Pricing Internet Services: Approaches and Challenges

Lee W. McKnight and Jahangir Boroumand

Mixing Internet protocol packets with economic theories and the real-world—but almost surreal—business practices that provide continually changing services to virtual markets gets complicated very fast. In part one of this two-part column, we review some of the current approaches and basic challenges to pricing Internet services. In part two, we will introduce new technical and economic approaches for overcoming these challenges.

FLAT-RATE PRICING AND CONGESTION

Currently, the predominant form of Internet retail pricing in the US is flat-rate pricing—charging a fixed fee for a set amount of bandwidth to access the network. This fixed fee does not vary with actual bandwidth usage, Internet congestion, the technical requirements of the information being transmitted, or the user's subjective valuation of the priority or importance of that information.

Flat-rate pricing's main appeal for users and service providers is its simplicity. Predictable fees reduce risks for both users and service providers, and they avoid administrative costs for tracking, allocating, and billing for usage. Having known expectations for payments also facilitates planning and budgeting.

Internet users who pay a fixed fee have no incentive to limit their use of the network.

However, some argue that the inevitable consequence of flat-rate pricing is the emergence of congestion. Internet users who have paid a fixed fee for network access are only concerned with their own costs and benefits, and they log on without the incentive to limit their usage of the network. Congestion eventually occurs, and another potential user has to wait for access.

For users who pay a fixed access fee, bandwidth is allocated by time rather than through pricing. Subscribers who access the Internet through service providers like AOL find that the system rations bandwidth resources according to user patience rather than social value. Impatient users who voluntarily leave the congested network are not necessarily the low-value users. The implication of this is the loss of social welfare when low-value users access the Internet while high-value users are denied access.

BANDWIDTH OVERPROVISIONING

Internet service providers have dealt with Internet congestion by overprovisioning of bandwidth, rather than through market mechanisms that allocate existing bandwidth among users and direct investments in new capacity in an economically efficient way. The steep decline in the cost of bandwidth in recent years has made the ISPs' overprovisioning response to the congestion problem possible, and that response seems likely to continue, at least in the near future. Overprovisioning of bandwidth is not sufficient to avoid Internet congestion because there is no intrinsic upper limit on bandwidth use.

The availability of more bandwidth will lead to the increasing use of bandwidth-intensive applications and services such as audio- and videostreaming. Even if capacity is expanded to handle real-time video, other more advanced applications such as 3D imaging and virtual reality will demand even more bandwidth.

Bandwidth prices vary dramatically in different regions of the world depending primarily on whether the market structure is competitive or noncompetitive. With the growing number of nations with competitive telecommunications markets and the increasing viability of both satellite and undersea fiber-delivered bandwidth, we foresee a general decrease in bandwidth prices over the coming decade. Therefore, any pricing model must support both increasing demand and supply of bandwidth.

Unless we envision a world in which there will be an unlimited supply of

This column is based on a paper titled "Pricing Internet Services: After Flat Rate," which Lee W. McKnight and Jahangir Boroumand presented at the MIT/Tufts Internet Quality of Service Workshop, sponsored by the National Science Foundation, Defense Advanced Research Projects Agency, and the French Embassy, 2-3 December 1999.

bandwidth at or near zero cost, we need a service model and pricing mechanism for allocating network resources that reflect consumers' valuations and the social cost of congestion. Ideally, this mechanism would give consumers choices in the selection of services and prices, and it would efficiently allocate investments for capacity expansion while allowing ISPs to recover their operating costs. However, this mechanism would not replace the flat-pricing scheme if consumers regard the pricing structure as too complicated or if service providers find the implementation and administrative costs to be too high.

INTEGRATED SERVICES

Internet congestion has prevented deployment of applications that are sensitive to packet delivery delays. This has led to development of the Integrated Services model, which provides extensions to the best-effort service model to allow some control over end-to-end packet delays. A key assumption of the IntServ model is that we must explicitly manage resources such as bandwidth to deliver quality of service (QoS). This implies that resource reservation and admission control are the service's key building blocks.

The IntServ model uses the RSVP protocol to implement resource reservation. But RSVP's deployment has faced many challenges. In particular, the resource requirements for running RSVP on a router increase proportionally with the number of separate RSVP reservations. This well-known scalability problem makes using RSVP on the public Internet impractical.

To put the IntServ model's approach to addressing Internet congestion in the proper perspective, we need to distinguish two very different notions of efficiency: network efficiency and economic efficiency, which the economics literature more accurately refers to as Pareto optimality.

If a network can maintain a target level of service while minimizing the resources it needs to provide this service, its operation is network efficient. On the other hand, economic efficiency refers to the relative valuations users attach to their network service. If no user currently receiving a particular service values it less than another user who is being denied that service, the operation is economically efficient. An example of an economically efficient operation is when no passenger of a railroad network values the transportation service less than someone who has been denied access to the same transportation service.

We need a mechanism for allocating network resources that gives consumers choices in services and prices and allows ISPs to recover their operating costs.

With this distinction between network and economic efficiency in mind, it is clear that the IntServ model's design philosophy and its approach to addressing the congestion problem are motivated more by achieving network efficiency than by gaining economic efficiency. Nevertheless, IntServ's design recognizes the scarcity of bandwidth and takes a significant step beyond providing QoS merely by overprovisioning.

However, problems with the RSVP protocol implementation have prevented deployment of the IntServ model over the public Internet. There has been no improvement over the best-effort service model, which has been in place since the early days of Internet development. The best-effort service model does not offer any QoS guarantee and provides no service differentiation to Internet users.

The flat-rate pricing that ISPs currently offer for accessing the Internet is compatible with the Internet's lack of service differentiation. There is no rationale for differential pricing, and the flat rate that ISPs charge to access the Internet provides the same best-effort service to all users.

TRAFFIC DELAYS

The public Internet's lack of control over QoS is preventing deployment of new applications. Today, depending on the state of congestion in various parts of the network, packets traversing the Internet can encounter significant delays or they can even be dropped. Existing as well as potential Internet applications have different requirements for timely delivery of their data over the network.

The so-called elastic applications can work without guarantees of timely delivery since they can stretch in the face of increased delay and still perform. Examples of elastic applications are Web browsing, e-mail, and FTPs. But some real-time applications such as audio- and videostreaming are essentially sensitive to timely data delivery. In particular, while the technology for voice over IP is now available, its deployment on the public Internet has been severely limited because of the congestion problem.

Researchers who are concerned with the consequences of Internet congestion—and who are not satisfied with the overprovisioning of bandwidth as a solution—have proposed other service and pricing models, which we will discuss in the next Internet Watch column.*

Lee W. McKnight is an associate professor of international communication and director of the Edward R. Murrow Center at the Fletcher School of Law and Diplomacy, Tufts University, and president of Marengo Research LLC, a consultancy. Contact him at lmcknigh@emerald.tufts.edu.

Jahangir Boroumand is a visiting professor at the R.H. Smith School of Business, University of Maryland. In 1999, he was a visiting scholar at the Edward R. Murrow Center at the Fletcher School of Law and Diplomacy, Tufts University, where the research for this column was carried out. Contact him at jborouma@wam.umd.edu.

Editor: Ron Vetter, University of North Carolina at Wilmington, Department of Computer Science, 601 South College Rd., Wilmington, NC 28403; voice +1 910 962 3667, fax +1 910 962 7107; vetterr@uncwil.edu