Why are pricing issues the key to infrastructure investment and future growth of the Internet?

The Economics of network management

Alok Gupta, Dale O. Stahl, and Andrew B. Whinston

It is not surprising that researchers in network technology are utilizing ideas from the field of economics since it provides the conceptual understanding of underlying constructs such as usage and resource allocation. Proper resource allocation plays a key role in improving network performance. There are two primary approaches to economic resource allocation: quantity limits, and pricing. The former has the advantage of low accounting costs. However, it requires central administration and, for network users wishing to purchase more quantity than allocated, a pricing schedule has to be developed. The pricing approach does not enforce any quantity limits. Instead, users self-select the quantity they are willing to purchase at prevailing prices. The pricing approach decentralizes the resource allocation problem but has higher accounting costs since usage needs to be monitored and billed.

The role of pricing as a resource allocation mechanism will be examined here, since central administration of quotas would be difficult in a global network. Furthermore, a pricing mechanism may still be needed to meet user demands in e-commerce. While the short-term role of economic pricing is to manage the network resources more efficiently, in the long run it could play an important role in the design of new network protocols. For example, new technological developments with priority classes in the IPv6 Internet protocol [3] will present the challenge of defining an economic basis for priority allocation.

The Internet traffic-pricing problem for e-commerce subsumes traditional telecommunication pricing policies and is much more complex. Traditional telecommunication systems offer a single quality level for their services on a given type of network, for example, the ability to carry voice data at a desired rate on a telephone network. However, in e-commerce there is a rich collection of services that require different levels of service quality over the same data communication network. For example, during a telnet session a user requires a high-speed and low-volume data transfer for real-time interaction with a server. On the other hand, a video conference needs clarity and coordination in picture and sound, necessitating a much more complex set of service quality attributes, such as synchronization of information sent via different interoperable applications (video, audio, and text). In addition, audio and video applications require high-volume and high-speed transfers with low variability in data transfer rates. These diverse application requirements lead us to ask: Can pricing facilitate a given quality of service (QoS) desired by a user in the highly distributed e-commerce infrastructure?

The diversity of application and user QoS requirements leads to a second question: How can we best provide different QoS and what role can pricing play in facilitating different QoS?

The final question we consider deals with the structure of physical ownership of network traffic rights and presents an argument against the efforts for a standardized pricing policy. Different parts of the network are owned by different entities and a data stream has to travel across a path that involves physical links owned by different infrastructure providers. Given the diverse nature of the organizations that own the traffic rights and their diverse cost structures, it is unlikely that they will agree to a single approach for pricing. Therefore, we believe that providing a global "ideal" pricing approach is not sufficient. We present a research agenda that explores the interaction among different e-commerce players with diverse pricing strategies. Specifically, we address the following question: What pricing policies will sustain the competitive advantage and market presence in a competitive environment in which different market entities might be selling similar services under different pricing policies?

Ideally, we would like to have reliable answers to these key questions before joining the e-commerce bandwagon. The rate of technological developments makes such perfect solutions a fantasy. However, since economic theory can provide principles in situations in which there is no reliable precedent, the economic theory-based approaches, used imaginatively, can still provide valuable guidelines and analysis.

Quality of Service and Pricing

Users' QoS requirements are driven by their usage context and may be defined by the characteristics of the end product the users receive such as the window size, audio quality, or the time to download a certain item. This QoS can be translated into technical requirements with the help of the following three attributes: probability of packet loss, data transfer rate, and consistency of delay suffered by packets in a data stream. In an e-commerce context, the QoS has to be defined even more carefully; specifically, two factors have to be considered in order to define complete QoS. First, there is application-based QoS, that is, different applications may inherently require different QoS. For example, an email message can be delivered without any loss in quality no matter how long it takes for all of its packets to arrive or in which order they arrive. An interactive game or audio conversation, however, requires a minimum data transfer rate and an appropriate ordering of packets. The second factor that must be considered is that consumers' desired QoS depends upon their usage context. For example, even though email does not require no-delay transfer at the application layer, a consumer may want to have access to instant email for urgent messages. Other examples of consumer-desired QoS may include a transaction in which a customer wants to download a movie to view immediately versus another customer who orders the same movie during his lunch break and wants to watch it at a later time. A pricing mechanism can be used as a tool to differentiate among different traffic requirements depending upon customers' desired QoS. Since e-commerce is primarily a service industry, consumer-desired QoS considerations are as important as application-based QoS considerations.

