Proceedings of the SPIE Conference on Internet Multimedia Management, 2004

Middleware-based connection management for QoS-enabled networks

Errin W. Fulp

Department of Computer Science, Wake Forest University, Winston-Salem, NC USA

ABSTRACT

Many applications require network performance bounds, or Quality of Service (QoS), for their proper operation. This is achieved through the appropriate allocation of network resources; however, providing end-to-end QoS is becoming more complex, due to the increasing heterogeneity of networks. For example, end-to-end QoS can be provided through the concatenation of services across multiple networks (domains), but each domain may employ different network technologies as well as different QoS methodologies. As a result, management strategies are needed to provide QoS across multiple domains in a scalable and economically feasible manner.

This paper describes a microeconomic-based middleware architecture that allows the specification and acquisition of QoS and resource policies. The architecture consists of users, bandwidth brokers, and network domains. Executing applications, users require network QoS obtained via middleware from a bandwidth broker. Bandwidth brokers then interact with one another to provide end-to-end QoS connections across multiple domains. This is done in a BGP manner which recursively provides end-to-end services in a scalable fashion. Using this framework, this paper describes management strategies to optimally provision and allocate end-to-end connections. The methods maintain a low blocking probability, and maximize utility and profit, which are increasingly important as network connectivity evolves as an industry.

Keywords: middleware, Quality of Service, resource management, microeconomics

1. INTRODUCTION

An increasing number of applications rely on computer networks to provide advance services, such as Quality of Service (QoS) guarantees, for their operation. Typical QoS components include bounds on the end-to-end delay, delay variation, and packet loss. These QoS assurances can be provided with the proper management (provisioning and allocation) of network resources, such as processor time, link bandwidth, and buffer space. 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.[1,2] These management issues occur within a single domain (intra-domain) as well as across multiple domains (inter-domain). Unfortunately, providing end-to-end QoS is complex due to the increasing ubiquity and heterogeneity of networks (different underlying QoS protocols and technologies). Therefore, management strategies are needed to provide QoS across multiple domains in a scalable and economically feasible manner.

Microeconomic theory has been utilized as an efficient mechanism for resource management, optimal allocations, and revenue generation.[3–9] For example, pricing and managing QoS-enabled networks in a retail market is described by Briscoe et al.[3] Network resource provisioning has also been investigated, where bandwidth contracts are bought and sold among network brokers and service providers.[10–13] 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.[13] While this paper provided important insight into provisioning and peering, it did not address resource allocation (pricing). Pricing was investigated in a companion paper[14]; however, these management issues are best answered simultaneously, since provisioning and allocation are interdependent.

Further author information: E-mail: `fulp@wfu.edu`, Telephone: 1 336 758 3752

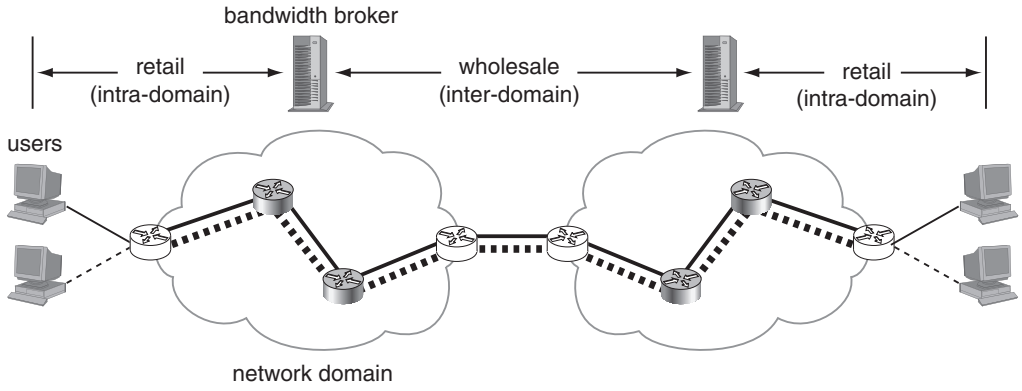


Figure 1. QoS enabled network consisting of users, bandwidth brokers, and network domains.

