated using simulation. For each simulation, a random number of users, uniformly distributed between 200 and 500, interacted with a single service provider during a day. For each experiment, we are interested in measuring the percent change in profit as the number of intervals increases during a day[3]. A single experiment consisted of six simulations, each simulating a different number of pricing intervals. User arrivals and demands were randomly generated at the beginning of each experiment and were used for the six simulations. For each experiment, the first simulation was performed to estimate demand curve parameters and to determine the smallest aggregate demand interval. An aggregate demand interval is the amount of time between successive demand changes. Changing prices more frequently than the shortest aggregate demand interval should not increase profits. Let T_* represent the number of pricing intervals based on the shortest aggregate demand interval. Five additional simulations were then performed, where the number of intervals were equal to 1, $T_*/100$, $T_*/10$, T_*, and $2 \cdot T_*$. Note T_* was greater than 100 for every experiment. Finally, 5000 independent experiments were conducted (30000 simulations total) to provide averages and 95% confidence intervals.

[3] For a given day, we assume all intervals are equal in duration.

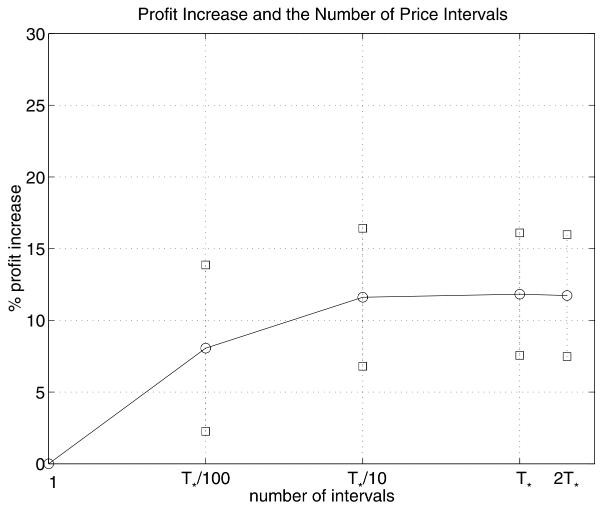


Fig. 7. Simulation results showing the profit change as the number of pricing intervals in a day increases. T_* is the number of pricing intervals based on the shortest aggregate demand interval. Percent change in profit is based on the profit of a single interval. For each data point, the 95% confidence intervals are shown.

The experimental results are shown in figure 7, which depicts the percent change in profit as the number of intervals increases. Note, the percent change in profit is compared to the profit obtained from a single interval. As expected, profit increased as the number of intervals increased. Initially, the profit increased rapidly. At $T_*/100$ the change was 0.081, while the profit increase was 0.116 at $T_*/10$ (a difference of 0.035) . However, the profit increase for T_* intervals was 0.118, which is negligible as compared to the profit obtained from $T_*/10$ intervals. Using more than T_* intervals did not increase profits as compared to T_* intervals, which was expected. Furthermore, the standard deviation of the average-profit-change reduced as the number of intervals increased, from 0.058 at $T_*/100$ to 0.043 at $2 \cdot T_*$. No users were blocked in any of the simulations performed. In summary, a service provider can increase profits by using more intervals; however, there is limited monetary benefit associated with using intervals lengths that approach the smallest change in aggregate user demand.

5 Conclusions

Service providers must provision connections (large bandwidth amounts over long periods of time) then allocate to individual users (smaller bandwidth amounts over short periods of time). Connection management is complex due to the different time scales and the unpredictable nature of users; however, as QoS becomes more ubiquitous, management becomes more difficult.

This paper described a scalable method for optimally provisioning and pricing network connections, based on a earlier model presented in [17,18]. Connections were provisioned over the long-term, then priced based on user demand over the short-term. Demand estimation methods were also described that are suitable for dynamic conditions. Using the management technique, the service provider was able to maximize profit given the estimated user demand and the connection duration (SLA term). Using simulation, shorter SLA terms were shown to yield higher profits, since the service provider is able to precisely provision based on the ToD statistics. In addition, QoS class promotion was described to increase utilization and profits. Class promotion can occur when demand for a lower class is greater than the amount provisioned (SLA agreement) and higher class bandwidth can be made available through higher prices. If this situation occurs, 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 15% as compared to not allowing QoS class promotion. Finally, this paper investigated the economic impact of price interval duration. Shorter intervals can provide higher profits; however, experimental results indicated these gains are modest. For example, experiments conducted indicate a profit increase no larger than 5% was achieved when smaller intervals were used as opposed to larger intervals (for example 100 times longer). Therefore, given users preferences toward fewer price changes, smaller price intervals hold few economic benefits.

Future areas of research include billing and accounting, and quantifying user's antipathy to price changes. The management techniques developed in this paper will rely on a billing and accounting infrastructure; therefore, more research is needed to develop scalable accounting procedures. In addition, the impact of interval durations on profit was demonstrated; yet, more research is needed to quantify users preferences for intervals.

References

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

[2] B. Briscoe, V. Darlagiannis, O. Heckman, H. Oliver, V. Siris, D. Songhurst, B. Stiller, A market managed mulit-service Internet (M3I), Computer Communications 26 (4) (2003) 404 – 414.

[3] C. Courcoubetis, V. A. Siris, G. D. Stamoulis, Integration of pricing and flow control for available bit rate services in ATM networks, in: Proceedings of the IEEE GLOBECOM, 1996, pp. 644 – 648.

[4] D. F. Ferguson, C. Nikolaou, J. Sairamesh, Y. Yemini, Economic models for allocating resources in computer systems, in: S. Clearwater (Ed.), Market Based Control of Distributed Systems, World Scientific Press, 1996.