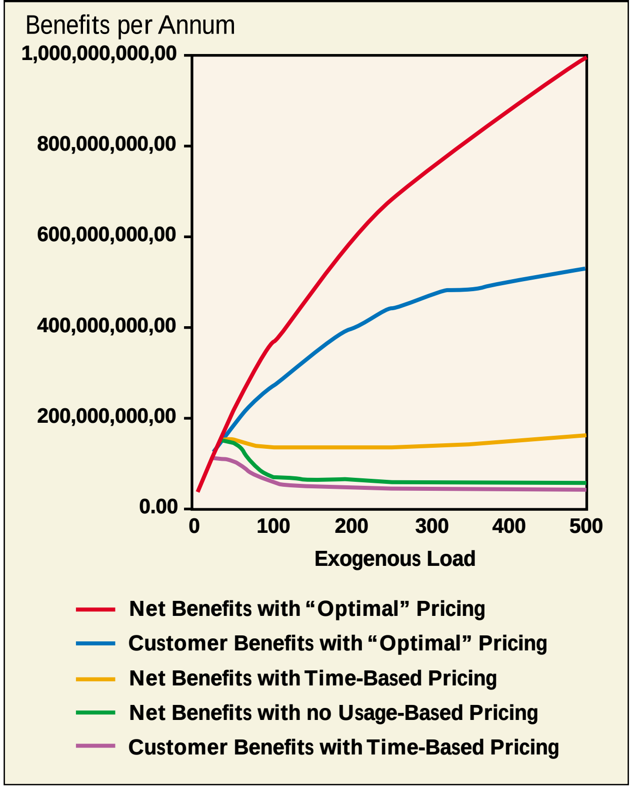
From a service management perspective, prices provide a way to facilitate differential services so that those who need high QoS can be provided with that—this simply is not a possibility with TCP/IP's best-effort service. Furthermore, there is a question of incentives; with purely application-based QoS approaches there is nothing stopping customers from developing software "masking" solutions in which an application requiring low QoS can behave like an application requiring a higher QoS. An example would be customers sending their email masked as a multimedia application. From a network management perspective, QoS-based pricing provides incentives for maintaining an appropriate range of QoS.

We contend, due to the previously mentioned factors, that to sustain an e-commerce environment in which each application and user will require a different QoS, pricing network traffic based on usage will be a necessity, perhaps combined with fixed access charges, for recovering costs.

Pricing Internet Traffic

There have been several pricing proposals in recent literature (see, for example [2, 5, 7, 11]). A comprehensive review of these pricing proposals is presented in [6, 12]. The basic idea in economic pricing of network traffic is to charge the incremental cost of providing passage. Since the short-term incremental cost of providing passage through fixed-capacity computer networks is essentially zero, researchers have focused on externality (or congestion) pricing. Under this framework, a job is priced according to what impact it has on the QoS of other jobs. For example, the price for a video stream would be based on how much potential additional delay it would cause to other users who want network access for their applications

Figure 1. Benefits from using different pricing policies.



during the time the video is being transferred, and how much loss these users would suffer due to this delay. The idea of externality pricing is not new. In fact, the rationale behind the time-of-day pricing of telephone services is an example of externality pricing. However, in the context of e-commerce, the problem becomes much more complex due to diverse QoS demands. It is important to realize that externality prices are not designed to meet other goals such as cost recovery. However, externality charges can be used in two-part tariff systems when fixed connection charges meet the goal of cost recovery and externality charges address the resource allocation problem.

One way to satisfy different QoS requirements is to provide multiple service classes characterized by different performance levels, such as the multiple levels of best effort service in which jobs in a higher class are transmitted before the jobs in a lower class. Other service classes may provide guaranteed service such as maximum delay (or minimum level of service). In a multiple service class network, base service can be thought of as a lowest best effort class with users paying just a fixed access fee and no usage-based fee. However, for a higher level of service, the users would pay according to a usage-based pricing scheme. A dynamic priority pricing mechanism, where prices vary based on the level of congestion, for multiservice class networks is presented in [5]. Furthermore, it presents a computational approach to calculate externality prices in real time.