In[2] a framework was described 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. The objective was to maximize profit and bandwidth utilization, while reducing the blocking experienced by users. However, the problem of efficiently managing connections that span multiple domains was not directly addressed. This paper builds on the previous hierarchical model to provide a scalable strategy for inter-domain connection management. This is achieved using middleware services[15, 16] and a BGP-oriented signaling method. Middleware hides the protocol specifics associated with different QoS implementations, thereby simplifying how connections are established. The BGP-oriented signaling only requires domains to advertise transport services with their immediate neighbors, yielding a scalable method for providing end-to-end connections. Given this framework for managing end-to-end network connections, this paper introduces techniques to determine the proper provisioning and allocation amounts that maximize profit and utilization, and reduce the blocking experienced by users.

The remainder of this paper is structured as follows. Section 2 describes the network model used, consisting of users, bandwidth brokers and network domains. The market model is given in section 3, where users and bandwidth brokers interacting via the middleware layer to provide end-to-end QoS service. Section 4 introduces optimal strategies for bandwidth provisioning and allocation (pricing). A numerical example is given in section 5 to demonstrate the potential gains of managing the retail and wholesale markets simultaneously. Finally, section 6 provides a summary of middleware connection management and discusses some areas of future research.

2. NETWORK MODEL

As seen in figure 1, the network model consists of two entities (users and bandwidth brokers) and two different markets (retail and wholesale). Users require a certain QoS and bandwidth amount along a path, for example effective bandwidth,[17] for their network applications. Users may request different levels of QoS and have varying session lengths. In addition, users can start a session at any time and desire immediate network access (minimal reservation delay). In contrast, the bandwidth broker owns large amounts of bandwidth (or rights to bandwidth).[18] The bandwidth broker purchases link bandwidth from other bandwidth brokers (end-to-end provisioning), then offers smaller bandwidth portions to individual users and/or larger amounts to neighboring domains. Therefore a session is a small amount of bandwidth, appropriate for a single user or application, while a connection is a large aggregate.

Link bandwidth will be the primary resource that is bought and sold in the markets, 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. Furthermore, this paper defines a *service* as a low-level network service that a bandwidth broker offers to its retail and wholesale customers. Services could include the Internet Integrated Services (IntServ) guaranteed service[19] 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

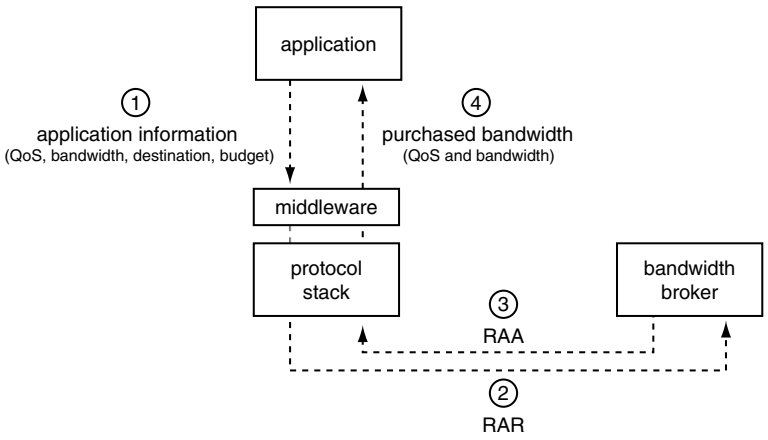


Figure 2. Resource request and answer through middleware services.

the customer will be shielded from such specifics via middleware.[3, 15] Once a QoS connection is established (provisioned), portions of the connection are sold (allocated) to individual users and neighboring domains at a price; therefore, it is expected that multiple users (local and other domains) will share a single QoS connection.

3. MARKET MODEL FOR RESOURCE MANAGEMENT

Using the network model described in the previous section, buying and selling of bandwidth will occur in two different types of markets: the retail market and the wholesale market. The retail market consists of a domain broker (acting as a service provider) selling portions of established connections to users. In contrast, the wholesale market consists of bandwidth brokers buying and selling large connection rights from each other. Bandwidth will be traded in these two types of markets in a scalable and economically feasible fashion to provide end-to-end service.

3.1. Retail Market and Resource Allocation

The user interacts in the retail market via the middleware layer, which provides a simple interface to purchase the necessary QoS.[16] As seen in figure 2, the user application provides the middleware layer the desired bandwidth amount, QoS specification, destination, and budget*. The QoS specification could be a simple ordinal value indicating the desired service type (e.g., gold, silver, or bronze service[20]) or a more complex QoS specification that is mapped to a QoS profile by the middleware layer.[21] As described by Maniatis et. al,[22] no standard exists for specifying QoS requirements; however, such standards would simplify the processing required by the middleware layer. The bandwidth broker periodically distributes the current price schedule and available QoS levels to the middleware layer of each user.