[5] E. W. Fulp, M. Ott, D. Reininger, 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, 1998, pp. 75 – 84.

[6] F. Kelly, A. K. Maulloo, D. K. H. Tan, Rate control for communication networks: Shadow prices, proportional fairness and stability, Journal of the Operational Research Society 49 (1998) 237 – 252.

[7] J. Murphy, L. Murphy, E. C. Posner, Distributed pricing for ATM networks, ITC-14 (1994) 1053 – 1063.

[8] P. Reichl, P. Flury, J. Gerke, B. Stiller, How to overcome the feasibility problem for tariffing Internet services: The cumulus pricing scheme, in: In Proceedings of the ICC, 2001.

[9] M. C. Chan, Y.-J. Lin, X. Wang, A scalable monitoring approach to service level agreement validation, in: IEEE International Conference on Network Protocols, 2000.

[10] C. Courcoubetis, V. A. Siris, Managing and pricing service level agreements for differentiated services, in: Proceedings of the IEEE Seventh International Workshop on Quality of Service, 1999.

[11] G. Fankhauser, D. Schweikert, B. Plattner, Service level agreement trading for the differentiated services architecture, Tech. Rep. 59, TIK (1999).

[12] Øystein Foros, B. Hansen, Competition and compatibility among Internet service providers, presented at the Second Berlin Internet Economics Workshop (1999).

[13] J. Hwang, H.-J. Kim, M. B. Weiss, Interprovider differentiated service interconnection management models in the Internet bandwidth commodity markets, Telematics and Informatics, Special Issues of Electronic Commerce 19 (4) (2002) 351–369.

[14] R. R.-F. Liao, A. T. Campbell, Dynamic core provisioning for quantitative differentiated service, in: Proceedings of the International Workshop on Quality of Service, 2001.

[15] N. Semret, R. R.-F. Liao, A. T. Campbell, A. A. Lazar, Peering and provisioning of differentiated Internet services, in: Proceedings of the IEEE INFOCOM, 2000.

[16] N. Semret, R. R.-F. Liao, A. T. Campbell, A. A. Lazar, Market pricing of differentiated Internet services, in: Proceedings of the 7th International Workshop on Quality of Service, 1999.

[17] E. W. Fulp, D. S. Reeves, Optimal provisioning and pricing of differentiated services using QoS class promotion, in: Proceedings of the Advanced Internet Charging and QoS Technology (ICQT'01), 2001.

[18] E. W. Fulp, D. S. Reeves, Optimal provisioning and pricing of Internet differentiated services in hierarchical markets, in: Proceedings of the IEEE International Conference on Networking, 2001.

[19] A. I. Elwalid, D. Mitra, Effective bandwidth of general markovian traffic sources and admission control of high speed networks, IEEE/ACM Transactions on Networking 1 (3) (1993) 329–343.

[20] K. Nichols, V. Jacobson, L. Zhang, A two-bit differentiated services architecture for the Internet, Internet Draft `http://ds.internic.net/internet-drafts/draft-nichols-diff-svc-arch-00.txt` (November 1997).

[21] J. Martin, A. Nilsson, On service level agreements for IP networks, in: Proceedings of the IEEE INFOCOM, 2002.

[22] R. Braden, L. Zhang, S. Berson, S. Herzog, S. Jamin, Resource reservation protocol (RSVP) - version 1 functional specifications, IETF RFC 2205 (September 1997).

[23] B. Aiken, J. Strassner, B. Carpenter, I. Foster, C. Lynch, J. Mambretti, R. Moore, B. Teitelbaum, A report of a workshop on middleware, IETF RFC 2768 (February 2000).

[24] R. Morris, D. Lin, Variance of aggregated web traffic, in: Proceedings of the IEEE INFOCOM, 2000.

[25] A. Odlyzko, The economics of the Internet: Utility, utilization, pricing, and quality of service, Tech. Rep. 99-08, DIMACS (Feb. 1999).

[26] I. C. Paschalidis, J. N. Tsitsiklis, Congestion-Dependent pricing of network services, IEEE/ACM Transactions on Networking 8 (2) (2000) 171–184.

[27] E. W. Fulp, D. S. Reeves, The economic impact of network pricing intervals, in: Proceedings of the Advanced Internet Charging and QoS Technology (ICQT'02), 2002.

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

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

[30] H. R. Varian, Estimating the demand for bandwidth, available through `http://www.INDEX.Berkeley.EDU/public/index.phtml` (1999).

[31] J. Nocedal, S. J. Wright, Numerical Optimization, Springer-Verlag, 1999.

[32] S. Yakowitz, F. Szidarovszky, An Introduction to Numerical Computations, 2nd Edition, Macmillan, 1989.

[33] E. W. Fulp, Resource allocation and pricing for qos management in computer networks, Ph.D. thesis, North Carolina State University (1999).

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

[35] O. Kyas, G. Crawford, ATM Networks, Prentice Hall, 2002.

[36] J. S. Shih, R. H. Katz, A. D. Joseph, Pricing experiments for a computer-telephony-service usage allocation, in: Proceedings of the IEEE Globecom, 2001.

[37] M. Yuksel, S. Kalyanaraman, Effect of pricing intervals on the congestion-sensitivity of network service prices, in: Proceedings of the IEEE INFOCOM, 2001.

[38] A. Bouch, M. A. Sasse, The case for predictable network service, in: K. Nahrstedt, W. Feng (Eds.), Proc. MMCN'2000. San Jose, CA, 2000, pp. 188–195.