The important characteristics of the computational approach used in [5] is that it uses approximations of performance parameters estimated based on short-term historical data collected at individual network nodes, and that it requires no network-wide information to compute the prices at a particular node. This computational approach has been evaluated by using a simulation platform that simulates a network with 50 servers and 100 distinct services. Using this model, we showed that significant performance enhancement and monetary benefits can be achieved by using dynamic prices as compared to two other pricing approaches: fixed charges and time-based charges. Figure 1 graphically represents projected Internet-wide benefits of using dynamic prices, time-based prices such as hourly connection charges,[1] and nonusage-based pricing, such as fixed monthly charges at different demand levels. Figure 1 presents system-wide benefits and consumer surplus.[2] It should be observed that even consumer surplus is higher than system-wide benefits from the other two pricing policies. Additional details of these computations and simulation model are provided in [6].

Desired Characteristics of an Internet Pricing Mechanism

Traditionally, economists look at pricing from an abstract point of view and do not consider the computational viability of pricing mechanisms. However, we stress the need for computationally viable mechanisms. Furthermore, since a single pricing approach may not be enforceable, it is important to recognize the characteristics of a desired pricing mechanism. We identify the following eight characteristics necessary for any viable pricing mechanism in multiservice class data communication networks [4]:

- Prices should encourage users to use the network when it is less congested by shifting their demands across time;
- Prices should take into account the impact of current load on future demand;
- Pricing should preferably be coarser than packet level pricing so that it is easier and less costly to implement;

[1]Time-based prices attempt an extremely coarse approximation of usage-based prices because connection time is not a good indicator of network usage.
[2]System-wide Benefit = Consumers' aggregate value – aggregate value loss due to delay; Consumer Surplus = System-wide Benefit – aggregate payment due to prices.

Time of the Day	Arrival Rate (requests/sec)
12:00 a.m. – 5:00 a.m.	20
5:00 a.m. – 8:00 a.m.	30
8:00 a.m. – 11:00 a.m.	75
11:00 a.m. – 5:00 p.m.	200
5:00 p.m. – 9:00 p.m.	100
9:00 p.m. – 12:00 a.m.	50

Table 1. Hypothetical arrival rates during a day.

- Prices should reflect the load status of the network nodes (routers, gateways);
- The pricing scheme should be implemented in a completely decentralized manner, for example, by requiring performance information at an individual node to set prices at that node but not requiring any system-wide information. Otherwise, the overhead costs involved in computing the prices may negate any potential benefits of the pricing method;
- Prices should yield effective load management by redistributing the load from highly loaded nodes to lightly loaded nodes;
- There should be multiple priorities in order to take into account the different QoS required by different applications and users;[3] and
- The pricing scheme should be implemented in such a way that users have incentives to make decisions based on the price they pay and service providers have incentives to provide the required QoS based on the profits they derive from pricing methods.

Shenker et al. [12] argue for adaptive pricing strategies based on characteristics similar to those listed here, and provide an excellent review of most pricing proposals and evaluate them based on the criteria mentioned here. They argue that most economic approaches depend upon knowing the demand function and cannot work appropriately when demand fluctuations are not known in advance. While the pricing models presented in [5, 7] derive theoretical prices for optimal resource allocation, the intention of price computation mechanism presented there is to provide a dynamic price computation methodology in which prices shift with changes in demand. These prices are approximations of "optimal" prices and are computed based on observable system performance. The computations presented in [4–7] do not require knowledge of demand characteristics. Instead, the process uses actual, observable demands at each network node to compute prices. We believe it is important to have a theoretical basis for any computational price-setting mechanism so that the theory can provide a basis for price computations under dynamic environments.

Dynamic Price Adjustment

The Internet exhibits two types of demand variations: the long-term demand pattern is different at different times of a day, and the short-term demand pattern has much larger variability than the long-term pattern—fractal demand. We present simulation results that show the robustness of our dynamic pricing mechanism with respect to both types of demand variations. Figure 2 illustrates the dynamic properties of our computational mechanism.[4] The

Figure 2. Prices with changing exogenous demand.