The retail 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 retail price is important issue. Bandwidth prices could remain fixed for long periods of time or continually change based on current congestion levels.[2] As a compromise, the management technique will use retail prices based on slowly varying parameters, such as Time of Day (ToD) statistics.[23] A day will be divided into equal periods of time called a retail price interval. During each retail price interval every QoS class will have a fixed bandwidth price. To provide predictability, prices are known a priori by the users via a price schedule.

Using the user and retail price information, the middleware layer determines if the desired bandwidth and QoS are affordable. If they are affordable, the middleware layer contacts the bandwidth broker and makes a Resource Allocation Request (RAR), identifying the bandwidth amount, QoS class, and destination. The bandwidth broker receives the request and issues a Resource Allocation Answer (RAA). If the amount or class

*Budget could be a single amount or a rate.

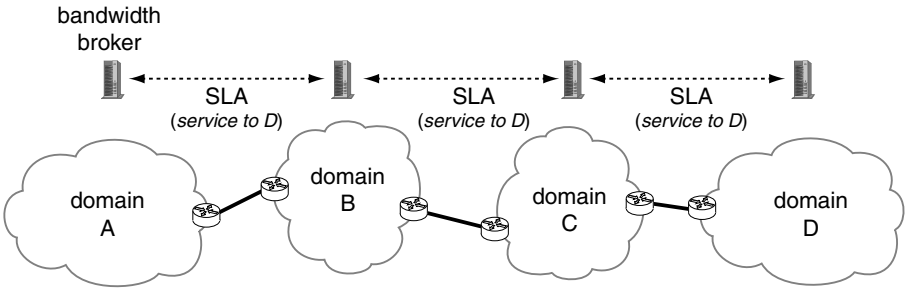


Figure 3. Multiple domain network, where each domain is managed by a separate bandwidth broker. Bandwidth broker A obtains service to domain D via SLA with neighboring bandwidth broker B.

is not available to the destination, a rejection is sent back to the user, signifying a blocked user. Otherwise, the bandwidth broker makes the appropriate reservation[†] and sends an acceptance back to the user. For a reservation, the bandwidth broker must signal to the appropriate ingress router: the connection identification, the allowable rate, QoS level, and the identity of the user. This policy information could be signaled using a variety of protocols, such as COPS or DIAMETER. The ingress router is then responsible for monitoring and policing user traffic. Some of the protocol issues related to user and bandwidth broker interactions are addressed by the IETF Authentication, Authorization, and Accounting (AAA) working group.

The retail 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 the retail price are **not** considered blocked.

3.2. Wholesale Market and Inter-Domain Provisioning

Users can only experience end-to-end QoS when there is consistent support from sender to receiver; therefore, bandwidth brokers must adequately provision bandwidth over multiple domains. This includes bandwidth for local users and neighboring bandwidth brokers. Management issues are similar to those that arise with intra-domain management (proper allocation and provisioning decisions). However, inter-domain management is more complex, due to the need for inter-domain policy management. As seen in figure 3, bandwidth brokers must negotiate and coordinate QoS connection rights with other bandwidth brokers to provide QoS connectivity across multiple domains. Connection rights have an associated Service Level Agreement (SLA), which specifies the location, delivery date, QoS, price, and term (connection duration), where the term is typically larger than a retail price interval. Note, the standardization of such agreements is the subject of current research.[22] In the absence of such standards, the middleware layer must match the different QoS descriptors offered per domain to provide end-to-end QoS.

Historically, agreements between domains (national backbone carriers) were bilateral.[1,24] Two neighboring domains agreed to carry each other's traffic as long as the amounts sent and received were comparable. However, such agreements are inherently cumbersome and not feasible between different-sized network operators (e.g., small Internet service provider and a major telecommunication carrier).[24] Bilateral agreements are further complicated by the advent of QoS, since bandwidth brokers must establish agreements for each service class. However, the use of multi-lateral agreements does not provide an elegant solution. In this scenario, bandwidth brokers would establish SLAs at each domain along every path. End-to-end QoS would then be provided through the concatenation of these agreements. Establishing agreements with multiple bandwidth brokers is problematic, considering the size of the global Internet, the number of individual SLAs needed by a single service provider, and the logistics (term and delivery location) of coordinating these agreements. For this reason, the proposed inter-domain management will rely on concepts from the Internet routing Border Gateway Protocol (BGP) to provide end-to-end QoS in a scalable fashion.