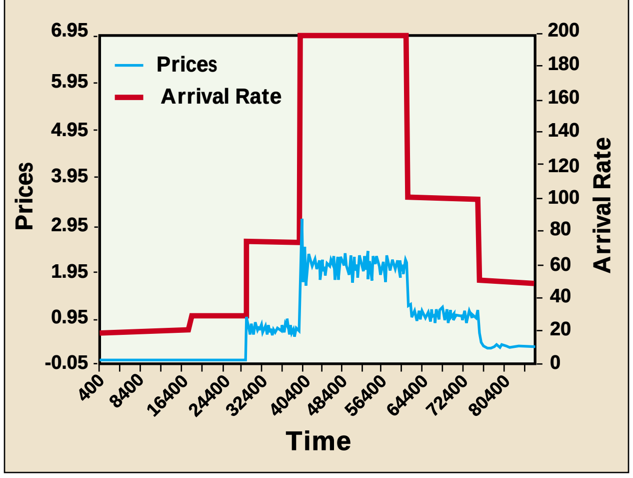


figure represents a demand pattern that varies during the course of the day according to the data presented in Table 1. These demand patterns cover the times when the network is underutilized (12:00 a.m.–5:00 a.m.) and the times when the network is significantly overloaded (11:00 a.m.–5:00 p.m.). We then superimpose the demand over the price fluctuations, the left y-axis represents the prices, the right y-axis represents the arrival rates, and the x-axis represents time in seconds. As the table indicates, the prices adjust quickly as the demand changes, with prices being close to 0 when the network is uncongested and being significantly higher when the network is very congested. Note that this adaptation is automatic and does not require any explicit knowledge regarding demand changes.

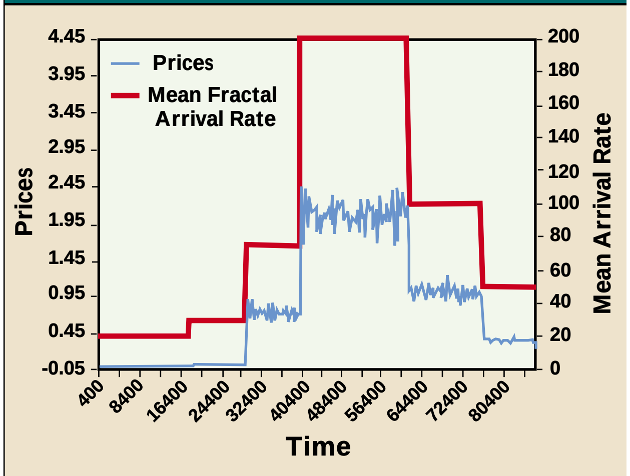
In our models, the demand is inherently bursty

[3] There need not be as many priority classes as the different types of QoS requirements, several different QoS requirements could be fulfilled by a single priority class.

[4] The details of the simulation model are provided in [5] and [8].

since there are 100 differently sized services and since each time a service request is made the amount of work could be in a range of 0–15 megabits. We introduced further variations in the demand process to simulate fractal demand patterns by generating each new arrival from a different distribution point, but the mean long-term arrival rates remained the same as those presented in Figure 2. For example, to achieve a fractal demand pattern with a long-term mean of 100 (or interarrival time of 0.01), we generated each new arrival time using an exponential distribution with parameter value of 0.02*U(0,1) for interarrival time, where U(0,1) is a uniformly distributed random variable between 0 and 1. Thus, each interarrival time is equally likely to be in the range of (0,0.02). Coupled with our job size distribution, this gives us a fractal demand pattern. Figure 3 presents the dynamic variation in prices using this approach; remarkably the prices are very similar to those presented in Figure 2. Both Figures 2 and 3 indicate that prices quickly increase as the load increases and decrease as the load decreases. This suggests that our computational approach works very well in dynamic environments and adjusts quickly to changing demand patterns.