[†]Note, the aggregate connection should already be established.

BGP assumes a network consists of multiple Autonomous Systems (AS).[25] To transmit data from one AS to another, each AS contracts with its neighbors for transport service. Likewise, these neighbors contract with their neighbors for service. This procedure repeats until all autonomous systems are interconnected. Therefore, each AS makes an agreement (bilateral) with its direct neighbors for data transport, eliminating the need for multi-lateral agreements. A similar approach is used by the Border Gateway Reservation Protocol (BGRP)[26] to reserve bandwidth across domains. Neighboring bandwidth brokers will buy and sell SLAs in the wholesale market. However, unlike a bilateral agreement, the SLA (forward contract) will specify the final destination domain, the bandwidth amount, QoS, cost, and term. Note that a bandwidth broker must be prevented from purchasing an SLA that contains its domain as an intermediary (or transit) domain from the source to the final destination. Given this scenario, the neighboring bandwidth broker (seller) would have established an SLA with another bandwidth broker for the same service and final destination domain. This pattern would repeat until a bandwidth broker purchases an SLA directly, with the bandwidth broker controlling the destination domain. For example, in figure 3, assume bandwidth broker C purchases an SLA from bandwidth broker D for service to domain D. Bandwidth broker C could then sell an SLA to domain D by creating a connection across its domain connecting to the ingress router specified by the SLA agreement with D. If bandwidth broker B purchases the SLA from bandwidth broker C for service to domain D, it could sell an SLA for service to domain D. At each stage the bandwidth broker would purchase an SLA large enough for its internal traffic plus additional bandwidth to sell to its neighbors in the wholesale market. A similar two-tier management approach based on BGP is described in[27]; however, agreements between domains are bilateral, do not specify the final destination, and lack any economic cost component.

Using the model described in the previous section, assume a bandwidth broker needs to establish a QoS connection to another domain. The bandwidth broker could query each of its neighbors for the price of an appropriate SLA to this destination. The neighboring bandwidth brokers would respond with the cost of the SLA (price per unit bandwidth per unit time). Once the quotes have been gathered from the neighbors, the bandwidth broker could then determine the appropriate SLA to purchase. This request-quote method provides a simple mechanism for trading bandwidth; however, it does not easily scale to larger networks (large number of domains). A better alternative for large networks is the bandwidth commodity market, where bandwidth brokers offer and bid on connection rights across domains at a single site. The market attempts to match buyers and sellers, and communicates the aggregate pricing to the rest of the market to encourage competitive behavior.[24]

4. STRATEGY FOR RESOURCE PROVISIONING AND ALLOCATION

The previous section introduced a framework for allocating and provisioning end-to-end connections over different domains. Each bandwidth broker seeks to provision and allocate connections that minimize blocking and maximize profit, while providing end-to-end QoS. The bandwidth broker must decide which connections (destination domains) and what QoS classes are required. Furthermore, the bandwidth broker must determine how much bandwidth to purchase for each Service Level Agreement (SLA). Decisions must account for local traffic (inter-domain) needs and demands from the wholesale market (neighboring domains).

Assume an end-to-end path requires multiple connections, each with a different QoS that belongs to the set Q. Each connection has an associated SLA that specifies the maximum bandwidth (provisioned amount), QoS class, location (ingress and egress routers), cost, and the term (connection duration). Let i uniquely identify a connection and QoS class. For the wholesale market, divide the SLA duration into M periods, where $m = 1...M$. This represents points in time where the wholesale market price will change. Similarly for the retail market, let each day be divided into N equal periods (retail price intervals), where $n = 1...N$. Therefore, an SLA would span several consecutive wholesale and retail time periods, as seen in figure 4. The bandwidth broker is interested in maximizing the profit of the connection, which occurs when the difference between the sum of the retail and wholesale revenues generated and the cost is maximized. This is given in the following formula.

$$\max \left\{ \sum_{i \in Q} \sum_{m=1}^{M} r_{i,m}(y_{i,m}) + \sum_{n=1}^{N} r_{i,n}(x_{i,n}) - c_i(s_i) \right\} \tag{1}$$

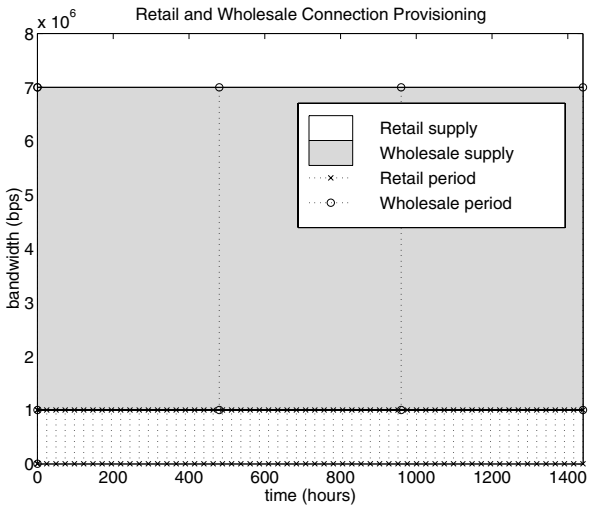


Figure 4. Example connection supply of 7 Mbps over 1440 hours, where retail periods are 24 hours in duration ($N = 60$) and wholesale periods are 360 hours in duration ($M = 3$).

The retail revenue generated for connection i during retail price interval n is $r_{i,n}(x_{i,n})$, which is based on the user demand $x_{i,n}$ for this connection. Similarly, the wholesale revenue generated for connection i during the time period m is $r_{i,m}(y_{i,m})$, which is based on the bandwidth broker demand $y_{i,m}$ for this connection. The cost of the connection i is $c_i(s)$ and is based on s_i the amount purchased in the wholesale market and the local connection cost. Note the profit maximization is over the SLA term (M consecutive wholesale and N consecutive retail periods which occur simultaneously, as seen in figure 4). The first-order conditions of the optimization problem (given in equation 1) are

$$\sum_{i \in Q} \sum_{m=1}^{M} \frac{\partial r_{i,m}(y_{i,m})}{\partial y_{i,m}} + \sum_{n=1}^{N} \frac{\partial r_{i,n}(x_{i,n})}{\partial x_{i,n}} = \sum_{i \in Q} \frac{\partial c(s_i)}{\partial s_i} \tag{2}$$

Note the supply (SLA provisioning amount) for the connection, s_i, is constant and must be sufficient for user and bandwidth broker demand. The left-hand side of equation 2 is referred to as the marginal revenue, which is the additional revenue obtained if the bandwidth broker 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. This relationship between revenue and cost can be depicted graphically, as seen in figure 5. 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.

4.0.1. Aggregate Retail and Wholesale Demand

The Cobb-Douglas demand function will be used to model aggregate demand in the retail and wholesale markets.[28] This function is commonly used in economics, because it is continuous, convex, and has a constant elasticity. A constant elasticity assumes consumers respond to proportional instead of absolute changes in price, which is more realistic. Therefore, this demand function is popular for empirical work. The INDEX Project used the Cobb-Douglas demand function to describe user demand for different Internet access speeds.[29] For this reason, this function is also appropriate for Internet QoS demand. The Cobb-Douglas function for the retail market has the following form,

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

where $x_{i,n}$ is the aggregate retail user demand for class i during retail price interval n; $p_{j,n}$ is the retail price for class j during retail price interval n; and the approximate aggregate wealth of users requiring class i is denoted by $\beta_{i,n}$. The cross-price elasticity during retail interval n is $\alpha_{ij,n}$, if $j = i$ then $\alpha_{ij,n}$ is the own-price elasticity.

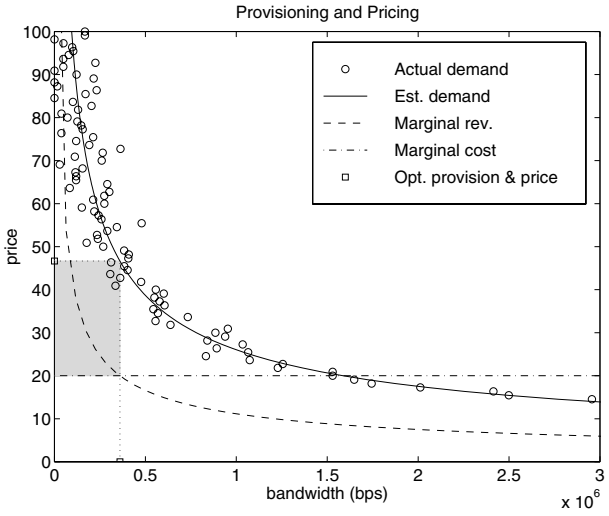


Figure 5. The bandwidth broker seeks the point where the marginal revenue equals the marginal cost. Profit is given in the shaded area, if optimal provisioning and pricing occurs.