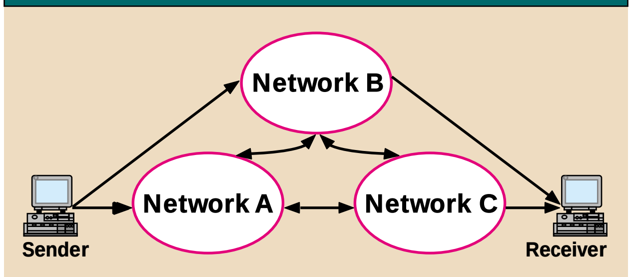
Figure 3. Prices with fractal demand pattern.



Other critiques of externality pricing are based on the fact that this method requires the knowledge of users' cost of delay. Most research on externality pricing assumes that users' cost of delay is known. Shenker et al. [12], appropriately point out that knowing the users' utility loss as a result of service degradation due to delay is "fundamentally unknowable." However, we have shown that reasonable estimates of users' delay-based utility loss can be computed by a Baysian computational approach based on current prices and observable user actions [10].

We do not believe that efficiencies similar to those provided by dynamic pricing can be achieved with static pricing approaches that do not consider the variation in demand or assume the variation in demand to be a known quantity. For static usage pricing approaches to be effective, demand patterns need to be continuously monitored and analyzed. This makes computing static pricing as difficult as computing dynamic prices. Furthermore, the approach presented in [5, 7] is completely decentralized and can be implemented based on information available at each individual network node and does not require system-wide information. We believe the key area of research is to continually improve the methods for short-term performance predictions. Such efforts will play a major role in improving the efficiency of any computable decentralized pricing mechanism.

Figure 4. Options for sending data.



Market Competition and Prices

The approach discussed earlier is an "ideal" solution for pricing traffic and services on an intranet or even on large networks with minor outside traffic [9]. However, most public data networks are a collection of networks owned by different entities such as Sprint, MCI, and AT&T. One of the major hurdles in evaluating any pricing mechanism for appropriateness is that it is unclear how a particular pricing approach will perform in the presence of competition in which different owners use different pricing strategies. In this section we explore the issues arising from competition and present a research agenda that can be used to evaluate market strategies and implications in e-commerce.

The importance of a decentralized pricing mechanism becomes evident when one considers the infrastructure of data communication networks. Consider the example depicted in Figure 4. The figure shows a transaction that may require traversing three separate networks A, B, and C. Within each network the data stream may go through several routers. The three networks may adopt different pricing policies. However, none of the network

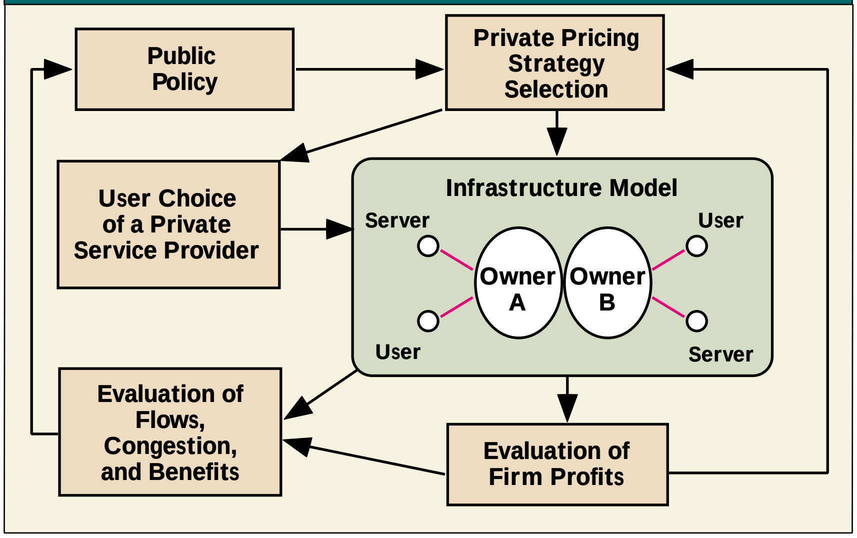


Figure 5. Model to explore economic incentives with multiple ownership.

providers can implement a pricing policy that requires complete path information and performance information at each router in the path. In a pricing framework such as [5, 7, 8], the network providers can price their cost of their servers based on the demand they observe and provide information regarding the total price and expected performance to users' decision-making agents or adjacent networks. This hierarchical approach provides a powerful analysis tool that can be used with edge-pricing[5]—where the price and QoS are agreed upon at the beginning of the transaction at the entry node.

Several types of existing interconnection agreements are described in [1]. These agreements are based on the relative size of the customer base and the size of the network owned by parties engaged in an agreement. Sizes of networks and customer bases reflect a coarse economic approximation of the costs and benefits of the interconnection agreement. However, given the growth in commercial traffic, increasing traffic management concerns, and the possibility of opportunistic resale of capacity, we believe that economic interconnection agreements will soon be negotiated based on cross-traffic flows. Usage-based pricing will play a key role in constructing these agreements.

For interconnection agreements to work, all the different entities involved must decide on pricing policies and QoS guidelines for their customers. It is unlikely that all the infrastructure providers are going to adopt the same pricing policy. Thus, while the study of optimal pricing mechanisms is important from the perspective of management, it is even more important to study the robustness of different pricing mechanisms under a multiple ownership environment. The study of different pricing mechanisms can provide valuable insights into the future of e-commerce as it evolves from a marginal marketplace to a substantial economic entity. Furthermore, such studies will provide valuable insights on public policy issues such as whether regulatory approaches should be used to force companies to adopt an optimal pricing scheme or what effect taxation of network traffic pricing will have on pricing policies and ultimately network performance. The current understanding of the behavior of public data communication networks with multiple owners, from an economic perspective, is limited. The majority of relevant studies focus on the unrealistic case of identical users and single service types, in which two-part tariffs (a combination of fixed fee and optimal congestion fee) can support optimal resource allocation. Unfortunately, these results are not valid in the case of heterogeneous user characteristics.

In [6] we present a comprehensive approach to studying a multiple ownership scenario that includes a rich realm of pricing policies such as fixed pricing, time-based pricing, volume-based pricing, monopoly pricing, and adaptive pricing strategies such as following the competitor and undercutting. Figure 5 presents the interacting components of this model.

Since the analytical models for such a complex system are not mathematically tractable, we are using simulation methodology to explore multiple ownership. At present we are calibrating the models for different pricing strategies. Our preliminary model includes two identical infrastructure providers that provide access to identical services. In this model, the owners adapt their pricing strategies according to their profits. This approach will provide us with a set of pricing policies that may coexist with each other while providing a consistent QoS for e-commerce applications.

In future work we will present the tradeoffs of these pricing policies in terms of performance, profits, and user satisfaction. We believe the economic stakes are sufficiently high so we can safely predict that considerable research will be devoted to devising

[5]Edge-pricing refers to pricing strategies in which a network is treated as a black box by users and all the prices and network latencies are considered by users only at the network entry points.

profitable strategies to price network traffic. One of the issues of interest is to anticipate how the different players might behave, and what public policies should be adopted to protect the common resource aspect of the Internet while providing the required and desired QoS. Some other key questions we intend to investigate are: Will the e-commerce infrastructure become dominated by a few powerful players and if so, will that be more efficient from a resource management perspective? Will there be incentives for or against interconnection agreements? How desirable is a multiservice class interoperable network with multiple QoS levels as compared to specialized networks for specialized services?

Conclusion

Pricing computer network traffic will be essential to manage network traffic and to control and provide different QoS required by different applications and users. A growing number of researchers are focusing on this initially unpopular idea. The realization comes from market forces, changing network infrastructure, projected Internet application base, projected use of the network, and the performance requirements of network users.

This article outlined the requirements, characteristics, and performance of an ideal price-setting mechanism. These prices could be computed in a completely decentralized manner. Decentralized price computation provides a high degree of protection against disruption due to network failure and congestion. Furthermore, with decentralized computing, the informational and computational requirements of price-computing mechanisms are minimal. One of the most significant benefits of this approach is that it provides an economic rationale for multiple levels of QoS. It will also provide incentives to maintain required QoS levels and will prevent the misuse of the network by redistributing the user demand patterns. Such a pricing mechanism can be used when the network is using different protocols for different types of applications, for example, ATM for real-time applications and TCP/IP for applications requiring only best-effort service.