Own-price elasticity represents the percent change in demand for class i in response to a percent change in the price of class i. The cross-price elasticity is the percentage change in the quantity demanded in response to a percent change in the price of another resource. If two resources are substitutes, the cross-price elasticity will be positive, since the price of one resource and the demand for another resource move in the same direction.

4.0.2. Optimal Provisioning

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 retail and wholesale demand functions, the revenue earned is the price multiplied by the demand. For example consider the revenue for the retail market,

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

Taking the derivative of equation 4 with respect to demand yields the retail marginal revenue for retail interval n. Similarly, taking the derivative of the cost function yields the marginal cost. Using this same procedure for the wholesale market and 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, $x_{i,n} + y_{i,m} = s_i, \quad \forall n, m$, which is the appropriate amount to provision for QoS class i.

The optimization problem can be solved using two different approaches. First, the optimization problem can be solved by treating the two markets independently. Using the techniques described in this section, determine the optimal amount for each market separately. The total amount to provision is the sum of the retail and wholesale supply. The amount provision per market is fixed for the duration of the connection; therefore, this method cannot take advantage of periods where the there is a drop in wholesale demand and an increase in retail demand (or vice versa). In contrast, the retail and wholesale market provisioning and allocation can be solved simultaneously. Although this is a more difficult optimization problem, this approach allows the relative bandwidth amount to change across wholesale periods, which may increase profits. However, the total amount provisioned is constant for both methods.

Since the marginal equations (revenue and possibly cost) are non-linear for both approaches, direct solutions cannot be found; however, gradient methods (e.g., Newton) can be used to determine the optimal provisioning

	Price Interval											
Wholesale	1.98 5×10^5 $M = 1$			1.95 1×10^5 $M = 2$			1.75 2×10^4 $M = 3$			1.95 5×10^4 $M = 4$		
Retail	1.5 1×10^4 $N = 1$	1.10 1×10^4 $N = 2$	1.90 2×10^4 $N = 3$	1.75 1×10^5 $N = 4$	1.15 4×10^5 $N = 5$	1.59 4×10^5 $N = 6$	1.90 1×10^5 $N = 7$	1.50 4×10^5 $N = 8$	1.55 4×10^5 $N = 9$	1.82 4×10^4 $N = 10$	1.10 2×10^4 $N = 11$	1.02 2×10^4 $N = 12$

Table 1. Wholesale and retail demand parameters for the numerical example given in section 5. For each interval, the top value is α, the middle value is β, and the bottom value is the interval index.

amounts.[30,31] Due to the time typically associated with negotiating an SLA,[2] calculations can be performed off-line, since convergence time is not critical.

4.0.3. Allocation per Time Period

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

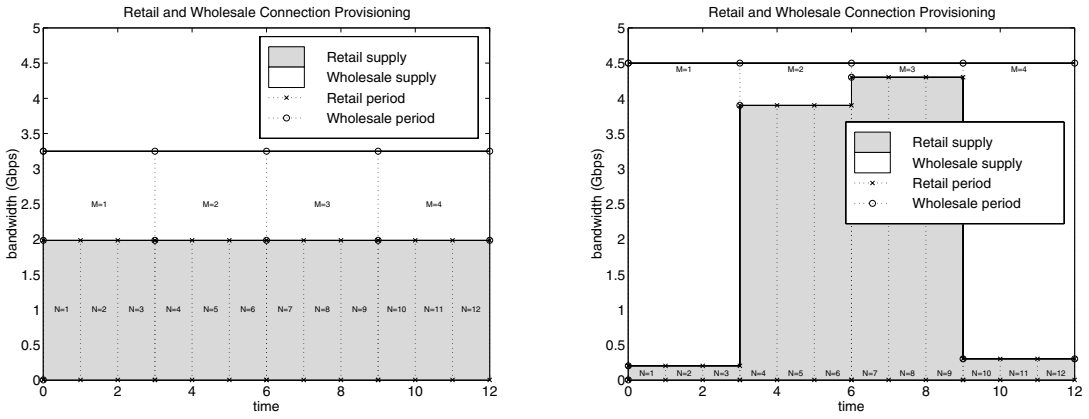
$$p_{i,n} = \frac{x_{i,n}}{\beta_{i,n} \cdot \prod_{j \in Q, j \neq i} p_{j,n}^{\alpha_{ij,n}}} \tag{5}$$

Substituting the retail supply for $x_{i,n}, n = 1 \ldots N$ gives the retail price per interval. Prices will form a price schedule, which is given to the user middleware layer. Using this schedule and the application requirements, the middleware layer will be able to determine the cost of a session and purchase the bandwidth that maximizes the QoS. The same procedure is used to determine the wholesale market prices. Since the system is based on the competitive market, the resulting allocations (at the equilibrium prices) are fair and efficient.[32]

5. A NUMERICAL EXAMPLE

In this section, a numerical example of the optimal resource management techniques described in section 4 is provided. Specifically, this section demonstrates the potential gains of optimizing the retail and wholesale markets simultaneously as opposed to treating the markets independently. Note for the example, bandwidth will be measured in bits per second while money will be measured in generic *tokens*.

The example consisted of one connection with 12 retail price intervals ($N = 12$) and 4 wholesale price intervals ($M = 4$); therefore, there were 3 retail price intervals per wholesale interval. The demand parameters for each interval is given in table 1. The provisioning amounts given by the two optimization methods are depicted in figure 6. The total connection bandwidth provision was 3.25 Gbps for the independent optimization, where 1.98 Gbps was for retail and 1.27 Gbps for wholesale. As seen in figure 6(a), the provisioning amounts for the markets were constant for the connection term which resulted in a profit of 393870 tokens. The provisioning amounts for the simultaneous optimization are given in figure 6(b). The total bandwidth provisioned was 4.5 Gbps for the duration of the connection; however, the amounts provisioned to the retail and wholesale markets varied. The retail market was provisioned between 200 Mbps (intervals $N = 1, 2, 3$) to 4.3 Gbps (intervals $N = 7, 8, 9$), while the wholesale market was provisioned between 200 Mbps (interval $M = 3$) and 4.3 Gbps (interval $M = 1$). The different provisioning amounts were in response to the changing demand characteristics of the retail and wholesale markets. For example during first wholesale period, the retail market demand was low while the wholesale market demand was high (given by the β parameters); as a result the provisioned amounts reflected these values. In contrast during the second wholesale interval, the wholesale demand was lower than the retail demand; thus the amount provisioned for the retail market increased while the wholesale market amount decreased. As a result of the additional provisioning flexibility, the wholesale market profits were 450762 tokens, which is a 15% increase over the independent optimization method.



(a) Provisioning independently.

(b) Provisioning simultaneously.

Figure 6. Retail and wholesale provisioning for the same connection using independent and simultaneous optimization. Simultaneous optimization method resulted in a 15% increase in profits, as compared to independent optimization.

6. CONCLUSIONS

QoS is only realized through the appropriate acquisition of resources or services from the source to the destination. However, providing end-to-end QoS is increasingly complex given the various QoS support offered by domains that comprise the connection. This paper introduced a scalable method for optimally provisioning and pricing end-to-end QoS connections, based on a earlier model presented by Fulp et. al.[2] The system consists of users and bandwidth brokers interacting in retail and wholesale bandwidth markets. Users purchase link bandwidth for their applications in the retail market via the middleware layer. The middleware layer hides the implementation details of the underlying services, thereby simplifying how bandwidth is obtained. Bandwidth brokers, each controlling a network domain, interact with each other in a wholesale market to provide end-to-end connections. These large aggregate connections are offered in a BGP-oriented fashion which requires bandwidth brokers to contract only with neighboring domains. This eliminates the need for multilateral contracts resulting in a scalable technique for establishing end-to-end connections. Using this framework, this paper described methods to optimally provision and allocate the end-to-end connections (capacity offered to neighboring domains and local users). This is done to maximize profits and utilization, while reducing the blocking experienced by users. In addition, optimization techniques were described that allows the domain brokers to change the amounts provisioned to the retail and wholesale markets, which can result in higher profits.

Future areas of research include billing, accounting, and standardizing QoS specifications. 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. In addition, better optimization methods are needed to determine the allocation and provisioning amounts. The introduction of agent-based trading will occur on a smaller time scale. As a result, efficient numerical methods are needed to solve this optimization problem.