Finally, public policy issues need to be addressed in the context of e-commerce. However, we still have relatively little insight into the operation of computer networks in a market-based economy. We have proposed a comprehensive model to analyze e-commerce and multiple ownership that could enhance our understanding and anticipate the problems and/or opportunities in e-commerce. These models will also be useful in analyzing the effects of taxation and regulation.

However, a significant amount of research and development is needed for the implementation of pricing mechanisms for network traffic. The challenges arise from network infrastructure, integration with network protocols, and ownership structure. We believe that a concerted effort is needed from academia, the computer industry, network service providers, and businesses involved in electronic commerce to design new mechanisms for network operations that will be suitable for a new generation of e-commerce applications. C

References

1. Bailey, J.P. The economics of Internet interconnection agreements. In L. McKnight and J. Bailey, Eds., *Internet Economics.* MIT Press, Cambridge, Mass., 1997, 155–168.
2. Cocchi, R., Estrin, D., Shenker, S., and Zhang, L. Pricing in computer networks: Motivation, formulation, and example. *ACM/IEEE Transactions on Networking 1* (1993), 614–627.
3. Deering, S. and Hinden, R. Internet protocol, version 6 (IPv6) specification. Technical report, IETF, Dec. 1995. www.globecom.net/(nocl,sv)/ietf/rfc/rfc1883.shtml
4. Gupta, A., Stahl, D.O., and Whinston, A.B. A priority pricing approach to manage multiservice class networks in real time. *Journal of Electronic Publishing*, 1995; www.press.umich.edu:80/jep/econTOC.html.
5. Gupta, A., Stahl, D.O., and Whinston, A.B. An economic approach to network computing with priority classes. *Journal of Organizational Computing and Electronic Commerce 6*, 1 (1996), 71–95.
6. Gupta, A., Stahl, D.O., and Whinston, A.B. Economic issues in electronic commerce. In R. Kalakota and A.B. Whinston, Eds., *Readings in Electronic Commerce.* Addison Wesley, MA, 1996, 197–227.
7. Gupta, A., Stahl, D.O., and Whinston, A.B. A stochastic equilibrium model of Internet pricing. *Journal of Economic Dynamics and Control 21,* (1997), 697–722.
8. Gupta, A., Stahl, D.O., and Whinston, A.B. Priority pricing of integrated services networks. In L. McKnight and J. Bailey Eds., *Internet Economics.* MIT Press, Cambridge, 1997, 323–352.
9. Gupta, A., Stahl, D.O., and Whinston, A.B. Managing computing resources in intranets: An electronic commerce perspective. *Decision Support Systems*; cism.bus.utexas.edu/alok/intranet.ps.Z.
10. Gupta, A., Jukic, B., Stahl, D.O., and Whinston, A.B. Designing an incentive compatible mechanism for Internet traffic pricing. Paper presented at the *DIMACS Workshop on Economics, Game Theory, and the Internet* (Apr. 18–19, Rutgers University), 1997; www.opim.uconn.edu/users/alokpapers/
11. MacKie-Mason, J., and Varian, H. Pricing the Internet. In B. Kahin and J. Keller, Eds., *Public Access to the Internet*. Prentice-Hall, NJ, 1995.
12. Shenker, S., Clark, D., Estrin, D., and Herzog, S. Pricing in computer networks: Reshaping the research agenda. *J. Telecommunications Policy 20*, 3 (1996), 183–201.

Alok Gupta (alok@sbaserv.sba.uconn.edu) is an assistant professor in the Department of OPIM at the University of Connecticut.

Dale O. Stahl (stahl@mundo.eco.utexas.edu) is Malcolm Forsman Professor of Economics in the Department of Economics at the University of Texas at Austin.

Andrew B. Whinston (abw@uts.cc.utexas.edu) is Director of the Center for Research in Electronic Commerce at the University of Texas.

This research was funded in part by National Science Foundation #IRI-9225010, but does not necessarily reflect the views of the NSF. Partial support was also provided by Texas Advanced Research Program.