REFERENCES

1. G. Huston, *ISP Survival Guide: Strategies for Running a Competitive ISP*, John Wiley & Sons, 1999.
2. E. W. Fulp and D. S. Reeves, "Bandwidth provisioning and pricing for networks with multiple classes of service," *Computer Networks* **46**(1), pp. 41–52, 2004.
3. B. Briscoe, V. Darlagiannis, O. Heckman, H. Oliver, V. Siris, D. Songhurst, and B. Stiller, "A market managed mulit-service Internet (M3I)," *Computer Communications* **26**(4), pp. 404 – 414, 2003.

4. C. Courcoubetis, V. A. Siris, and G. D. Stamoulis, "Integration of pricing and flow control for available bit rate services in ATM networks," in *Proceedings of the IEEE GLOBECOM*, pp. 644 – 648, 1996.
5. 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.
6. 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.
7. 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**, pp. 237 – 252, 1998.
8. J. Murphy, L. Murphy, and E. C. Posner, "Distributed pricing for ATM networks," *ITC-14* , pp. 1053 – 1063, 1994.
9. P. Reichl, P. Flury, J. Gerke, and B. Stiller, "How to overcome the feasibility problem for tariffing Internet services: The cumulus pricing scheme," in *In Proceedings of the ICC*, 2001.
10. M. C. Chan, Y.-J. Lin, and X. Wang, "A scalable monitoring approach to service level agreement validation," in *IEEE International Conference on Network Protocols*, November 2000.
11. J. Hwang, H.-J. Kim, and 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), pp. 351–369, 2002.
12. R. R.-F. Liao and A. T. Campbell, "Dynamic core provisioning for quantitative differentiated service," in *Proceedings of the International Workshop on Quality of Service*, 2001.
13. 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.
14. 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.
15. B. Aiken, J. Strassner, B. Carpenter, I. Foster, C. Lynch, J. Mambretti, R. Moore, and B. Teitelbaum, "A report of a workshop on middleware." IETF RFC 2768, February 2000.
16. K. Geihs, "Middleware challenges ahead," *Computer* **34**, pp. 24 – 31, June 2001.
17. A. I. Elwalid and D. Mitra, "Effective bandwidth of general markovian traffic sources and admission control of high speed networks," *IEEE/ACM Transactions on Networking* **1**, pp. 329–343, June 1993.
18. D. Verma, *Supporting Service Level Agreements on IP Networks*, Macmillan Technical Publishing, 1999.
19. R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource reservation protocol (RSVP) - version 1 functional specifications." IETF RFC 2205, September 1997.
20. K. Nichols, V. Jacobson, and 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. K. Nahrstedt, D. Xu, D. Wichadakul, and B. Li, "QoS-aware middleware for ubiquitous and heterogeneous environments," *IEEE Communications Magazine* **39**, pp. 140 – 148, Nov 2001.
22. S. I. Maniatis, E. G. Nikolozou, and I. Venieris, "End-to-end QoS specification issues in the converged all-IP wired and wireless environment," *IEEE Communications Magazine* **42**(6), pp. 80 – 86, 2004.
23. A. Odlyzko, "The economics of the Internet: Utility, utilization, pricing, and quality of service," Tech. Rep. 99-08, DIMACS, Feb. 1999.
24. W. Lehr and L. McKnight, "Next generation bandwidth markets," in *Communications and Strategies*, October 1998.
25. Y. Rekhter and T. Li, "A border gateway protocol 4 (BGP-4)." RFC1771, 1995.
26. P. Pan, E. L. Hahne, and H. Schulzrinne, "BGRP: A tree-based aggregation protocol for inter-domain reservations," *Journal of Communications and Networks* **2**(2), pp. 175 – 167, 2000.
27. A. Terzis, L. Wang, J. Ogawa, and L. Zhang, "A two-tier resource management model for the internet," in *Proceedings of Global Internet*, December 1999.
28. H. R. Varian, *Microeconomic Analysis*, W. W. Norton & Co., 1992.

29. H. R. Varian, "Estimating the demand for bandwidth." Available through `http://www.INDEX.Berkeley.EDU/public/index.phtml`, 1999.
30. J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, 1999.
31. S. Yakowitz and F. Szidarovszky, *An Introduction to Numerical Computations*, Macmillan, second ed., 1989.
32. E. W. Fulp, *Resource Allocation and Pricing for QoS Management in Computer Networks*. PhD thesis, North Carolina State University, 1999.